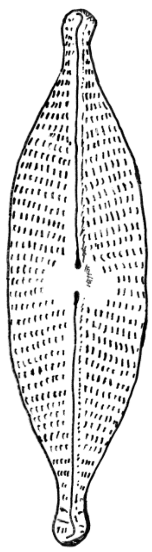
DNA Barcoding and eDNA Barcoding in Diatoms

Jonas Zimmermann



Dissertation zur Erlangung des Doktorgrades der Naturwissenschaftlichen Fachbereiche der Justus-Liebig-Universität Gießen

November 2014

DEKAN
Prof. Dr. Holger Zorn

GUTACHTERIN
PD Dr. Birgit Gemeinholzer
PD Dr. Gernot Glöckner

Jonas Zimmermann: DNA BARCODING AND eDNA BARCODING IN DIATOMS, Dissertation zur Erlangung des Doktorgrades der Naturwissenschaftlichen Fachbereiche der Justus-Liebig-Universität Gießen, 2014

COVER: *Navicula rhynchotella* (LANGE-BERTALOT, 1993), drawing of a specimen from the River Spree

The world is full of magical things patiently waiting for our wits to grow sharper

Bertrand Russel

To NE, my friends and parents

CONTENTS

PART I

SYNOPSIS

RATIONALE

Water is covering over 70 % of our planet's surface – not counting ice formations and phreatic waters – and it is a key resource for survival of all organisms, among them human beings. The oceans are one of the major driving forces for climatic conditions on earth and at the same time responsible for buffering climatic extremes. The complex network of ocean and fresh water organisms provide a crucial food source for mankind. Clean inland fresh water is already a heavily fought over resource and its availability is further decreasing due to e.g. pollution through anthropogenic influences and climate change. Many countries have now installed systems for monitoring the environmental health of their inland fresh water bodies and coastal waters.

The use of aquatic organisms as indicators of the status of water bodies has become a common approach many decades ago. Among the most important indicator organism groups are diatoms as implemented e.g. in the European Water Framework Directive. The occurrence and composition of diatom taxa – each correlated to specific environmental conditions - is currently commonly assessed via light microscopy. This approach is prevalent, however, it is also limited because it is prone to subjectivity in classification of specimen phenotypes and also requires expert taxonomic knowledge.

This dissertation aims to develop an alternative and more objective system for the identification of diatom taxa and the assessment of diatom community structure for water quality analysis using a DNA sequence based approach. The presented alternative – called environmental DNA barcoding – uses a short, but taxon specific DNA fragment to unambiguously identify diatom taxa from water samples. The publications in this work not only introduce an appropriate DNA barcoding marker along with a standard laboratory protocol, but demonstrate the potential of the new approach to capture diatom diversity. Furthermore, this dissertation includes a critical examination of necessary standards for the build-up of reference databases. It emphasises the necessity of integrative science in the development of new methodologies, as it combines DNA metabarcoding, taxonomy, phylogeny and scientific documentation.

INTRODUCTION

AN INTRODUCTION TO DIATOMS

Diatoms were first mentioned as adhering to roots of pond-weed in 1703 by an anonymous gentleman (Anonymous 1703). His findings were nonetheless published in the Philosophical Transactions of the Royal Society of London (Ford 1991; Round *et al.* 1990). In the second half of the 18th century O.F. Müller described several species of diatoms (Müller 1783, 1786), and one of these species, namely *Vibrio paxillifer*, later on became the type taxon for the first diatom genus *Bacillaria*, described by Gmelin (1791). Müller called diatoms "Animalcula infusoria" (Müller 1786), the German translation "Infusionsthierchen" was also used by Ehrenberg (1838), an early pioneer of micropaleontology and diatom research. "Infusionsthierchen" as part of the protozoa were considered to belong to the animal kingdom. As unicellular motile organisms, diatoms were classified as animals by most authors (e.g. Bory (1822)), but it remained a topic of discussion whether this classification was accurate. In 1844 Kützing published a monograph on "kieselschalige Bacillarien", and he treated diatoms as plants inter alia due to the presence of chlorophyll and photosynthetic activity in diatom cells (Kützing 1844). After this publication, diatoms were almost always treated as algae (with the exception of Ehrenberg); however, "algae" is a paraphyletic taxon, pooling all aquatic photosynthetic organisms.

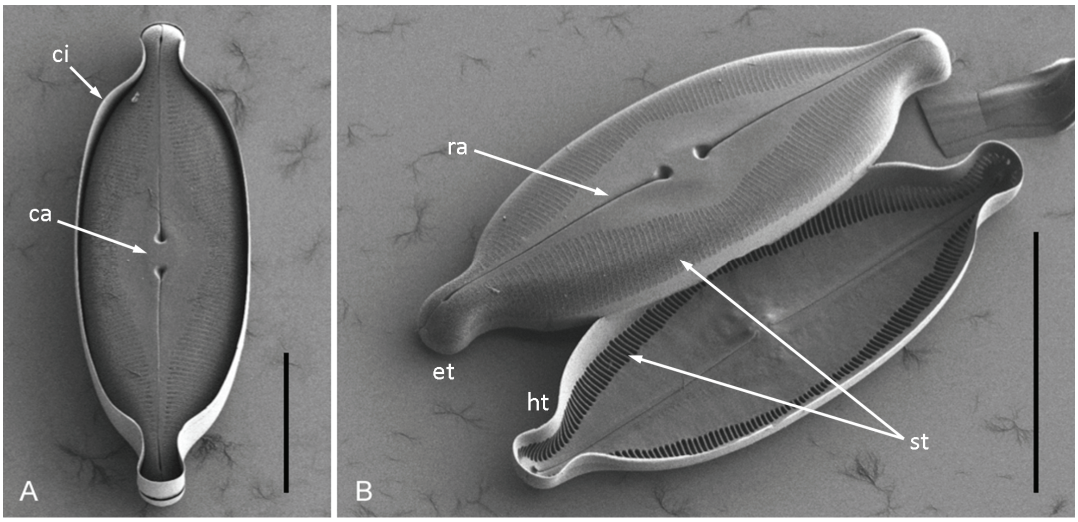


Figure 1: SEM pictures of *Caloneis amphisbaena* (Bory) Cleve frustules (Lusatian Neisse, Germany). A Top view; ca central area, ci cingulum (girdle band); scale 20 µm. B External view of epitheca and internal view of hypotheca; et epitheca, ht hypotheca, ra raphe, st striae; scale 30 µm. Photographs taken by the Research Group Diatoms, Botanic Garden and Botanical Museum Berlin-Dahlem.

Diatom plastids are usually greenish-brown, as carotinoid pigments (e.g. ß-carotene, diatoxanthin, diadinoxanthin and fucoxanthin) conceal the green colour of the chlorophylls (Goodwin 1974). Diatoms possess chlorophyll *a*, c_2 and c_1 or c_3 (Stauber & Jeffrey 1988). The diatom plastids arose from an ancient secondary endosymbiosis with a photosynthetic eukaryote of red algal origin, but new studies showed that the plastid genome of the diatoms also contains genes of green algal derivation, whereupon different scenarios of this process are discussed (Dorrell & Smith 2011; Moustafa *et al.* 2009). However, the ancestry of these genes is still disputed due to large scale phylogenetic analyses (Burki *et al.* 2012; Woehle *et al.* 2011).

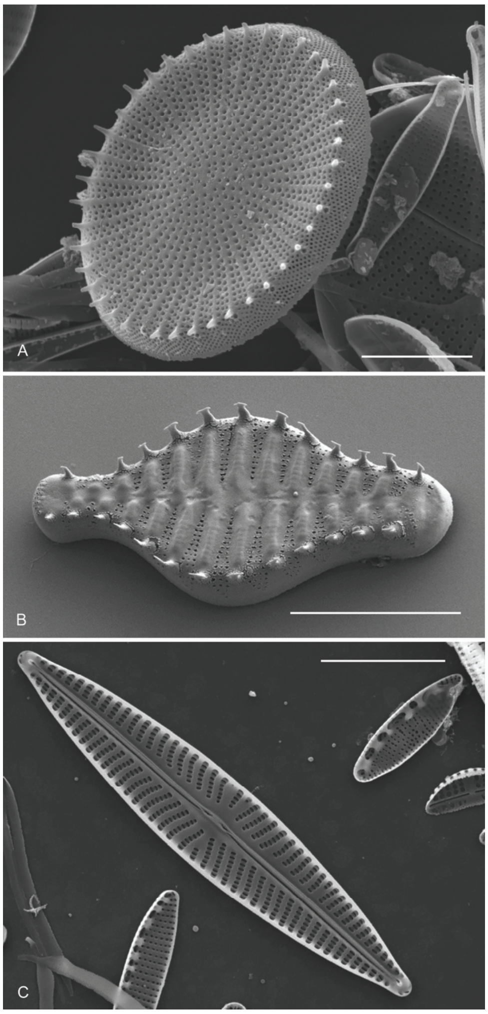


Figure 2: SEM pictures of valves from the main diatom groups. A Centric diatom genus *Stephanodiscus* Ehrenberg (Lake Tegeler See, Germany); scale 6 µm. B Pennate, araphid diatom genus *Staurosira* Ehrenberg (Lake Balaton, Hungary); scale 5 µm. C Pennate, raphid diatom genus *Navicula* Bory (River Spree, Germany); scale 8 µm. Photographs taken by the Research Group Diatoms, Botanic Garden and Botanical Museum Berlin-Dahlem.

Today, the mostly photoautotrophic diatoms or Bacillariophyta are classified as Stramenopiles, formerly belonging to the kingdom Chromalveolata, a group that also includes brown algae, golden algae and several non-photosynthetic relatives (Andersen 2004). A more recent classification places the Stramenopiles into the SAR supergroup (Burki *et al.* 2007), including Stramenopiles, Alveolates and Rhizaria.

The traditional delimitation of diatom taxa is based on the morphology of the silica (SiO_2; nH_2O) cell wall, a unique feature of diatoms. The wall of each cell is composed of two valves (shells), the epi- and hypotheca, which are connected by several girdle bands or cinctures (Fig. 1). The valves feature several elaborate and complex structures that can vary distinctively between different taxa (Round *et al.* 1990). Typical structures are areolae, ocelli, portulae, pores, spines, slits (e.g. raphe and striae), and ridges. The siliceous valves void of other organic compounds are referred to as frustules (Round *et al.* 1990). The size, shape and sculpturing of the valves are used for the taxonomic subdivision of diatoms (Stoermer & Smol 1999).

Traditionally diatoms were arranged into round or actinomorphic “centric” (Fig. 2a) and elongate, bilateral symmetric “pennate” cells (Fig. 2b,c). Some of the pennate cells are characterised by one or two longitudinal slits, the raphes. This group of diatoms is called raphid pennates (Fig. 2c) and it is discriminated from a second group, the araphids (Fig. 2b), which lack a raphe. The traditional division into two orders Centrales (centrics) and Pennales (pennates) has been refined to the classification into three classes, splitting the Pennales based on raphe morphology (Guiry 2013+; Round *et al.* 1990). The centrics are combined in the Class Coscinodiscophyceae Round & R.M. Crawford, the araphid pennate diatoms in the Class Fragilariophyceae Round and the raphid pennate diatoms in the Class Bacillariophyceae Haeckel. Whereas the Centrales (corresponding to the Coscinodiscophyceae) are a paraphyletic taxon, the pennates (including Fragilariophyceae and Bacillariophyceae) are considered a monophyletic group (Sims *et al.* 2006). However, this classification is still in discussion and under revision (Medlin & Kaczmarska 2004), especially because molecular data has considerably altered the view on diatom systematics and phylogeny (Alverson & Theriot 2005; Theriot *et al.* 2010).

According to e.g. Coyne and Orr (Coyne & Orr 2004) the most commonly used species concept is the so called “biological” species definition introduced by Mayr (1942). However, this is not practicable for the definition of species in protists (Brown & Lomolino 1998; Wilson 2003) and therefore also in diatoms. The mostly descriptive species concept in diatoms is based on morphological discontinuities in closely related taxa; and utilises therefore a “phenetic” species definition (Hutchinson 1968).

The number of extant diatom species is estimated to be between 30,000 and 100,000 (Mann & Vanormelingen 2013; Norton *et al.* 1996). Therefore they are often considered as one of the most species rich lineages in the Eukaryotes (Mann 1999; Mann & Droop 1996), albeit diatoms are better documented than most other protist groups, which renders numerical comparisons difficult. So far only 12,000 species have been formally presented to science, therefore the majority of species still awaits discovery and description (Mann & Vanormelingen 2013; Smetacek 1999).

The diatom fossil record is very good due to their siliceous cell walls which survive aeons. Many taxa have been and still are being described from sediment cores and other fossilised material. For a long time the first paleontological occurrence of diatoms was considered to be derived from the Lower Cretaceous (early Albian, 111 Ma) from the Weddell Sea, Antarctica (Gersonde & Harwood 1990). But more recent fossil data from the early Jurassic (185 Ma ago, Pliensbachium) had proven that diatoms are much “older” (Kooistra & Medlin 1996) and molecular clock data (Kooistra & Medlin 1996) and sedimentary evidence (Schieber *et al.* 2000) suppose that they have an earlier origin. There are also ideas that the origin of diatoms could be correlated to the period after the Permian mass extinction (P-Tr event) (Medlin *et al.* 1997). The lack of fossil data within this period is explained by the hypothesis that diatoms first were not silicified (Raven & Waite 2004).

Diatoms occur in virtually all aquatic habitats – freshwater as well as marine - but also in soil or as aeroplankton. Life strategies include planktonic forms and benthic species adhering to many different kinds of substrate (e.g. mud, sand, stones, plants, algae, animals). Benthic diatoms are mostly raphid and their raphe seems to be an essential innovation for the very successful colonisation of benthic habitats in both marine and freshwater ecosystems, as it allows directional motility (Sims *et al.* 2006). Diatoms also appear in interstitial and intertidal communities. Psychrophilic diatoms populate e.g. the upper and lower side of Arctic (Van Baalen & O'Donnell 1983; Van de Vijver *et al.* 2004) and Antarctic (Aletsee & Jahnke 1992; Van de Vijver *et al.* 2004) ice formations. To endure different extreme environmental conditions, e.g. desiccation during droughts, many diatom species can form resting stages, either in the form of auxospores or dormant vegetative cells (Round *et al.* 1990).

Most diatoms secrete polysaccharides which can form gelled capsules around the cells in the form of threads, pads or stalks (Daniel *et al.* 1987) which are sometimes used to firmly attach to all kind of substrates. This polysaccharide mucus also allows these single celled organisms to live in many different types of colonies (Round *et al.* 1990). These complex polysaccharides are a photosynthetic storage product as are a number of long-chain omega-3 polyunsaturated fatty acids. Both organic compounds are responsible for diatoms being at the base of most aquatic food webs and a preferred food source for invertebrate grazers (Armbrust 2009). The omega-3 fatty acids bio-accumulate in fish after passage through the aquatic food web, and are here harvested as dietary "fish oil" supplements (Lebeau & Robert 2003).

Not only because of these polysaccharides and fatty acids do diatoms play an important role in most ecosystems, but also due to their capability to fixate carbon dioxide and silica. They are responsible for at least 25 % of global carbon dioxide fixation (Falkowski *et al.* 1998; Field *et al.* 1998; Mann 1999; Smetacek 1999). Furthermore, they contribute approximately between 20 % (Mann 1999) and 45 % (Yool & Tyrrell 2003) of the global net primary production. A very interesting unique feature for photosynthetic organisms is a metazoan-like urea-cycle in diatoms, which is an important player in effective nitrogen fixation and is believed to be responsible for the dominating role of diatoms especially in marine environments (Allen *et al.* 2011).

DIATOMS AS BIOLOGICAL INDICATORS

Water bodies are complex ecosystems that are influenced by many different factors, such as e.g. climate, land use and geomorphology in the watershed (Stoermer & Smol 1999). These influences can change the chemical, physical and biological composition of the water bodies on a spatial and temporal scale (Richards *et al.* 1996). Quality control of water bodies is crucial as they account for over 70 % of the Earth's surface (not counting phreatic water systems and ice formations) and represent a large economic value for human society (Zinger *et al.* 2012).

To assess the ecological status of a water body often physical and chemical measurements are used. In an ideal approach all chemical and physical factors that could influence the integrity of the ecosystem should be continuously measured (Barbour *et al.* 1995; Norris & Morris 1995); this, however, is highly impractical (Stoermer & Smol 1999). An alternative is the use of biological indicators. Biological indicators show specific reactions to their environment. Thus, complex chemical and physical changes which might be difficult to measure can be assessed with such indicators (Stoermer & Smol 1999). They provide a holistic assessment not only of the current status in a water system, but also often of past events (Stoermer & Smol 1999). Beyond the assessment of water quality many monitoring programs are also concerned with the management and protection of the organisms that dwell in it. Chemical and physical data cannot provide information on the status of organisms or their condition, while biological indicators can. Commonly used groups of indicator organisms are e.g. fish, macrozoobenthos, phytoplankton and phytobenthos; a major part of the latter being diatoms (European Parliament 2000).

Diatoms have one of the shortest generation times of all biological indicator groups and therefore have the potential to also provide short term indications of environmental change (Rott 1991). Diatoms as bioindicators were first used to assess water quality in lakes, but this application was later extended to streams and rivers (Round 1991). Periphytic and phytobenthic diatoms are used for the assessment of ecological status in rivers and streams, whereas planktonic diatoms provide valuable information in large lowland rivers and lakes (Stoermer & Smol 1999), as well as in marine environments.

As diatoms are well preserved in fossil sediments and serve as important stratigraphic indicators (Spaulding & McKnight 1999; Stevenson & Pan 1999; Stoermer & Smol 1999), their analysis enables status assessments of water bodies not only in the present but also in the past. This is achieved through the combined analysis of diatom assemblages in the sediment and extant samples, which allows for the reconstruction of past water chemistry and periods of increased or decreased moisture, resulting in data on climatic variability through time (Davies *et al.* 2002).

The presence/absence as well as the abundance of individual diatom taxa is a direct respond to short term shifts in biological, chemical and physical factors, such as herbivory (McCormick & Stevenson 1989; Steinman *et al.* 1987), nutrient concentration (Pan *et al.* 1996; Pringle & Bowers 1984), contaminants (Cox 1991) and temperature (Descy & Mouvet 1984; Squires *et al.* 1979). However, the specific composition of diatom communities is also a sensitive indicator of mid term and long term changes in their environment that are due to climatic change and an increasing anthropogenic pressure (Eiler *et al.* 2013), the latter resulting in e.g. acidification or eutrophication (Spaulding & McKnight 1999; Stevenson & Pan 1999). While some diatom species have a broad ecological plasticity, others are adapted to specific environmental conditions. Those species with narrow optima and tolerances for different environmental variables often react quickly to an alteration of their environment with changes in their growth rate, which can be quite considerable in diatom populations. This leads to differences in biomass and species composition of diatom assemblages (Patrick 1961). These species can function with a high level of certainty as indicators for shifts in environmental conditions (Round *et al.* 1990). Closely related species are often found under vastly different ecological conditions (Poulíčková *et al.* 2008; Vanelslander *et al.* 2009).

In many countries (Adler 1995; Dell'Uomo 1999; Descy & Coste 1991; Kelly 1998; Prygiel *et al.* 2002; Stevenson & Pan 1999; Zalack *et al.* 2010) diatom indices are being used for water quality assessments. In the European Union, most countries have by now implemented the standardised European Water Framework Directive (WFD) (2000) into their legislation. Two main types of diatom indices for water quality assessments have been used. The first type are autecological indices, based on the work of Kolkwitz and Marsson (1908). These infer levels of pollution utilising the composition of the species and the ecological preferences and tolerances of the individual taxa present in the assemblage (Butcher 1947; Dell'Uomo 1999; Lange-Bertalot 1979a; Lange-Bertalot 1979b; Lowe *et al.* 1986). A second type of indices mainly relies on the diversity of the diatoms as indicator of ecosystem health (Patrick 1949; Patrick & Strawbridge 1963; Patrick *et al.* 1954).

The WFD issues standardised guidelines on how to handle and analyse diatom assemblages to infer the ecological status of water bodies (European Committee for Standardization 2003, 2004; Kelly *et al.* 1998). The analysis demands unambiguous identification of diatoms down to the species level via their frustules using light microscopy. However, even though diatoms possess more morphological characters than other protist groups, it is still difficult to identify diatoms morphologically below the genus level. It needs expert taxonomic knowledge and often expensive infrastructure as many of the morphological characters can only be detected by scanning electron microscopy or similar high resolution technologies. Furthermore, the morphological characters can substantially vary even within populations or clonal cultures (Bailey-Watts 1976; Jahn 1984; Medlin *et al.* 1991; Van de Vijver *et al.* 2013; Zinger *et al.* 2012), e.g. due to frustule size changes during vegetative reproduction. And even if all necessary characters for identification are available, trained taxonomists sometimes come to different conclusions, either due to differing taxonomic concepts or because the descriptions available allow for different interpretations (Mann *et al.* 2010). All these factors can contribute to misleading ecological assessments due to misidentifications (Archibald 1984; Morales *et al.* 2001). Furthermore, it renders these assessments inherently operator dependent (Zinger *et al.* 2012).

DNA BARCODING AND EDNA BARCODING

The concept of DNA barcoding (Hebert *et al.* 2003) introduced a DNA sequence based identification alternative/complement to traditional morphological approaches. The DNA barcode consists of a marker fragment easily sequenced in one read that unambiguously identifies a given taxon (Hebert *et al.* 2003; Moritz & Cicero 2004; Stoeckle 2003; Zimmermann *et al.* 2011), independent of its life stage (Eiler *et al.* 2013; Hajibabaei *et al.* 2005; Hajibabaei *et al.* 2012; Mann *et al.* 2010; Shokralla *et al.* 2012) by comparison with a reference database. Environmental DNA (eDNA) barcoding (Kermarrec *et al.* 2014) or metabarcoding compares the sequence of an unidentified specimen from an environmental sample to sequences in a reference database. The unassigned sequence is matched to an identified specimen with the highest sequence similarity in the reference database, thus allowing the placement of this sequence in a relationship tree (Ratnasingham & Hebert 2007).

DNA barcoding can serve the purpose of sequenced based characterisation of genera, species or even taxa on the subspecific level (molecular taxonomy), whereas eDNA barcoding evaluates the community composition in a target environment (Schindel & Miller 2005; Zimmermann *et al.* 2011). Especially the latter requires high throughput approaches. Collins & Cruickshank (2013) recommend the concept of specimen identification and species discovery to elucidate the fact that the identified organism is an individual representing a taxon rather than delimiting one. Species discovery on the other hand can only be aided by DNA barcoding, as it cannot be decided on sequence data alone whether a sequence dissimilar to already known ones correlates to a novel species; it needs a formal taxonomic description and the substantiation by different kinds of data: e.g. morphology, DNA sequence data of several markers, ecology (DeSalle *et al.* 2005; DeSalle 2006; Moritz & Cicero 2004).

However, as the eukaryotic diversity of small and microbial organisms has a formative role in almost all ecosystems and because its traditional taxonomic discovery poses a big challenge due to the lack of accessible discriminating morphological characters, new and accelerated methods of discovery and identification are needed (Godfray *et al.* 1999; Monaghan *et al.* 2009; Novotny *et al.* 2006). Generally, DNA based approaches could overcome some of the challenges (Blaxter & Floyd 2003; Tautz *et al.* 2003), but only in the case that the variation in the used marker regions would correspond to species level taxa (Monaghan *et al.* 2009).

Impediments for DNA based taxonomical approaches could be incomplete lineage sorting and introgression, because divergence of a single marker can antedate or lag behind species divergence (Hudson & Coyne 2002).

A third concern is the lack of knowledge about the relation between inter- and intraspecific genetic variation of DNA barcoding regions in several lineages which makes it difficult to establish quantitative boundaries between taxa. Thresholds of genetic divergence, derived from uncorrected p-distances, have been used to delimit species, the best-known (and controversial) being the 10x rule in metazoans (Hebert *et al.* 2004). The rule claims that a genetic variation of 10× the average intraspecific difference indicates a new species in metazoans (Hebert *et al.* 2004). For prokaryotes a similar concept has been introduced with the 97 % rule by Forney *et al.* (2004); stating that sequences with less than 97 % similarity are most likely to be derived from different species. Threshold based approaches yielded mixed success in many organism groups (Meyer & Paulay 2005; Moritz & Cicero 2004), among them diatoms (Zimmermann *et al.* 2011). Tree based approaches, such as the general mixed Yule-coalescent (GMYC) model (Fontaneto *et al.* 2007; Pons *et al.* 2006) implement evolutionary models in the assessment of species boundaries by inferring these from a shift in branching rates on the tree (Monaghan *et al.* 2009).

In summary, next generation DNA sequencing techniques have the potential to accelerate species discovery from environmental samples, but only under the condition that operational taxonomic units (OTUs) – a "cloud" of similar sequences – could be related to species (Monaghan *et al.* 2009).

Using DNA barcoding for the identification of specimen rather than species discovery has advantages over morphological approaches (Packer *et al.* 2009). Not only because the interpretation of morphological characters is prone to subjectivity, but also because there is only a limited number of taxonomists that are qualified to do morphology based identifications (Archibald 1984). Furthermore, there are cases where morphological features are insufficient to discriminate between taxa or are not easily accessible, but where the DNA sequence shows distinct nucleotide differences for organismal delimitation. These taxa are often referred to as cryptic or semi-cryptic species. In diatoms for only about 15 species the relationship between diagnostic morphology and biological species boundary has been investigated in detail (Amato *et al.* 2007; Behnke *et al.* 2004; Casteleyn *et al.* 2008; Kaczmarska *et al.* 2008; Vanormelingen *et al.* 2007). Molecular taxonomy could make an important addition to the so far used phenetic species concept in diatoms as it already has in other micro-eukaryotic groups e.g. (Barth *et al.* 2006; Chantangsi *et al.* 2007; Scicluna *et al.* 2006).

If DNA barcoding is used for identification of diatom species, it does not rely on a pre-existing concept of species boundaries, the identified DNA sequence based units can be transferred to any taxonomic concept (Rach *et al.* 2008). This is due to the fact that the sequence of any given specimen remains the same, regardless which taxon the specimen is assigned to. Of course different alleles or multiple and varying copies of the same sequence locus can occur in a single specimen and this should be taken into account when choosing and using barcode makers for identification.

Identification of specimen using DNA barcoding has also been criticised; partly because it is said to use a single complex character for identification rather than multiple morphological (Dasmahapatra & Mallet 2006; Moritz & Cicero 2004; Rubinoff *et al.* 2006) and/or molecular traits (Dasmahapatra *et al.* 2010; Dupuis *et al.* 2012). Also it has been noted that because often only very few individuals of a given taxon or closely related taxa have been molecularly studied, intraspecific variation has likely been underestimated, while interspecific variation has been overestimated (Dasmahapatra & Mallet 2006). Collins and Cruickshank (2013) suggest an increased diligence in the a priori identification of specimen via DNA barcoding due to these and other shortcomings, such as the influence of misidentifications of reference taxa (see below). Despite all these problems DNA barcoding has proven useful in many organismal lineages, especially in the investigation of environmental samples (Behnke *et al.* 2011; Eiler *et al.* 2013; Hajibabaei *et al.* 2011; Kermarrec *et al.* 2014; Pawlowski *et al.* 2011; Zimmermann *et al.* 2014b).

Several studies demonstrated the use of DNA barcoding for the identification of diatoms from environmental samples e.g. (Kermarrec *et al.* 2014; Stoof-Leichsenring *et al.* 2012), but there is an ongoing discussion about the ideal barcoding marker. The first proposed DNA barcode marker has been the mitochondrial *cytochrome c oxidase* subunit 1 (*cox*1) (Hebert *et al.* 2003). However, in diatoms it has been proven to be not practicable, mostly due to the impossibility of designing universal diatom primers e.g. (Moniz & Kaczmarska 2009; Zimmermann *et al.* 2011). As in most organism groups many alternatives have been discussed (Hamsher *et al.* 2011; Luddington *et al.* 2012; Moniz & Kaczmarska 2009, 2010; Zimmermann *et al.* 2011), among them the *ribulose-1,5-bisphosphate-carboxylase/oxygenase* large subunit (*rbc*L) and the 18S V4 region. The latter is also the pre-barcode for protists (Pawlowski *et al.* 2012).

Another critical point for eDNA barcoding, besides the selection of the appropriate barcode marker and the establishment of a working analysis protocol, is the creation of the reference library (in form of a reference database) against which the environmental sequences are compared e.g. (Becker *et al.* 2011; Bortolus 2008; Collins & Cruickshank 2013; Kvist 2013). The most commonly used reference databases are the barcoding of life database BOLD (Ratnasingham & Hebert 2007) and GenBank (http://www.ncbi.nlm.nih.gov/genbank) respectively EMBL/ENA (http://www.ebi.ac.uk/ena) and DDJB (http://www.ddbj.nig.ac.jp/). But the taxonomic classification of the sequence data in these databases is often of questionable quality (Kvist 2013; Zimmermann *et al.* 2014a). The dubious quality of the reference data has a direct influence on the quality of eDNA barcoding analysis of environmental samples. Therefore it is necessary to establish standards on how to deposit and document reference data. Better sequence documentation - facilitating the traceability of sequence data to the original specimen - does not only increase the quality of eDNA barcoding but also increase the re-usability of deposited data (Enke *et al.* 2012).

Research Objectives

To establish eDNA barcoding as an alternative and/or an addition to classical microscopical analysis of diatom communities for water quality assessments, this study explores several methodological aspects of eDNA barcoding:

- The choice of an appropriate marker for eDNA barcoding in diatoms
- The establishment of a working laboratory protocol for the analysis of diatom communities from environmental samples
- A critical assessment of reference libraries in diatoms
- The taxonomical discriminatory power of eDNA barcoding in diatoms in comparison to light microscopy

PAPER OUTLINE

This dissertation comprises four papers that are presented in chronological as well as logical order. The first two papers are concerned with the establishment of an universal barcoding marker system in protists (paper 2) and the appropriate marker region in diatoms and the establishment of a suitable laboratory protocol for eDNA barcoding in the latter group (paper 1). The third paper critically discusses the current state of sequence associated taxonomic information in the available reference databases for diatoms and proposes guidelines for future data deposition. The fourth paper shows the applicability of eDNA barcoding for diatom community analysis within the scope of water quality assessments in the European water framework directive.

PUBLICATION 1

Zimmermann, J., Jahn, R., Gemeinholzer, B. (2011). Barcoding diatoms: evaluation of the V4 subregion on the 18S rRNA gene, including new primers and protocols. Organism Diversity & Evolution 11 (4): 173-192.

Outline – This study evaluates several regions of the SSU 18S rRNA gene for DNA barcoding in diatoms. With universal applicability, easy amplification and sequencing as well as taxonomical discriminatory power as desired prerequisites for a barcoding marker in diatoms. The V4 region of the 18S gene is identified to have the best properties in relation to these criteria. It is also short enough to be suitable for next generation sequencing approaches. A laboratory protocol for the amplification and sequencing of the 18S V4 region diatoms is established including universal primers. The intra- and interspecific variation of the marker region is tested on 35 representative taxa of diatom diversity. The taxonomical discriminatory power of the 18S V4 regions has been further tested *in silico* on 123 taxa as well as in a cryptic species complex within *Sellaphora*.

Contribution – Lead author, molecular/lab work, data analysis.

PUBLICATION 2

Pawlowski, J., Audic, S., Adl, S., Bass, D., Belbahri, L., Berney, C., Bowser, S.S., Cepicka, I., Decelle, J., Dunthorn, M., Fiore-Donno, A.M., Gile, G.H., Holzmann, M., Jahn, R., Jirků, M., Keeling, P.J., Kostka, M., Kudryavtsev, A., Lara, E., Lukeš, J., Mann, D.G., Mitchell, E.A., Nitsche, F., Romeralo, M., Saunders, G.W., Simpson, A.G., Smirnov, A.V., Spouge, J.L., Stern, R.F., Stoeck, T., Zimmermann, J., Schindel, D., de Vargas, C. (2012). CBOL Protist Working Group: Barcoding Eurakyotic Richness beyond the Animal, Plant, and Fungal Kingdoms. PLoS Biol.;10(11): e1001419.

Outline – This publication summarises the discussion on DNA barcoding within the protist research community. As the protist richness is vastly unknown, the paper proposes a two step DNA barcoding approach to assess and describe protist diversity: in a first step the universal pre-barcode 18S V4 should document the basic diversity in all protist groups. This universal pre-barcode is especially important in the analysis of environmental communities where many kinds of protists live along each other and need to be accessed with the same marker. In a second step group-specific barcodes are discussed to provide further information on taxon diversity and delimitation within each individual group. To document the protist diversity also cell cultures need to be established and the sequence information derived from those cultures need to be deposited in a protistan reference library to make protist diversity less anonymous.

Contribution – Analysis of diatom metadata, active contributions to content, text and figures of manuscript.

PUBLICATION 3

Zimmermann, J., Abarca, N., Skibbe, O., Kusber, W.-H., Enk, N. & Jahn, R. (2014). Taxonomic Reference Libraries for Environmental Barcoding: a Best Practice Example from Diatom Research. PLoS ONE 9(9): e108793.

Outline – This paper critically assesses the consistency of sequence associated taxonomic information for the two commonly used barcoding markers 18S V4 and *rbc*L in the International Nucleotide Sequence Database Collaboration (INSDC, including GenBank, EMBL/ENA and DDBJ) for naviculoid diatoms. As the taxonomic consistency is found to be very low, this paper develops best practice guidelines in a mulitproxy data approach on how to document sequence associated information on the example of 37 freshwater from 15 naviculoid diatom genera. These guidelines include that each reference sequence needs to be linked to a voucher specimen – and also to photographic documentation – in order to provide a complete chain of evidence back to the formal taxonomic literature. It is demonstrated that the critical revision of collection material for the taxonomic reference database opens up the possibility of reassessing phylogenetic and taxonomical data and in the present paper also results in the description of 4 new species.

Contribution – Lead authorship, design and performance of molecular experiments, data analysis.

PUBLICATION 4

Zimmermann, J., Glöckner, G., Jahn, R., Enke, N. & Gemeinholzer, B. (2014). Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies. Molecular Ecology Resources .

Outlook – This publication comprises a methodological comparison of eDNA barcoding (marker 18S V4) and light microscopy for the assessment of diatom communities according to the European Water Framework Directive. Seven environmental samples of benthic diatoms from the course of the Lusatian Neisse/Odra river system are simultaneously analysed via light microscopy and by using a next generation sequencing eDNA barcoding approach. The latter includes, besides the usage of BLAST algorithms for taxon identification, also a phylogenetic tree based approach. The taxon lists for the classic morphological approach are compared to the taxon lists derived from eDNA barcoding and it is shown that the latter always results in a finer description of diatom diversity (a total of 103 taxa by light microscopy vs. 270 taxa identified by eDNA barcoding). Therefore this study demonstrates that eDNA barcoding is a competitive approach to describe diatom communities and holds the potential for a large application spectrum. Methodological challenges in eDNA barcoding are analysed and discussed.

Contribution – Lead authorship, sample collection, data collection, morphological and molecular/NGS data analysis, development of NGS data analysis approach.

CONCLUSIONS AND OUTLOOK

The here assembled studies demonstrate the feasibility of environmental DNA (eDNA) barcoding via the 18S V4 marker region for the description of diatom diversity for water quality assessments. Important for the applicability of eDNA barcoding has been the invention of next generation sequencing (NGS) technology. It enables scientists to characterise community compositions via DNA sequences by producing large numbers of sequences from environmental samples (Andersson *et al.* 2009; Eiler *et al.* 2013; Shokralla *et al.* 2012; Sogin *et al.* 2006).

The handling of large sequence datasets from environmental samples already requires insights into informatic procedures. But the implementation and standardisation of bioinformatic pipelines, however, demands a profound expertise in both biology and informatics (Collins & Cruickshank 2013; Smith & Staetsky 2007; Yoccoz 2012), a combination still rarely found. If the bioinformatic pipelines for the analysis of sequence datasets are further standardised and automated, eDNA barcoding as an objective sequence-based method for specimen identification can increase the comparability of datasets from different studies and laboratories. A standardised eDNA barcoding approach will also facilitate monitoring programs with large spatiotemporal dimensions (Hajibabaei *et al.* 2011; Hajibabaei *et al.* 2012). Due to the high number of sequences obtained from an investigated community, eDNA barcoding also allows for deeper sampling compared to microscopic approaches (Stoeck *et al.* 2009). It furthermore facilitates an improved insight into diatom and other protist diversity and could therefore be the basis for a global projection of their biodiversity (Stoeck *et al.* 2009). With further development of high throughput sequencing the future of eDNA barcoding will lie in sequencing multiple markers in parallel (multiplexing) (Collins & Cruickshank 2013; Dupuis *et al.* 2012; Fujita *et al.* 2012) or even whole genomes of at least small organelles (Eiler *et al.* 2013; Taberlet *et al.* 2012a; Taberlet *et al.* 2012b).

Another critical point for the accurateness of diversity descriptions via eDNA barcoding is the establishment of taxonomically well curated reference databases and the introduction of community standards for the deposition of reference data. Curated reference databases will adopt the role for DNA sequence based approaches that printed identification literature has for morphology based taxon assignment. This is due to the fact that reference libraries are algorithmically searchable for sequence similarities. The presently available reference databases for diatoms often include large percentages of sequences with doubtful taxonomic annotation on the species level. Furthermore, these databases currently do not provide standard pathways to trace the taxonomic annotation back to the original voucher specimen. This means that many eDNA barcoding studies are limited to identifications to the genus and family level; and while for some studies (e.g. large scale biodiversity assessments) this level of taxonomic depth is sufficient, for others such as species inventories and monitoring programs it is not. The here presented studies could show that with a combination of well curated databases and tree based delimitation methods sequence identification even to the sub-specific level are generally possible.

SUMMARY

The microscopical analysis of diatom diversity is a well established approach in the assessment of the ecological status of water bodies and implemented e.g. in the European Water Framework Directive. This dissertation presents environmental DNA (eDNA) barcoding of diatom communities as an alternative methodology to the routine use of light microscopy.

The V4 locus of the 18S rRNA gene is introduced in this work as an adequate DNA barcoding marker for diatoms: the locus is flanked by regions that are conserved throughout the diatoms, thus serving as universal primer binding sites. The 18S V4 region also contains sufficient sequence variation for diatom discrimination on a stretch of only approximately 390 bp; therefore it is also suitable for next generation sequencing. A standard laboratory protocol for the amplification of this marker from diatoms in culture as well as from environmental samples is specified in the present dissertation. Furthermore, this work includes the agreement of the CBOL Protist Working Group (ProWG) on the 18S V4 locus as pre-barcode for the assessment of general protist diversity. In addition to that each protist group could apply a second marker region to further examine taxon delimitation and diversity, if necessary.

eDNA barcoding relates unidentified sequences from an environmental sample to identified sequences in a reference database to obtain an assignment of the environmental sequences to a respective taxon. The present study investigates the taxonomic discriminatory power of eDNA barcoding in comparison to light microscopy. It is demonstrated that the DNA based approach provides a more refined taxon detection than the microscopical approach. As the simple BLAST algorithm proved to be insufficient for the assignment of sequences to taxonomic entities, the phylogenetic-based coalescent model approach (PCMA) is introduced. The PCMA combines the general mixed Yule-coalescent (GMYC) model – a tree based approach - with the statistic evaluation of genetic clusters via bootstrap support.

It is also shown that the quality of sequence assignment to taxonomic entities is directly related to the quality of the taxonomic treatment of the reference sequences; and the information in the most commonly used reference database INSDC (International Nucleotide Sequence Database Collaboration; incl. Genbank, EMBL/ENA and DDBJ) is of questionable quality.

This work therefore includes best practice guidelines for the deposition of sequence information in taxonomic reference libraries. In diatoms reference sequences are almost always obtained from clonal cultures permitting good documentation possibilities. Along with the reference sequences, physical vouchers in form of herbarium specimen and the DNA extract need to be

deposited in scientifically curated collections. Also minimum requirements for metadata are suggested: these include e.g. collection data, cultivation data, primer and amplification details, pherograms, as well as photographic documentation of microstructures important for identification. Even though these specifications are proposed for diatoms, they are transferable to other protist groups and also beyond that. As parts of the specified metadata cannot be deposited in the INSDC along with the sequence data, AlgaTerra and the database of the DNA Bank Network are discussed as complements.

The publications assembled in the present dissertation contribute towards the establishment of the standard application of eDNA barcoding in water quality assessments via diatom community analysis by introducing (a) a methodological approach and (b) best practice guidelines for the deposition of reference sequences.

ZUSAMMENFASSUNG

Die lichtmikroskopische Bewertung der Diversität von Diatomeen ist eine erfolgreich praktizierte Methode zur Gewässergüteanalyse, die z.B. auch in der Wasserrahmenrichtlinie der Europäischen Union umgesetzt wird. Die vorliegende Arbeit stellt environmental DNA (eDNA) Barcoding als Alternative zur üblichen Untersuchung von Diatomeengesellschaften mit Hilfe von Lichtmikroskopie vor.

Die auf dem 18S rRNA Gen befindliche V4 Region wird in dieser Dissertation als adäquater Marker für Diatomeen DNA Barcoding eingeführt: die Region wird von innerhalb der Diatomeen konservierten Sequenzabschnitten flankiert, die als universelle Primer-Bindungsstellen fungieren können. Des Weiteren zeigt die 18S V4 Region auf einem relativ kurzen Sequenzabschnitt (ca. 390 bp) ausreichend Variation für die Unterscheidung verschiedener Taxa. Die geringe Abschnittslänge ermöglicht zusätzlich die Sequenzierung durch Next-Generation-Sequencing-Technologie. Die Arbeit beinhaltet ein Standardprotokoll für die Amplifizierung der 18S V4 Region sowohl aus Kulturmaterial als auch aus Umweltproben. Außerdem ist in der vorliegenden Studie die Entscheidung der CBOL Protist Working Group (ProWG) für die 18S V4 Region als Pre-Barcode für alle Gruppen der Protisten enthalten. Der Marker wird also zur generellen Erfassung der Protistendiversität vorgeschlagen. Sollte darüber hinaus ein weiterer genetischer Marker für die Untersuchung von Taxongrenzen und Diversität innerhalb einzelner Gruppen nötig werden, wird dieser für jede Protistengruppe einzeln festgelegt.

Beim eDNA Barcoding werden unidentifizierte Sequenzen aus einer Umweltprobe identifizierten Sequenzen in einer Referenzdatenbank zugeordnet, um die Sequenzen aus der Umweltprobe einem entsprechenden Taxon zuzuweisen. Die hier vorgestellte Untersuchung prüft die Unterschiede von Lichtmikroskopie und eDNA Barcoding hinsichtlich ihrer Möglichkeiten der taxonomischen Unterscheidung. Die sequenzbasierte Methodik lieferte eine feinere taxonomische Aufgliederung als die Lichtmikroskopie. Da sich eine einfache Zuordnung von Sequenzen zu taxonomischen Einheiten mittels des BLAST Algorithmus als nicht ausreichend erwies, stellt diese Arbeit den phylogenetic-based coalescent model Ansatz (PCMA) vor. Dieser kombiniert das general mixed Yule-coalescent (GMYC) Modell, einen auf Baumberechnungen beruhenden Ansatz, mit der statistischen Unterstützung einzelner genetischer Gruppen durch den Bootstrap Support.

Zusätzlich wird der direkte Zusammenhang von der Qualität der Sequenzzuordnung zu taxonomischen Einheiten zur Qualität der taxonomischen Bearbeitung von Referenzsequenz gezeigt. In der gebräuchlichsten Referenzdatenbank INSDC (International Nucleotide Sequence Database Collaboration; inkl. Genbank, EMBL/ENA and DDBJ) ist die vorhandene taxonomische Information zu Diatomeensequenzen fragwürdig.

Deshalb schlägt die vorliegende Arbeit Richtlinien zur Hinterlegung von Referenzsequenzen in taxonomischen Referenzdatenbanken vor.

Referenzsequenzen werden im Falle der Diatomeen meist aus klonalen Kulturen erzeugt. Hierdurch ist eine umfangreiche Dokumentation möglich. Neben der Sequenzinformation sollten Herbarbelege und die physische DNA-Probe in wissenschaftlich betreuten Sammlungen hinterlegt werden. Minimalanforderungen für begleitende Metadaten werden ebenfalls vorgestellt; hierzu gehören neben Sammlungsdaten, Daten zur Kultivierung, Information zu Primern und PCR sowie den Pherogrammen auch die fotografische Dokumentation zur Bestimmung relevanter Mikrostrukturen. Diese Richtlinien für Diatomeen können ohne weiteres in anderen Protistengruppen oder auch darüber hinaus angewandt werden. Da ein Teil der aufgeführten Metadaten in INSDC nicht hinterlegt werden können, werden AlgaTerra und die Datenbank des DNA-Bank-Netzwerks ergänzend vorgeschlagen.

Die in der vorliegenden Dissertation zusammengefassten Arbeiten tragen zur Etablierung von eDNA Barcoding als Standardanwendung zur Gewässergüteanalyse mittels der Untersuchung von Diatomeengesellschaften bei. Sie stellen erstens einen methodischen Ansatz vor und zweitens formulieren sie Richtlinien zur guten wissenschaftlichen Praxis hinsichtlich der Hinterlegung von Referenzsequenzen.

References

Adler RW (1995) Filling the gaps in water quality standards: Legal perspectives on biocriteria. In Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making, ed. W. S. Davis & T. P. Simon, pp. 31-47. Boca Raton, Florida: Lewis Publishers.

Aletsee L, Jahnke J (1992) Growth and productivity of the psychrophilic marine diatoms *Thalassiosira antarctica* (Comber) and *Nitzschia frigida* (Grunow) in batch cultures at temperatures below the freezing point of sea water. *Polar Biology* **11**, 643-647.

Allen AE, Dupont CL, Obornik M*, et al.* (2011) Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. *Nature* **473**, 203-207.

Alverson AJ, Theriot EC (2005) Comments on recent progress toward reconstructing the diatom phylogeny. *Journal of Nanoscience and Nanotechnology* **5**, 57-62.

Amato A, Kooistra WHCF, Levialdi Ghiron JH*, et al.* (2007) Reproductive Isolation among Sympatric Cryptic Species in Marine Diatoms. *Protist* **158**, 193-207.

Andersen RA (2004) Biology and systematics of heterokont and haptophyte algae. *Am J Bot* **91**, 1508-1522.

Andersson AF, Riemann L, Bertilsson S (2009) Pyrosequencing reveals contrasting seasonal dynamics of taxa within Baltic Sea bacterioplankton communities. *The ISME Journal* **4**, 171-181.

Anonymous (1703) Two letters from a gentleman in the country, relating to Mr. Leuwenhoeck’s letter in Transaction, No. 283. *Philosophical Transactions of the Royal Society of London* **23**, 1494-1498.

Archibald REM (1984) Diatom illustrations—an appeal. *Bacillaria 7*, 173-178.

Armbrust EV (2009) The life of diatoms in the world's oceans. *Nature* **459**, 185-192.

Bailey-Watts AE (1976) Planktonic diatoms and some diatom-silica relations in a shallow eutrophic Scottish loch. *Freshwater Biology* **6**, 69-80.

Barbour MT, Stribling JB, Karr JR (1995) Multimetric approach for establishing biocriteria and measuring biological condition. In Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making, ed. W. S. Davis & T. P. Simon, pp. 63-77· Boca Raton, Florida: Lewis publishers.

Barth D, Krenek S, Fokin SI, Berendonk TU (2006) Intraspecific Genetic Variation in *Paramecium* Revealed by Mitochondrial *Cytochrome c Oxidase* I Sequences. *Journal of Eukaryotic Microbiology* **53**, 20-25.

Becker S, Hanner R, Steinke D (2011) Five years of FISH-BOL: Brief status report. *Mitochondrial DNA* **22**, 3-9.

Behnke A, Engel M, Christen R*, et al.* (2011) Depicting more accurate pictures of protistan community complexity using pyrosequencing of hypervariable SSU rRNA gene regions. *Environmental Microbiology* **13**, 340-349.

Behnke A, Friedl T, Chepurnov VA, Mann DG (2004) Reproductive compatibility and rDNA sequence analyses in the *Sellaphora pupula* species complex (Bacillariophyta). *Journal of Phycology* **40**, 193-208.

Blaxter M, Floyd R (2003) Molecular taxonomics for biodiversity surveys: already a reality. *Trends in Ecology & Evolution* **18**, 268-269.

Bortolus A (2008) Error Cascades in the Biological Sciences: The Unwanted Consequences of Using Bad Taxonomy in Ecology. *AMBIO: A Journal of the Human Environment* **37**, 114-118.
Bory JBM (1822) Dictionnaire classique d'Histoire Naturelle, 1. Paris.
Brown JH, Lomolino MV (1998) Biogeography. 2nd ed. Sunderland (MA): Sinauer.
Burki F, Flegontov P, Obornik M*, et al.* (2012) Re-evaluating the green versus red signal in eukaryotes with secondary plastid of red algal origin. *Genome Biol Evol* **4**, 626-635.
Burki F, Shalchian-Tabrizi K, Minge M*, et al.* (2007) Phylogenomics Reshuffles the Eukaryotic Supergroups. *PLoS ONE* **2**, e790.
Butcher RW (1947) Studies in the ecology of rivers. IV. The algae of organically enriched water. *Journal of Ecology* **35**, 186-191.
Casteleyn G, Chepurnov VA, Leliaert F*, et al.* (2008) *Pseudo-nitzschia pungens* (Bacillariophyceae): A cosmopolitan diatom species? *Harmful Algae* **7**, 241-257.
Chantangsi C, Lynn DH, Brandl MT*, et al.* (2007) Barcoding ciliates: a comprehensive study of 75 isolates of the genus *Tetrahymena*. *International Journal of Systematic and Evolutionary Microbiology* **57**, 2412-2423.
Collins RA, Cruickshank RH (2013) The seven deadly sins of DNA barcoding. *Molecular Ecology Resources* **13**, 969-975.
Cox E (1991) What is the basis for using diatoms as monitors of river quality? in: B .A. Whitton, E. Rott & G.Friedrich (eds), Use of algae for monitoring rivers. E. Rott, Innsbruck, Austria: 33-40.
Coyne JA, Orr H (2004) Speciation. Sinauer, Sunderland, Massuchussets, USA. 545 pp.
Daniel GF, Chamberlain AHL, Jones EBG (1987) Cytochemical and electron microscopical observations on the adhesive materials of marine fouling diatoms. *British Phycological Journal* **22**, 101-118.
Dasmahapatra KK, Elias M, Hill RI, Hoffman JI, Mallet J (2010) Mitochondrial DNA barcoding detects some species that are real, and some that are not. *Molecular Ecology Resources* **10**, 264-273.
Dasmahapatra KK, Mallet J (2006) Taxonomy: DNA barcodes: recent successes and future prospects. *Heredity* **97**, 254-255.
Davies SJ, Metcalfe SE, Caballero ME, Juggins S (2002) Developing diatom-based transfer functions for Central Mexican lakes. *Hydrobiologia* **467**, 199-213.
Dell'Uomo A (1999) Use of algae for monitoring rivers in Italy: current situation and perspectives. in: Use of Algae for Monitoring Rivers III - Douai, France. l'Agence de l'Eau Artois-Picardie.
DeSalle R, Egan MG, Siddall M (2005) The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360**, 1905-1916.
DeSalle ROB (2006) Species Discovery versus Species Identification in DNA Barcoding Efforts: Response to Rubinoff. *Conservation Biology* **20**, 1545-1547.
Descy J, Coste M (1991) A test of methods for assessing water quality based on diatoms. *Negotiations of the International Society of Limnology* **24**, 2112-2116.

Descy JP, Mouvet C (1984) Impact of the Tihange nuclear power plant on the periphyton and the phytoplankton of the Meuse River (Belgium). *Hydrobiologia* **119**, 119-128.
Dorrell RG, Smith AG (2011) Do Red and Green Make Brown?: Perspectives on Plastid Acquisitions within Chromalveolates. *Eukaryotic Cell* **10**, 856-868.
Dupuis JR, Roe AD, Sperling FAH (2012) Multi-locus species delimitation in closely related animals and fungi: one marker is not enough. *Molecular Ecology* **21**, 4422-4436.
Ehrenberg CG (1838) Die Infusionsthierchen als vollkommene Organismen. Leipzig: Leopold Voss. i-xviii +548 pp.
Eiler A, Drakare S, Bertilsson S*, et al.* (2013) Unveiling Distribution Patterns of Freshwater Phytoplankton by a Next Generation Sequencing Based Approach. *PLoS ONE* **8**, e53516.
Enke N, Thessen A, Bach K*, et al.* (2012) The user's view on biodiversity data sharing — Investigating facts of acceptance and requirements to realize a sustainable use of research data —. *Ecological Informatics* **11**, 25-33.
European Committee for Standardization (2003) European Standard. EN 14407. Water Quality – Guidance Standard for the Identification, Enumeration and Interpretation of Benthic Diatom Samples from Running Waters. *CEN*, 14 pp.
European Committee for Standardization (2004) European Standard. EN 14407. Water Quality – Guidance Standard for the Identification, Enumeration and Interpretation of Benthic Diatom Samples from Running Waters. *CEN* **Brussels**, 12 pp.
European Parliament (2000) Directive 2000/60/EC of the European Parliament and of the Council. *Official Journal of the European Communities* **327**.
Falkowski PG, Barber RT, Smetacek V (1998) Biogeochemical Controls and Feedbacks on Ocean Primary Production. *Science* **281**, 200-206.
Field CB, Behrenfeld MJ, Randerson JT, Falkowski P (1998) Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science* **281**, 237-240.
Fontaneto D, Herniou EA, Boschetti C*, et al.* (2007) Independently Evolving Species in Asexual Bdelloid Rotifers. *PLoS Biol* **5**, e87.
Ford BJ (1991) The Leeuwenhoek legacy. Biopress and Farrand Press, 185 pp.
Forney LJ, Zhou X, Brown CJ (2004) Molecular microbial ecology: land of the one-eyed king. *Current Opinion in Microbiology* **7**, 210-220.
Fujita MK, Leaché AD, Burbrink FT, McGuire JA, Moritz C (2012) Coalescent-based species delimitation in an integrative taxonomy. *Trends in Ecology & Evolution* **27**, 480-488.
Gersonde R, Harwood DM (1990) Lower Cretaceous diatoms from ODP Leg 113 site 693 (Weddell Sea) Part 1: vegetative cells. *Proceedings of the Ocean Drilling Program, Scientific Results* **113**, 365-402.
Gmelin JF (1791) In Systema Naturae Edition 13 1(6). Lipsiae.
Godfray HCJ, Lewis OT, Memmott J (1999) Studying insect diversity in the tropics. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **354**, 1811-1824.

Goodwin TW (1974) Carotenoids and biliproteins. In: Algal Physiology and Biochemistry. Botanical Monographs No. 10 (ed. Stuart, W. D. P.). Oxford, UK, Blackwells. 176-205.

Guiry MD (2013+) "Bacillariophyceae". http://www.algaebase.org.

Hajibabaei M, deWaard JR, Ivanova NV*, et al.* (2005) Critical factors for assembling a high volume of DNA barcodes. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360**, 1959-1967.

Hajibabaei M, Shokralla S, Zhou X, Singer G, Baird D (2011) Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS ONE* **6**, e17497.

Hajibabaei M, Spall J, Shokralla S, van Konynenburg S (2012) Assessing biodiversity of a freshwater benthic macroinvertebrate community through non-destructive environmental barcoding of DNA from preservative ethanol. *BMC Ecology* **12**, 1-10.

Hamsher SE, Evans KM, Mann DG, Poulíčková A, Saunders GW (2011) Barcoding Diatoms: Exploring Alternatives to COI-5P. *Protist* **162**, 405-422.

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **270**, 313-321.

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM (2004) Identification of Birds through DNA Barcodes. *PLoS Biol* **2**, e312.

Hudson RR, Coyne JA (2002) Mathematical consequences of the genealogical species concept. *Evolution* **56**, 1557-1565.

Hutchinson GE (1968) When are species necessary? In Population biology and evolution (R.C. Lewontin, ed.), 177-186. Syracuse University Press, New York.

Jahn R (1984) The structure of the frustule and its variability in clones and populations. *Proceedings of the 8th International Diatom Symposium 1984* **Proceedings of the 8th International Diatom Symposium 1984**, 191-204.

Kaczmarska I, Reid C, Martin JL, Moniz MBJ (2008) Morphological, biological, and molecular characteristics of the diatom *Pseudo-nitzschia delicatissima* from the Canadian MaritimesThis paper is one of a selection of papers published in the Special Issue on Systematics Research. *Botany* **86**, 763-772.

Kelly MG (1998) Use of the trophic diatom index to monitor eutrophication in rivers. *Water Research* **32**, 236-242.

Kelly MG, Cazaubon A, Coring E*, et al.* (1998) Recommendations for the routine sampling of diatoms for water quality assessments in Europe. *Journal of Applied Phycology* **10**, 215-224.

Kermarrec L, Franc A, Rimet F*, et al.* (2014) A Next-Generation Sequencing Approach to River Biomonitoring Using Benthic Diatoms. *Freshwater Science* **33**, 349-363.

Kolkwitz R, Marsson M (1908) Ökologie der pflanzlichen Saprobien. *Berichte der Deutschen Botanischen Gesellschaft* **26**, 505-519.

Kooistra WHCF, Medlin LK (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**, 391-407.

Kützing FT (1844) Die kieselschaligen Bacillarien oder Diatomeen, Nordhausen. 152 pp.

Kvist S (2013) Barcoding in the dark?: A critical view of the sufficiency of zoological DNA barcoding databases and a plea for broader integration of taxonomic knowledge. *Molecular Phylogenetics and Evolution* **69**, 39-45.
Lange-Bertalot H (1979a) Pollution tolerance of diatoms as a criterion for water quality estimation. *Nova Hedwigia* **64**, 285-304.
Lange-Bertalot H (1979b) Toleranzgrenzen und Populationsdynamik benthischer Diatomeen bei unterschiedlich starker Abwasserbelastung. *Algological Studies/Archiv für Hydrobiologie, Supplement Volumes* **23**, 184-219.
Lebeau TL, Robert JMR (2003) Diatom cultivation and biotechnologically relevant products. Part II: Current and putative products. *Applied Microbiology and Biotechnology* **60**, 624-632.
Lowe RL, Golladay SW, Webster JR (1986) Periphyton response to nutrient manipulation in streams draining clearcut and forested watersheds. *Journal of the North American Benthological Society* **5**.
Luddington IA, 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**, e45664.
Mann D, Sato S, Trobajo R, Vanormelingen P, Souffreau C (2010) DNA barcoding for species identification and discovery in diatoms. *Cryptogamie. Algologie* **31**, 557-577.
Mann DG (1999) The species concept in diatoms. *Phycologia* **38**, 437-495.
Mann DG, Droop SJM (1996) 3. Biodiversity, biogeography and conservation of diatoms. *Hydrobiologia* **336**, 19-32.
Mann DG, Vanormelingen P (2013) An Inordinate Fondness? The Number, Distributions, and Origins of Diatom Species. *Journal of Eukaryotic Microbiology* **60**, 414-420.
Mayr E (1942) Systematics and the origin of species. New York: Columbia University Press. 334 pp.
McCormick PV, Stevenson RJ (1989) Effects of snail grazing on benthic algal community structure in different nutrient environments. *Journal of the North American Benthological Society* **82**, 162-172.
Medlin LK, Elwood HJ, Stickel S, Sogin ML (1991) Morphological and genetic variation within the diatom *Skeletonema costatum* (Bacillariophyta): evidence for a new species *Skeletonema pseudocostatum*. *Journal of Phycology* **27**, 514-524.
Medlin LK, Kaczmarska I (2004) Evolution of the diatoms: V. Morphological and cytological support for the major clades and a taxonomic revision. *Phycologia* **43**, 245-270.
Medlin LK, Kooistra WH, Gersonde R, Wellbrock U (1997) Is the origin of diatoms related to the end-Permian mass extinction? *Nova Hedwigia* **65**, 1-11.
Meyer CP, Paulay G (2005) DNA Barcoding: Error Rates Based on Comprehensive Sampling. *PLoS Biol* **3**, e422.
Monaghan MT, Wild R, Elliot M*, et al.* (2009) Accelerated Species Inventory on Madagascar Using Coalescent-Based Models of Species Delineation. *Systematic Biology* **58**, 298-311.
Moniz MBJ, Kaczmarska I (2009) Barcoding diatoms: Is there a good marker? *Molecular Ecology Resources* **9**, 65-74.

Moniz MBJ, Kaczmarska I (2010) Barcoding of Diatoms: Nuclear Encoded ITS Revisited. *Protist* **161**, 7-34.
Morales EA, Siver PA, Trainor FR (2001) Identification of diatoms (Bacillariophyceae) during ecological assessments: Comparison between Light Microscopy and Scanning Electron Microscopy techniques. *Proceedings of the Academy of Natural Sciences of Philadelphia* **151**, 95-103.
Moritz C, Cicero C (2004) DNA Barcoding: Promise and Pitfalls. *PLoS Biol* **2**, e354.
Moustafa A, Beszteri B, Maier UG*, et al.* (2009) Genomic Footprints of a Cryptic Plastid Endosymbiosis in Diatoms. *Science* **324**, 1724-1726.
Müller OF (1783) Strand Parlebandet och Armbandet, tvanne microscopiska Strandvaxter. Kongl. Vetenskaps Aeademiens, Nya Handlingar, 4: 80-5 and Tab. III.
Müller OF (1786) Animalcula infusoria fluviatilia et marina. Lvi + 367 pp. Havniae: N. Moller.
Norris RH, Morris KR (1995) The need for biological assessment of water quality: Australian perspective. *Australian Journal of Ecology* **20**, 1-6.
Norton TA, Melkonian M, Andersen RA (1996) Algal biodiversity. *Phycologia* **35**, 308-326.
Novotny V, Drozd P, Miller SE*, et al.* (2006) Why Are There So Many Species of Herbivorous Insects in Tropical Rainforests? *Science* **313**, 1115-1118.
Packer L, Gibbs J, Sheffield C, Hanner R (2009) DNA barcoding and the mediocrity of morphology. *Molecular Ecology Resources* **9 Suppl s1**, 42-50.
Pan Y, Stevenson RJ, Hill BH, Herlihy AT, Collins GB (1996) Using diatoms as indicators of ecological conditions in lotic systems: A regional assessment. *Joumal of the North American Benthological Society* **15**, 481-495.
Patrick R (1949) A proposed biological measure of stream conditions based on a survey of the Conestoga Basin, Lancaster County, Pennsylvania. *Proceedings of the Academy of Natural Sciences of Philadelphia* **101**, 277-341.
Patrick R (1961) A study of the numbers and kinds of species found in rivers of the Eastern United States. *Proceedings of the Academy of Natural Sciences of Philadelphia* **113**, 215-258.
Patrick R, Strawbridge D (1963) Variation in the structure of natural diatom communities. *The American Naturalist* **97**, 51-57.
Patrick R, Wallace JH, Hohn MH (1954) A new method for determining the pattern of the diatom flora. *Notulae Naturae* **259**, 2-12.
Pawlowski J, Audic S, Adl S*, et al.* (2012) CBOL Protist Working Group: Barcoding Eukaryotic Richness beyond the Animal, Plant, and Fungal Kingdoms. *PLoS Biol* **10**, e1001419.
Pawlowski J, Christen R, Lecroq B*, et al.* (2011) Eukaryotic Richness in the Abyss: Insights from Pyrotag Sequencing. *PLoS ONE* **6**, e18169.
Pons J, Barraclough TG, Gomez-Zurita J*, et al.* (2006) Sequence-Based Species Delimitation for the DNA Taxonomy of Undescribed Insects. *Systematic Biology* **55**, 595-609.
Poulíčková A, Špačková J, Kelly M, Duchoslav M, Mann D (2008) Ecological variation within *Sellaphora* species complexes (Bacillariophyceae): specialists or generalists? *Hydrobiologia* **614**, 373-386.

Pringle CM, Bowers JA (1984) An in situ substratum fertilization technique: diatom colonization on nutrient-enriched, sand substrata. *Canadian Journal of Fisheries and Aquatic Sciences* **41**, 1247-1251.
Prygiel J, Carpentier P, Almeida S*, et al.* (2002) Determination of the biological diatom index (IBD NF T 90–354): results of an intercomparison exercise. *Journal of Applied Phycology* **14**, 27-39.
Rach J, DeSalle R, Sarkar IN, Schierwater B, Hadrys H (2008) Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata. *Proceedings of the Royal Society B: Biological Sciences* **275**, 237-247.
Ratnasingham S, Hebert PDN (2007) BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes* **7**, 355-364.
Raven JA, Waite AM (2004) The evolution of silicification in diatoms: inescapable sinking and sinking as escape? *New Phytologist* **162**, 45-61.
Richards C, Johnson LB, Host GE (1996) Landscape-scale influences on stream habitats and biota. *Canadian Journal of Fisheries and Aquatic Sciences* **53**, 295-311.
Rott E (1991) Methodological aspects and perspectives in the use of periphyton for monitoring and protecting rivers. In Use of Algae for Monitoring Rivers, ed. RA. Whitton, E. Rott & G. Friedrich, pp. 9-16. Institut für Botanik, Universitat Innsbruck, Innsbruck, Austria: E. Rott, Publisher.
Round FE (1991) Diatoms in river water-monitoring studies. *Journal of Applied Phycology* **3**, 129-145.
Round FE, Crawford RM, Mann DG (1990) The Diatoms: Biology and Morphology of the Genera. Cambridge University Press.
Rubinoff D, Cameron S, Will K (2006) A Genomic Perspective on the Shortcomings of Mitochondrial DNA for "Barcoding" Identification. *Journal of Heredity* **97**, 581-594.
Schieber J, Krinsley D, Riciputi L (2000) Diagenetic origin of quartz silt in mudstones and implications for silica cycling. *Nature* **406**, 981-985.
Schindel DE, Miller SE (2005) DNA barcoding a useful tool for taxonomists. *Nature* **435**, 17-17.
Scicluna SM, Tawari B, Clark CG (2006) DNA Barcoding of *Blastocystis*. *Protist* **157**, 77-85.
Shokralla S, Spall J, Gibson J, Hajibabaei M (2012) Next-generation sequencing technologies for environmental DNA research. *Mol Ecol* **21**, 1794 - 1805.
Sims PA, Mann DG, Medlin LK (2006) Evolution of the diatoms: insights from fossil, biological and molecular data. *Phycologia* **45**, 361-402.
Smetacek V (1999) Diatoms and the Ocean Carbon Cycle. *Protist* **150**, 25-32.
Smith TMF, Staetsky L (2007) The teaching of statistics in UK universities. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **170**, 581-622.
Sogin ML, Morrison HG, Huber JA*, et al.* (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences* **103**, 12115-12120.
Spaulding SA, McKnight DM (1999) Assessing ecological conditions in rivers and streams with diatoms. In: *The diatoms: Applications to the environmental and earth sciences* eds. Stoermer EP, Smol JP), pp. 245-260, Cambridge.

Squires L, Rushforth S, Brotherson J (1979) Algal response to a thermal effluent: study of a power station on the provo river, Utah, USA. *Hydrobiologia* **63**, 17-32.

Stauber JL, Jeffrey SW (1988) Photosynthetic pigments in fifty-one species of marine diatoms. *Journal of Phycology* **24**, 158-172.

Steinman AD, McIntire CD, Gregory SV, Lamberti GV, Ashkenas L (1987) Effect of herbivore type and density on taxonomic structure and physiognomy of algal assemblages in laboratory streams. *Canadian Journal of Fisheries and Aquatic Sciences* **44**, 1640-1648.

Stevenson RJ, Pan Y (1999) Assessing ecological conditions in rivers and streams with diatoms. In: *The diatoms: Applications to the environmental and earth sciences* eds. Stoermer EP, Smol JP), pp. 11-40. Cambridge University Press, Cambridge.

Stoeck T, Behnke A, Christen R*, et al.* (2009) Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities. *BMC Biology* **7**, 72.

Stoeckle M (2003) Taxonomy, DNA and the barcode of life. *Bioscience* **53**, 2-3.

Stoermer EF, Smol JP (1999) The diatoms: Applications for the environmental and earth sciences. - Cambridge, Cambridge University Press.

Stoof-Leichsenring KR, Epp LS, Trauth MH, Tiedemann R (2012) Hidden diversity in diatoms of Kenyan Lake Naivasha: a genetic approach detects temporal variation. *Molecular Ecology* **21**, 1918-1930.

Taberlet P, Coissac E, Hajibabaei M, Rieseberg L (2012a) Environmental DNA. *Mol Ecol* **21**, 1789 - 1793.

Taberlet P, Prud'homme S, Campione E*, et al.* (2012b) Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies. *Mol Ecol* **21**, 1816 - 1820.

Tautz D, Arctander P, Minelli A, Thomas RH, Vogler AP (2003) A plea for DNA taxonomy. *Trends in Ecology & Evolution* **18**, 70-74.

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

Van Baalen C, O'Donnell R (1983) Isolation and Growth of Psychrophilic Diatoms from the Ice-edge in the Bering Sea. *Journal of General Microbiology* **129**, 1019-1023.

Van de Vijver B, Beyens L, Lange-Bertaltot H (2004) The genus *Stauroneis* in the Arctic and (Sub-)Antarctic Regions. Bibliotheca Diatomologica, 51, p. 317.

Van de Vijver B, Jarlman A, Haan MD, Compere P (2013) The Ultrastructure of *Gomphonema augur* and *Gomphonema gautieri* (Bacillariophyta). *Cryptogamie, Algologie* **34**, 103-116.

Vanelslander B, Créach V, Vanormelingen P*, et al.* (2009) Ecological differantiation between sympatric pseudocryptic species in the estuarine benthic diatom *Navicula phyllepta* (Bacillariophyceae). *Journal of Phycology* **45**, 1278-1289.

Vanormelingen P, Chepurnov VA, Mann DG, Cousin S, Vyverman W (2007) Congruence of morphological, reproductive and ITS rRDNA sequence data in some Australasian *Eunotia bilunaris* (Bacillariophyta). *European Journal of Phycology* **42**, 61-79.

Wilson EO (2003) The encyclopedia of life. *Trends in Ecology & Evolution* **18**, 77-80.

Woehle C, Dagan T, Martin WF, Gould SB (2011) Red and Problematic Green Phylogenetic Signals among Thousands of Nuclear Genes from the Photosynthetic and Apicomplexa-Related Chromera velia. *Genome Biol Evol* **3**, 1220-1230.

Yoccoz NG (2012) The future of environmental DNA in ecology. *Molecular Ecology* **21**, 2031-2038.

Yool A, Tyrrell T (2003) Role of diatoms in regulating the ocean's silicon cycle. *Global Biogeochemical Cycles* **17**, 1103.

Zalack JT, Smucker NJ, Vis ML (2010) Development of a diatom index of biotic integrity for acid mine drainage impacted streams. *Ecological Indicators* **10**, 287-295.

Zimmermann J, Abarca N, Enk N*, et al.* (2014a) Taxonomic Reference Libraries for Environmental Barcoding: A Best Practice Example from Diatom Research. *PLoS ONE* **9**, e108793.

Zimmermann J, Glöckner G, Jahn R, Enke N, Gemeinholzer B (2014b) Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies. *Molecular Ecology Resources*, n/a-n/a.

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 & Evolution* **11**, 173-192.

Zinger L, Gobet A, Pommier T (2012) Two decades of describing the unseen majority of aquatic microbial diversity. *Molecular Ecology* **21**, 1878-1896.

Part II

Publications

BARCODING DIATOMS: EVALUATION OF THE V4 SUBREGION ON THE 18S rRNA GENE, INCLUDING NEW PRIMERS AND PROTOCOLS

JONAS ZIMMERMANN

REGINE JAHN

BIRGIT GEMEINHOLZER

2011

Organism Diversity & Evolution 11 (4): 173-192

Org Divers Evol
DOI 10.1007/s13127-011-0050-6

ORIGINAL ARTICLE

ORGANISMS DIVERSITY & EVOLUTION

Barcoding diatoms: evaluation of the V4 subregion on the 18S rRNA gene, including new primers and protocols

Jonas Zimmermann · Regine Jahn · Birgit Gemeinholzer

Received: 5 July 2010 / Accepted: 15 July 2011

Abstract Diatoms are present in all types of water bodies and their species diversity is influenced greatly by environmental conditions. This means that diatom occurrence and abundances are suitable indicators of water quality. Furthermore, continuous screening of algal biodiversity can provide information about diversity changes in ecosystems. Thus, diatoms represent a desirable group for which to develop an easy to use, quick, efficient, and standardised organism identification tool to serve routine water quality assessments. Because conventional morphological identification of diatoms demands specialised in-depth knowledge, we have established standard laboratory procedures for DNA barcoding in diatoms. We (1) identified a short segment (about 400 bp) of the SSU (18S) rRNA gene which is applicable for the identification of diatom taxa, and (2) elaborated a routine protocol including standard primers for this group of microalgae. To test the universality of the primer binding sites and the discriminatory power of the proposed barcode region, 123 taxa, representing limnic diatom diversity, were included in the study and identified at species level. The effectiveness of the barcode was also scrutinised within a closely related species group, namely the *Sellaphora pupula* taxon complex and relatives.

J. Zimmermann (✉) · R. Jahn
Botanic Garden and Botanical Museum Berlin-Dahlem,
Freie Universität Berlin,
Königin-Luise-Str. 6-8,
14195 Berlin, Germany
e-mail: j.zimmermann@bgbm.org

J. Zimmermann · B. Gemeinholzer
Justus Liebig University,
Heinrich-Buff-Ring 38,
35392 Giessen, Germany

Keywords 18S (SSU) rRNA gene · Bacillariophyta · DNA barcoding · Diatoms · Standard laboratory procedure

Introduction

Diatoms are unicellular photoautotrophic eukaryotes which are responsible for at least 25% of the global carbon dioxide fixation (Falkowski et al. 1998; Field et al. 1998; Mann 1999; Smetacek 1999). They are an important part of benthic and planktonic biocoenoses and occur nearly ubiquitously in limnic, marine, and terrestrial ecosystems as well as in aerosols (Jahn et al. 2007). Therefore, diatoms are often used as biodindicators in water monitoring assessments and ecological studies (Stevenson and Pan 1999; Stoermer and Smol 1999). Even closely related taxa (excluding cryptic species) are often indicative of different ecological conditions (Poulíčková et al. 2008; Vanelslander et al. 2009). Hence, unambiguous identification of organisms down to species level is crucial for the quality of these studies. Archibald (1984) and Morales et al. (2001) have pointed out that many ecological and monitoring studies are misleading, because identifications have not been verified by experienced diatom taxonomists. To identify diatoms morphologically beyond the genus level is difficult and requires expert knowledge, especially because frustule morphology can vary considerably even within a population (Babanazarova et al. 1996; Bailey-Watts 1976; Jahn 1986; Medlin et al. 1991).

In cases of groups with poor morphological resolution, Hebert et al. (2003) promoted the concept of a DNA barcode to help with the identification of taxa. A DNA barcode is an instrument for the correlation of a taxonomically undetermined individual to a taxon with similar genetic sequence in a given reference database (Ratnasingham and Hebert 2007).

Published online: 26 July 2011

However, a suitable barcode marker has to meet three requirements. The ideal barcode marker (1) consists of a short sequence that can be easily amplified and sequenced in one read following a standardised laboratory protocol, (2) is flanked by a conserved region in which universal primers can be placed, and (3) has the power to resolve organisms at species level (e.g. Hebert et al. 2003; Moritz and Cicero 2004; Stoeckle 2003). Therefore, as in any environmental sampling approach, the quality of the method is not only related to the extent and quality of the reference database but also to the number of taxa that can be identified unambiguously (Erickson et al. 2008), and to the rate at which taxa are retrieved from environmental samples.

Applying the DNA barcoding concept to diatoms promises great potential to resolve the problem of inaccurate species identification and thus facilitate analyses of the biodiversity of environmental samples. In particular, the use of DNA barcodes in diatoms can serve various purposes, such as (1) DNA-based species characterisation and (2) surveying the genetic diversity in an environment of interest. Each of these goals implies different requirements with respect to sequence characteristics. Whereas species characterisation needs sequences with high discriminatory power for defining and identifying even cryptic species, it is not necessarily dependent on fast and universal laboratory protocols. A survey of genetic diversity in environmental samples, however, often relies on high-throughput techniques and therefore needs universal primers and standard protocols where taxa do not have be resolved on the finest scale (e.g. subspecies, cryptic species) (e.g. Hamsher et al. 2011).

Various gene regions have been proposed as barcode markers for diatoms. The mitochondrial cytochrome oxidase I gene (*cox*1) has been widely used for barcoding animals and other organism groups (e.g. Blaxter 2004; Blaxter et al. 2004; Hajibabaei et al. 2006a; Hebert et al. 2004; Robba et al 2006; Saunders 2005; Seifert et al. 2007; Ward et al. 2005). Evans et al. (2007, 2008) successfully tested *cox*1 as a barcoding marker in 22 *Sellaphora* species and three other raphid genera of diatoms. Their study also included a test of the chloroplast ribulose-1,5-bisphosphate carboxylase oxygenase gene (*rbcL*), which was less variable than *cox*1 within the species sampling. However, in other organism groups such as red algae (e.g. Robba et al. 2006; Saunders 2005, 2008), brown algae (Kucera and Saunders 2008) and some green algae (e.g. Lewis and Flechtner 2004; McManus and Lewis 2005), the *rbcL* gene proved to be a promising barcode marker. Moniz and Kaczmarska (2009, 2010) proposed a combination of the nuclear 5.8S rRNA gene and ITS2 upon screening the most species-rich classes of diatoms including mainly marine taxa of the Mediophyceae and Bacillariophyceae. Furthermore, binary characteristics, such as presence/absence of compensatory base changes (CBCs) in the secondary structure of ITS2 or the presence/absence of certain indels have been used to resolve species level diversity in all kind of organisms, including diatoms (Müller et al. 2007). This, however, includes the additional procedural step of calculating and analysing the secondary structure and, therefore, is too laborious for standard high-throughput analyses of environmental samples.

In existing sequence databases, the most extensive data record available for diatoms concerns the nuclear small ribosomal subunit (SSU-rRNA gene), as the latter has been used widely for phylogenetic and taxonomic purposes (e.g. Behnke et al. 2004; Beszteri et al. 2001; Friedl and O'Kelly 2002; Kooistra and Medlin 1996; Medlin et al. 1996; Medlin and Kaczmarska 2004; Sarno et al. 2005; Sorhannus 2007). This means that a substantial reference volume is already available (Hajibabaei et al. 2007), even though identification quality often is not verifiable and therefore does not meet DNA barcoding requirements. The 18S rRNA gene has been suggested as a potential barcoding marker for various organism groups, e.g. nematodes, tardigrades, and diatoms (Bhadury et al. 2006; Blaxter 2004; Blaxter et al. 2004; Floyd et al. 2002; Jahn et al. 2007; Powers 2004). The 18S region has been tested for diatoms in a pilot study by Jahn et al. (2007) and has been used as a marker in other protist groups (Scicluna et al. 2006; Utz and Eizirik 2007).

The present study proposes a 390–410 bp long fragment of the 1800 bp long 18S rRNA gene locus as a barcode marker for the analysis of environmental samples with high-throughput technologies such as 454 sequencing or microarrays, and discusses its use and limitations for diatom identification. The partial 18S region includes a section that is termed V4 in the nomenclature of Nelles et al. (1984) and represents the largest and most complex of the highly variable regions within the 18S locus (Nickrent and Sargent 1991).

Using newly designed universal primers for the V4 region that are introduced below, the region is identified as the most applicable one for barcoding on the 18S locus. Furthermore, an optimised standard laboratory protocol (including DNA extraction, PCR amplification and sequencing) is provided which was developed using diatoms from various limnic genera across many families to represent the freshwater diatom diversity. The study includes taxa from the three major divisions of diatoms: Coscinodiscophyceae (e.g. *Aulacoseira* spp.), Mediophyceae (e.g. *Cyclotella* spp., *Stephanodiscus* spp.) and Bacillariophyceae, with both raphid (e.g. *Nitzschia* spp.) and araphid representatives (e.g. *Fragilaria* spp.) (Table 1).

Table 1 List of all taxa, including strains, EMBL accession numbers (18 s rRNA), voucher identification codes in the Herbarium Berolinense (B), and sampling localities

Taxon	Strain	EMBL no.	B voucher	Locality
Achnanthidium cf. *longipes* Kisselev	–	AY485500	–	–
Achnanthidium minutissimum (Kützing) Czarnecki	D05_008	FR873231	B40 0040686	Germany, Iffeldorf
	–	AJ866992	–	–
	–	AY485500	–	–
Aulacoseira alpigena (Grunow) Krammer	–	AY569578	–	–
Aulacoseira ambigua (Grunow) Simonsen	–	AY569579	–	–
	–	AY569580	–	–
	–	AY569581	–	–
	–	AY569582	–	–
	–	AY569583	–	–
	–	X85404	–	–
Aulacoseira granulata (Ehrenberg) Simonsen	Aula2	FR873232	–	Germany, Berlin
	–	AY569584	–	–
	–	AY569585	–	–
Aulacoseira crenulata (Ehrenberg) Thwaites	–	AY569586	–	–
Aulacoseira granulata var. angustissima (O.Müller) Simonsen	–	AY485493	–	–
Aulacoseira nyassensis (O.Müller) Simonsen	–	AJ535187	–	–
	–	AY121819	–	–
Aulacoseira skvortzowii Edlund, Stoermer & C.M.Taylor	–	AJ535184	–	–
Aulacoseira subarctica (O.Müller) E.Y.Haworth	–	AY569573	–	–
	–	AY569574	–	–
	–	AY569576	–	–
	–	AY569577	–	–
Aulacoseira valida (Grunow) Krammer	–	AY569586	–	–
Cocconeis cf. molesta Kützing	–	AJ535148	–	–
Cocconeis pediculus Ehrenberg	D36_020[a]	FR873233	B40 0040644	Germany, Berlin
	Coco1	FR873234	B40 0040687	Germany, Berlin
	LuCoc03	FR873235	B40 0040734	Germany; Lubmin
	–	AM502010	–	–
Cocconeis placentula Ehrenberg	D36_012[a]	FR873236	B40 0040647	Germany, Berlin
	D17_011	FR873237	B40 0040735	Faroe Islands, Eysturo
	D26_016	FR873238	B40 0040736	South Korea, Tae An Sa
	WiCoc01	FR873239	B40 0040737	Germany, Wismar
	–	AM502013	–	–
Cyclotella atomus Hustedt	–	DQ514858	–	–
Cyclotella bodanica Grunow	–	DQ514901	–	–
Cyclotella choctawhatcheeana A.K.S.Prasad	–	AM712618	–	–
Cyclotella distinguenda Hustedt	–	DQ514859	–	–
Cyclotella gamma Sovereign	–	DQ514852	–	–
Cyclotella meneghiniana Kützing	D15_036	FR873240	B40 0040688	Faroe Islands, Eiði
	–	AM236073	–	–
	–	AJ535172	–	–
	–	AY496206	–	–
	–	AY496207	–	–
	–	AY496210	–	–
	–	AY496211	–	–
	–	AY496212	–	–

Table 1 (continued)

Taxon	Strain	EMBL no.	B voucher	Locality
	–	AY496213	–	–
	–	DQ514853	–	–
	–	DQ514860	–	–
	–	DQ514861	–	–
	–	GQ148714	–	–
	–	GQ148716	–	–
Cyclotella ocellata Pant.	–	DQ514904	–	–
Diatoma hyemalis (Roth) Heib.	–	AB085829	–	–
Diatoma tenuis C. Agardh	D03_036	FR873241	B40 0040689	Germany, Berlin
	–	AJ535143	–	–
	–	AM497730	–	–
	–	AM497731	–	–
	–	EF423403	–	–
	–	EU260466	–	–
Diatoma vulgare var. *linearis* Grunow	–	EF465466	–	–
Encyonema caespitosum Kützing	–	AM502035	–	–
Encyonema minutum (Hilse) D.G.Mann	D36_007	FR873242	B40 0040690	Germany, Berlin
	–	AM501961	–	–
Encyonema silesiacum (Bleisch) D.G.Mann	D16_038	FR873243	B40 0040691	Faroe Islands, Viðareiði
Encyonema triangulum (Ehrenberg) Kützing	–	AJ535157		
Eolimna minima (Grunow) Lange–Bertalot[b]	D03_030	FR873244	B40 0040692	Germany, Berlin
	–	AJ243063	–	–
Eolimna subminuscula (Manguin) Gerd Moser, LangeBertalot & Metzeltin	–	AJ243064	–	–
Fragilaria bidens Heib.	–	AM497732	–	–
Fragilaria capucina Desmazières	–	EF465492	–	–
Fragilaria mesolepta Rabenhorst	D03_041	FR873245	B40 0040693	Germany, Berlin
Fragilaria nanana Lange–Bertalot	–	AM497738	–	–
	–	AM497739	–	–
Fragilaria pinnata Ehrenberg	–	EF192988	–	–
Fragilaria striatula Ehrenberg	–	AY485474	–	–
	–	EU090016	–	–
	–	EU090017	–	–
Fragilaria ulna (Nitzsch) Lange–Bertalot	–	AJ866993	–	–
Fragilaria vaucheriae (Kützing) J.B.Petersen	–	AM497733	–	–
	–	AM497735	–	–
	–	AM497741	–	–
Gomphonema acuminatum Ehrenberg	–	AM502019	–	–
Gomphonema micropus Kützing	–	AM501965	–	–
Gomphonema parvulum Kützing	D33_006	FR873246	B40 0040694	Mexico, Ixtlán de los Hervores
	–	AJ243062	–	–
Gomphonema productum (Grunow) Lange–Bertalot & E.Reichardt	–	AM501993	–	–
Gomphonema truncatum Ehrenberg	–	AM501956	–	–
Hantzschia amphioxys (Ehrenberg) Grunow	D28_006	FR873247	B40 0040695	Germany, Berlin
	D27_008	FR873248	B40 0040696	Germany, Berlin
Mayamaea atomus (Kützing) Lange–Bertalot var. *atomus*	–	AM501968	–	–
Mayamaea atomus var. *permitis* (Hustedt) Lange–Bertalot	D06_107	FR873249	B40 0040697	Germany, Berlin

Table 1 (continued)

Taxon	Strain	EMBL no.	B voucher	Locality
	–	AM501968	–	–
	–	AM501969	–	–
Melosira cf. *octogona* (Grunow) Hustedt	–	AY485518	–	–
Melosira varians C. Agardh	D34_009	FR873250	B40 0040698	Spain, Mallorca, Sa Calobra
	–	AJ243065	–	–
	–	AY569590	–	–
Navicula brockmanni Hustedt	–	AM50202	–	–
Navicula capitatoradiata H.Germ.	–	AM502012	–	–
Navicula cari Ehrenberg	–	GU295220	–	–
Navicula cryptocephala Kützing	D06_059	FR873251	B40 0040700	Germany, Berlin
	–	AM501973	–	–
	–	AM501996	–	–
Navicula cryptotenella Lange–Bertalot	–	AM502011	–	–
	–	AM502029	–	–
Navicula diserta Hustedt	–	AJ535159	–	–
Navicula glaciei van Heurck	–	EF106788	–	–
Navicula gregaria Donkin	D08_002	FR873252	B40 0040703	Sweden; Ängelholm
	–	AM501974	–	–
Navicula lanceolata (C.Agardh) Ehrenberg	–	AY485484	–	–
Navicula pelliculosa (Bréb. ex Kützing) Hilse	–	EU260468	–	–
Navicula phyllepta Kützing	–	FJ624253	–	–
Navicula radiosa Kützing	–	AM501972	–	–
	–	AM502027	–	–
	–	AM502034	–	–
Navicula ramosissima (C.Agardh) Cleve	–	AY485512	–	–
Navicula reinhardtii (Grunow) Grunow	–	AM501976	–	–
Navicula saprophila Lange–Bertalot & Bonik	–	AJ867025	–	–
Navicula slesvicensis Grunow	D06_038	FR873253	B40 0040705	Germany, Berlin
Navicula sp.	D08_051	FR873254	B40 0040699	Sweden; Ängelholm
Navicula subminuscula Manguin	–	AJ867026	–	–
Navicula tripunctata (O.F.Müller) Bory	D03_093	FR873255	B40 0040706	Germany, Berlin
	–	AM502028	–	–
Nitzschia acicularis (Kützing) W.Smith	D03_095	FR873256	B40 0040707	Germany, Berlin
	–	AJ867000	–	–
Nitzschia cf. *fonticola* Grunow	D15_023	FR873257	B40 0040712	Faroe Islands, Eiði
Nitzschia cf. *semirobusta* Lange–Bertalot	D32_012	FR873258	B40 0040708	Mexico, Ojo de Agua
	–	AJ867277	–	–
Nitzschia cf. *supralitorea* Lange–Bertalot	–	AJ867020	–	–
Nitzschia communis Rabenhorst	–	AJ867014	–	–
	–	AJ867278	–	–
Nitzschia closterium (Ehrenberg) W.Smith	–	AY485455	–	–
	–	HQ121419	–	–
Nitzschia dissipata (Kützing) Rabenhorst	D03_120	FR873259	B40 0040710	Germany,Berlin
	–	AJ867018	–	–
Nitzschia inconspicua Grunow	D10_009	FR873260	B40 0040709	Hungary, Sofiok
	–	AJ867021	–	–
Nitzschia filiformis (W.Sm.) van Heurck	–	AJ866999	–	–
Nitzschia linearis (C.Agardh) W. Smith	D05_031	FR873261	B40 0040711	Germany, Iffeldorf

Table 1 (continued)

Taxon	Strain	EMBL no.	B voucher	Locality
	–	AJ867011	–	–
	–	AJ867012	–	–
	–	AJ867013	–	–
Nitzschia palea (Kützing) W.Smith	–	AJ867006	–	–
	–	AJ867009	–	–
	–	AJ867010	–	–
Nitzschia palecea (Grunow) Grunow	–	AJ866996	–	–
Nitzschia paleaformis Hustedt	–	AJ866997	–	–
	–	AJ866998	–	–
Nitzschia pusilla Grunow	–	AJ867015	–	–
	–	AJ867016	–	–
Nitzschia sigma (Kützing) W.Sm.	–	AJ867279	–	–
Nitzschia sigmoidea (Nitzsch) W.Smith	Nitz1	FR873262	B40 0040713	Germany, Berlin
Nitzschia supralitorea Lange–Bertalot	–	AJ867019	–	–
Nitzschia thermalis (Kützing) Auerswald	–	AY485458	–	–
Nitzschia vitrea G.Norman	–	AJ867280	–	–
Pinnularia acrosphaeria W.Smith	–	AM502021	–	–
Pinnularia anglica Krammer	–	AM501980	–	–
Pinnularia appendiculata (C.Agardh) Cleve	–	AM743095	–	–
Pinnularia brauniana (Grunow) Studnicka	–	AM743097	–	–
Pinnularia cf. *gibba* (Ehrenberg) Ehrenberg	–	EF151977	–	–
Pinnularia divergens W.Smith	D31_023	FR873263	B40 0040714	Mexico, Barranca del Aguacate
Pinnularia gentilis (Donkin) Cleve	–	AM743099	–	–
Pinnularia mayeri Krammer	–	AM743102	–	–
Pinnularia mesolepta (Ehrenberg) W.Smith	–	AM501994	–	–
	–	AM502024	–	–
Pinnularia microstauron (Ehrenberg) Cleve	–	AM501981	–	–
	–	AM501982	–	–
	–	AM501983	–	–
	–	AM501984	–	–
Pinnularia obscura Krasske	–	AM743104	–	–
	–	AM743105	–	–
Pinnularia rupestris Hantzsch	–	AM501992	–	–
Pinnularia subcapitata W.Gregory	–	AM501979	–	–
Pinnularia substreptoraphe Krammer	–	AM502036	–	–
Pinnularia viridis (Nitzsch) Ehrenberg	–	AM502023	–	–
Stauroneis anceps Ehrenberg	–	AM502008	–	–
Stauroneis constricta Ehrenberg	–	AY485521	–	–
Stauroneis gracilior E.Reichardt	–	AM501988	–	–
Stauroneis kriegeri R.M.Patrick	–	AM501990	–	–
	–	AM502037	–	–
Stauroneis phoenicenteron (Nitzsch) Ehrenberg	Stau1	FR873264	B40 0040715	Germany, Berlin
	–	AM501987	–	–
	–	AM502031	–	–
Stephanodiscus agassizensis Håkansson. & H.J.Kling	–	DQ514895	–	–
Stephanodiscuc binderanus (Kützing) Willi Krieger	–	DQ514896	–	–
Stephanodiscus hantzschii Grunow	–	DQ514914	–	–
Stephanodiscus minutulus (Kützing) Cleve & Möller	D03_012	FR873265	B40 0040716	Germany, Berlin

Table 1 (continued)

Taxon	Strain	EMBL no.	B voucher	Locality
	–	DQ514900	–	–
	–	DQ514911	–	–
	–	DQ514915	–	–
	–	DQ514916	–	–
Stephanodiscus neoastraea Håkansson & B.Hickel	–	DQ514906	–	–
Stephanodiscus niagarae Ehrenberg	–	DQ514907	–	–
	–	DQ514908	–	–
Stephanodiscus reimeri Theriot	–	DQ514909	–	–
Stephanodiscus yellowstonensis Simonsen	–	DQ514910	–	–

[a] Epitype strain

[b] Treated as *Sellaphora minima* (Grunow) D.G. Mann by some authors

Methods

Taxon sampling

One hundred twenty three taxa from a wide range of genera throughout Bacillariophyta were used to test the universal applicability of different primer pairs of the 18S rRNA gene. The taxa sampled, the sample origins and/or corresponding EMBL numbers are listed in Tables 1 and 2. Vouchers of sequenced material are deposited in the Herbarium of the Botanic Garden and Botanical Museum Berlin-Dahlem (B), and described in more detail in AlgaTerra (Jahn and Kusber 2002+).

To specifically test the power of the proposed barcode region to distinguish between closely related species, the genus *Sellaphora* (incl. *Sellaphora pupula*-group) was chosen as a test case (Table 2). This is a diatom genus with well-defined biological species concepts (Evans et al. 2007, 2008) as well as vouchered sequences.

Cultivation

DNA was isolated from non-axenic unialgal cultures derived from single cells isolated from environmental samples. The cultures were raised on a modified WC medium (Guillard and Lorenzen 1972) with salt concentrations of 28 g/l of $CaCl_2$, 21 g/l of Na_2SiO_3 and 0.01 g/l of $CuSO_4$. The cultures were stored in petri dishes sealed with Parafilm® M (American National Can Group; Chicago, IL) at 15–17°C and a 12 hday/night rhythm, or at room temperature and the ambient day/night cycle.

DNA isolation

The harvested cultures were transferred to 1.5 ml tubes. DNA was isolated using either Dynal® DynaBeads (Invitrogen Corporation; Carlsbad, CA, USA) or Qiagen® Dneasy Plant Mini Kit (Qiagen Inc.; Valencia, CA) following the respective product instructions. DNA concentrations were checked using gel electrophoresis (1.5% agarose gel) and Nanodrop® (PeqLab Biotechnology LLC; Erlangen, Germany). DNA samples were stored at −20°C until further use.

Secondary structure analysis

The secondary structure of the V4 region was analysed using Mfold (Zuker 2003) running under standard RNA settings (default), and compared to the secondary structure of a consensus sequence (Alverson et al. 2006) to identify possible primer regions within the 18S locus. Primers were designed manually. To assess the variability of the fragment within any given primer pairing, the consensus sequence of Alverson et al. (2006) was used.

Primer testing

All primers given in Table 3 were also tested for amplification and sequencing success at annealing temperatures of 50–54°C under the PCR regime mentioned below. Melting temperature, dimerisation between primer pairs and within single primers, as well as GC content were determined using SeqState under default settings (Müller 2005).

PCR amplification

The V4 region of the 18S locus was amplified using different primer combinations (Table 3). The polymerase chain reaction (PCR) mix (25 µl) consisted of 14.65 µl HPLC H_2O, 2.5 µl 10× buffer S, 1.5 µl $MgCl_2$, 2.5 µl pecGOLD dNTPs, 0.5 µl BSA, 1 µl of each primer (20 pm/µl), 0.35 µl pecGOLD Pur Taq® (all products by PeqLab Biotechnology), and 1 µl DNA sample. The PCR regime included an initial denaturation at 94°C (2 min), then five

Table 2 List of the tested *Sellaphora* taxa with corresponding phenodemes, clone names (both after Evans et al. 2007, 2008) and EMBL accession numbers (18S rRNA)

Taxon	Corresponding phenodeme	Clone	EMBL no.
Sellaphora auldreekie D.G.Mann & S.M.McDonald	–	DUN1	EF151965
Sellaphora bacillum (Ehrenberg) D.G.Mann	–	BLA3	EF151980
Sellaphora blackfordensis D.G.Mann & S.Droop	–	BLA6	EF151969
Sellaphora capitata D.G.Mann & S.M.McDonald	–	BLA10	EF151971
Sellaphora cf. *minima* (Grunow) D.G.Mann	–	BM42	EF151966
Sellaphora cf. *seminulum* (Grunow) D.G.Mann	–	TM37	EF151967
Sellaphora lanceolata D.G.Mann & S.Droop	–	BLA13	EF151978
Sellaphora laevissima (Kützing) D.G.Mann	–	SCOT	AJ544655
	–	THR1	EF151979
	–	THR4	EF151981
Sellaphora pupula (Kützing) Mereschkovsky	Afro	AFR1	EF151975
	cf. capitate	THR7	EF151976
	cf. obese	BEL2	EF151973
	Elliptical	RGB1	EF151962
	Elliptical	THR9	EF151972
	Europa	GER1	EF151984
	Large	THR11	EF151970
	Little	RGB2	EF151964
	Small lanceolate	THR14	EF151963
	Spindle	BLA16	EF151974
	Southern capitate	AUS1	EF151982
	Southern pseudocapitate	AUS4	EF151983

Table 3 Tested primer sequences for 18S amplificates (M13 tails shown in italics; → =forward, ← =reverse)

Primer	Sequence 5′-3′	Direction	Source	GC content	MT in °C
D1NS1	CTA GTC ATA CGC TCG TCT C	→	Brinkmann et al. (pers. comm.)	52.0%	64.6
D2NS1	GTA GTC ATA CCC TCG TCT C	→	Brinkmann et al. (pers. comm.)	52.0%	64.6
D1800R	GCT TGA TCC TTC TGC AGG T	←	Brinkmann et al. (pers. comm.)	52.0%	64.6
Algen F	CTG GTT GAT CCT GCC AGT AG	→	This paper	53.8%	66.4
Primer IR3	GGT AAT TTA CGC GCC TGC T	←	This paper	52.0%	64.6
D1294for 18S	TTY CGT TAA CGA ACG AGA CC	→	This paper	48.1%	64.0
D514for 18S	TCC AGC TCC AAT AGC GTA	→	This paper	50.0%	61.8
D356rev 18S	GGT AAT YTA CGC GCS TGC	←	This paper	54.3%	63.4
D514rev 18S a	ATA CGC TAT TGG AGC TGG	←	This paper	50.0%	61.8
D1069rev 18S a	TCT TTA AGT TTC AGC CTT GC	←	This paper	42.3%	61.6
D512for 18S	ATT CCA GCT CCA ATA GCG	→	This paper	50.0%	61.8
D978rev 18S	GAC TAC GAT GGT ATC TAA TC	←	This paper	42.3%	61.6
M13F-D512for 18S	*TGT AAA ACG ACG GCC AGT* ATT CCA GCT CCA ATA GCG	→	This paper; M13 tail after Messing (1983)	–	–
M13R-D978rev 18S	*CAG GAA ACA GCT ATG AC* GAC TAC GAT GGT ATC TAA TC	←	This paper; M13 tail after Messing (1983)	–	–
M13F (-21)	TGT AAA ACG ACG GCC AGT	→	Messing (1983)	–	–
M13R (-27)	CAG GAA ACA GCT ATG AC	←	Messing (1983)	–	–

MT melting temperature

cycles consisting of denaturation at 94°C (45 s), annealing at 52/54°C (45 s), respectively, and elongation at 72°C (1 min), followed by 35 cycles in which the annealing temperature was lowered to 50/52°C, and a final elongation at 72°C (10 min). PCR products were visualised in a 1.5% agarose gel and cleaned with MSB Spin PCRapace® (Invitek LLC; Berlin, Germany) following standard procedure. DNA content was measured using Nanodrop (PeqLab Biotechnology).

A second PCR following the same protocol and primers (modified with 6 bp long 454 primertails for sample identification) was run to produce samples for the 454 sequencing. After PCR they were also cleaned with MSB Spin PCRapace® (Invitek LLC) following standard procedure. The samples were normalised to a total DNA content >200 ng using Nanodrop (PeqLab Biotechnology).

Sequencing

Sanger sequencing was used for the establishment of reference sequences, whereas 454 sequencing was conducted to establish intragenomic diversity. The Sanger sequencing was conducted by Starseq® (GENterprise LLC; Mainz, Germany). As sequencing primers the M13 tails were used (Table 3), following Ivanova et al. (2007). M13 tails consist of 17–18 bases that are attached at the 5′ end of the regular PCR primer during oligo synthesis. The M13 sequences become amplified at both ends of the PCR product and subsequently can be used as sequencing primers. This prevents loss of sequence information compared to the use of normal internal sequencing primers. As M13 tails can be attached to any primer, only one pair of sequencing primers are necessary regardless of the PCR primers used.

The sequences were edited in ChromasPro (Technelysium Pty. Ltd.; Tewantin, Australia), aligned using ClustalW (Larkin et al. 2007), and manually improved in BioEdit (Hall 1999).

Sequences for intragenomic comparisons were generated with a 454 sequencer (454 Life Sciences; Roche Company; Branford, CT) using GS FLX Titanium® chemistry, following the manufacturer's instructions. All sequences were compared against the reference sequence database created via Sanger sequencing. Only sequences with a complete primer sequence and longer than 250 bp were included.

Statistics

For analysis of the intraspecific and intrageneric variation, sequences from Sanger sequencing (35 sequences; Table 1, EMBL accession numbers FR873231 to FR873265) were used and complemented with sequences downloaded from EMBL (164 sequences; Table 1, all remaining EMBL accession numbers).

Uncorrected *p*-distances were computed using both DOINK (J. Ehrman, Digital Microscopy Facility, Mount Allison University, Sackville, NB, Canada) and PAUP 4.0b10 (Swofford 2002), as the former program cannot interpret ambiguity coding, whereas the latter does not distinguish between gaps and missing data. The significance of the divergence between intraspecific and intrageneric genetic distances was tested with the Wilcoxon rank-sum test using R (R Development Core Team 2005).

Results

DNA isolation

Non-destructive DNA isolation with the Dynal® DynaBeads generally yielded more DNA (up to 50%; details available from the authors upon request) than isolation with the Qiagen® Dneasy Plant Mini Kit for which the diatom frustules were crushed before the extraction procedure was started.

PCR protocol

First the entire 18S rRNA gene was screened for genetic variability between several diatom taxa for barcoding purposes. Then different fragments of high variability, short enough to be sequenced in one read (454 and Sanger), were tested for universal primer binding sites, PCR amplification and sequencing success. A summary of amplification and sequencing success, fragment lengths and variable positions within a fragment is given in Table 4. Among the tested primer pairs, D512for 18S and D978rev 18S as well as their M13 derivates were successful in 100% of the tested taxa in both amplification and sequencing, with the PCR regime given below gaining the most PCR products. All other primer pairings were less suitable as barcoding primers due to poorer amplification and sequencing success and/or to a worse fragment length/variability ratio (Table 4, Fig. 1). Furthermore, the fragment enclosed by the D512for/D978rev primer pair is short enough to be sequenced in one read and has at least 60 putatively variable basepair (bp) positions. The automated primer design software SeqState (Müller 2005) also favoured the application of this primer pair.

DNA sequencing

Sanger sequencing produced sequences of 35 taxa from unialgal cultures (Table 1, EMBL accession numbers FR873231 to FR873265).

The number of generated sequences (454 sequencing) for calculating the intragenomic variation varies between 16 and 112 per taxon (total 1010; Table 5). All sequences >250 bp

Table 4 Percentage of successful amplifications (annealing temperature regime 1 (52–50°C) / regime 2 (54–52°C)) and percentage of successful sequences of the amplificates from PCR regimes 1 and 2 in all 35 taxa; fragment length and number of variable positions on the given fragment following Alverson et al. (2006) for each primer pair

Primer pair	PCR success (regimes 1 / 2)	Sequencing success (regimes 1 / 2)	Fragment length in bp	Positions >20% variable
D514for 18S, D1069rev 18S	93% / 100%	100% / 100%	ca. 500	70
AlgenF, Primer IR3	66% / 90%	68% / 85%	ca. 400	43
D1294for 18S, D1800R	79% / 76%	96% /100%	ca. 700	69
(D1NS1, D2NS1), D514rev 18S	72% / 72%	71% / 71%	ca. 450	43
(D1NS1, D2NS1), D356rev 18S	66% / 62%	68% / 68%	ca. 400	42
D512for 18S, D978rev 18S	100% / 100%	100% / 100%	ca. 390–410	60
M13F D512for 18S, M13R D978rev 18S	100% / 100%	100% / 100%	ca. 390–410	60

from the 454 run could be assigned unambiguously to one of the reference sequences from the Sanger sequencing.

Genetic distances and statistics

To analyse genetic distances between and within strains (several sequences analysed for one unialgal culture), species and genera for the proposed 18S rRNA gene fragment (V4), uncorrected *p*-distances were calculated. The average, minimum and maximum *p*-distance values are given in Table 5. Average genetic distance within one strain varied between p=0.000 (*Nitzschia acicularis*, *N. linearis*) and p=0.005 (*Hantzschia amphioxys*). Intraspecific variation also ranged between p=0.000 (e.g. *Achnanthidium minutissimum*) and p=0.005 (*Nitzschia pusilla*, *Pinnularia mesolepta*, *Stauroneis kriegeri*). Intrageneric distance varied between p=0.011 (*Mayamaea* spp.) and p=0.174 (*Melosira* spp.), except for *Stephanodiscus* spp., in which the average intrageneric variation was only p=0.001 (Table 5). Except for *Stephanodiscus*, intrageneric (heterospecific) variation was always higher than both, intraspecific variation and the variation within each strain (for example, intraspecific variation in *Aulacoseira* varied between p=0.000 and p=0.001 while intrageneric distance was p=0.048; Table 5). The Wilcoxon rank-sum test showed that the genetic distances within the species of the 16 tested genera (Table 5) is significantly lower than between the single species in these genera ($p = 2.2 \times 10^{-16}$; Fig. 2).

Genetic distance among taxa in *Sellaphora* ranged between p=0.003 (*Sellaphora blackfordensis*/*Sellaphora pupula* phenodeme southern pseudocapitate) and p=0.087 (*Sellaphora* cf. *minima*/*Sellaphora pupula* phenodeme europa), with an average p=0.039 (Table 6). The average intraspecific genetic distance within *Sellaphora laevissima* is p=0.005 (min. p=0.000, max. p=0.007; number of sequences=3; Table 6); within *Sellaphora pupula* phenodeme elliptical it is p=0.000 (number of sequences=2; Table 6).

Discussion

The analysis of environmental samples via DNA barcoding needs to facilitate the detection of—in this case diatom—diversity as well as the identification of species present in the respective sample. For the first part a standard laboratory protocol (including universal primers) is essential, for the second a critical assessment of intra- versus interspecific variation is needed.

Standard laboratory protocol

The development of a standard laboratory protocol considered DNA extraction as well as fragment amplification and sequencing including primer design. The DNA extraction using Dynal® DynaBeads is a non-destructive process that leaves the frustules intact and available for microscopic examination and taxonomic determination, e.g. if species have not yet been deposited in a reference database and morphological vouchers have to be cross-checked after sequencing or if mixed samples have to be analysed microscopically and valves have to be counted for quantification. Even if the Qiagen® Dneasy Plant Mini Kit is used non-destructively it includes more centrifuging steps that could damage especially the larger diatom frustules or fragile frustule characteristics that can be crucial for identification.

Concerning the Dynal® DynaBeads method it has to be noted that after the extraction the residue containing the frustules has to be centrifuged, the supernatant removed, and replaced by pH neutral storing buffer. Otherwise the frustules might be dissolved. The DNA yield is higher than with the Qiagen® Dneasy Plant Mini Kit. Because of the better performance and the conservation of the frustules, the non-destructive DNA isolation was chosen.

Of the six different primer pairs that were tested, D512for 18S and D978rev 18S, as well as their M13

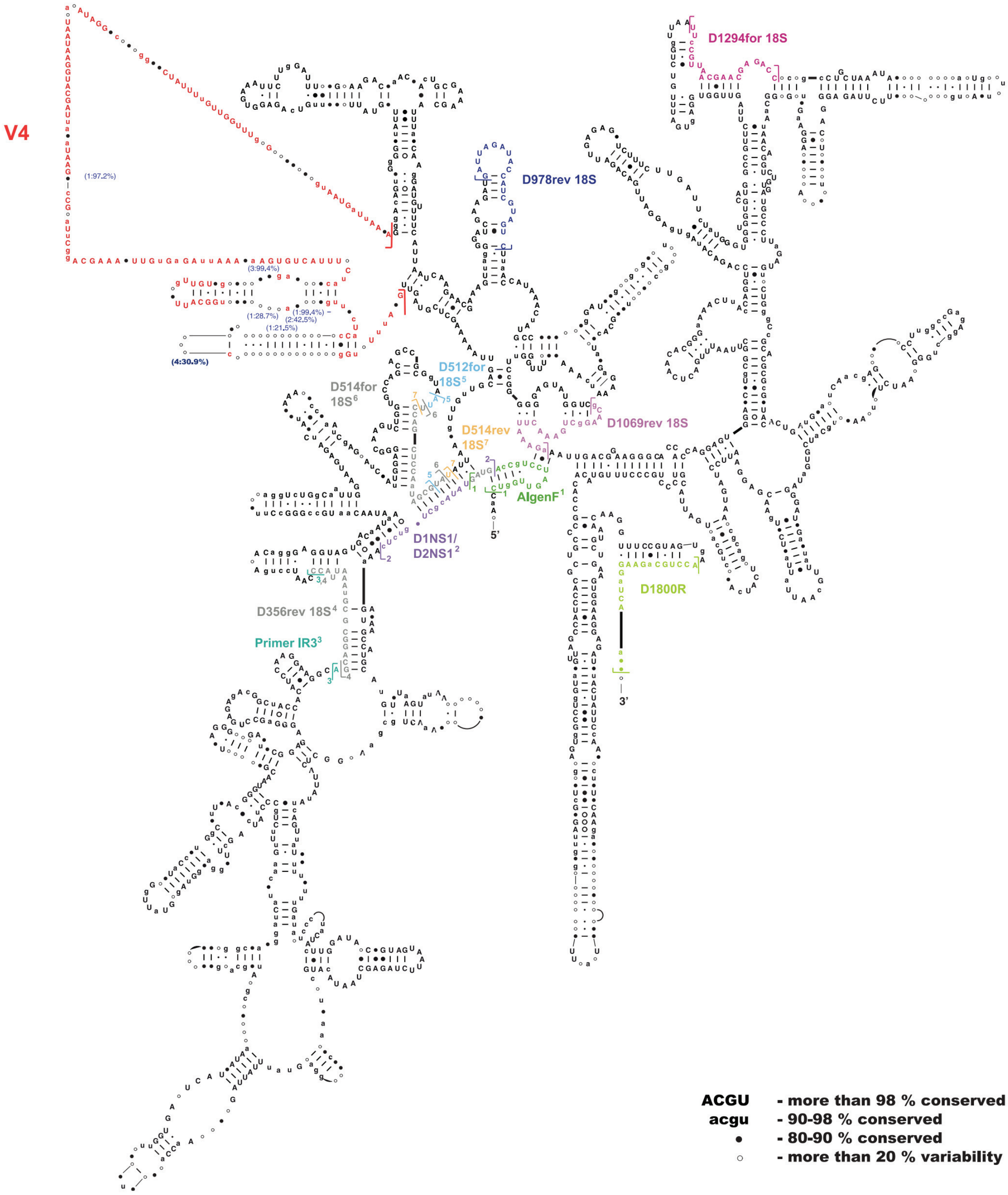


Fig. 1 Consensus secondary structure of the 18S locus (SSU rRNA gene) in diatoms (181 sequences), based on the *Toxarium undulatum* 18S secondary structure model as reference sequence. Upper-case letters indicate that nucleotides at corresponding positions are conserved in 98–100% of sequences, lower-case letters indicate 90–98% conservation, dots 80–90% conservation, circles indicate greater than 20% variability. V4 region and primer binding sites (see Table 3) shown highlighted and in brackets. Where primers overlap their names and brackets are numbered accordingly. Tags at V4 region indicate indels relative to *Toxarium undulatum* sequence; tag format is (maximum length of indel: percentage of sequences showing length polymorphisms). Figure modified after Alverson et al. (2006) and Gillespie et al. (2006)

Table 5 Uncorrected *p*-distances given as average, minimum and maximum values; *n* = number of sequences (per strain) or number of individuals (per species/genus), respectively

Taxon	*n*	*p*-distance		
		Average	Minimum	Maximum
Within strains				
Achnanthidium minutissimum	74	0.003	0.000	0.007
Aulacoseira granulata	24	0.001	0.000	0.002
Cyclotella meneghiniana	104	0.001	0.000	0.004
Diatoma tenue	20	0.003	0.000	0.007
Encyonema minutum	112	0.002	0.000	0.006
Fragilaria melosepta	29	0.003	0.000	0.006
Gomphonema parvulum	17	0.003	0.000	0.005
Hantzschia amphioxys	50	0.005	0.000	0.008
Mayamaea atomus var. *permitis*	67	0.002	0.000	0.008
Melosira varians	93	0.001	0.000	0.008
Navicula slesvicensis	48	0.002	0.000	0.007
Navicula tripunctata	16	0.001	0.000	0.003
Nitzschia acicularis	96	0.000	0.000	0.002
Nitzschia cf. *semirobusta*	54	0.001	0.000	0.006
Nitzschia linearis	95	0.000	0.000	0.000
Pinnularia divergens	24	0.002	0.000	0.008
Stephanodiscus minutulus	87	0.001	0.000	0.004
Within species				
Achnanthidium minutissimum	2	0.000	–	–
Aulacoseira ambigua	6	0.001	0.000	0.002
Aulacoseira baicalensis	2	0.000	–	–
Aulacoseira granulata	3	0.001	0.000	0.002
Aulacoseira nyassensis	2	0.001	0.000	0.002
Aulacoseira islandica	3	0.000	0.000	0.000
Aulacoseira subarctica	4	0.001	0.000	0.002
Cocconeis pediculus	4	0.000	0.000	0.000
Cocconeis placentula	5	0.000	0.000	0.000
Cyclotella meneghiniana	2	0.004	0.000	0.007
Diatoma tenue	6	0.001	0.002	0.000
Encyonema minutum	2	0.000	–	–
Eolimna minima	2	0.000	–	–
Fragilaria nanana	2	0.000	–	–
Fragilaria striatula	3	0.003	0.002	0.005
Fragilaria vaucheriae	3	0.000	0.000	0.000
Gomphonema parvulum	2	0.008	–	–
Hantzschia amphioxys	2	0.001	–	–
Mayamaea atomus var. *permitis*	3	0.000	0.000	0.000
Melosira varians	3	0.002	0.000	0.003
Navicula cryptocephala	3	0.000	0.000	0.000
Navicula cryptotenella	2	0.002	0.002	0.002
Navicula gregaria	2	0.000	–	–
Navicula radiosa	3	0.000	0.000	0.000
Navicula tripunctata	2	0.000	–	–
Nitzschia acicularis	2	0.000	–	–
Nitzschia cf. *semirobusta*	2	0.000	–	–

Table 5 (continued)

Taxon	*n*	*p*-distance		
		Average	Minimum	Maximum
Nitzschia closterium	2	0.000	–	–
Nitzschia communis	2	0.000	–	–
Nitzschia dissipata	2	0.000	–	–
Nitzschia inconspicua	2	0.000	–	–
Nitzschia linearis	4	0.003	0.000	0.005
Nitzschia palea	3	0.002	0.000	0.003
Nitzschia paleaformis	2	0.000	–	–
Nitzschia pusilla	2	0.005	–	–
Pinnularia mesolepta	2	0.005	–	–
Pinnularia microstauron	4	0.000	0.000	0.000
Pinnularia obscura	2	0.000	–	–
Stauroneis kriegeri	2	0.005	–	–
Stauroneis phoenicenteron	3	0.000	0.000	0.000
Stephanodiscus minutulus	5	0.000	0.000	0.000
Within genera				
Achnanthidium	2	0.129	–	–
Aulacoseira	9	0.048	0.018	0.085
Cocconeis	3	0.134	0.044	0.181
Cyclotella	7	0.059	0.018	0.098
Diatoma	3	0.029	0.021	0.038
Encyonema	4	0.032	0.019	0.056
Eolimna	2	0.104	–	–
Fragilaria	8	0.074	0.018	0.124
Gomphonema	5	0.072	0.042	0.101
Mayamaea	2	0.011	–	–
Melosira	2	0.174	–	–
Navicula	20	0.069	0.0155	0.186
Nitzschia	19	0.093	0.013	0.212
Pinnularia	17	0.168	0.015	0.329
Stauroneis	5	0.064	0.015	0.129
Stephanodiscus	8	0.001	0.000	0.005

variants, were the most successful with respect to amplification and sequencing success, and exhibited the best fragment length/variability ratio (Table 4, Fig. 1). PCR amplification with primers D512for 18S and D978rev 18S was successful in all taxa in our study and in many other taxa (e.g. *Skeletonema* spp., *Phaeodactylum* spp., *Surirella* spp., *Campylodiscus* spp.; authors' unpublished data). This high amplification efficiency is due to the placement of the primers in highly conserved stemloop sections of the 18S rRNA gene (Fig. 1) that exhibit low mutation rates and are conserved across a wide range of diatom taxa, therefore make ideal binding sites for universal primers. The M13 tails were used as universal sequencing primers (Ivanova et al. 2007), which contributed to the high sequencing success.

Importantly, the primer combination D512for 18S and D978rev 18S includes the highly variable V4 region of the 18S rRNA gene (Fig. 1) which encloses many indel regions that contribute to the increased information level on this short fragment (Alverson et al. 2006). The other tested primer pairs also result in short variable segments, but with lower universality concerning the laboratory success. The fragments are also less variable, thus do not allow species-level identification within diatoms (Fig. 1).

Besides the primer universality, the V4 region has another promising feature for barcoding environmental samples: The association of the sequences produced by 454 sequencing to the reference data generated via Sanger sequencing was always unambiguously possible—due to the systematic selection procedure—without much computing

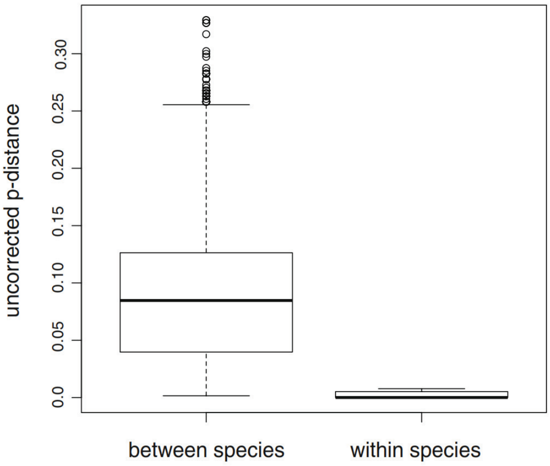


Fig. 2 Box-and-whisker plot of intraspecific and intrageneric (x-axis) genetic distances measured in uncorrected *p*-distances (y-axis). Thick black lines indicate median values, boxes represent upper and lower quartiles, whiskers indicate value ranges, circles represent outliers

and editing effort after sequencing. In addition, no problems emerged in the present study concerning homopolymer errors in the sequences as are often encountered when applying pyrosequencing (Huse et al. 2007).

For high-throughput studies it is also important that the barcode does not exceed a certain length, currently around 400 bp. This length keeps increasing along with the development of sequencing techniques and computation capacity (Schloss 2010), but the cost of sequencing increases accordingly. This is one reason why Hajibabaei et al. (2006b) proposed a 100 bp barcode, which would also work with high-throughput technologies that only produce shorter read length such as Illumina. The V4 region (Fig. 1) in itself is only about 60 bp long, so that it could qualify as such a short barcode without losing its resolving power. Some studies already use very short sequences to evaluate prokaryotic diversity in environmental samples (Huber et al. 2009; Huse et al. 2007; Schloss 2010).

For these reasons, standard laboratory protocols, primer universality, informational indels on a short fragment, the V4 region—maybe only a 60 bp part of it—show high potential for the use in fast, high-throughput approaches to environmental barcoding using next-generation sequencing.

Species identification

For the assessment of the 18S fragment's power to resolve taxa at species level, uncorrected *p*-distances were used. All species tested in this study feature uniform sequences allowing unambiguous resolution at species level, with the only exception concerning *Stephanodiscus*. This genus is well known as problematic in morphological discriminations due to small size of the individuals and to valve plasticity which is often overlapping between species (Håkansson and Kling 1989; 1990; Kobayasi et al. 1985; Spamer and Theriot 1997; Teubner 1997; Wolf et al. 2002). Molecular species identification in *Stephanodiscus* is also difficult (Moniz and Kaczmarska 2009, 2010), possibly because some taxa have diverged only very recently, e.g. *S. niagarae* and *S. yellowstonensis* about 12.000 to 8.000 years ago (Zechman et al. 1994).

Intraspecific variation was very low in general, not exceeding $p=0.005$ (*Hantzschia amphioxys*, Table 5). Intrageneric variation was significantly higher than intraspecific variation in all cases (Table 5). This leads to the assumption that, even though the *p*-distances are comparatively low compared to other markers (e.g. Huang et al. 2007; Wu et al. 2008; Xia et al. 2003), the 18S fragment (V4) used in the present study still has informative value as a barcoding marker to resolve taxa at the species level.

So far, the resolving ability of a given barcode marker has been assessed using either a fixed threshold or the concept of the "barcode gap" (Hollingsworth et al. 2009), meaning a well-defined difference between the levels of intra- and interspecific variation, often calculated by means of a ratio. Initially some studies used a 10-fold increase to gauge the applicability of a certain marker (Hebert et al. 2003). More recently, however, it has been shown that taxa differ considerably in their genetic variation, so that different studies now use very different ratios and thresholds depending on the respective organism group and marker (e.g. Cywinska et al. 2006; Hajibabaei et al. 2006a; b; Hebert et al. 2004; Hickerson et al. 2006; Meyer and Paulay 2005; Ward et al. 2005). For the *cox*1 gene a threshold of $p=0.04$ is considered sufficient in red algae (Saunders 2005), for the ciliate genus *Tetrahymena* $p=0.11$ (Chantangsi et al. 2007), and for *Paramecium* $p=0.20$ (Barth et al. 2006). Moniz and Kaczmarska (2009) give a minimum intrageneric distance of $p=0.07$ for a combination of the 5.8S rRNA gene and ITS2 within diatoms.

The variation in the 18S rRNA gene has been considered as too low for a barcoding marker in diatoms (Moniz and Kaczmarska 2009, 2010). This, however, refers to the complete 18S locus, which is much longer (1800 bp) than the one used in the present study (ca. 390–410 bp). As most of the 1800 bp fragment comprises extremely conserved regions, the genetic distance between species is reduced if the complete 18S rRNA gene locus is used. In the present study the region responsible for species identification is mainly the only ca. 60 bp long V4 region (Fig. 1). As mentioned above, the V4 region comprises not only many variable character sites but also many inversions, insertions and deletions, resulting in a highly concentrated information content on a very short fragment (Alverson et al. 2006).

Table 6 Uncorrected *p*-distances among tested taxa (clones and/or phenodemes) in the genus *Sellaphora*. Values in boxes labeled a–e are discussed in detail

Taxon (clone and/or phenodeme)	*p*-distance																					
S. blackfordensis clone BLA6	-																					
S. bacillum clone BLA3	0.026	-																				
S. lanceolata clone BLA13	0.033	0.008	-																			
S. capitata clone BLA10	0.020	0.038	0.046	-																		
S. cf. *seminulum* clone TM37	0.038	0.036	0.038	0.054	-																	
S. cf. *minima* clone BM42	0.074	0.069	0.069	0.084	0.056	-																
S. auldreekie clone DUN1	0.026	0.026	0.028	0.036	0.041	0.071	-															
S. laevissima clone THR4	0.031	0.031	0.038	0.038	0.048	0.071	0.041	-														
S. laevissima strain SCOT	0.031	0.031	0.038	0.038	0.048	0.071	0.041	0.000														
S. laevissima clone THR1	0.036	0.038	0.046	0.036	0.054	0.077	0.048	0.008	0.008	-												
S. pupula clone GER1 europa	0.043	0.043	0.046	0.051	0.056	0.087	0.038	0.046	0.046	0.051	-											
S. pupula clone THR9 elliptical	0.031	0.031	0.033	0.038	0.043	0.074	0.026	0.031	0.031	0.038	0.018	-										
S. pupula clone BLA16 spindle	0.031	0.033	0.036	0.038	0.043	0.079	0.028	0.036	0.036	0.043	0.026	0.010	-									
S. pupula clone RBG1 elliptical	0.031	0.031	0.033	0.038	0.043	0.074	0.026	0.031	0.031	0.038	0.018	0.000	0.010	-								
S. pupula clone RBG2 little	0.036	0.036	0.038	0.043	0.041	0.071	0.031	0.036	0.036	0.043	0.023	0.008	0.013	0.008	-							
S. pupula clone AUS4 southern pseudocapitate	0.003	0.028	0.036	0.020	0.041	0.077	0.028	0.028	0.028	0.033	0.043	0.031	0.031	0.031	0.036	-						
S. pupula clone AUS1 southern capitate	0.008	0.028	0.036	0.015	0.041	0.071	0.028	0.031	0.031	0.033	0.046	0.033	0.033	0.033	0.038	0.010	-					
S. pupula clone THR7 cf. capitata	0.015	0.033	0.041	0.015	0.048	0.079	0.033	0.038	0.038	0.038	0.051	0.038	0.038	0.038	0.043	0.018	0.010	-				
S. pupula clone THR11large	0.041	0.041	0.043	0.048	0.054	0.084	0.036	0.041	0.041	0.048	0.026	0.015	0.018	0.015	0.018	0.041	0.043	0.048	-			
S. pupula clone THR14 small lanceolate	0.048	0.036	0.043	0.059	0.064	0.079	0.048	0.054	0.054	0.059	0.054	0.041	0.048	0.041	0.046	0.048	0.054	0.059	0.054	-		
S. pupula clone AFR1 afro	0.020	0.038	0.046	0.000	0.054	0.084	0.036	0.038	0.038	0.036	0.051	0.038	0.038	0.038	0.043	0.020	0.015	0.015	0.048	0.059	-	
S. pupula clone BEL2 cf. obese	0.010	0.036	0.043	0.020	0.048	0.084	0.036	0.038	0.038	0.043	0.048	0.036	0.036	0.036	0.041	0.010	0.018	0.026	0.046	0.054	0.020	-
	S. blackfordensis clone BLA6	*S. bacillum* clone BLA3	*S. lanceolata* clone BLA13	*S. capitata* clone BLA10	*S.* cf. *seminulum* clone TM37	*S.* cf. *minima* clone BM42	*S. auldreekie* clone DUN1	*S. laevissima* clone THR4	*S. laevissima.* strain SCOT	*S. laevissima* clone THR1	*S. pupula* clone GER1 europa	*S. pupula* clone THR9 elliptical	*S. pupula* clone BLA16 spindle	*S. pupula* clone RBG1 elliptical	*S. pupula* clone RBG2 little	*S. pupula* clone AUS4 southern pseudocapitate	*S. pupula* clone AUS1 southern capitate	*S. pupula* clone THR7 cf. capitata	*S. pupula* clone THR11 large	*S. pupula* clone THR14 small lanceolate	*S. pupula* clone AFR1 afro	*S. pupula* clone BEL2 cf. obese

The V4 region appears to allow discrimination between species to a degree sufficient for environmental DNA barcoding. Therefore, to further test the power of this region for species identification in a closely related taxon complex, an exclusive in silico analysis within the *Sellaphora pupula*-group and sister taxa was performed. The genus *Sellaphora* is a genus with well-established species concepts and extensive data on mating behaviour, morphology, ecology and DNA sequence variation within the genus (Evans et al. 2007, 2008). The *Sellaphora pupula*-group consists of very closely related species, thus provides a strong test of the reliability of the proposed barcode region. The V4 region was able to discriminate between all the included taxa (following Evans et al. 2008).

There are some taxon pairs with very low genetic distances (Table 6, b–d), one of them comprising *Sellaphora blackfordensis* and *S. pupula* clone AUS4 phenodome southern pseudocapitate, (Table 6, b), the second *S. blackfordensis* and *S. pupula* clone AUS1 phenodome southern capitate (Table 6, c). These three taxa also form a well-supported clade in the *rbcL*-based phylogenetic tree provided by Evans et al. (2008). The third such pair contains *Sellaphora lanceolata* and *S. bacillum* (Table 6, d), showing a relationship which is consistent with the findings of Evans et al. (2008) as well. That the genomic variation between these pairs is lower than or similar to the variation within *Sellaphora laevissima* could indicate, for instance, that the V4 region is not powerful enough to distinguish between all cryptic species, or that the species circumscriptions do not necessarily reflect the genetic diversity.

Within the former *Sellaphora pupula* taxon there are two identical sequences (Table 6, e), both designated as *S. pupula* phenodeme elliptical by Evans et al. (2008). Whether the genetic distances between these phenodemes represent population differences or variation between cryptic species needs further consideration (e.g. Evans et al. 2008). This shows that the V4 region also may have some potential for identifying closely related species, even though it might not be enough for defining them.

The V4 region of the 18S locus as a barcode marker

Various other barcodes have been proposed for various groups of organisms, among them the plastid regions *rbcL*, *matK*, *trnH-psbA*, the 23S rRNA gene, the mitochondrial gene *cox*1, and the nuclear markers ITS, entire 18S (SSU) rRNA gene and 28S (LSU) rRNA gene (e.g. Bhadury et al. 2006; Fazekas et al. 2008; Hebert et al. 2004; Hollingsworth et al. 2009; Kress and Erickson 2007; Kress et al. 2005; Newmaster et al. 2008; Summerbell et al. 2005). However, *cox*1, ITS, 18S and *rbcL* are the only ones which have been applied to diatoms, with mixed results, i.e. *cox*1 was very variable but no universal primers could be found, ITS was variable but is not universally amplifiable with standard laboratory protocols, *rbcL* was less variable, and 18S (whole gene) was not variable enough (e.g. Evans et al. 2007, 2008; Jahn et al. 2007; Moniz and Kaczmarska 2009, 2010).

That the *cox*1 gene is variable enough to discriminate between very similar taxa (e.g. cryptic species) has been stated for many groups throughout the tree of life (Barth et al. 2006; Chantangsi et al. 2007; Hebert et al. 2003; Kucera and Saunders 2008; Lynn and Strüder-Kypke 2006; Saunders 2005). However, a preliminary study using a dataset of over 60 diatom species from various groups to design universal primers for the *cox*1 gene (unpublished data) showed that it is virtually impossible to do so, because the locus lacks sufficiently conserved regions for primer binding. Universal primers constitute an essential condition for environmental analyses. Various publications have shown that this problem occurs not only within diatoms (e.g. Evans et al. 2007, 2008; Moniz and Kaczmarska 2009) but also in many other eukaryotic organism groups, e.g. in land plants (Cowan et al. 2006), dinoflagellates (Ferrell and Beaton 2007), gastropods (Kane et al. 2008), and fungi (Seifert et al. 2007). Most studies on the use of the *cox*1 gene as a barcoding marker for protists are limited to very confined groups, e.g. genera, and use group-specific primers (Chantangsi et al. 2007; Evans et al. 2007, 2008). In diatoms this high variability of the *cox*1 locus could be due to the occurrence of intron events and introgression of bacterial genes, both common in diatoms (Armbrust et al. 2004; Bowler et al. 2008; Ehara et al. 2000; Imanian et al. 2007; Ravin et al. 2010).

The combination of the 5.8S rRNA gene and ITS2 has been suggested as an alternative barcoding locus (Moniz and Kaczmarska 2009, 2010). Its potential to identify species is promising and has been demonstrated in many protists, fungi and plant groups (e.g. Gemeinholzer et al. 2006; Kelly et al. 2010; Litaker et al. 2007; Taylor et al. 2008). There are, however, some problems, the main one being that ITS is not easy to amplify and sequence with standard laboratory protocols (unpublished data; see also Hamsher et al. 2011). Furthermore, studies in fungi using ITS suggested that errors in amplification/sequencing—especially in high throughput—could easily lead to overestimation of diversity in environmental samples (Bellemain et al. 2010).

Plastid markers such as the *rbcL* gene could be problematic for DNA barcoding, as the plastid inheritance in diatoms is not uniform but can be either uniparental or biparental (Casteleyn et al. 2009; Jensen et al. 2003; Levialdi Ghiron et al. 2008; Round et al. 1990), and there are rare reports of natural hybrids (Casteleyn et al. 2009).

The 18S rRNA gene locus is often used to estimate the relative abundances and diversities of species in environmental samples (Liao et al. 2007), due to its low intraspecific but high interspecific variation. It also has been used to define operational taxonomic units (OTUs) in various eukaryots (Ciliophora, Dinophyceae, Cercozoa und Fungi; Lefèvre et al. 2007). The analysis of water samples via a 550 bp long fragment of the 18S rRNA gene locus was able to resolve organisms of the metazoans (e.g. nematodes), the algae Prasinophyceae, Cryptophyceae, Dinophyceae and Prymnesiophyceae, as well as heterotrophic Cercozoa, Choanoflagellates, Stramenopiles, and Cilitates (Romari and Vaulot 2004). It has been shown that the 18S rRNA gene can also discriminate diatoms in most cases of environmental samples, often to the species level (Jahn et al. 2007; Savin et al. 2004).

The main advantage of the V4 fragment of the 18S locus is that it is very easy to amplify with the proposed universal primers using our documented standard laboratory protocol, while it still has considerable power to resolve taxa on the species level. Both of these characteristics are crucial for its successful use in environmental studies. The potential of the V4 fragment to discriminate between (semi-)cryptic species has to be further evaluated. However, while this aspect is desirable it is not necessary for its use in environmental studies, as the members of cryptic-species complexes generally seem to have similar ecology (Beszteri et al. 2005a,b, 2007).

A further advantage of the 18S locus is its high representation in databases. A good retrieval rate for correct identifications strongly depends on the reference data. But even though the reference database for the 18S rRNA gene is more extensive than for many other proposed barcode regions, it nevertheless has to be extended, especially with voucher-based sequences.

Conclusions

The crucial problem in selecting an applicable barcode is the balance between variability and primer-binding universality. For the analysis of environmental samples primer universality and reproducible laboratory protocols are of high importance, whereas for the detection and delimitation of cryptic species these aspects are often secondary.

For the detection of cryptic species other, more variable barcodes might be more feasible. But as discussed in many other studies, some problems, such as species delimitation and α-taxonomy, presumably cannot be solved with only one barcode (e.g. Chase et al. 2007; Cowan et al. 2006; Kress and Erickson 2007). A single barcode represents only a fraction of an organism's variation; therefore its power to define a taxon should not be overestimated. Consequently, a combination of the V4 region with other barcodes such as ITS should be discussed.

The 18S rRNA gene fragment proposed in the present study shows enough variation to unambiguously identify almost all tested taxa. Furthermore, the highly conserved primer binding sites allow amplification following a standard procedure. Due to its relatively short length it is also feasible for time- and cost-saving high-throughput analysis methods. The V4 region of the 18S locus therefore is a good candidate for barcoding diatoms in environmental samples.

Acknowledgements The authors wish to thank Martin Pfannkuchen and Daniela Maric, Monica Moniz and Irena Kaczmarska, James Ehrman, Neela Enke, Nelida Abarca, Daniel Lauterbach, Wolf-Henning Kusber and Weliton da Silva for fruitful discussions, Oliver Skibbe and Jana Bansemer for diatom cultivation. We also thank Michael Kube and Richard Reinhardt (MPI for Molecular Genetics, Berlin) for providing time and guidance at the 454 sequencer. The Association of the Friends of the Botanic Garden and Botanical Museum Berlin-Dahlem and the Academic Senate of the Freie Universität Berlin have provided financial support.

References

Alverson, A. J., Cannone, J. J., Gutell, R. R., & Theriot, E. C. (2006). The evolution of elongate shape in diatoms. *Journal of Phycology, 42*, 655–668.

Archibald, R. E. M. (1984). Diatom illustrations—an appeal. *Bacillaria, 7*, 173–178.

Armbrust, E. V., Berges, J. A., Bowler, C., Green, B. R., Martinez, D., Putnam, N. H., et al. (2004). The genome of *Thalassiosira pseudonana*: Ecology, evolution, and metabolism. *Science, 306*, 79–86.

Babanazarova, O. V., Likhoshway, Y. V., & Sherbakov, D. Y. (1996). On the morphological variability of *Aulacoseira baicalensis* and *Aulacoseira islandica* (Bacillariophyto) of Lake Baikal, Russia. *Phycologia, 35*, 113–123.

Bailey-Watts, A. E. (1976). Planktonic diatoms and some diatom-silica relations in a shallow eutrophic Scottish loch. *Freshwater Biology, 6*, 69–80.

Barth, D., Krenek, S., Fokin, S. I., & Berendonk, T. (2006). Intraspecific genetic variation in *Paramecium* revealed by mitochondrial cytochrome *c* oxidase I sequences. *Journal of Eukaryotic Microbiology, 53*, 20–25.

Behnke, A., Friedl, T., Chepurnov, V. A., & Mann, D. G. (2004). Reproductive compatibility and rDNA sequence analyses in the *Sellaphora pupula* species complex (Bacillariophyta). *Journal of Phycology, 40*, 193–208.

Bellemain, E., Carlsen, T., Brochmann, C., Coissac, E., Taberlet, P., & Kauserud, H. (2010). ITS as an environmental DNA barcode for fungi: An in silico approach reveals potential PCR biases. *BMC Microbiology, 10*, 189. doi:10.1186/1471-2180-10-189.

Beszteri, B., Acs, E., Makk, J., Kovács, G., Márialigeti, K., & Kiss, K. T. (2001). Phylogeny of six naviculoid diatoms based on 18S rDNA sequences. *International Journal of Systematic and Evolutionary Microbiology, 51*, 1581–1586.

Beszteri, B., Ács, É., & Medlin, L. K. (2005a). Conventional and geometric morphometric studies of valve ultrastructural variation in two closely related *Cyclotella* species (Bacillariophyta). *European Journal of Phycology, 40*, 89–103.

Beszteri, B., Ács, É., & Medlin, L. K. (2005b). Ribosomal DNA sequence variation among sympatric strains of the *Cyclotella meneghiniana* complex (Bacillariophyceae) reveals cryptic diversity. *Protist, 156*, 317–333.

Beszteri, B., John, U., & Medlin, L. K. (2007). An assessment of cryptic genetic diversity within the *Cyclotella meneghiniana* species complex (Bacillariophyta) based on nuclear and plastid genes, and amplified fragment length polymorphism. *European Journal of Phycology, 42*, 47–60.

Bhadury, P., Austen, M. C., Bilton, D. T., Lambshead, P. J. D., Rogers, A. D., & Smerdon, G. R. (2006). Development and evaluation of a DNA-barcoding approach for the rapid identification of nematodes. *Marine Ecology Progress Series, 320*, 1–9.

Blaxter, M. L. (2004). The promise of a DNA taxonomy. *Philosophical Transactions of the Royal Society of London, Biological Sciences, 359*, 669–679.

Blaxter, M., Elsworth, B., & Daub, J. (2004). DNA taxonomy of a neglected animal phylum: An unexpected diversity of tardigrades. *Proceedings of the Royal Society of London, Biological Sciences, 271*, 189–192.

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

Casteleyn, G., Adams, N. G., Vanormelingen, P., Debeer, A. E., Sabbe, K., & Vyverman, W. (2009). Natural hybrids in the marine diatom *Pseudo-nitzschia pungens* (Bacillariophyceae): Genetic and morphological evidence. *Protist, 160*, 343–354.

Chantangsi, C., Lynn, D. H., Brandl, M. T., Cole, J. C., Hetrick, N., & Ikonomi, P. (2007). Barcoding ciliates: A comprehensive study of 75 isolates of the genus *Tetrahynema*. *International Journal of Systematic and Evolutionary Microbiology, 57*, 2412–2425.

Chase, M. W., Cowan, R. S., Hollingsworth, P. M., van den Berg, C., Madriñán, S., Petersen, G., et al. (2007). A proposal for a standardised protocol to barcode all land plants. *Taxon, 56*, 295–299.

Cowan, R. S., Chase, M. W., Kress, W. J., & Savolainen, V. (2006). 300,000 species to identify: Problems, progress, and prospects in DNA barcoding of land plants. *Taxon, 55*, 611–616.

Cywinska, A., Hunter, F. F., & Hebert, P. D. N. (2006). Identifying Canadian mosquito species through DNA barcodes. *Medical and Veterinary Entomology, 20*, 413–424.

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

Erickson, D. L., Spouge, J., Resch, A., Weigt, L. A., & Kress, W. J. (2008). DNA barcoding in landplants: Developing standards to quantify and maximise success. *Taxon, 57*, 1304–1316.

Evans, K. M., Wortley, A. H., & Mann, D. G. (2007). An assessment of potential diatom "barcode" genes (*cox*1, *rbcL*, 18S and ITS rDNA) and their effectiveness in determining relationships in *Sellaphora* (Bacillariophyta). *Protist, 158*, 349–364.

Evans, K. M., Wortley, A. H., Simpson, G. E., Chepurnov, V. A., & Mann, D. G. (2008). A molecular systematic approach to explore diversity within the *Sellaphora pupula* species complex (Bacillariophyta). *Journal of Phycology, 44*, 215–231.

Falkowski, P. G., Barber, R. T., & Smetacek, V. (1998). Biogeochemical controls and feedbacks on ocean primary production. *Science, 281*, 200–207.

Fazekas, A. J., Burgess, K. S., Kesanakurti, P. R., Graham, S. W., Newmaster, S. G., Husband, B. C., et al. (2008). Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. *PloS One, 3*(7), e2802. doi:10.1371/journal.pone.0002802.

Ferrell, J., & Beaton, M. (2007). The evaluation of DNA barcoding for identification of dinoflagellates: A test using *Prorocentrum*. In: *Canadian barcode of life network 2007 science symposium* (pp. 37). Guelph: Blackwell.

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

Floyd, R., Abebe, E., Papert, A., & Blaxter, M. L. (2002). Molecular barcodes for soil nematode identification. *Molecular Ecology, 11*, 839–850.

Friedl, T., & O'Kelly, C. J. (2002). Phylogenetic relationships of green algae assigned to the genus *Planophila* (Chlorophyta): Evidence from 18S rDNA sequence data and ultrastructure. *European Journal of Phycology, 37*, 373–384.

Gemeinholzer, B., Oberprieler, C., & Bachman, K. (2006). Using GenBank data for plant identification: Possibilities and limitations using the ITS1 of Asteraceae species belonging to the tribes Lactuceae and Anthemideae. *Taxon, 55*, 173–187.

Gillespie, J. J., Johnston, J. F., Cannone, J., & Gutell, R. R. (2006). Characteristics of the nuclear (18S, 5.8S, 28S and 5S) and mitochondrial (12S and 16S) rRNA genes of *Apis mellifera* (Insecta: Hymenoptera): Structure, organization, and retrotransposable elements. *Insect Molecular Biology, 15*, 657–686.

Guillard, R. R. L., & Lorenzen, C. J. (1972). Yellow green algae with chlorophyllide. *Journal of Phycology, 8*, 10–14.

Hajibabaei, M., Janzen, D. H., Burns, J. M., Hallwachs, W., & Hebert, P. D. N. (2006). DNA barcodes distinguish species of tropical Lepidoptera. *Proceedings of the National Academy of Sciences of the USA, 103*, 968–971.

Hajibabaei, M., Smith, A., Janzen, D. H., Rodriguez, J. J., Whitfield, J. B., & Hebert, P. D. N. (2006). A minimalist barcode can identify specimens whose DNA is degraded. *Molecular Ecology, 6*, 959–964.

Hajibabaei, M., Singer, G. A. C., Hebert, P. D. N., & Hickey, D. A. (2007). DNA barcoding: How it complements taxonomy, molecular phylogenetics and population genetics. *Trends in Genetics, 23*, 167–172.

Håkansson, H., & Kling, H. (1989). A light and electron microscope study of previously described and new *Stephanodiscus* species (Bacillariophyceae) from central and northern Canadian lakes, with ecological notes on the species. *Diatom Research, 4*, 269–288.

Håkansson, H., & Kling, H. (1990). The current status of some very small freshwater diatoms of the genera *Stephanodiscus* and *Cyclostephanos*. *Diatom Research, 5*, 273–287.

Hall, T. A. (1999). BioEdit: A user friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleid Acids Symposium Series, 41*, 95–98.

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

Hebert, P. D. N., Cywinska, A., Ball, S. L., & de Waard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London, Biological Sciences, 270*, 313–321.

Hebert, P. D. N., Stoeckle, M. Y., Zemlak, T. S., & Francis, C. M. (2004). Identification of birds through DNA barcodes. *PLoS Biology, 2*, 1657–1663.

Hickerson, M. J., Meyer, C. P., & Moritz, C. (2006). DNA barcoding will often fail to discover new animal species over broad parameter space. *Systematic Biology, 55*, 729–739.

Hollingsworth, M. L., Clark, A. A., Forrest, L. L., Richardson, J., Pennington, R. T., Long, D. G., et al. (2009). Selecting barcoding loci for plants: Evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. *Molecular Ecology Resources, 9*, 439–457.

Huang, J., Xu, Q., Sun, Z. J., Tang, G. L., & Su, Z. Y. (2007). Identifying earthworms through DNA barcodes. *Pedobiologia, 51*, 301–309.

Huber, J. A., Morrison, H. G., Huse, S. M., Neal, P. R., Sogin, M. L., & Welch, D. M. (2009). Effect of PCR amplicon size on assessments of clone library microbial diversity and community structure. *Environmental Microbiology, 11*, 1292–1302.

Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L., & Welch, D. M. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology, 8*, R143.

Imanian, B., Carpenter, K. J., & Keeling, P. J. (2007). Mitochondrial genome of a tertiary endosymbiont retains genes for electron transport proteins. *Journal of Eukaryotic Microbiology, 54*, 146–153.

Ivanova, N. V., Zemlak, T. S., Hanner, R. H., & Hebert, P. D. N. (2007). Universal primer cocktails for fish DNA barcoding. *Molecular Ecology Notes, 7*, 544–548.

Jahn, R. (1986). A study of *Gomphonema augur* Ehrenberg: The structure of the frustule and its variability in clones and populations. In M. Ricard (Ed.), *Proceedings of the 8th International Diatom Symposium 1984* (pp. 191–204). Paris: Koeltz Scientific Books.

Jahn, R., & Kusber, W. H. (2002+). *AlgaTerra Information System (online)*. Botanic Garden and Botanical Museum Berlin-Dahlem, Freie Universität Berlin. http://www.algaterra.org. Accessed 30 December 2010.

Jahn, R., Zetzsche, H., Reinhardt, R., & Gemeinholzer, B. (2007). Diatoms and DNA barcoding: A pilot study on an environmental sample. In W. H. Kusber & R. Jahn (Eds.), *Proceedings of the 1st Central European Diatom Meeting 2007* (pp. 63–68). Berlin: Botanic Garden and Botanical Museum Berlin-Dahlem.

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*, 98–105.

Kane, R. A., Stothard, J. R., Emery, A. M., & Rollinson, D. (2008). Molecular characterization of freshwater snails in the genus *Bulinus*: a role for barcodes? *Parasites & Vectors, 1*(15). doi:10.1186/1756-3305-1-15.

Kelly, L. J., Ameka, G. K., & Chase, M. W. (2010). DNA barcoding of African Podostemaceae (river-weeds): A test of proposed barcode regions. *Taxon, 10*, 251–260.

Kobayasi, H., Kobayashi, H., & Idei, M. (1985). Fine structure and taxonomy of the small and tiny *Stephanodiscus* (Bacillariophyceae) species in Japan. 3. Co-occurrence of *Stephanodiscus minutulus* (Kütz.) Round and *S. parvus* Stoerm. & Håk. *Japanese Journal of Phycology, 33*, 293–300.

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*, 391–407.

Kress, W. J., & Erickson, D. L. (2007). A two-locus global DNA barcode for landplants: the coding *rbcL* gene complements the non-coding *trnH-psbA* spacer region. *PLoS Biology, 2*, e508. doi:10.1371/journal.pone.0000508.

Kress, W. J., Wurdack, K. J., Zimmer, E. A., Weigt, L. A., & Janzen, D. H. (2005). Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences of the USA, 102*, 8369–8374.

Kucera, H., & Saunders, G. W. (2008). Assigning morphological variance of *Fucus* (Fucales, Phaeophyceae) in Canadian waters to recognized species using DNA barcoding. *Botany, 86*, 1065–1079.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics, 23*, 2947–2948.

Lefèvre, E., Bardot, C., Noël, C., Carrias, J., Viscogliosi, C., Amblard, C., et al. (2007). Unveiling fungal zooflagellates as members of freshwater picoeukaryotes: Evidence from a molecular diversity study in a deep meromictic lake. *Environmental Microbiology, 9*, 61–71.

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

Lewis, L. A., & Flechtner, V. R. (2004). Cryptic species of *Scenedesmus* (Chlorophyta) from desert soil communities of western North America. *Journal of Phycology, 40*, 1127–1137.

Liao, P. C., Huang, B. H., & Huang, S. (2007). Microbial community composition of the Danshui River estuary of northern Taiwan and the practicality of the phylogenetic method in microbial barcoding. *Microbial Ecology, 54*, 497–507.

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

Lynn, D. H., & Strüder-Kypke, M. C. (2006). Species of *Tetrahymena* identical by small subunit rRNA gene sequences are discriminated by mitochondrial cytochrome c oxidase I gene sequences. *Journal of Eukaryotic Microbiology, 53*, 385–387.

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

McManus, H. A., & Lewis, L. A. (2005). Molecular phylogenetics, morphological variation, and colony-form evolution in the family Hydrodictyaceae (Sphaeropleales, Chlorophyta). *Phycologia, 44*, 582–595.

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*, 245–270.

Medlin, L. K., Elwood, H. J., Stickel, S., & Sogin, M. L. (1991). Morphological and genetic variation within the diatom *Skeletonema costatum* (Bacillariophyta): Evidence for a new species, *Skeletonema pseudocostatum*. *Journal of Phycology, 27*, 514–524.

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*, 67–75.

Messing, J. (1983). New M13 vectors for cloning. *Methods in Enzymology, 101*, 20–78.

Meyer, C. P., & Paulay, G. (2005). DNA barcoding: Error rates based on comprehensive sampling. *PLoS Biology, 3*, e422. doi:10.1371/journal.pbio.0030422.

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

Moniz, M. B. J., & Kaczmarska, I. (2010). Barcoding of diatoms: Nuclear encoded ITS revisited. *Protist, 161*, 7–34.

Morales, E. A., Siver, P. A., & Trainor, F. R. (2001). Identification of diatoms (Bacillariophyceae) during ecological assessments: Comparison between light microscopy and scanning electron microscopy techniques. *Proceedings of the Academy of Natural Sciences of Philadelphia, 151*, 95–103.

Moritz, C., & Cicero, C. (2004). DNA barcoding: Promise and pitfalls. *PLoS Biology, 2*, 1529–1531.

Müller, K. (2005). SeqState – primer design and sequence statistics for phylogenetic DNA data sets. *Applied Bioinformatics, 4*, 65–69.

Müller, T., Philippi, N., Dandekar, T., Schultz, J., & Wolf, M. (2007). Distinguishing species. *RNA, 13*, 1469–1472.

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

Newmaster, S. G., Fazekas, A., Steeves, R., & Janovec, J. (2008). Testing candidate plant barcode regions in the Myristicaceae. *Molecular Ecology Resources, 8*, 480–490.

Nickrent, D. L., & Sargent, M. L. (1991). An overview of the secondary structure of the V4 region of eukaryotic small-subunit ribosomal RNA. *Nucleic Acid Research, 19*, 227–235.

Poulíčková, A., Špačková, J., Kelly, M. G., Duchoslav, M., & Mann, D. G. (2008). Ecological variation within *Sellaphora* species complexes (Bacillariophyceae): Specialists or generalists? *Hydrobiologia, 614*, 373–386.

Powers, T. (2004). Nematode molecular diagnostics: From bands to barcodes. *Annual Review of Phytopathology, 42*, 367–38.

R Development Core Team. (2005). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Ratnasingham, S., & Hebert, P. D. N. (2007). BOLD: The barcode of life data system. *Molecular Ecology Notes, 7*, 355–364.

Ravin, N. V., Galachyants, Y. P., Merdanov, A. V., Beletsky, A. V., Petrova, D. P., Sherbakova, T. A., et al. (2010). Complete sequence of the mitochondrial genome of a diatom alga *Synedra acus* and comparative analysis of diatom mitochondrial genomes. *Current Genetics, 56*, 215–223.

Robba, L., Russell, S. J., Barker, G. L., & Brodie, J. (2006). Assessing the use of the mitochondrial *cox*1 marker for use in DNA barcoding of red algae (Rhodophyta). *American Journal of Botany, 93*, 1101–1108.

Romari, K., & Vaulot, D. (2004). Composition and temporal variability of picoeukaryote communities at a coastal site of the English Channel from 18S rDNA sequences. *Limnology and Oceanography, 49*, 784–798.

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

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*, 151–176.

Saunders, G. W. (2005). Applying DNA barcoding to red macroalgae: a preliminary appraisal holds promise for future application. *Philosophical Transactions of the Royal Society of London, Biological Sciences, 360*, 1879–1888.

Saunders, G. W. (2008). A DNA barcode examination of the red algal family Dumontiaceae in Canadian waters reveals substantial cryptic species diversity. 1. The foliose *Dilsea-Neodilsea* complex and *Weeksia*. *Botany, 86*, 773–789.

Savin, M. C., Martin, J. L., Giewat, M., & Rooney-Varga, J. (2004). Plankton diversity in the Bay of Fundy as measured by morphological and molecular methods. *Microbial Ecology, 48*, 51–65.

Schloss, P. D. (2010). The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. *PLoS Computational Biology, 6*, e1000844. doi:10.1371/journal.pcbi.1000844.

Scicluna, S. M., Tawari, B., & Clark, C. G. (2006). DNA barcoding of *Blastocystis*. *Protist, 157*, 77–85.

Seifert, K. A., Samson, R. A., de Waard, J. R., Houbraken, J., Lévesque, C. A., Moncalvo, J. M., et al. (2007). Prospects for fungus identification using COI DNA barcodes, with *Penicillium* as a test case. *Proceedings of the National Academy of Sciences of the USA, 104*, 3901–3906.

Smetacek, V. (1999). Diatoms and the carbon ocean cycle. *Protist, 150*, 25–32.

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

Spamer, E. E., & Theriot, E. C. (1997). "*Stephanodiscus minutulus*", "*S. minutus*", and similar epithets in taxonomic, ecological, and evolutionary studies of modern and fossil diatoms (Bacillariophyceae: Thalassiosiraceae)—A century and a half of uncertain taxonomy and nomenclatural hearsay. *Proceedings of the Academy of Natural Sciences of Philadelphia, 148*, 231–272.

Stevenson, R. J., & Pan, Y. (1999). Assessing ecological conditions in rivers and streams with diatoms. In E. P. Stoermer & J. P. Smol (Eds.), *The diatoms: Applications to the environmental and earth sciences* (pp. 11–40). Cambridge: Cambridge University Press.

Stoeckle, M. (2003). Taxonomy, DNA and the barcode of life. *Bioscience, 53*, 2–3.

Stoermer, E. P., & Smol, J. P. (1999). *The diatoms: Applications to the environmental and earth sciences*. Cambridge: Cambridge University Press.

Summerbell, R. C., Lévesque, C. A., Seifert, K. A., Bovers, M., Fell, J. W., Diaz, M. R., et al. (2005). Microcoding: The second step in DNA barcoding. *Philosophical Transactions of the Royal Society of London, Biological Sciences, 360*, 1897–1903.

Swofford, D. L. (2002). *PAUP*: Phylogenetic Analyses Using Parsimony (* and other methods). 4.0 beta*. Sunderland: Sinauer Associates.

Taylor, J., Bruns, T., & Lutzoni, F. (2008). *ITS as the fungal barcode*. http://www.allfungi.com/its-barcode.php. Accessed 30 December 2010.

Teubner, K. (1997). Merkmalsvariabilität bei planktischen Diatomeen in Berlin-Brandenburger Gewässern. *Nova Hedwigia, 65*, 233–250.

Utz, L. R., & Eizirik, E. (2007). Molecular phylogenies of subclass Peritrichia (Ciliophora: Oligohymenophorea) based on expanded analyses of 18S rRNA sequences. *Journal of Eukaryotic Microbiology, 54*, 303–305.

Vanelslander, B., Créach, V., Vanormelingen, P., Ernst, A., Chepurnov, V. A., Sahan, E., et al. (2009). Ecological differentiation between sympatric pseudocryptic species in the estuarine benthic diatom *Navicula phyllepta* (Bacillariophyceae). *Journal of Phycology, 45*, 1278–1289.

Ward, R., Zemlack, T. S., Innes, B. H., Last, P. R., & Hebert, P. D. N. (2005). DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society of London, Biological Sciences, 360*, 1847–1857.

Wolf, M., Scheffler, W., & Nicklisch, A. (2002). *Stephanodiscus neoastraea* and *Stephanodiscus heterostylus* (Bacillariophyceae) are one and the same species. *Diatom Research, 17*, 445–451.

Wu, S. G., Wang, G. T., Xi, B. W., Gao, D., & Nie, P. (2008). Molecular characteristics of *Camallanus* spp. (Spirurida: Camallanidae) in fishes from China based on its rDNA sequences. *Journal of Parasitology, 94*, 731–736.

Xia, X. H., Xie, Z., & Kjer, K. M. (2003). 18S ribosomal RNA and tetrapod phylogeny. *Systematic Biology, 52*, 283–295.

Zechman, F. W., Zimmer, E. A., & Theriot, E. C. (1994). Use of ribosomal DNA internal transcribed spacers for phylogenetic studies in diatoms. *Journal of Phycology, 30*, 507–512.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research, 31*, 3406–3415.

CBOL PROTIST WORKING GROUP: BARCODING EURAKYOTIC RICHNESS BEYOND THE ANIMAL, PLANT, AND FUNGAL KINGDOMS

JAN PAWLOWSKI
STÉPHANE AUDIC
SINA ADL
DAVID BASS
LASSAÂD BELBAHRI
CÉDRIC BERNEY
SAMUEL S BOWSER
IVAN CEPICKA
JOHAN DECELLE
MICAH DUNTHORN
ANNA M FIORE-DONNO
GILLIAN H GILE
MARIA HOLZMANN
REGINE JAHN
MILOSLAV JIRKŮ
PATRICK J KEELING
MARTIN KOSTKA
ALEXANDER KUDRYAVTSEV
ENRIQUE LARA
JULIUS LUKEŠ
DAVID G MANN
EDWARD A MITCHELL
FRANK NITSCHE
MARIA ROMERALO
GERY W SAUNDERS
ALASTAIR G SIMPSON
ALEXEY V SMIRNOV
JOHN L SPOUGE
ROWENA F STERN
THORSTEN STOECK
JONAS ZIMMERMANN
DAVID SCHINDEL
COLOMBAN DE VARGAS

2012
PLoS Biology;10(11):e1001419

OPEN ACCESS Freely available online

PLOS | BIOLOGY

Community Page

CBOL Protist Working Group: Barcoding Eukaryotic Richness beyond the Animal, Plant, and Fungal Kingdoms

Jan Pawlowski[1]*, Stéphane Audic[2], Sina Adl[3], David Bass[4], Lassaâd Belbahri[5], Cédric Berney[4], Samuel S. Bowser[6], Ivan Cepicka[7], Johan Decelle[2], Micah Dunthorn[8], Anna Maria Fiore-Donno[9], Gillian H. Gile[10], Maria Holzmann[1], Regine Jahn[11], Miloslav Jirků[12], Patrick J. Keeling[13], Martin Kostka[12,14], Alexander Kudryavtsev[1,15], Enrique Lara[5], Julius Lukeš[12,14], David G. Mann[16], Edward A. D. Mitchell[5], Frank Nitsche[17], Maria Romeralo[18], Gary W. Saunders[19], Alastair G. B. Simpson[20], Alexey V. Smirnov[15], John L. Spouge[21], Rowena F. Stern[22], Thorsten Stoeck[8], Jonas Zimmermann[11,23], David Schindel[24], Colomban de Vargas[2]*

1 Department of Genetics and Evolution, University of Geneva, Geneva, Switzerland, **2** Centre National de la Recherche Scientifique, Unité Mixte de Recherche 7144 and Université Pierre et Marie Curie, Paris 6, Station Biologique de Roscoff, France, **3** Department of Soil Science, University of Saskatchewan, Saskatoon, Saskatchewan, Canada, **4** Department of Life Sciences, Natural History Museum, London, United Kingdom, **5** Laboratory of Soil Biology, University of Neuchâtel, Neuchâtel, Switzerland, **6** Wadsworth Center, New York State Department of Health, Albany, New York, United States of America, **7** Department of Zoology, Charles University in Prague, Prague, Czech Republic, **8** Department of Ecology, University of Kaiserslautern, Kaiserslautern, Germany, **9** Institute of Botany and Landscape Ecology, University of Greifswald, Greifswald, Germany, **10** Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, Canada, **11** Botanischer Garten und Botanischer Museum Berlin-Dahlem, Freie Universität Berlin, Berlin, Germany, **12** Institute of Parasitology, Biology Centre, Czech Academy of Sciences, České Budějovice, Czech Republic, **13** Canadian Institute for Advanced Research, Botany Department, University of British Columbia, Vancouver, British Columbia, Canada, **14** Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic, **15** Department of Invertebrate Zoology, St-Petersburg State University, St-Petersburg, Russia, **16** Royal Botanic Garden Edinburgh, Edinburgh, United Kingdom, **17** Allgemeine Ökologie, Universität zu Köln, Köln, Germany, **18** Department of Systematic Biology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden, **19** Department of Biology, University of New Brunswick, Fredericton, New Brunswick, Canada, **20** Department of Biology, Life Sciences Centre, Halifax, Nova Scotia, Canada, **21** National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Computational Biology Branch, Bethesda, Maryland, United States of America, **22** Sir Alister Hardy Foundation for Ocean Science, Citadel Hill, Plymouth, United Kingdom, **23** Justus-Liebig-University, Giessen, Germany, **24** Smithsonian Institution, National Museum of Natural History, Washington, DC, United States of America

Animals, plants, and fungi—the three traditional kingdoms of multicellular eukaryotic life—make up almost all of the visible biosphere, and they account for the majority of catalogued species on Earth [1]. The remaining eukaryotes have been assembled for convenience into the *protists*, a group composed of many diverse lineages, single-celled for the most part, that diverged after Archaea and Bacteria evolved but before plants, animals, or fungi appeared on Earth. Given their single-celled nature, discovering and describing new species has been difficult, and many protistan lineages contain a relatively small number of formally described species (Figure 1A), despite the critical importance of several groups as pathogens, environmental quality indicators, and markers of past environmental changes. It would seem natural to apply molecular techniques such as DNA barcoding to the taxonomy of protists to compensate for the lack of diagnostic morphological features, but this has been hampered by the extreme diversity within the group. The genetic divergence observed between and within major protistan groups greatly exceeds that found in each of the three multicellular kingdoms. No single set of molecular markers has been identified that will work in all lineages, but an international working group is now close to a solution. A universal DNA barcode for protists coupled with group-specific barcodes will enable an explosion of taxonomic research that will catalyze diverse applications.

The undiscovered species diversity among protists may be orders of magnitude greater than previously thought. Surveys of protistan environmental diversity usually based on Sanger sequencing of polymerase chain reaction-amplified 18S rDNA clone libraries revealed an extremely high proportion of sequences that could not be assigned to any described species and in some cases even suggested the presence of several new eukaryotic kingdoms [2,3]. Although some of these sequences have since been shown to be chimeric or long-branch attraction artefacts (caused by

The Community Page is a forum for organizations and societies to highlight their efforts to enhance the dissemination and value of scientific knowledge.

Citation: Pawlowski J, Audic S, Adl S, Bass D, Belbahri L, et al. (2012) CBOL Protist Working Group: Barcoding Eukaryotic Richness beyond the Animal, Plant, and Fungal Kingdoms. PLoS Biol 10(11): e1001419. doi:10.1371/journal.pbio.1001419

Published November 6, 2012

Funding: The initial phase of the Protist Working Group activities presented in this paper were supported by the Consortium for the Barcoding of Life, the European ERA-net program BiodivErsA, under the BioMarKs project, the French ANR project 09-BLAN-0348 POSEIDON, and the Swiss National Science Foundation 31003A-140766. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: Jan.Pawlowski@unige.ch (JP); vargas@sb-roscoff.fr (CdV)

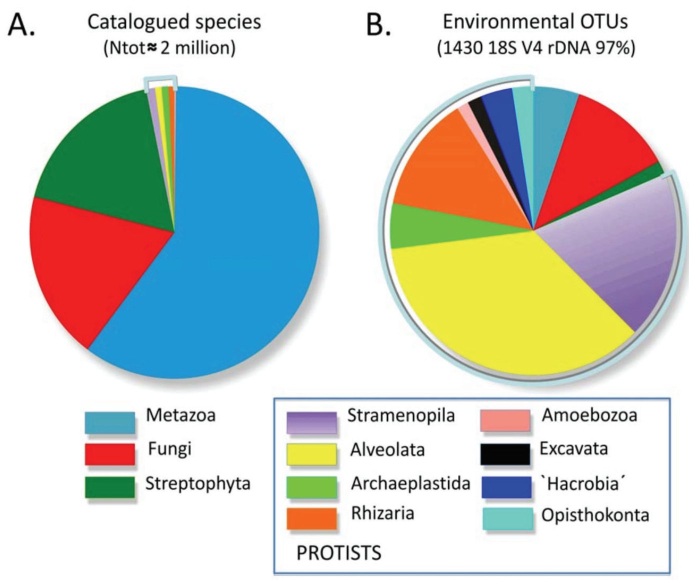


Figure 1. Morphological versus genetic views of total eukaryotic diversity. (A) Relative numbers of described species per eukaryotic supergroup—see Table S1 for a detailed count per division/class. (B) Relative number of V4 18S rDNA Operational Taxonomic Units (97%) per eukaryotic supergroup, based on 59 rDNA clone library surveys of marine, fresh-water, and terrestrial total eukaryotic biodiversity (as listed in [55]).
doi:10.1371/journal.pbio.1001419.g001

heterogeneity of evolutionary rates) [4], novel protistan phyla continue to be discovered (e.g., [5,6]). More recently, the growing number of Next Generation Sequencing (NGS) studies of eukaryotic diversity [6–9] has confirmed that the evolutionary and ecological importance of protists is much higher than traditionally thought (Figure 1B) and suggest that the number of protist species may easily exceed one million, although the correct estimation depends on many factors discussed below. The flow of eukaryotic sequence data produced by NGS from environmental DNA extracts is exponentially increasing, but there is currently no way to interpret these sequences in terms of species diversity and ecology.

DNA barcoding is a technique that uses a short standardized DNA region to identify species [10]. Large public reference libraries of DNA barcodes are being developed for animals, plants, and fungi, but there is no general agreement on which region to use for protists. Identifying the standard barcode regions for protists and assembling a reference library are the main objectives of the Protist Working Group (ProWG), initiated by the Consortium for the Barcode of Life (CBOL, http://www.barcodeoflife.org/). The ProWG unites a panel of international experts in protist taxonomy and ecology, with the aim to assess and unify the efforts to identify the barcode regions across all protist lineages, create an integrated plan to finalize the selection, and launch projects that would populate the reference barcode library. Here, we discuss the advantages and limitations of DNA barcodes currently in use and introduce a two-steps barcoding approach to assess protistan biodiversity.

The Unknown Vastness of Protist Richness

The first task of the protist barcoding initiative is to assess species richness in all protistan supergroups. In historically well-studied and biologically well-known taxa, such as higher plants or vertebrates, the number of predicted and described species is relatively similar. The situation is diametrically different for the fungi, for which *catalogued* species comprise ~7% of the predicted species number [1]. It is even worse for protists. The number of catalogued protistan species is very low in comparison to the diversity of animals, plants, and fungi, ranging from ~26,010 excluding marine nonphotosynthetic protists [1] to ~43,000 [11] and ~74,400 for the novel *ProWG* estimates presented herein (Table S1). Among the seven protistan supergroups (Figure 2A), the most diverse are Stramenopiles, with ~25,000 morphospecies. Over 10,000 described species are also found in Alveolata, Rhizaria, and Archaeplastida (excluding land-plants). Much fewer species have been catalogued for Amoebozoa (~2,400), Excavata (~2,300), and for the unicellular Opisthokonta (~300)—this latter group being dominated by animals and fungi.

The *predicted* richness of protistan species ranges from 1.4×10^5 to 1.6×10^6 [12]. In several groups, the number of predicted species has been arbitrarily estimated to be twice the number of described species [12]. But the true number of species could be several orders of magnitude higher. For example, the Apicomplexa are obligatory parasites, including the malaria agent *Plasmodium* and omnipresent *Toxoplasma*, and thus could reach up to 1.2×10^6 species if we assume a strict specificity to their metazoan hosts. The same argument can be applied to predict extreme species richness in protistan parasites of fishes (e.g., Mesomycetozoa) and plants (e.g., Oomycetes). However, most of these predictions are highly subjective.

Moreover, just like in Bacteria and Archaea [13,14], there is no general agreement on how to define species in protists, and no single species concept can be applied unequivocally to all protistan groups. Molecular studies typically reveal a multitude of genotypes hidden within protist species that have been discovered and described using traditional methods based on morphological criteria (often referred to as "morphospecies"). Reproductive isolation could theoretically be used in differentiating eukaryotic species, but data on the very existence of a sexual phase are very sparse in protists. Mating studies in some "model" systems (e.g., [15,16]) are consistent with the evidence from molecular data that protistan species diversity is greatly underestimated by classical morphological approaches. Overall intraspecific and intragenomic variabilities in environmental protistan populations are still largely unknown, because most genetic studies are carried out on clonal strains maintained in laboratory cultures.

Protist Barcoding: State of the Art

Although the term *DNA barcoding* appeared only recently in the protistological

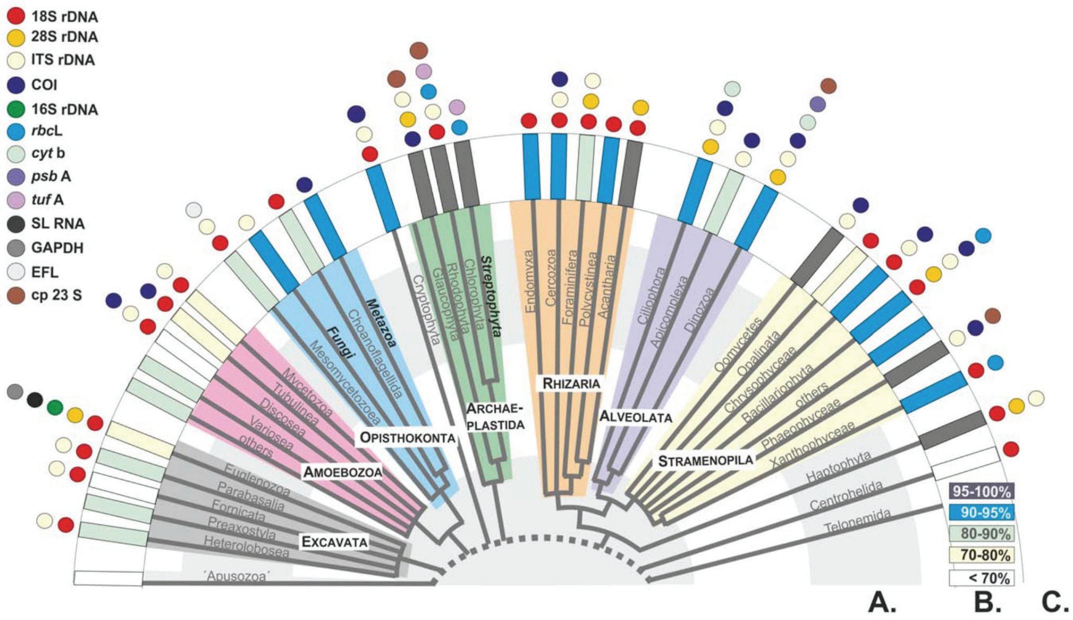


Figure 2. Current state-of-the art phylogeny and barcode markers for the main protistan lineages. (A) A recent phylogeny of eukaryotic life, after [56]. (B) Mean V4 18S rDNA genetic similarity between all congeneric species within each lineage, available in GenBank. (C) Currently used group-specific barcodes. The dashed line indicates the incertitude concerning the position of the root in the tree of eukaryotic life. The unresolved relationships between eukaryotic groups are indicated by polytomies. The names of the three multicellular classical "kingdoms" are highlighted.
doi:10.1371/journal.pbio.1001419.g002

literature, the identification of protistan taxa using molecular markers has a long history. The most commonly used markers have been parts of the genes coding for ribosomal RNAs, in particular 18S rDNA (e.g., [17]). The advantages of 18S rDNA are many: found in all eukaryotes, it occurs in many copies per genome, allowing genetic work at the individual (single-cell) level; it is highly expressed, permitting molecular ecological investigation at the RNA level; and it includes a mosaic of highly conserved and variable nucleotide sequences allowing combined phylogenetic reconstruction and biota recognition at various taxonomic levels. Different 18S rDNA variable regions have been used in clone libraries and NGS-based environmental surveys [3,7,18]. 18S rDNA barcodes have been shown to effectively distinguish species in some groups, such as foraminifera [19,20] and some diatoms [21], however they are not sufficiently variable to resolve interspecies relationships in several other taxa (Figure 2B).

Various alternative protistan DNA barcodes have been proposed (Figure 2, Table S2). The D1–D2 and/or D2–D3 regions at the 5′ end of 28S rDNA have been positively tested in ciliates [22], haptophytes [23], and acantharians [24] and are also promising for diatoms [25,26]. Ribosomal internal transcribed spacers (ITS1 and/or ITS2 rDNA), which are the main fungal barcodes [27], are also commonly utilized in oomycetes [28], chlorarachniophytes [29], and green algae [30] and have also been suggested for dinoflagellates [31,32] and diatoms [33] with some reserve [34]. The mitochondrial gene coding for cytochrome oxidase 1 (COI), which has been proposed as the universal barcode for animals [10], also allows morpho-species identification in red [35–37] and brown [38,39] algae, dinoflagellates [40], some raphid diatoms [41], Euglyphida [42], lobose naked [43] and shelled [44] amoebae, coccolithophorid haptophytes [45], and some ciliates [46,47]. Other group-specific barcodes include the large subunit of the ribulose-1,5-biphosphate carboxylase–oxygenase gene (*rbc*L) and the chloroplastic 23S rRNA gene for photosynthetic protists [25,48–50], and Spliced Leader RNA genes for trypanosomatids [51]. Clearly, the choice of group-specific barcodes is often a question of tradition or ease of use, and studies systematically comparing the resolution power of different protistan DNA barcodes are rare [25,42,43,52].

ProWG Objectives and Perspectives

The ultimate objective of the CBOL ProWG is to establish universal criteria for barcode-based species identification in protists. The DNA barcoding approach has several well-known limitations related to the standardization of species identification [53,54], and addressing some of the challenges raised by genetic identification of protists will certainly require more fundamental research on protistan speciation. The ProWG will organise workshops and seminaries that will provide opportunity to discuss general questions concerning species definition, genetic variations, and applications of DNA barcodes in all protistan groups.

From a practical perspective, the ProWG mission is to establish the genetic standards that will allow recognition of protistan taxa exclusively on the basis of DNA sequence data. Our goal is not to exclude morphological identification but to propose alternative tools that will be more efficient in dealing with the immense protistan biodiversity and more objective and accessible to nonspecialists. In most protistan groups, morphological characters are unreliable for identification at the species level but do provide guides for higher level taxonomic assignments, as well as valuable information about the biology, ecology, and evolution of organisms. Therefore, every protistan reference DNA barcode must be associated with voucher material and/or illustrations providing phenotypic data from the barcoded specimen.

Because of their long, independent, and complex evolutionary histories, protists are so genetically variable that it is virtually impossible to find a single universal DNA barcode suitable for all of them. The ProWG consortium therefore recommends a two-step barcoding approach, comprising a preliminary identification using a universal eukaryotic barcode, called the pre-barcode, followed by a species-level assignment using a group-specific barcode (Figure 3). In this nested strategy, the ~500 bp variable V4 region of 18S rDNA is proposed as the universal eukaryotic pre-barcode. Group-specific barcodes (Figure 2C) will then have to be defined separately for each major protistan group, based on comparative studies using the CBOL selection criteria, and much of this work is still to be done. Depending on the type of material (isolates and cultures) and whether or not DNA extraction is destructive for the analysed species, the morphological appearance of each barcoded protist will be preserved as microphotographs, fixed cells, or live and/or cryopreserved cultures. This voucher would be deposited in a public collection, just as type specimens are required for new taxa by the nomenclatural codes. Collection details including locality, date, and (as far as possible) habitat characteristics must also be provided, accompanied in parasitic and symbiotic taxa by an accurately identified host voucher or its DNA/tissue sample wherever this is available. Moreover, the extracted DNA must be deposited in a recognized DNA bank or museum collection and cited with a unique identifier to allow checks and further genetic analyses.

Most of these recommendations are already followed where newly described protistan species are based on cultured strains deposited in collections. However, the large majority of protists are currently uncultivable by known means or not available in culture collections, and genetic data only exist for a very small fraction of described species. Therefore, it is imperative to establish standard barcoding protocols for future protist barcoding projects that will substantially increase the number of collected, described, but uncultivable protists. A combination of novel high-throughput imaging/sorting with newer genetic technologies—including single-amplified-genome methods—opens exciting avenues in protistan metabarcoding. A protist barcoding protocol such as that outlined in Figure 3 will allow collection of the data necessary to set up a representative protist species reference library. The protocols and recommendations concerning protist barcoding will be available at the *ProWG* website (under construction at www.protistbarcoding.org), and a platform dedicated to protist multi-locus barcodes will be accessible at the Barcode of Life Database.

Given the ongoing DNA sequencing revolution, the 21st-century exploration of biodiversity must do more than document the higher macrofaunal and macrofloral branches on the Tree of Life. Amongst other microbes, protists are key but poorly known elements of the ecosystems we see in Nature, including the complex microbiomes hidden within individual plants, animals, and fungi. Ecological models must include protists based on the new knowledge of their species-level diversity that will mostly come from the billions of NGS-generated environmental barcodes. The reference library of standard protistan barcodes will be the Rosetta stone that makes protist diversity less anonymous.

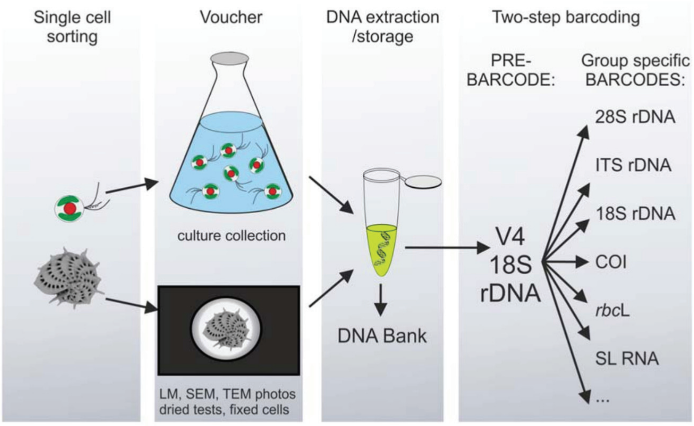


Figure 3. Two-step protist barcoding pipeline. Protistan species, spanning four orders of cell-size magnitude (from <1 μm to >10,000 μm), are individually sorted from the environment, phenotyped either directly or after culture growth, DNA extracted, and barcoded using a two-step, nested strategy.
doi:10.1371/journal.pbio.1001419.g003

Supporting Information

Table S1 Number of catalogued morphospecies and V4 18S rDNA OTU-97% among the 60 main eukaryotic lineages.
(PDF)

Table S2 Group-specific barcodes for selected genera representing all eukaryotic supergroups (in brackets, number of corresponding sequences in the GenBank). NM, nucleomorph origin. Variable regions used in 18S and 28S genes are indicated in some cases.
(PDF)

References

1. Mora C, Tittensor DP, Adl S, Simpson AG, Worm B (2011) How many species are there on Earth and in the ocean? PLoS Biol 9: e1001127. doi:10.1371/journal.pbio.1001127
2. Dawson SC, Pace NR (2002) Novel kingdom-level eukaryotic diversity in anoxic environments. Proc Natl Acad Sci U S A 99: 8324–8329.

3. López-García P, Rodríguez-Valera F, Pedrós-Alió C, Moreira D (2001) Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. Nature 409: 603–607.
4. Berney C, Fahrni J, Pawlowski J (2004) How many novel eukaryotic "kingdoms"? Pitfalls and limitations of environmental DNA surveys. BMC Biol 2: 13.
5. Kim E, Harrison JW, Sudek S, Jones MD, Wilcox HM, et al. (2011) Newly identified and diverse plastid-bearing branch on the eukaryotic tree of life. Proc Natl Acad Sci U S A 108: 1496–1500.
6. Behnke A, Engel M, Christen R, Nebel M, Klein RR, et al. (2011) Depicting more accurate pictures of protistan community complexity using pyrosequencing of hypervariable SSU rRNA gene regions. Environ Microbiol 13: 340–349.
7. Edgcomb V, Orsi W, Bunge J, Jeon S, Christen R, et al. (2011) Protistan microbial observatory in the Cariaco Basin, Caribbean. I. Pyrosequencing vs Sanger insights into species richness. ISME J 5: 1344–1356.
8. Lecroq B, Lejzerowicz F, Bachar D, Christen R, Esling P, et al. (2011) Ultra-deep sequencing of foraminiferal microbarcodes unveils hidden richness of early monothalamous lineages in deep-sea sediments. Proc Natl Acad Sci U S A 108: 13177–13182.
9. Logares R, Audic S, Santini S, Pernice MC, de Vargas C, et al. (2012) Diversity patterns and activity of uncultured marine heterotrophic flagellates unveiled with pyrosequencing. ISME J 6:1823–1833.
10. Hebert PDN, Cywinska A, Ball SL, DeWaard JR (2003) Biological identifications through DNA barcodes. Proceedings of the Royal Society of London Series B-Biological Sciences 270: 313–321.
11. Hoef-Emden K, Kupper FC, Andersen RA (2007) Meeting report: Sloan Foundation Workshop to resolve problems relating to the taxonomy of microorganisms and to culture collections arising from the barcoding initiatives; Portland ME, November 6–7, 2006. Protist 158: 135–137.
12. Adl SM, Simpson AG, Farmer MA, Andersen RA, Anderson OR, et al. (2007) The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. J Eukaryot Microbiol 57: 189–196.
13. Doolittle WF, Papke RT (2006) Genomics and the bacterial species problem. Genome Biology 7: 116.
14. Caro-Quintero A, Konstantinidis KT (2012) Bacterial species may exist, metagenomics reveal. Environ Microbiol 14: 347–355.
15. Coleman AM (1977) Sexual and genetic isolation in the cosmopolitan algal species *Pandorina morum*. American Journal of Botany 64: 361–368.
16. Amato A, Kooistra WHCF, Ghiron JHL, Mann DG, Proschold T, et al. (2007) Reproductive isolation among sympatric cryptic species in marine diatoms. Protist 158: 193–207.
17. Stothard DR, Schroeder-Diedrich JM, Awwad MH, Gast RJ, Ledee DR, et al. (1998) The evolutionary history of the genus Acanthamoeba and the identification of eight new 18S rRNA gene sequence types. J Eukaryot Microbiol 45: 45–54.
18. Stoeck T, Bass D, Nebel M, Christen R, Jones MD, et al. (2010) Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. Mol Ecol 19: 21–31.
19. Pawlowski J, Lecroq B (2010) Short rDNA barcodes for species identification in foraminifera. Journal of Eukaryotic Microbiology 57: 197–205.
20. de Vargas C, Norris R, Zaninetti L, Gibb SW, Pawlowski J (1999) Molecular evidence of cryptic speciation in planktonic foraminifers and their relation to oceanic provinces. Proc Natl Acad Sci U S A 96: 2864–2868.
21. Zimmermann J, Jahn R, Gemeinholzer B (2011) Barcoding diatoms: evaluation of the V4 subregion on the 18S rRNA gene, including new primers and protocols. Org Divers Evol 11(3): 173–192.
22. Gentekaki E, Lynn DH (2009) High-level genetic diversity but no population structure inferred from nuclear and mitochondrial markers of the peritrichous ciliate *Carchesium polypinum* in the Grand River basin (North America). Appl Environ Microbiol 75: 3187–3195.
23. Liu H, Probert I, Uitz J, Claustre H, Aris-Brosou S, et al. (2009) Extreme diversity in noncalcifying haptophytes explains a major pigment paradox in open oceans. Proc Natl Acad Sci U S A 106: 12803–12808.
24. Decelle J, Suzuki N, Mahe F, de Vargas C, Not F (2012) Molecular phylogeny and morphological evolution of the acantharia (radiolaria). Protist 163: 435–450.
25. Hamsher SE, Evans KM, Mann DG, Poulickova A, Saunders GW (2011) Barcoding diatoms: exploring alternatives to COI-5P. Protist 162: 405–422.
26. Trobajo R, Mann DG, Clavero E, Evans KM, Vanormelingen P, et al. (2010) The use of partial cox1, rbcL and LSU rDNA sequences for phylogenetics and species identification within the *Nitzschia palea* complex (Bacillariophyceae). Eur J Phycol 45: 413–425.
27. Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, et al. (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proc Natl Acad Sci U S A 109: 6241–6246.
28. Robideau GP, De Cock AW, Coffey MD, Voglmayr H, Brouwer H, et al. (2011) DNA barcoding of oomycetes with cytochrome c oxidase subunit I and internal transcribed spacer. Mol Ecol Resour 11: 1002–1011.
29. Gile GH, Stern RF, James ER, Keeling PJ (2010) DNA barcoding of Chlorarachniophytes using nucleomorph ITS sequences. J Phycol 46: 743–750.
30. Coleman AW (2000) The significance of a coincidence between evolutionary landmarks found in mating affinity and a DNA sequence. Protist 151: 1–9.
31. Litaker RW, Vandersea MW, Kibler SR, Reece KS, Stokes NA, et al. (2007) Recognizing Dinoflagellate species using its rDNA sequences. J Phycol 43: 344–355.
32. Stern RF, Andersen R.A., Jameson I., Küpper F.C., Coffroth M.A., et al. (2012) Evaluating the ribosomal internal transcribed spacer (ITS) as a candidate dinoflagellate barcode marker. PLoS One 7: e42780. doi:10.1371/journal.pone.0042780
33. Moniz MB, Kaczmarska I (2010) Barcoding of diatoms: nuclear encoded ITS revisited. Protist 161: 7–34.
34. Mann DG, Sato S, Trobajo R, Vanormelingen P, Souffreau C (2010) DNA barcoding for species identification and discovery in diatoms. Cryptogamie: Algologie 31: 557–577.
35. Le Gall L, Saunders G.W. (2010) DNA barcoding is a powerful tool to uncover algal diversity: a case study of the Phyllophoraceae (Gigartinales, Rhodophyta) in the Canadian flora. J Phycol 46: 374–389.
36. Saunders GW (2005) Applying DNA barcoding to red macroalgae: a preliminary appraisal holds promise for future applications. Philos Trans R Soc Lond B Biol Sci 360: 1879–1888.
37. Saunders GW (2008) A DNA barcode examination of the red algal family Dumontiaceae in Canadian waters reveals substantial cryptic species diversity. 1. The foliose Dilsea-Neodilsea complex and Weeksia. Botany 86: 773–789.
38. Kucera E, Saunders G.W. (2008) Assigning morphological variants of Fucus (Fucales, Phaeophyceae) in Canadian waters to recognized species using DNA barcoding. Botany 86: 1065–1079.
39. McDevit DC, Saunders GW (2009) On the utility of DNA barcoding for species differentiation among brown macroalgae (Phaeophyceae) including a novel extraction protocol. Phycological Research 57: 131–141.
40. Stern RF, Horak A, Andrew RL, Coffroth MA, Andersen RA, et al. (2010) Environmental barcoding reveals massive dinoflagellate diversity in marine environments. PLoS One 5: e13991. doi:10.1371/journal.pone.0013991
41. Evans KM, Wortley AH, Mann DG (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: 349–364.
42. Heger TJ, Pawlowski J, Lara E, Leander BS, Todorov M, et al. (2011) Comparing potential COI and SSU rDNA barcodes for assessing the diversity and phylogenetic relationships of cyphoderiid testate amoebae (Rhizaria: Euglyphida). Protist 162: 131–141.
43. Nassonova E, Smirnov A, Fahrni J, Pawlowski J (2010) Barcoding amoebae: comparison of SSU, ITS and COI genes as tools for molecular identification of naked lobose amoebae. Protist 161: 102–115.
44. Kosakyan A, Heger TJ, Leander BS, Todorov M, Mitchell EA, et al. (2012) COI barcoding of nebelid testate amoebae (Amoebozoa: Arcellinida): extensive cryptic diversity and redefinition of the hyalospheniidae schultze. Protist 163: 415–434.
45. Hagino K, Bendif El M., Young JR, Kogame K, Probert I, et al. (2011) New evidence for morphological and genetic variation in the cosmopolitan coccolithophore Emiliana huxleyi (Prymnesiophyceae) from the cox1b-ATP4 genes. J Phycol 47: 1164–1176.
46. Barth D, Krenek S, Fokin SI, Berendonk TU (2006) Intraspecific genetic variation in Paramecium revealed by mitochondrial cytochrome C oxidase I sequences. J Eukaryot Microbiol 53: 20–25.
47. Chantangsi C, Lynn DH, Brandl MT, Cole JC, Hetrick N, et al. (2007) Barcoding ciliates: a comprehensive study of 75 isolates of the genus Tetrahymena. Int J Syst Evol Microbiol 57: 2412–2425.
48. MacGillivary ML, Kaczmarska I (2011) Survey of the efficacy of a short fragment of the rbcL gene as a supplemental DNA barcode for diatoms. J Eukaryot Microbiol 58: 529–536.
49. Santos S, Gutierrez-Rodriquez C, Coffroth MA (2003) Phylogenetic identification of symbiotic dinoflagellates via length heteroplasmy in domain V of chloroplast large subunit (cp23S)—ribosomal DNA sequences. Mar Biotechnol (NY) 5(2): 130–140.
50. Sherwood AR, Prestling G.G. (2007) Universal primers amplify a 23S rDNA plastid marker in eukaryotic algae and cyanobacteria. J Phycol 43(3): 605–608.
51. Votypka J, Maslov DA, Yurchenko V, Jirku M, Kment P, et al. (2010) Probing into the diversity of trypanosomatid flagellates parasitizing insect hosts in South-West China reveals both endemism and global dispersal. Mol Phylogenet Evol 54: 243–253.
52. Saunders GW, Kucera H. (2010) An evaluation of rbcL, tufA, UPA, LSU and ITS as DNA barcode markers for the marine green macroalgae. Cryptogamie, Algologie 31: 487–528.
53. Moritz C, Cicero C (2004) DNA barcoding: promise and pitfalls. PLoS Biol 2: 1529–1531. doi:10.1371/journal.pbio.0020354
54. Taylor HR, Harris WE (2012) An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. Mol Ecol Resour 12: 377–388.
55. del Campo J, Massana R (2011) Emerging diversity within chrysophytes, choanoflagellates and bicosoecids based on molecular surveys. Protist 162(3): 435–448.
56. Burki F, Okamoto N, Pombert JF, Keeling PJ (2012) The evolutionary history of haptophytes and cryptophytes: phylogenomic evidence for separate origins. Proc Biol Sci 279:2246–2254.

TAXONOMIC REFERENCE LIBRARIES FOR ENVIRONMENTAL BARCODING: A BEST PRACTICE EXAMPLE FROM DIATOM RESEARCH

JONAS ZIMMERMANN

NÉLIDA ABARCA

OLIVER SKIBBE

WOLF-HENNING KUSBER

NEELA ENK

REGINE JAHN

2014

PLoS ONE 9(9): e108793

Taxonomic Reference Libraries for Environmental Barcoding: A Best Practice Example from Diatom Research

Jonas Zimmermann[1,2*¶], Nelida Abarca[1¶], Neela Enk[1¶], Oliver Skibbe[1,3], Wolf-Henning Kusber[1], Regine Jahn[1]

1 Botanic Garden and Botanical Museum Berlin-Dahlem, Freie Universität Berlin, Berlin, Germany, 2 Justus-Liebig-Universität Giessen, AG Spezielle Botanik, Giessen, Hessen, Germany, 3 Larger than life - micro and nature photography, http://www.larger-than-life.de, Berlin Germany

Abstract

DNA barcoding uses a short fragment of a DNA sequence to identify a taxon. After obtaining the target sequence it is compared to reference sequences stored in a database to assign an organism name to it. The quality of data in the reference database is the key to the success of the analysis. In the here presented study, multiple types of data have been combined and critically examined in order to create best practice guidelines for taxonomic reference libraries for environmental barcoding. 70 unialgal diatom strains from Berlin waters have been established and cultured to obtain morphological and molecular data. The strains were sequenced for 18S V4 rDNA (the pre-Barcode for protists) as well as *rbc*L data, and identified by microscopy. LM and for some strains also SEM pictures were taken and physical vouchers deposited at the BGBM. 37 freshwater taxa from 15 naviculoid diatom genera were identified. Four taxa from the genera *Amphora*, *Mayamaea*, *Planothidium* and *Stauroneis* are described here as new. Names, molecular, morphological and habitat data as well as additional images of living cells are also available electronically in the AlgaTerra Information System. All reference sequences (or reference barcodes) presented here are linked to voucher specimens in order to provide a complete chain of evidence back to the formal taxonomic literature.

Citation: Zimmermann J, Abarca N, Enk N, Skibbe O, Kusber W-H, et al. (2014) Taxonomic Reference Libraries for Environmental Barcoding: A Best Practice Example from Diatom Research. PLoS ONE 9(9): e108793. doi:10.1371/journal.pone.0108793

Editor: Bernd Schierwater, University of Veterinary Medicine Hanover, Germany

Received May 13, 2014; **Accepted** August 14, 2014; **Published** September 29, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are within the paper. The web link to AlgaTerra for all taxa, the DNA Bank number and the EMBL accession numberss are available in Table 2.

Funding: Deutsche Forschungsgemeinschaft (http://www.dfg.de/) grant INST 130/839-1 FUGG. Institution was funded. Deutsche Forschungsgemeinschaft (http://www.dfg.de/) grant GE 12 42/11-1 RJ; JZ is employed on this grant. German Federal Ministry of Education and Research (http://www.bmbf.de/) funding for microscopy data within the GBIF-D project 01 LI 1001 A. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: j.zimmermann@bgbm.org

¶ These authors are shared first authors on this work.

Introduction

Diatoms are unicellular and usually photoautotroph micro algae which are responsible for about 25% of global CO_2 fixation [1–3] and contribute approximately 20% of the global net primary production [4].

Diatoms are important bioindicators for monitoring water quality because they are sensitive to changes in pollution, nutrient availability, acidity and salinity, e.g. [5,6]. They are the most ubiquitous group within the microscopic algae as they occur in all types of water bodies and play an important part in benthic and planktonic biocoenoses [7]. They are routinely used as bioindicators within the EU Water Framework Directive (WFD) as well as in water quality monitoring worldwide [8–13].

Each diatom cell is encased in two siliceous shells (frustules) that are connected by girdle bands [1–3]. Current identification of diatoms is based on a morphological and mostly descriptive species concept (Zimmermann et al. subm.) and relies exclusively on micro-characters of the frustule such as size, symmetry, shape, and sculpture which can be seen by light microscopy [14]; more detailed analyses of the siliceous structures lead to more and more refined differentiation of species, which is possible through the development of higher resolution techniques, e.g. electron microscopy.

Identification via microscopy is challenging and time consuming, especially for routine use [15], and relies on individual taxonomic expertise. Therefore different taxonomists could arrive at different conclusions, depending i.a. on the taxonomic concept, species with limited diagnostic morphological features, cryptic species, available reference floras and quality of microscopes used by each individual researcher [15] as well as unavailability of adequate descriptions.

The application of molecular markers for taxon identification – DNA barcoding – is an emerging method which has the potential to be faster, universally applicable and generate reliable identification. Furthermore, as it uses DNA sequences for identification, it is independent of pre-existing morphological species concepts and can be linked to any taxonomic concept [16]. However, correct identification relies fundamentally on the quality of the reference library the DNA barcodes are checked against. DNA barcoding is based on the assumption that sequences of a certain marker locus exhibit enough variation between species to be discriminative for unambiguous species discovery [17,18]. DNA barcoding is also a useful tool to access concealed diversity e.g. [19–25]. DNA barcoding in combination with next generation sequencing techniques also allows for the description of community compositions through the large numbers of sequences generated by this approach e.g. [26,27]. A schematic overview on environmental DNA barcoding of diatoms and the establishment of a reference library is given in Fig. 1.

The requirement for reliable taxon identification by DNA barcode(s) is an unambiguous link between the genotype and the phenotype (or morphotype) to which the name of the species is attached. This means that a reference library consisting of taxon names belonging to specimens that have been identified by experts as well as providing descriptions together with barcode sequences, which were derived from well documented strains (e.g. voucher deposition, sampling localities and collectors, basic environmental data, high-resolution LM pictures, morphometrics, taxonomy and nomenclature, maps, literature and references to databases where this data is deposited) for every single species is necessary. For unicellular diatoms, clone cultures (strains) need to be established which offer enough material for sequencing as well as for identification by light and electron microscopy. Once established and linked to a taxonomic reference library, the DNA barcoding method could offer a time and cost efficient alternative/extension to microscopic identification for routine applications by limiting morphological taxonomy to critical groups which feature a distinct genetic aberration to known and identified organisms in the library.

Recently, the CBOL Protist Working Group [28] has designated the 18S V4 rDNA marker region as first or pre-barcode for Protist organisms. In this paper, we follow the 18S V4 protocols designed for diatoms by Zimmermann et al. [19], and present 70 strains for which this pre-barcode (18S V4) as well as a second widely used barcode, *rbc*L [20,21,29], has been generated. The reference library includes these two DNA barcodes, the respective taxon name, images, morphometric and geographic data as well as vouchers for further reference. Further data and additional images also of living cells are available electronically through the AlgaTerra Information System [30]. We demonstrate the benefits of a well documented reference library for DNA barcoding for identification, taxonomy, phylogeny, and further scientific analyses on an exemplary group. This paper focuses on naviculoid diatom strains from Berlin waters since its diatom flora has been well studied for almost two centuries by light microscopy [31] and a recent diatom flora is available for water quality assessments [32].

Materials and Methods

Sampling

Benthic samples from which the 70 strains were established were collected at 11 sites in the catchment area of Berlin (Fig. 2); one additional sample was from the River Elbe, downstream of the Berlin Rivers Spree and Havel. Conductivity of Berlin water ranges mostly between 400 to 900 $\mu S\ cm^{-1}$, pH is frequently 6,5 to 9 (80% respectively 88% of about 300 measurements of Berlin water samples, Kusber unpubl. data). For samples, sites, dates, collectors of the samples and isolators of the strains see Table 1. No specific permissions were required for the sampled locations/activities. The field studies did not involve endangered or protected species.

Cultivation

The diatom cells were isolated from environmental water samples observed under a stereo light microscope using capillary glass pipettes. The respective cell was then transferred to a 5 cm diameter plastic petri dish containing autoclaved habitat water and/or culture medium (WC [33], Chu [34], AlgaGrow, Plagron, Weert, Netherlands) of adequate salinity and pH. In order to remove unwanted particles, this treatment was repeated several times until microscopic inspection confirming that a culture derived from one cell, but not axenic had been established. The cultures were grown at a temperature between 18–22°C and a 12 h day/night cycle.

Preparation of frustules

By the time of harvesting the cultures, one fraction was used for obtaining DNA (see below) and the other part was cleaned with H_20_2 at 80°C and rinsed several times with H_20. A few drops of the resulting suspension of diatom frustules were dried on a cover slip and embedded as slides in Naphrax for study in LM or on stubs if for SEM. Vouchers of each strain were deposited in the Herbarium Berolinense (B) (see Table 2).

Light and electron microscopy

The LM pictures were acquired with a Zeiss Axio Imager.M2 with an implemented AxioCam HRc (Zeiss, Oberkochen, Germany). SEM pictures were produced with Philips SEM 515 operating at 30 KV (Philips, Eindhoven, The Netherlands), and Hitachi 8010 Field Emission Electron Microscope (Hitachi, Tokyo, Japan).

Identification

The taxa were identified with Hofmann et al. [32], Krammer & Lange-Bertalot (1997) [35], Ettl & Gärtner (2013) [36], Lange-Bertalot (2001), Levkov et al. (2009) [37], Levkov et al. (2014) [38] as well as particular papers (vide infra) for selected species. For strain numbers, taxon names, voucher codes in the Herbarium Berolinense (B), EMBL Accession Numbers, images, and morphometric data for all strains see Table 2.

DNA isolation

The harvested cultures were transferred to 1.5 ml tubes. DNA was isolated using Dynal DynaBeads (Invitrogen Corporation; Carlsbad, CA, USA), NucleoSpin Plant II Mini Kit (Machery and Nagel, Düren, Germany) or Qiagen Dneasy Plant Mini Kit (Qiagen Inc.; Valencia, CA) following the respective product instructions. DNA concentrations were checked using gel electrophoresis (1.5% agarose gel) and Nanodrop (PeqLab Biotechnology LLC; Erlangen, Germany). DNA samples were stored at −20°C until further use. DNA material was deposited in the Berlin collection of the DNA bank network [39].

PCR amplification

The V4 region of the 18S locus was amplified in all strains with the primer pair M13F-D512 for 18S/M13F-D978rev 18S [19]. The *rbc*L locus was amplified in two overlapping parts using two different primer pairs; Diat-rbcL-F and Diat-rbcL-iR as well as Diat-rbcL-iF and Diat-rbcL-R [40] for all strains. The polymerase chain reaction (PCR) for the V4 region was conducted after Zimmermann et al. (2011) [19] and for *rbc*L carried out after Abarca et al. (2014) [40]. PCR products were visualised in a 1.5% agarose gel and cleaned with MSB Spin PCRapace (Invitek LLC; Berlin, Germany) following standard procedure. DNA content was measured using Nanodrop (PeqLab Biotechnology). The samples

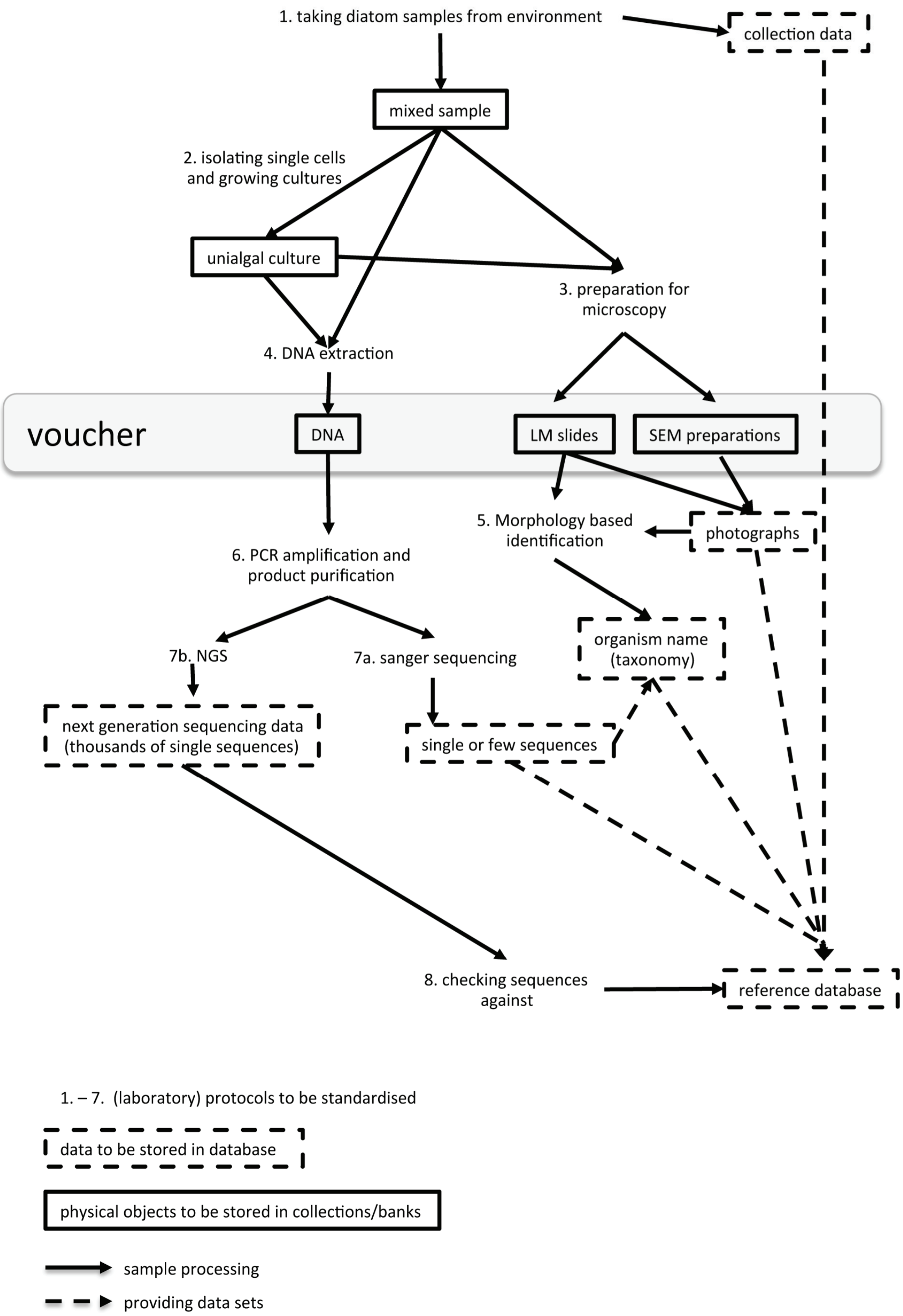

1. taking diatom samples from environment
collection data
mixed sample
2. isolating single cells
and growing cultures
unialgal culture
3. preparation for
microscopy
4. DNA extraction
voucher
DNA
LM slides
SEM preparations
5. Morphology based
identification
photographs
6. PCR amplification and
product purification
7b. NGS
7a. sanger sequencing
organism name
(taxonomy)
next generation sequencing data
(thousands of single sequences)
single or few sequences
8. checking sequences
against
reference database
1. – 7. (laboratory) protocols to be standardised
data to be stored in database
physical objects to be stored in collections/banks
sample processing
providing data sets

Figure 1. Schematic overview of sample processing and voucher as well as data production and deposition for environmental barcoding.
doi:10.1371/journal.pone.0108793.g001

were normalised to a total DNA content >100 ng/µl using Nanodrop (PeqLab Biotechnology) for further sequencing.

Sequencing

The Sanger sequencing was conducted by Starseq (GENterprise LLC; Mainz, Germany). As sequencing primers the M13 tails [19,41] were used for the V4 region, following [42]. The sequences were edited in PhyDE [43] aligned using MUSCLE [44], and alignments were manually improved in PhyDE [43].

Molecular analysis

The aligned sequences were compared to each other calculating uncorrected p distances in PAUP [45]. Then they were blasted against existing INSDC entries for the respective taxa (accessed July 2013). All INSDC accessions with references are given in Appendix S1. Base pair differences were counted in overlapping parts of the sequences in Mega 5 [46]. Results are summarised in Table 3.

Tree building

To identify molecular relations between the here presented strains, trees were calculated with Mega 5 using the Neighbour Joining algorithm with gamma distributed rates among sites followed by a statistical test of the tree topologies with 10 000 bootstrap replications. Trees for the individual alignments of 18S V4 and *rbc*L sets as well as a concatenated dataset were calculated.

Furthermore, we created 18S V4 as well as *rbc*L datasets including INSDC sequences for the genera *Amphora*, *Mayamaea*, *Planothidium* and *Stauroneis* to exemplarily test the taxonomic consistency of available sequences as well as the placement of our new taxa. Each of these eight datasets was analysed under the aforementioned conditions.

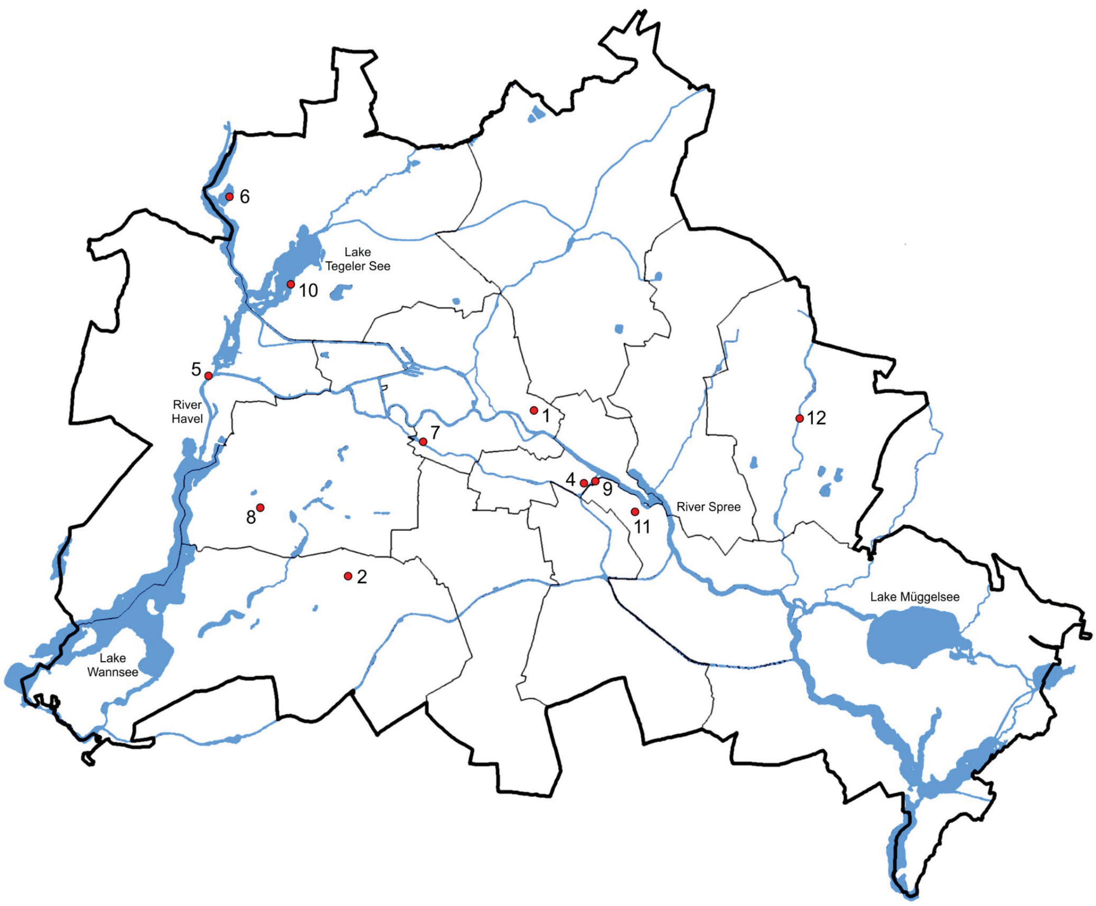


Figure 2. Map of the sampling localities in the Berlin region. For details to the numbered sampling sites see Table 1.
doi:10.1371/journal.pone.0108793.g002

Table 1. List of Localities: Strain Numbers, Geo-Reference, Habitat, Ecology.

Locality	Geo references*	Date	Habitat	Ecology	Strains	collected and isolated
1. Berlin**	N 52.518611°	2005	Freshwater	E	Amph5	O. Skibbe
	E 13.408056°					
2. Dahlem, Berlin	N 52.460833°	19 Oct. 2004	Agricultural soil	T, E	D27_003, D27_006, D27_009 D28_001 D28_004b, D28_007b D29_003b, D29_009b, D30_003 D30_006b, D30_009	L. Buhr (coll.), O. Skibbe (isol.)
	E 13.296944°					
3. Elbe near Schnackenburg, Lower Saxony	N 53.039634°	26 Oct. 2009	River	T, E	ElCal01a, ElPin01	D. Borgwardt (coll.), O. Skibbe (isol.)
	E 11.564806°					
4. Görlitzer Park, Berlin	N 52.494850°	May 2006	Pond	P, E	PinnB	O. Skibbe
	E 13.443891°					
5. Havel at Spandau, Berlin	N 52.534512°	June 2006	River	T, E	Amph4	O. Skibbe
	E 13.204309°					
6. Heiligensee, Berlin	N 52.60394°	August 2011	Shallow lake	D, E	HSB02	O. Skibbe
	E 13.21499°					
7. Landwehrkanal, Berlin	N 52.510733°	11 June 2005	Canal	T, E	D36_003, D36_012, D36_020	W.-H. Kusber (coll.), J. Bansemer (isol.)
	E 13.338638°					
8. Ökowerk, Berlin	N 52.49179°	June 2006	Artificial moorland pool	P, A	PinnC	O. Skibbe
	E 13.23589°					
9. Spree at Kreuzberg, Berlin	N 52.49491°	March/April 2004	River	T, E	Amph1, Coco1, Pinn1	O. Skibbe
	E 13.44729°					
		22 March 2004			D03_030, D03_034, D03_063,	O. Skibbe (coll.), J. Bansemer (isol.)
					D03_074, D03_082, D03_093,	
					D03_139,	
		2007			D45_03,	O. Skibbe
		20 Sept. 2007			D54_02,	O. Skibbe
		October 2011			SpCo1	W. da Silva (coll.), O. Skibbe (isol.)
10. Tegeler See	N 52.57023°	September 2009	Lake	D, E.	TeAm01, TeNav01	R. Jahn (coll.), O. Skibbe (isol.)
	E 13.25691°					
11. Treptower Park, Berlin	N 52.48445°	May 2004	Pond	T, E	Navi1	O. Skibbe
	E 13.47148°					
12. Wuhle, Berlin	N 52.52079	21 April 2004	Small river	T, E	D06_006, D06_014, D06_023, Navi2,	O. Skibbe (coll.), J. Bansemer (isol.)
	E 13.57781°					
					D06_029, D06_036, D06_038,	
					D06_047, D06_059, D06_060,	
					D06_067, D06_069, D06_074,	
					D06_077, D06_083, D06_087,	
					D06_093, D06_095, D06_096,	
					D06_102, D06_106, D06_107,	

Table 1. Cont.

Locality	Geo references*	Date	Habitat	Ecology	Strains	collected and isolated
					D06_110, D06_113, D06_122,	
					D06_138, D06_139,	
		April 2004			Navi4, Stau1	O. Skibbe
		June 2004			Navi5, Pinn2	O. Skibbe

(Ecology: D = dimictic, P = polymictic, T = turbid, E = eutrophic, alkaline, A = acidic, all running waters are part of Elbe catchment area).
*Uncertainty = ±50 m,
** Uncertainty = ±22500 m.
doi:10.1371/journal.pone.0108793.t001

Nomenclature

The electronic version of this article in Portable Document Format (PDF) in a work with an ISSN or ISBN will represent a published work according to the International Code of Nomenclature for algae, fungi, and plants, and hence the new names contained in the electronic publication of a PLOS ONE article are effectively published under that Code from the electronic edition alone, so there is no longer any need to provide printed copies. The online version of this work is archived and available from the following digital repositories: PubMed Central, LOCKSS. http://edocs.fu-berlin.de/docs/content/below/index.xml.

Results

Morphological analyses

The morphological identification of the 70 strains resulted in 37 taxa (see Table 2 and Figs. 3 and 4). 21 taxa were identified by only one strain but 10 taxa were represented by two strains, three taxa by three strains, one taxon by four strains, one taxon by five strains and one taxon by 11 strains.

DNA sequence analyses

PCR and sequencing success for 18S V4 and *rbc*L was 100% for all strains, resulting in 140 reference sequences for 70 strains. We established 129 novel sequences (INSDC accession numbers KM084866-KM084994) and an additional 11 sequences that had been previously published in Abarca et al. [40] and Zimmermann et al. [19].

There was little molecular variation within the here generated sequence data – only up to 0.5% in 18S V4 (representing 2 bp) and 0.3% in *rbc*L (corresponding to 3 bp) – between the different strains representing one taxon (Appendix S2). The highest in-taxon variation was found in e.g. *Mayamaea terrestris* 0.53% (18S V4), respectively *Navicula cryptocephala* e.g. 0.33% (*rbc*L). The uncorrected p distances for all genera and sequences are given in Appendix S2.

The results from sequence comparison with sequences published in the databases of the International Nucleotide Sequence Database Collaboration (INSDC, includes GenBank, EMBL and DDBJ) are shown in Table 3 and summarised in Fig. 5a, 5b. In the case of 18S V4, 22% of our taxa had entries with identical sequences in the INSDC whereas for *rbc*L this number was 21% (Fig. 5b). This was the case e.g. for *Caloneis silicula* and *Navicula cryptotenella* (Table 3). 22% (18S V4, Fig. 5a) respectively 25% (*rbc*L, Fig. 5b) of our taxa had no entry in the INSDC databases, e.g. *Amphora ovalis* and *Luticola sparsipunctata* (Table 3). For 15% of our taxa an identical 18S V4 sequence (Fig. 5a) with a different taxon name was found in the INSDC databases (e.g. *Gomphonema parvulum*); the number was considerably lower in *rbc*L with only 4% (Fig. 5b). The remaining taxa of which many showed sequence dissimilarities of over 15 bp were 41% for 18S V4 (Fig. 5a) and 50% for *rbc*L (Fig. 5b). The highest difference was found for *Pinnularia viridiformis* with 97 bp in 18S V4 (Table 3).

The tree derived from the concatenated data set and calculated by the Neighbour Joining (NJ) algorithm, including only the here presented strains, is shown in Fig. 6; the trees of the individual analysis of both markers are given in the Appendix S2. The molecular clades are congruent between 18S V4 and *rbc*L, the tree topology is partly differing between both markers (Appendix S3, S4); however, the conflicting nodes have bootstrap values below 0.85 and are therefore neglected.

In the tree derived from the combined dataset, the sampled genera are monophyletic and well supported (>0.98 bootstrap support BS, Fig. 6), except for *Caloneis*, *Craticula* and *Sellaphora*. *Craticula buderi* falls into a clade with the genera *Stauroneis* and *Karayevia* (0.48 BS; Fig. 6). *Sellaphora* falls into one group with *Eolimna* (0.98 BS; Fig. 6). The genus *Caloneis* is found in two distinct clades: *Caloneis silicula* is clustering with *Pinnularia* (0.61 BS; Fig. 6), *Caloneis amphisbaena* forms an independent clade on its own (1.00 BS; Fig. 6). The deeper bifurcations representing the relationship between the genera are generally not well supported by bootstrap values. All 37 subgeneric taxa included in this study are monophyletic (Fig. 6).

The trees for the genus *Amphora* including all available data from INSDC databases (this includes also accessions from the genus *Halamphora*) are shown in Fig. 7a (18S V4) and Fig. 7b (*rbc*L). The *Amphora ovalis* strains (Amph1, Amph4, Amph5, D45_003 and TeAm01) form a monophyletic clade, that is well supported in both 18S (0.99 BS) and *rbc*L (0.97 BS). The strain HSB02, identified as *Amphora berolinensis* appears to be rather isolated within the *Amphora* tree, except for an affiliation with the unidentified strain C10 (INSDC accession number FJ002132) in the *rbc*L tree (0.89 BS; Fig. 7b). All strains identified as *Amphora pediculus* cluster in one clade in 18S V4 (0.90 BS; Fig. 7a) and *rbc*L (Fig. 7b). This includes also the strain D54_002 named *Amphora* sp. aff. *atomoides*. The tree derived from *rbc*L sequences also includes the strain AT-21.206 (INSDC accession number AN502022) identified as *Amphora* cf. *fogediana* (Fig. 7b), which forms a branch with strain s0992 named *Amphora copulata* (INSDC accession number AB754831) in 18S V4 adjacent to the *Amphora pediculus* clade (Fig. 7a). In respect to the other strains available from the INSDC databases there is no topology consistent with the taxonomic identifications found in the trees (Fig. 7a, 7b). Several taxa, including the species *Amphora coffeaeformis*, *Amphora normannii* and *Amphora montana* were

Table 2. List of materials from Berlin localities: names of 37 Taxa and 70 strains, voucher codes in the Herbarium Berolinense (B), DNA Bank voucher numbers in the Berlin-Dahlem Plant DNA Bank, INSDC accession numbers, picture numbers in this publication and morphometrics of each strain (n = number of evaluated valves).

Taxon	Strain	B Voucher*	DNA Bank no.	INSDC no. 18S V4 rRNA	INSDC no. *rbc*L	Figures in this paper**	Morphology length µm	Morphology breadth µm	Morphology striae/ 10 µm [punctae/ 10 µm]	n
Achnanthidium saprophilum (H.Kobayasi & Mayama) Round & Bukht.	D06_036	B400040822	DB 8634	KM084866	KM084941	3.1	12.8–13.7	3.2–3.9	27–31	8
Amphora berolinensis N.Abarca & R. Jahn sp. nov.	HSB02	B400040823 Holotype	DB 8635	KM084914	KM084982	4.1a–i	9.5 18.9	4.7–5.2	15	5
Amphora ovalis (Kütz.) Kütz.	Amph1	B400040825	DB 8637	KM084867	KM084926	-	22.9–28.5	7.5–9.2	13–14	7
	Amph4	B400040826	DB 8638	KM084868	KM084927	3.11	41.5–48.6	10.9–11.1	13	7
	Amph5	B400040827	DB 8639	KM084869	KM084928		33.7–37	7.8–9.4	11–13	4
	D45_003	B400040824	DB 8636	KM084910	KM084978	-	78.5–79.1	13.5–16.1	11–12	4
	TeAm01	B400040828	DB 8640	KM084924	KM084993		61–67.3	13–17.5	11–13	5
Amphora pediculus (Kütz.) Grunow	D03_074	B400040829	DB 8641	KM084872	KM084933	3.8	11.9–12.6	3.0–3.8	20–21	7
Amphora cf. *pediculus* (Kütz.) Grunow	D03_063	B400040830	DB 8642	KM084871	KM084932		9.9–13.3	3.2–4.1	19–21	7
	D03_082	B400040831	DB 8643	KM084873	KM084934	3.10	11.5–14.7	2.7–3.6	20–21	8
Amphora sp. aff. atomoides Levkov	D54_002	B400040832	DB 8644	KM084911	KM084979	3.9	10.2–12.4	4.6–5	16	6
Caloneis amphisbaena (Bory) Cleve	ElCal01a	B400040834	DB 8646	KM084912	KM084980		72–74	23.5–24.7	15–17	7
	Navi1	B400040833	DB 8645	KM084915	KM084983	3.19	72.2–73.4	25.7–26.9	15–16	3
Caloneis silicula (Ehrenberg) Cleve	D06_074	B400040835	DB 8647	KM084883	KM084948		12.5–13	7–7.7	19–20	10
Cocconeis pediculus Ehrenberg	Coco1	B400040687	DB 8209	FR873234	KM084929		22.5–29.1	18.2–21-8	19	3
	D36_020	B400040644 [Epitype]	DB 8208	FR873233	KM084977	3.12	16.6–40	13–26	14–22	50
	SpCo1	B400040808	DB 8648	KM084923	KM084991		23–23.7	16.5–17.6	18–20	3
Cocconeis placentula Ehrenberg	D36_012	B400040647 [Epitype]	DB 8211	FR873236	KM084976	3.13	14.7–33.7	9.7_22.8	18–26	50
Craticula cuspidata (Kütz.) D.G.Mann	Navi4	B400040836	DB 8649	KM084917	KM084985	3.23	133.3	27.5	13	**3**
Craticula buderi (Hust.) Lange-Bert.	D06_069	B400040837	DB 8650	KM084882	KM084947	3.25	27.5–28.1	6.0–7.1	19–20	4
Eolimna minima (Grunow) Lange-Bert.	D03_030	B400040692	DB 8219	FR873244	KM084930	3.15	7.3–7.8	3.4–3.8	28	5
Eolimna sp. (teratological valves)	D06_023	B400040880	DB 8651	KM084877	KM084939		11.9–12.4	3.8–4.0	28	2
Gomphonema saprophilum (Lange-Bert. & Reichardt) N.Abarca, R.Jahn, J. Zimmermann & Enke	D36_003	B400040919 [Epitype]	DB 8652	HG530017	HG530056	3.34	24.5–26	6.5–8	14–20	10
Karayevia ploenensis var. *gessneri* (Hust.) Bukt.	D03_034	B400040838	DB 8653	KM084870	KM084931	3.6	13.5–14.0	4.9–5.4	16–18	6
Luticola sparsipunctata Levkov, Metzeltin & Pavlov	D06_029	B400040839	DB 8654	KM084878	KM084940	3.7	16.4–26.7	5.5–8.4	16–18	15
Mayamaea terrestris N.Abarca et R.Jahn sp. nov.	D27_003	B400040840	DB 8655	KM084897	KM084963		7.1–8.0	3.0–4.3	24–26	19
	D27_006	B400040841	DB 8656	KM084898	KM084964		7.5–7.9	3.5–4.1	22–24	9
	D27_009	B400040864	DB 8657	KM084899	KM084965		7.7–8.8	3.4–4	20–22	6

Table 2. Cont.

Taxon	Strain	B Voucher*	DNA Bank no.	INSDC no. 18S V4 rRNA	INSDC no. *rbc*L	Figures in this paper**	Morphology length µm	Morphology breadth µm	Morphology striae/ 10 µm [punctae/ 10 µm]	n
	D28_001	B400040843	DB 8658	KM084900	KM084966		7.0–8.7	3.7–4.5	22–24	17
	D28_004b	B400040844	DB 8659	KM084902	KM084968		7.4–7.6	3.6–4.1	20–22	9
	D28_007b	B400040845	DB 8660	KM084903	KM084969		7.5–8.6	3.5–4.2	21–22	12
	D29_003b	B400040846	DB 8661	KM084905	KM084971		7.7–8.5	3.6–4.4	22	5
	D29_009b	B400040847 Holotype	DB 8662	KM084906	KM084972	4.2a–g	7.2–8	3.3–3.7	22.5–27	17
	D30_003	B400040848	DB 8663	KM084907	KM084973		7.2–8.3	4.2–4.4	24	10
	D30_006b	B400040849	DB 8664	KM084908	KM084974		7.4–8.2	3.5–3.7	22–23	6
	D30_009	B400040850	DB 8665	KM084909	KM084975		7.3–8.0	3.8–4.3	22	5
Mayamaea permitis (Hust.) Bruder & Medlin	D06_106	B400040851	DB 8666	KM084891	KM084956	3.17;3.18	6.6–8.2	3.6–4.4	no data	17
	D06_107	B400040697	DB 8224	FR873249	KM084957		7.2–8.5	3.6–3.85	22	10
Navicula cryptotenella Lange-Bert.	TeNav01	B400040852	DB 8667	KM084925	KM084994		17.5–20.5	4.6–5.7	15–16	10
Navicula cryptocephala Kütz.	D06_059	B400040700	DB 8226	FR873251	KM084944	3.26	29.0–30.4	5.8–6.6	15–16	8
	D06_067	B400040853	DB 8668	KM084881	KM084946		28.0–30.4	5.7–6.8	15–17	10
Navicula gregaria Donkin	D06_077	B400040855	DB 8670	KM084884	KM084949		20.6–24.8	5.7–6.7	19	2
	D06_096	B400040854	DB 8669	KM084889	KM084954		27.5–27.7	6.5–6.9	18–19	3
	D06_122	B400040856	DB 8671	KM084894	KM084960	3.24	24.2–25.0	5.9–6.2	17–18	6
Navicula radiosa Kütz.	D06_102	B400040857	DB 8672	KM084890	KM084955	3.22	59.5–59.9	10.3–10.9	11–12 [24]	4
Navicula rhynchotella Lange-Bert.	D06_083	B400040860	DB 8676	KM084885	KM084950		42.2–43.0	11.0–11.8	9–10 [24]	4
	D06_087	B400040861	DB 8675	KM084886	KM084951		46.9–47.0	12.2–12.6	8–9 [24]	4
	D06_093	B400040858	DB 8673	KM084887	KM084952	3.21	43.2–45.8	11.7–12.4	8–10 [24]	6
	D06_095	B400040859	DB 8674	KM084888	KM084953		44.2–44.3	11.3–11.5	9–10 [24]	3
Navicula slesvicensis Grunow	D06_038	B400040705	DB 8228	FR873253	KM084942	3.27	42.2–44.0	8.9–9.4	9–10 [24]	5
Navicula tripunctata (O.F.Müll.) Bory	D03_093	B400040706	DB 8230	FR873255	KM084935		32.5–36.3	7.6–8.0	10–11	6
	D03_139	B400040862	DB 8677	KM084874	KM084936	3.20	42.0–43.2	7.6–8.3	11–12	5
	Navi5	B400040863	DB 8678	KM084918	KM084986		28.5–31.3	7.5–8.0	11–12	7
Pinnularia neomajor Krammer	Pinn1	B400040865	DB 8679	KM084919	KM084987		185.1–206.2	27.0–28.5	6–7	4
	PinnB	B400040866	DB 8680	KM084921	KM084989	3.29	196.8–202.0	26.7–27.7	7	3
Pinnularia viridiformis Krammer	ElPin01	B400040870	DB 8682	KM084913	KM084981	3.32	69.5–77.6	15–17.2	8–9	8
	Pinn2	B400040867	DB 8681	KM084920	KM084988	3.31	78.1–84.9	17.2–17.9	8–9	2
Pinnularia sp.	Navi2	B400040868	DB 8683	KM084916	KM084984	3.33	31.1–35.6	10.8–11.3	10.5–11	8
Pinnularia sp.	PinnC	B400040869	DB 8684	KM084922	KM084990	3.30	91.7–93.4	17.8–18.4	6–7	3
Planothidium caputium J.Zimmermann, & R.Jahn sp. nov.	D06_014	B400040871 Holotype	DB 8687	KM084876	KM084938	4.3a–h	20.7–22.9	6.0–6.4	13–14	4

Table 2. Cont.

Taxon	Strain	B Voucher*	DNA Bank no.	INSDC no. 18S V4 rRNA	INSDC no. *rbc*L	Figures in this paper**	Morphology length μm	Morphology breadth μm	Morphology striae/ 10 μm [punctae/ 10 μm]	n
	D06_113	B400040875	DB 8688	KM084893	KM084959		20.3–21	5.5–6.2?	13–15	8
Planothidium frequentissimum (Lange-Bert.) Lange-Bert.	D06_138	B400040872	DB 8685	KM084895	KM084961		14.7–16.7	5.4–5.9	14–15	9
	D06_139	B400040873	DB 8686	KM084896	KM084962	3.2;3.3	14–6-16	5.6–6.2	14–16	4
Planothidium lanceolatum (Bréb. ex Kütz.) Lange-Bert.	D06_047	B400040874	DB 8689	KM084879	KM084943	3.4;3.5	10.1–12.5	4.6–5.1	13–14	20
Sellaphora pupula (Kütz.) Mereschk.	D06_060	B400040877	DB 8690	KM084880	KM084945	3.16	17.7–18.2	6.7–7.0	21–24	5
	D06_110	B400040878	DB 8691	KM084892	KM084958		17.8–18.1	6.7–7.0	21–22	7
Sellaphora seminulum (Grunow) D.G.Mann	D06_006	B400040879	DB 8692	KM084875	KM084937	3.14	11.1–18.1	3.8–4.8	20–22	9
Stauroneis phoenicenteron (Nitzsch) Ehrenb.	Stau1	B400040715	DB 8239	FR873264	KM084992	3.28	161.2–164.1	31.2–31.6	13–14	3
Stauroneis schmidiae R.Jahn & N.Abarca sp. nov.	D28_002	B400040883	DB 8693	KM084901	KM084967		25.2–27.4	5.5–6	17–19	3
	D28_008	B400040882 Holotype	DB 8694	KM084904	KM084970	4.4a–h	27–28.2	5.5–6	15–18	7

INSDC accession numbers starting with KM are newly published here, those starting with FR have been published in Zimmermann et al. (2011) [19] and numbers starting with HG in Abarca et al. (2014) [39].
* URL for AlgaTerra http://herbarium.bgbm.org/object/plus the corresponding B Voucher.
**pictures of all strains are available online by AlgaTerra [30].
doi:10.1371/journal.pone.0108793.t002

Table 3. Sequence comparison of the here presented strains with corresponding accessions from INSDC databases.

Taxon	BGBM Strain	Accession INSDC (Strain)			
		18S V4	**bp diff.**	**rbcL**	**bp diff.**
Achnanthidium saprophilum (H.Kobayasi & Mayama) Round & Bukht.	D06_036	*Achnanthidium minutissima*			
		FR873231 (D05_008)[1]	15		
		AM502032 (AT-196Gel02)[1]	15	AM710499 (AT-196Gel02)[1]	21
		AJ866992 (AMIN)[1]	15		
Amphora berolinensis Abarca & R.Jahn	HSB02	*Amphora copulata*[2] (Kütz.)			
		Schoeman & R.E.M. Archibald			
		AM501959 (AT-117.10)	5	AM710425 (AT-117.10)	35
				JN162763 (IKCCMP0165)	42
Amphora ovalis (Kütz.) Kütz.	D45_003	-	-	-	-
	Amph1	-	-	-	-
	Amph4	-	-	-	-
	Amph5	-	-	-	-
	TeAm01	-	-	-	-
Amphora pediculus (Kütz.) Grunow	D03_074	AM501960 (AT-117.11)	3	AM710426 (AT-117.11)	14
				HQ912403 (L1030)	7
Amphora cf. *pediculus* (Kütz.) Grunow	D03_063	AM501960 (AT-117.11)	3	AM710426 (AT-117.11)	13
				HQ912403 (L1030)	2
	D03_082	AM501960 (AT-117.11)	3	AM710426 (AT-117.11)	14
				HQ912403 (L1030)	3
Amphora sp. aff. *atomoides* Levkov	D54_002	-	-	-	-
Caloneis amphisbaena (Bory) Cleve	Navi1	AM501954 (AT-177.07)	1	AM710507 (AT-177.07)	0
	ElCal01a	AM501954 (AT-177.07)	1	AM710507 (AT-177.07)	0
Caloneis silicula (Ehrenberg) Cleve	D06_074	JN418593 (Cal890TM)	0	JN418663 (Cal890TM)	0
Cocconeis pediculus Ehrenberg	D36_020	AM502010 (AT-212.07)	0	AM710477 (AT-212.07)	0
		FR873235 (LuCoc03)	0		
	Coco1	AM502010 (AT-212.07)	1	AM710477 (AT-212.07)	0
		FR873235 (LuCoc03)	1		
	SpCo1	AM502010 (AT-212.07)	2	AM710477 (AT-212.07)	0
		FR873235 (LuCoc03)	2		
Cocconeis placentula Ehrenberg	D36_012[1]	FR873239 (WiCoc01)	20		
		FR873237 (D17_011)	24		
		KC736616 (TCC501)	16	KC736591 (TCC501)	14
		HQ912592 (UTEXFD23)	23	HQ912456 (UTEXFD23)	40
		AM502013 (AT-212.Gel11)	19	AM710480 (AT-212.Gel11)	24
Craticula cuspidata (Kütz.) D.G.Mann	Navi4	HQ912581 (UTEX.FD35)	28	HQ912445 (UTEX.FD35)	10
Craticula buderi (Hust.) Lange-Bert.	D06_069	-	-	-	-
Eolimna minima (Grunow) Lange-Bert	D03_030	AM501962 (AT-70Gel18)	9	AM710427 (AT-70Gel18)	18
		AJ243063 (SNA15)	1		
		HM449712[3]	5		
Eolimna sp. (teratological valves)	D06_023	-	-	-	-
Gomphonema saprophilum Abarca et al.	D36_003	-	-	-	-
Karayevia ploenensis var. *gessneri* (Hust.) Bukt.	D03_034	-	-	-	-
Luticola sparsipunctata Levkov, Metzeltin & Pavlov	D06_029	-	-	-	-
Mayamaea terrestris Abarca & R. Jahn	D27_003	*Mayamaea atomus* (Kütz.) Lange-Bert var. atomus			
		AM501968 (AT-115Gel07)	5	AM710434 (AT-115Gel07)	25
				AM710510 (AT-199Gel01)	69

Table 3. Cont.

Taxon	BGBM Strain	Accession INSDC (Strain)			
		18S V4	bp diff.	rbcL	bp diff.
	D27_006	AM501968 (AT-115Gel07)	5	AM710434 (AT-115Gel07)	25
				AM710510 (AT-199Gel01)	69
	D27_009	AM501968 (AT-115Gel07)	5	AM710434 (AT-115Gel07)	25
				AM710510 (AT-199Gel01)	69
	D28_001	AM501968 (AT-115Gel07)	5	AM710434 (AT-115Gel07)	25
				AM710510 (AT-199Gel01)	69
	D28_004b	AM501968 (AT-115Gel07)	5	AM710434 (AT-115Gel07)	25
				AM710510 (AT-199Gel01)	69
	D28_007b	AM501968 (AT-115Gel07)	5	AM710434 (AT-115Gel07)	25
				AM710510 (AT-199Gel01)	69
	D29_003b	AM501968 (AT-115Gel07)	5	AM710434 (AT-115Gel07)	25
				AM710510 (AT-199Gel01)	69
	D29_009b	AM501968 (AT-115Gel07)	5	AM710434 (AT-115Gel07) AM710510 (AT-199Gel01)	25, 69
	D30_003	AM501968 (AT-115Gel07)	5	AM710434 (AT-115Gel07)	25
				AM710510 (AT-199Gel01)	69
	D30_006b	AM501968 (AT-115Gel07)	5	AM710434 (AT-115Gel07)	25
				AM710510 (AT-199Gel01)	69
	D30_009	AM501968 (AT-115Gel07)	5	AM710434 (AT-115Gel07)	25
				AM710510 (AT-199Gel01)	69
Mayamaea permitis (Hust.) Abarca & R.Jahn comb. nov.	D06_107	AM501969 (AT-101Gel04)	0	AM710435 (AT-101Gel04)	27
		JN418600 (Wes2f)	42	JN418670 (Wes2f)	34
	D06_106	AM501969 (AT-101Gel04)	0	AM710435 (AT-101Gel04)	27
		JN418600 (Wes2f)	35	JN418670 (Wes2f)	34
Navicula cryptocephala Kütz	D06_059	KC736631 (TCC515)	4	KC736601 (TCC515)	8
		HQ912603 (UTEX FD109)	0	HQ912467 (UTEX FD109)	0
				HQ337543 (CCMP2519)	19
		AM501996 (AT-176Gel05)	0	AM710463 (AT-176Gel05)	0
		AM501973 (AT-114Gel08c)	0	AM710439 (AT-114Gel08c)	0
	D06_067	KC736631 (TCC515)	4	KC736601 (TCC515)	13
		HQ912603 (UTEX FD109)	0	HQ912467 (UTEX FD109)	3
				HQ337543 (CCMP2519)	24
		AM501996 (AT-176Gel05)	0	AM710463 (AT-176Gel05)	3
		AM501973 (AT-114Gel08c)	0	AM710439 (AT-114Gel08c)	3
Navicula cryptotenella Lange-Bert.	TeNav01	AM502011 (AT-212Gel01)	1	AM710478 (AT-212Gel01)	0
		AM502029 (AT-202Gel03)	0	AM710496 (AT-202Gel03)	0
		AM502015 (AT-210Gel05)	0	AM710482 (AT-210Gel05)	0
Navicula gregaria Donkin	D06_096	FR873252 (D08_002)	1		
		HM805037 (BA102)	1		
		AM501974 (AT-117Gel05)	1	AM710440 (AT-117Gel05)	1
	D06_077	FR873252 (D08_002)	0		
		HM805037 (BA102)	0		
		AM501974 (AT-117Gel05)	0	AM710440 (AT-117Gel05)	1
	D06_122	FR873252 (D08_002)	0		
		HM805037 (BA102)	0		
		AM501974 (AT-117Gel05)	0	AM710440 (AT-117Gel05)	1
Navicula radiosa Kütz.	D06_102	AM502034 (AT-205.02b)	0	AM710501 (AT-205.02b)	0

Table 3. Cont.

Taxon	BGBM Strain	Accession INSDC (Strain)			
		18S V4	bp diff.	rbcL	bp diff.
		AM502027 (AT-200.04)	0	AM710494 (AT-200.04)	0
		AM501972 (AT-114Gel06)	0	AM710438 (AT-114Gel06)	0
Navicula rhynchotella Lange-Bert.	D06_093	-	-	-	-
	D06_095	-	-	-	-
	D06_087	-	-	-	-
	D06_083	-	-	-	-
Navicula slesvicensis Grunow	D06_038	-	-	-	-
Navicula tripunctata (O.F.Müll.) Bory	D03_093	AM502028 (AT-202.01)	0	AM710495 (AT-202.01)	0
	D03_139	AM502028 (AT-202.01)	0	AM710495 (AT-202.01)	0
	Navi5	AM502028 (AT-202.01)	0	AM710495 (AT-202.01)	0
Pinnularia neomajor Krammer	Pinn1	JN418585 (Corsea2)	0	JN418655 (Corsea2)	0
		JN418571 (Tor1a)	31	JN418641 (Tor1a)	15
	PinnB	JN418585 (Corsea2)	0	JN418655 (Corsea2)	0, 15
		JN418571 (Tor1a)	31	JN418641 (Tor1a)	
Pinnularia viridiformis Krammer	Pinn2	JN418589 (Pin870MG)	22	JN418659 (Pin870MG)	19
		JN418574 (Enc2a)	26	JN418644 (Enc2a)	24
		AM501985 (AT-70.10)	9	AM710451 (AT-70.10)	5
		AM743108 (L1716)	97		
	ElPin01	JN418589 (Pin870MG)	22	JN418659 (Pin870MG)	19
		JN418574 (Enc2a)	26	JN418644 (Enc2a)	24
		AM501985 (AT-70.10)	9	AM710451 (AT-70.10)	5
		AM743108 (L1716)	97		
Pinnularia sp.	Navi2	-	-	-	-
Pinnularia sp.	PinnC	-	-	-	-
Planothidium frequentissimum (Lange-Bert.) Lange-Bert.	D06_138	-	-	-	-
	D06_139	-	-	-	-
Planothidium caputium R.Jahn & Abarca sp. nov.	D06_014	-	-	-	-
	D06_113	-	-	-	-
Planothidium lanceolatum (Bréb. ex Kütz.) Lange-Bert.	D06_047	AJ535189 (L1249)	2	JQ610173 (LCR-S2-1-1)	17
Sellaphora pupula (Kütz.) Mereschk.	D06_060	EF151973 (Bel2)	1	EF143266 (Bel2)	0
		EF151983 (Aus4)	1	EF143317 (Aus4)	15
	D06_110	EF151973 (Bel2)	1	EF143266 (Bel2)	0
		EF151983 (Aus4)	1	EF143317 (Aus4)	15
Sellaphora seminulum (Grunow) D.G.Mann	D06_006	EF151967 (TM37)	0	EF143280 (TM37)	32
		KC736642 (TCC461)	22	KC736613 (TCC461)	16
Stauroneis phoenicenteron (Nitzsch.) Ehrenb.	Stau1	AM502031 (AT-182.07)	0	AM710498 (AT-182.07)	0
		AM501987 (AT-117.04)	2	AM710453 (AT-117.04)	0
Stauroneis schmidiae R.Jahn & Abarca sp. nov.	D28_002	-	-	-	-
	D28_008	-	-	-	-

Basepair differences (bp diff.) for each taxon and strain number specified for both markers 18S V4 and *rbc*L. – denotes missing representative for taxon in INSDC databases (accessed July 2013).
[1]*Achnanthidium minutissimum* (Kützing) Czarnecki.
[2]new name for the taxon formerly identified as *Amphora libyca* Ehrenberg.
[3]as *Navicula minima*.
doi:10.1371/journal.pone.0108793.t003

recently transferred to the genus *Halamphora* [37]; these taxa and also the two INSDC accessions listed as *Halamphora* in the (numbers AB754832, AB754833;) are forming a loose cluster in the upper part of the 18S V4 tree (Fig. 7a). The *rbc*L data set supports an independent clade for the taxa of the genus *Halamphora* (*Amphora coffeaeformis*, *Amphora normannii*, *Am-*

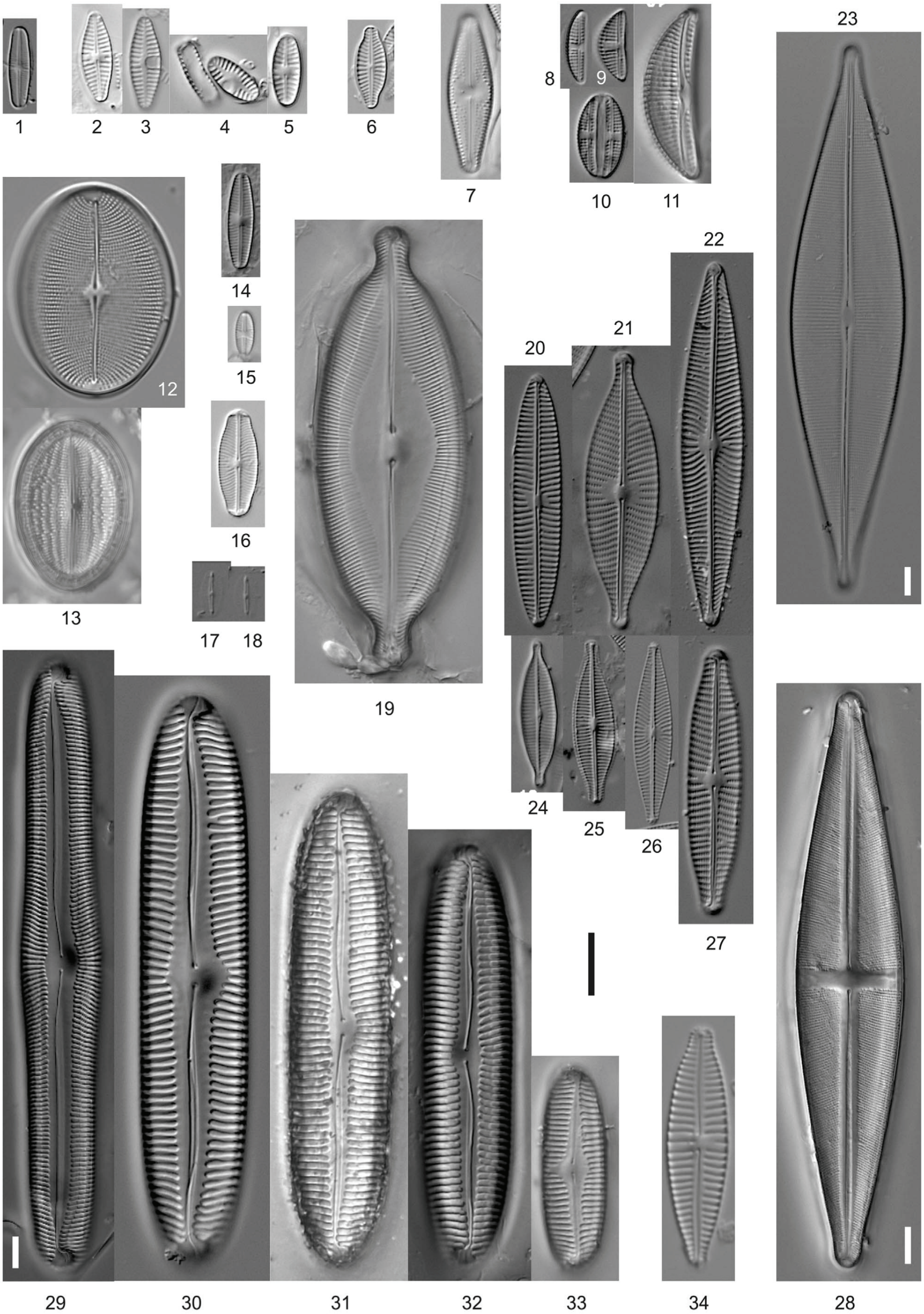

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34

Figure 3. LM photos of individual valves from strains. Fig. 3.1. *Achnanthidium saprophilum* (H.Kobayasi & Mayama) Round & Bukht., Strain D06_036. Fig. 3.2.–3. *Planothidium frequentissimum* (Lange-Bert.) Lange-Bert., Strain D06_139. Fig. 3.4.–5. *Planothidium lanceolatum* (Bréb. ex Kütz.) Lange-Bert., Strain D06_047. Fig. 3.6. *Karayevia ploenensis* var. *gessneri* (Hust.) Bukt., Strain D03_034. Fig. 3.7. *Luticola sparsipunctata* Levkov, Metzeltin & Pavlov, Strain D06_029. Fig. 3.8. *Amphora pediculus* (Kütz.) Grunow, Strain D03_074. Fig. 3.9. *Amphora* sp. aff. *atomoides* Levkov, strain D54_002. Fig. 3.10. *Amphora* cf. *pediculus* (Kütz.) Grunow, Strain D03_082. Fig. 3.11. *Amphora ovalis* (Kütz.) Kütz., Strain Amph4. Fig. 3.12. *Cocconeis pediculus* Ehrenberg, Epitype-Strain D36_020. Fig. 3.13. *Cocconeis placentula* Ehrenberg, Epitype-Strain D36_012. Fig. 3.14. *Sellaphora seminulum* (Grunow) D.G.Mann, Strain D06_006. Fig. 3.15. *Eolimna minima* (Grunow) Lange-Bert., Strain D03_030. Fig. 3.16. *Sellaphora pupula* (Kütz.) Mereschk., Strain D06_060. Fig. 3.17.–18. *Mayamaea permitis* (Hust.) Bruder & Medlin, Strain D06_106. Fig. 3.19. *Caloneis amphisbaena* (Bory) Cleve, Strain Navi1. Fig. 3.20. *Navicula tripunctata* (O.F.Müll.) Bory, Strain D03_139. Fig. 3.21. *Navicula rhynchotella* Lange-Bert., Strain D06_093. Fig. 3.22. *Navicula radiosa* Kütz., Strain D06_102. Fig. 3.23. *Craticula cuspidata* (Kütz.) D.G.Mann, Strain Navi4. Fig. 3.24. *Navicula gregaria* Donkin, Strain D06_122. Fig. 3.25. *Craticula buderi* (Hust.) Lange-Bert., Strain D06_069. Fig. 3.26. *Navicula cryptocephala* Kütz., Strain D06_059. Fig. 3.27. *Navicula slesvicensis* Grunow, Strain D06_038. Fig. 3.28. *Stauroneis phoenicenteron* (Nitzsch) Ehrenb., Strain Stau1. Fig. 3.29. *Pinnularia neomajor* Krammer, Strain PinnB. Fig. 3.30. *Pinnularia* sp., Strain PinnC. Fig. 3.31. *Pinnularia viridiformis* Krammer, Strain Pinn2. Fig. 3.32. *Pinnularia viridiformis* Krammer, Strain ElPin01. Fig. 3.33. *Pinnularia* sp., Strain Navi2. Fig. 3.34. *Gomphonema saprophilum* (Lange-Bert. & Reichardt) N.Abarca, R.Jahn, J. Zimmermann & Enke, Strain D36_003. Scale bar represents 10 μm.
doi:10.1371/journal.pone.0108793.g003

phora montana; 0.97 BS; Fig. 7b). However, within the *Halamphora* clade the strains identified as *Amphora coffaeaformis* are not monophyletic (Fig. 7b).

The trees for the genus *Mayamaea* including all available data from INSDC databases are given in Fig. 7c (18S V4) and Fig. 7d (*rbc*L). All strains identified as *Mayamaea terrestris* are forming an independent clade in both trees (0.84 BS in 18S V4, 1.00 BS in *rbc*L; Fig. 7c, 7d). The strains D06_106 and D06_107 representing *Mayamaea permitis* (Syn.: *Mayamaea atomus* var. *permitis*) cluster together in one clade (in *rbc*L 1.00 BS), however other strains named either *Mayamaea atomus*, *Mayamaea permitis* or *Mayamaea atomus* var. *permitis* show no clear pattern according to their names provided in the INSDC databases (Fig. 7c, 7d).

The trees for the genus *Planothidium* including all available data from INSDC databases are given in Fig. 8a (18S V4) and Fig. 8b (*rbc*L). 18S V4 supports three independent clades for the three including *Planothidium* taxa; namely *Planothidium caputium*, *Planothidium frequentissimum* and *Planothidium lanceolatum* (each taxon supported by 1.00 BS; Fig. 8a). Strain LCR-S18-1-1 (INSDC accession number JQ610164), listed in the INSDC databases as *Planothidium* sp., sits on another branch (Fig. 8a). The topology derived from *rbc*L sequences gives one clade (0.97 BS) for *Planothidium caputium* and strain LCR-S18-1-1 (INSDC accession number JQ610172) plus a second clade for a monophyletic group *Planothidium frequentissimum* (0.92 BS; Fig. 8b). The strains identified as *Planothidium lanceolatum* do not form an independent clade in the *rbc*L analysis (Fig. 8b).

The trees for the genus *Stauroneis* including all available data from INSDC databases are given in Fig. 8c (18S V4) and Fig. 8d (*rbc*L). The *Stauroneis schmidiae* strains (D28_002, D28_008) cluster in one clade, which is sister to the strain UTEX FD 51 (INSDC accession numbers HQ912579 (18S V4) and HQ912443 (*rbc*L)) in both analyses (1.00 BS; Fig. 8c, 8d). The strain Stau1 is identified as *Stauroneis phoenicenteron* and forms a monophyletic clade (1.00 BS for both markers) with all the other accessions with this name available from the INSDC databases (AT-18.207 (INSDC accession numbers AM502031 (18S V4) and AM710498 (*rbc*L)) and AT-11.704 (INSDC accession numbers AM501987 (18S V4) and AM710453 (*rbc*L))). The other taxa available from the INSDC databases also cluster taxonomically consistent, however there are difference in the overall topology recovered from 18S V4 respectively *rbc*L sequences (Fig. 8c, 8d).

Nomenclatural and taxonomical consequences

Two new taxa were first discovered by morphological means namely *Amphora berolinensis* and *Stauroneis schmidiae*. The analysis of molecular data suggested the existence of two more previously undetected taxa that could later be also morphologically confirmed (*Mayamaea terrestris*, *Planothidium caputium*). For yet another two taxa morphological data is incomplete (teratological outline, micro-morphological data missing) but the molecular data show that they both are different from an identified taxon in this genus; these strains are named sp. (*Amphora* sp. aff. *atomoides*); in one case we used the term cf. (*Amphora* cf. *pediculus*) to show that it is closely related to a known taxon.

Amphora cf. *pediculus*

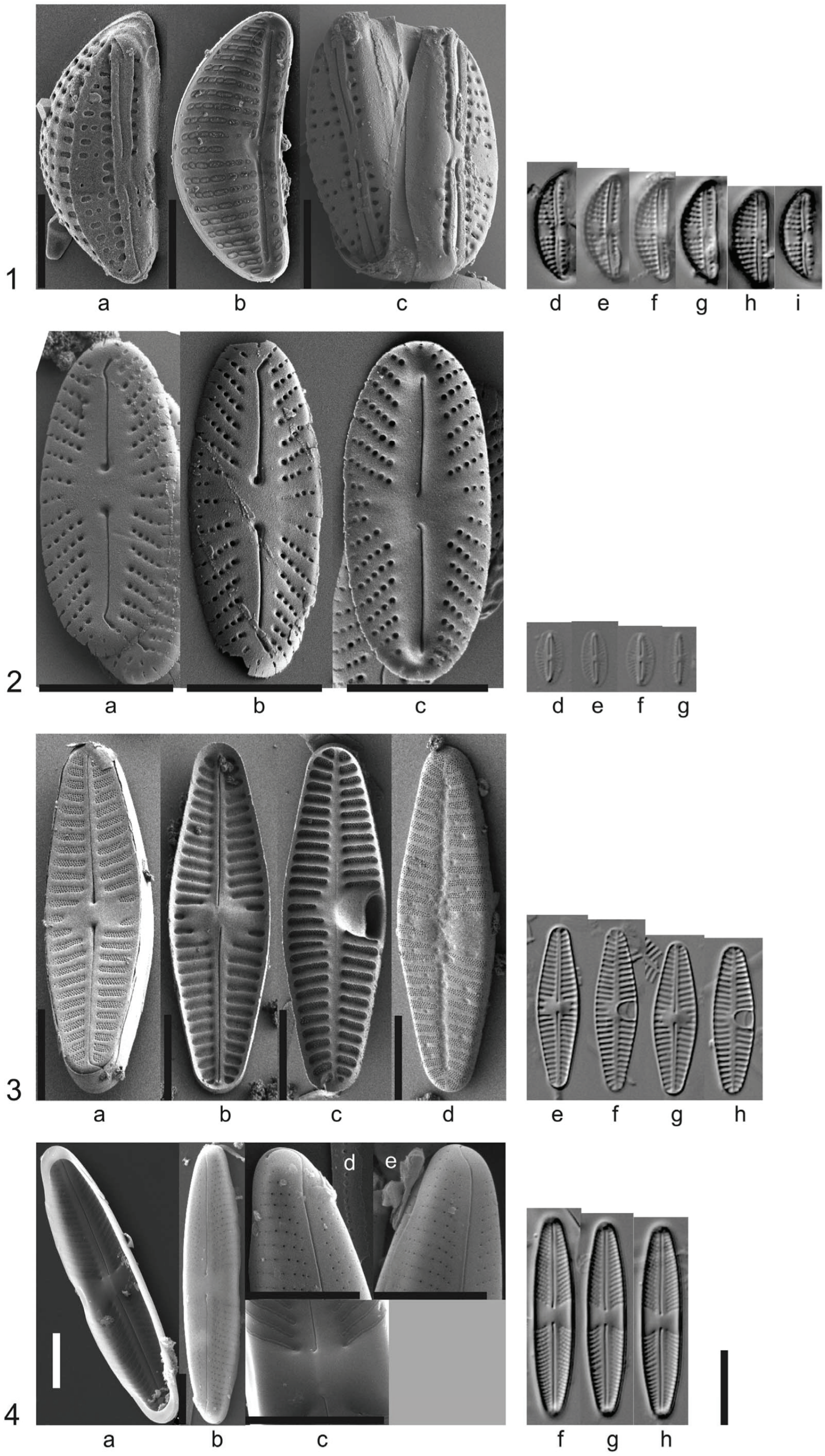
The strains D03_063 & D03_082 are morphologically very similar to our *Amphora pediculus* D03_074 but have double areolae in each ventral stria and not only a single elongated areola like *A. pediculus*. The specimens of these strains have a similar valve outline as *A. indistincta*, but in SEM the differences are more distinct because in *A. indistincta* the width of the central and dorsal side is almost equal and the striae are composed of elongated areolae.

Amphora sp. aff. *atomoides* Levkov

The strain D54_002 has a valve semi elliptical with arched dorsal margin, concave ventral margin and narrowly rounded valve ends. Valve length is 10–12.4 μm, breadth 4.6–5 μm. The central area on dorsal side is a rectangular fascia almost extending to the dorsal margin; on the ventral side the much broader fascia is expanding towards the valve margin. Raphe branches linear, filiform. Proximal raphe endings straight, distal raphe endings ventrally deflected. Dorsal striae radiate throughout, 16 in 10 μm.

This species closely resembles *A. atomoides* but differences can be observed in the shape of the central area and valve breadth (7–11 μm in *A. atomoides*). In *A. atomoides* the central area on the dorsal side is small or absent not extending to the valve margin, contrary to our Amphora sp. aff. *atomoides* where the central area presents a rectangular fascia almost extending to the dorsal margin. D54_002 also resembles *A. pediculus* with respect to its valve shape and size. However D54_002 can be differentiated by the valve width (*A. pediculus* is narrower with 2.5–4 μm) the central area (*A pediculus* has a distal raphe dorsally deflected and a central area with a rectangular facia, extended to the dorsal valve margin) and the stria density (*A. pediculus* has more striae 18–24/10 μm). D54_002 can also be differentiated from *A. minutissima* by the shape of valve apices (ventrally bent in *A. minutissima*). Additional observations of more specimens by SEM would be necessary to establish the proper identity of this population from Heiligensee, Berlin.

Four taxa in the genera *Amphora*, *Mayamaea*, *Planothidium*, and *Stauroneis* do not fall within the description of any previously known taxa and are therefore described here as new.



1
a
b
c
d
e
f
g
h
i
2
a
b
c
d
e
f
g
3
a
b
c
d
e
f
g
h
4
a
b
c
d
e
f
g
h

Figure 4. SEM and LM photos of the newly described species. Figs. 4.1a–i. *Amphora berolinensis* N.Abarca & R. Jahn sp. nov., Strain HSB02; Fig. 4.1d–i. Holotype B 40 0040823. Figs. 4.2a–g. *Mayamaea terrestris* N.Abarca et R.Jahn sp. nov., Strain D29_009b; Fig. 4.2d–g. Holotype B 40 0040847. Figs. 4.3a–h. *Planothidium caputium* J.Zimmermann & R.Jahn sp. nov., Strain D06_014; Fig. 4.3e–h Holotype B 40 0040871. Fig. 4.4a–h. *Stauroneis schmidiae* R.Jahn & N.Abarca sp. nov., Strain D28_008; Fig. 4.4f–h. Holotype B 40 0040882.
doi:10.1371/journal.pone.0108793.g004

Amphora berolinensis N.Abarca & R.Jahn (Figs. 4.1a–i)

Holotype: B 40 0040871 from strain HSBO2; the holotype is represented by Fig. 4.1d.

Type locality: Germany, Berlin, Heiligensee, N 52.60394°E 13.21499° leg. and isolated by O. Skibbe, August 2011.

Amphora berolinensis differs from *A. copulata* (Kützing) Schoeman & Archibald because the latter has bigger valves (19–42 µm length, 5–7.5 µm breadth). In SEM the differences are more distinct. Differences can be observed in the shape of the central area (bordered by striae close to the valve margin in *A. copulata*), the raphe (biarcuate in *A. copulata*) and the morphology of the dorsal striae (crossed by longitudinal bars in *A. copulata*). *A. berolinense* also differs from *A. neglectiformis* Levkov & Edlund by the larger valves of the later (18–53 µm length, 5–7 µm breadth) and the ventral striae which are composed of two areolae in *A. neglictiformis* near the valve ends.

The valves of *Amphora berolinensis* are semi-lanceolate to semi-elliptical, with smoothly arched dorsal margin and straight to slightly concave ventral margin, valve ends rounded. Valve length is 9.5–18.9 µm, breadth 4.7–5.2 µm. Axial area is narrow, slightly arched. The central area on the dorsal side has a rectangular fascia extending to the dorsal margin; on the ventral side the fascia is wider expanding towards the valve margin. Raphe is filiform and more or less straight, in some valves the proximal raphe endings are straight, in others they are dorsally bent and the distal raphe endings are straight and in some valves they are ventrally bent. Dorsal striae are coarsely punctated and radiate throughout, 12–14 in 10 µm. Ventrally striae are radiate, composed of one areola.

Amphora copulata (Kütz.) Schoeman & R.E.M. Archibald (concept syn. *Amphora libyca* Ehrenberg, sensu post auct.) is morphologically the closest fit to *Amphora berolinensis*, the latter forms a distinctly different clade according to both 18S V4 and *rbc*L (Fig. 7a, 7b). The sequence difference to strain AT-117.10 belonging to *Amphora libyca* sensu post auct. is e.g. 5 bp for 18S V4 and 35 bp for *rbc*L.

Mayamaea terrestris N.Abarca & R.Jahn sp. nov. (Fig. 4.2a–g)

Holotype: B 40 0040847 from strain D29_009b; the holotype is represented by Fig. 4.2e.

Type locality: Germany, Berlin-Dahlem, agricultural soil, N 52.460833°E 13.296944°, leg. L. Buhr, 21 April 2004, cultures isolated by J. Bansemer.

Mayamaea terrestris differs from *Mayamaea atomus* var. *atomus* [47] because the latter is longer and wider and has less striae (8.5–13 µm length, 4–5.5 µm breadth, 19–22/10 µm striae). Also the molecular data differ from *Mayamaea atomus* var. *atomus* entries in the INSDC databases in 5 bp for 18S V4 and 25 bp for *rbc*L of strain AT-115Gel07 and even in 69 bp for *rbc*L of strain AT-199Gel01 (AM710510) [48].

The valves of *Mayamaea terrestris* are narrow linear-elipical, ends obtusely rounded. Valve length is 7–8.7 µm, breadth 3–4.5 µm. Striae are radiate throughout, 22–24 (–26) in 10 µm with c. 50 areolae in 10 µm. Raphe is filiform, the two branches are gently arcuate with distinct central pores. Axial area is slightly broad, widening lanceolately towards the middle of the valve. Central raphe ends expanded by depressions around the central pores and deflected, while the ends of the terminal raphe fissures are deflected to the opposite side.

This new species lives in soil; this is signified by the epithet name.

10 further strains (D27_003 & D27_006 & D27_009 & D28_001 & D28_004b & D28_007b & D29_003b & D30_003 & D30_006b & D30_009) have only low sequence differences for 18S V4 and *rbc*L (Appendix S2) and form a clade clearly different from all the other available *Mayamaea* strains (Fig. 7c, 7d).

Planothidium caputium J.Zimmermann & R.Jahn sp. nov. (Fig. 4.3a–h)

Holotype: B 40 0040871; strain D06_014; the holotype is represented by Fig. 4.3f.

Type locality: Germany, Berlin, small river Wuhle, N 52.52079°E 13.57781°, leg. O. Skibbe, 21 April 2004, cultures isolated by J. Bansemer.

Morphologically, *Planothidium caputium* has a similar outline as *Planothidium lanceolatum* but differs from it by a hood over the depression on the rapheless valve as in *P. frequentissimum*. The difference to *P. frequentissimum* lies in the form and size of the hood; which is bigger, longer and wider in *P. caputium* than in *P. frequentissimum* and the hood has a wider opening; this results in a line-like instead of a horse shoe appearance when focusing through the hood. The uncorrected p-distances show that *Planothidium caputium* sequences differ at least 2.4% (18S V4) respectively 2% (*rbc*L) from *Planothidium frequentissimum*, and 6% (18S V4) respectively 4% (*rbc*L) from *Planothidium lanceolatum* (Appendix S2), this is also represented in the trees including all available *Planothidium* strains (Fig. 8a, 8b).

Valves are elliptical to elliptic-lanceolate, with rounded apices. Valve length is 20–22.9 µm, breadth 5.5–6.4 µm. The striae are radiate on both valves, becoming more radiate towards the apices, with 13–14 in 10 µm. Striae are multiseriate with three to five rows of areolae per stria. The axial area is narrow and linear to lanceolate in both valves. A weak central area on the raphe valve and a horseshoe-shaped collar on one side of the rapheless valve which by focusing in LM another line less arched can be recognized (see also Straub 1990 [49]).

Also strain D06_113 belongs to this species.

Stauroneis schmidiae R.Jahn & N.Abarca sp. nov. (Fig. 4.4a–h)

Holotype: B 40 0040883 from strain D28_008; the holotype is represented by Fig. 4.4f.

Type locality: Germany, Berlin-Dahlem, agricultural soil, N 52.460833°E 13.296944°, leg. L. Buhr, 21 April 2004, cultures isolated by J. Bansemer.

Morphologically, *Stauroneis schmidiae* differs from *Stauroneis borrichii* (Petersen) Lund, which has a similar valve outline but with protracted ends, because the latter is shorter and more slender and has more striae (18–25 µm length, 4.0–5.0 breadth, 20–22 striae and 25–28 punctae per 10 µm (see Van de Vijver et al 2004) and from *Stauroneis pseudomuriella* Van de Vijver & Lange-Bert. (2004) [50] which has similar morphometrics as our new species but more striae (21–42 µm length, 5–6.5 µm breadth,

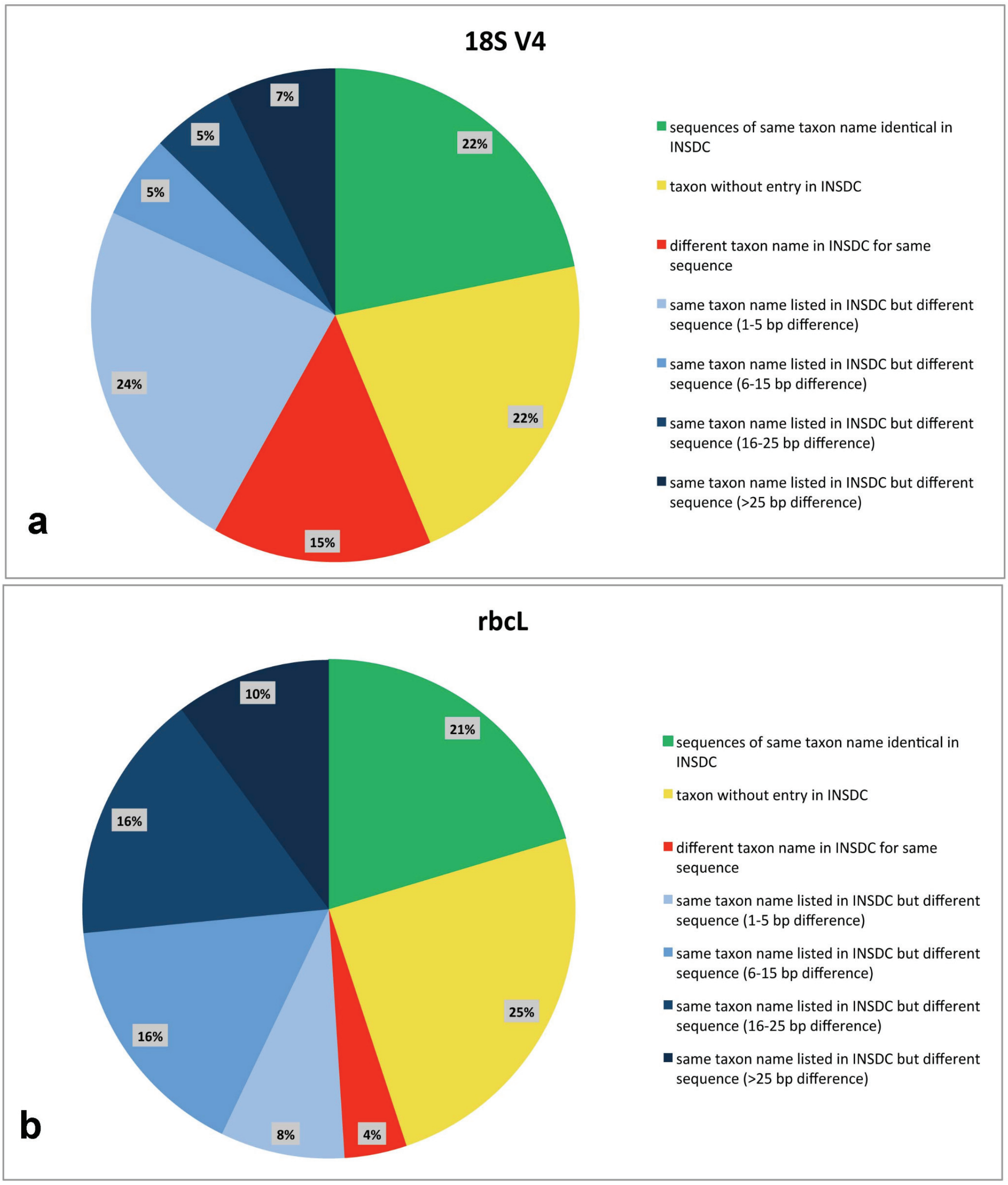


Figure 5. Chart giving classes of base pair (bp) differences for both markers (18S V4, *rbc*L) between here presented molecular data and corresponding data from INSDC databases. Inferred from data in Table 3.
doi:10.1371/journal.pone.0108793.g005

20–22 striae and 25 punctae per 10 µm) but this species has no pseudosepta.

Valves are linear-lanceolate with very slightly rounded non-protracted ends. Valve length is 27–28.2 µm, breadth 5.5–6 µm. Striae are radiate throughout the entire valve, 15–18 in 10 µm. Puncta of the striae are discernible in LM and are 24–28 in 10 µm. Pseudosepta present.

Also strain D28_002 belongs to this species.

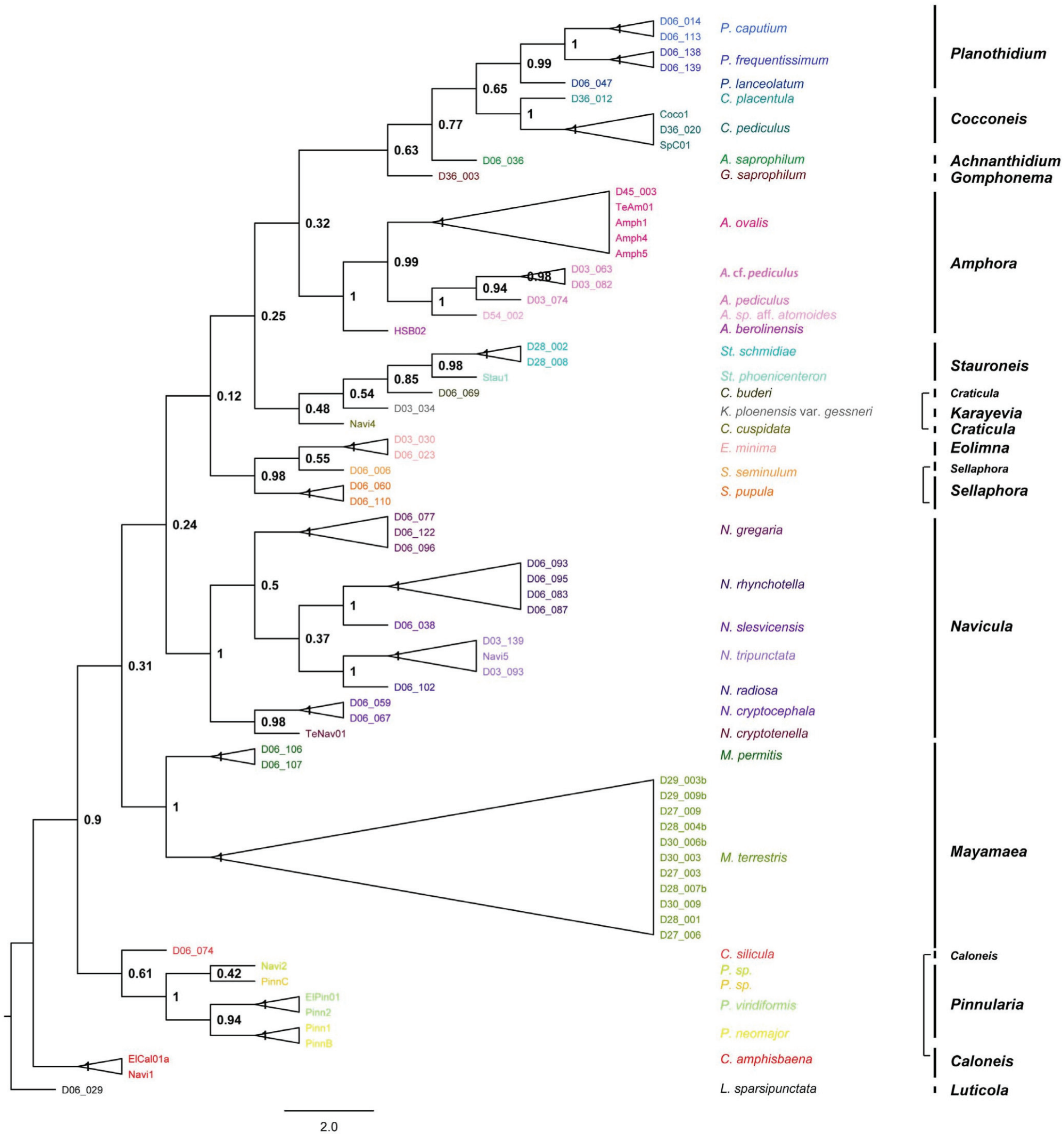


Figure 6. Neighbour Joining Tree (10 000 bootstrap replicates) derived from concatenated dataset (18S V4, *rbc*L) including all sequences from this study. All bootstrap support values given above branches.
doi:10.1371/journal.pone.0108793.g006

Compared to the other available *Stauroneis* strains *Stauroneis schmidiae* clusters independently for both markers (Fig. 8c, 8d).

This species is named in honor of Prof. Dr. AnnaMaria Schmid who was an inspiring diatom teacher to Regine Jahn.

Discussion

The 37 naviculoid diatom taxa, of which reference barcodes are published here, represent only about 7% of the total diatom flora which is 14% of the naviculoid taxa recorded for Berlin waters (539 taxa, see [31]). Nevertheless, it is a first milestone in characterising diatoms not only by morphological but also by molecular means, which represents the start of a taxonomic reference library for diatoms.

Identification via DNA sequences is an important tool, especially in microorganisms. Many of the large scale environmental DNA barcoding studies in protists so far rely on higher taxonomic levels of families and above; only rarely they reach a

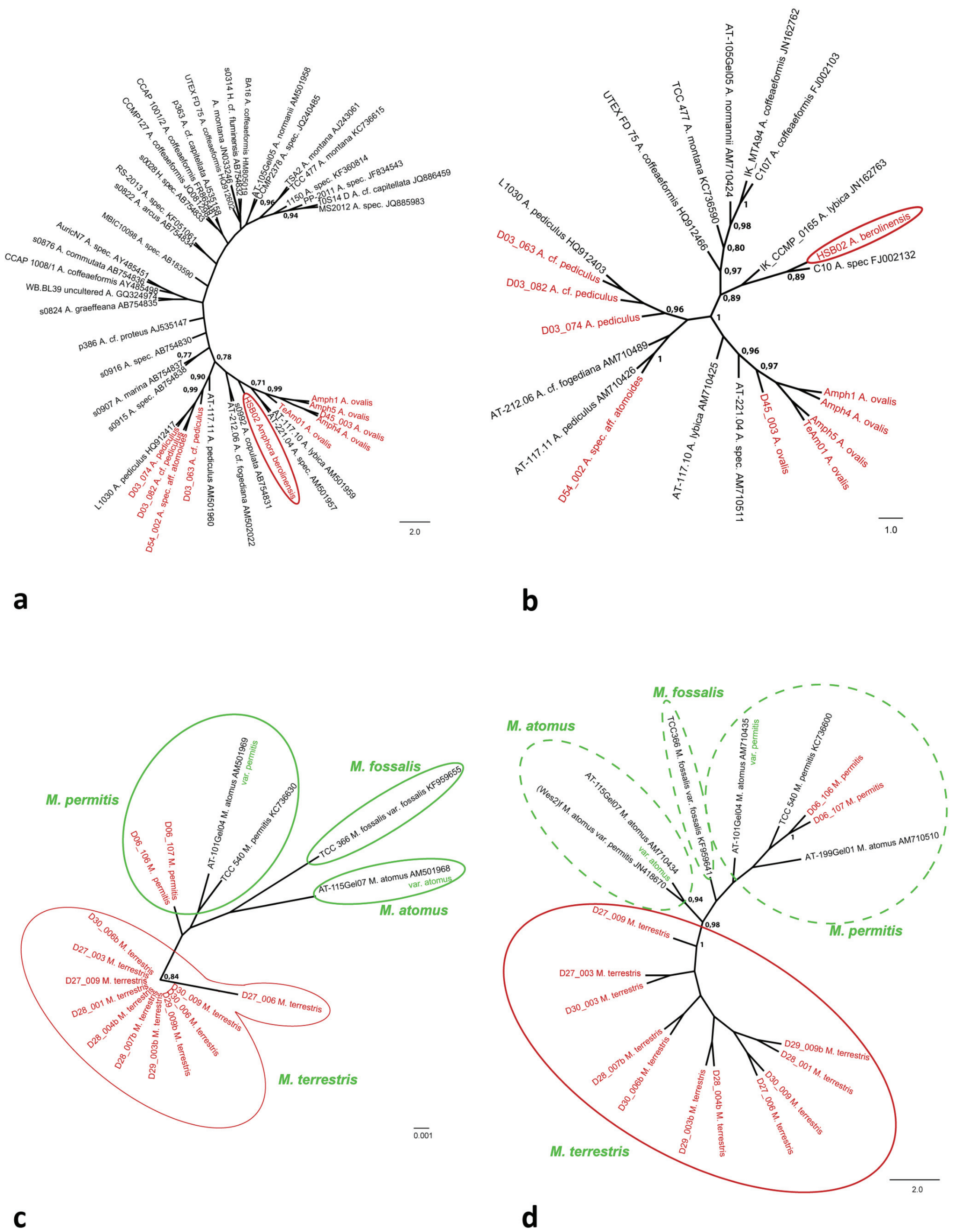


Figure 7. Neighbour Joining Tree (10 000 bootstrap replicates) including all sequences available from the INSDC databases for genus *Amphora* (a) 18S V4, (b) *rbc*L as well as *Mayamaea* (c) 18S V4, (d) *rbc*L. Bootstrap support values >0.75 given at nodes. Red indicates data from new species, green information and conlusions derived from data in the AlgaTerra Information System [30].
doi:10.1371/journal.pone.0108793.g007

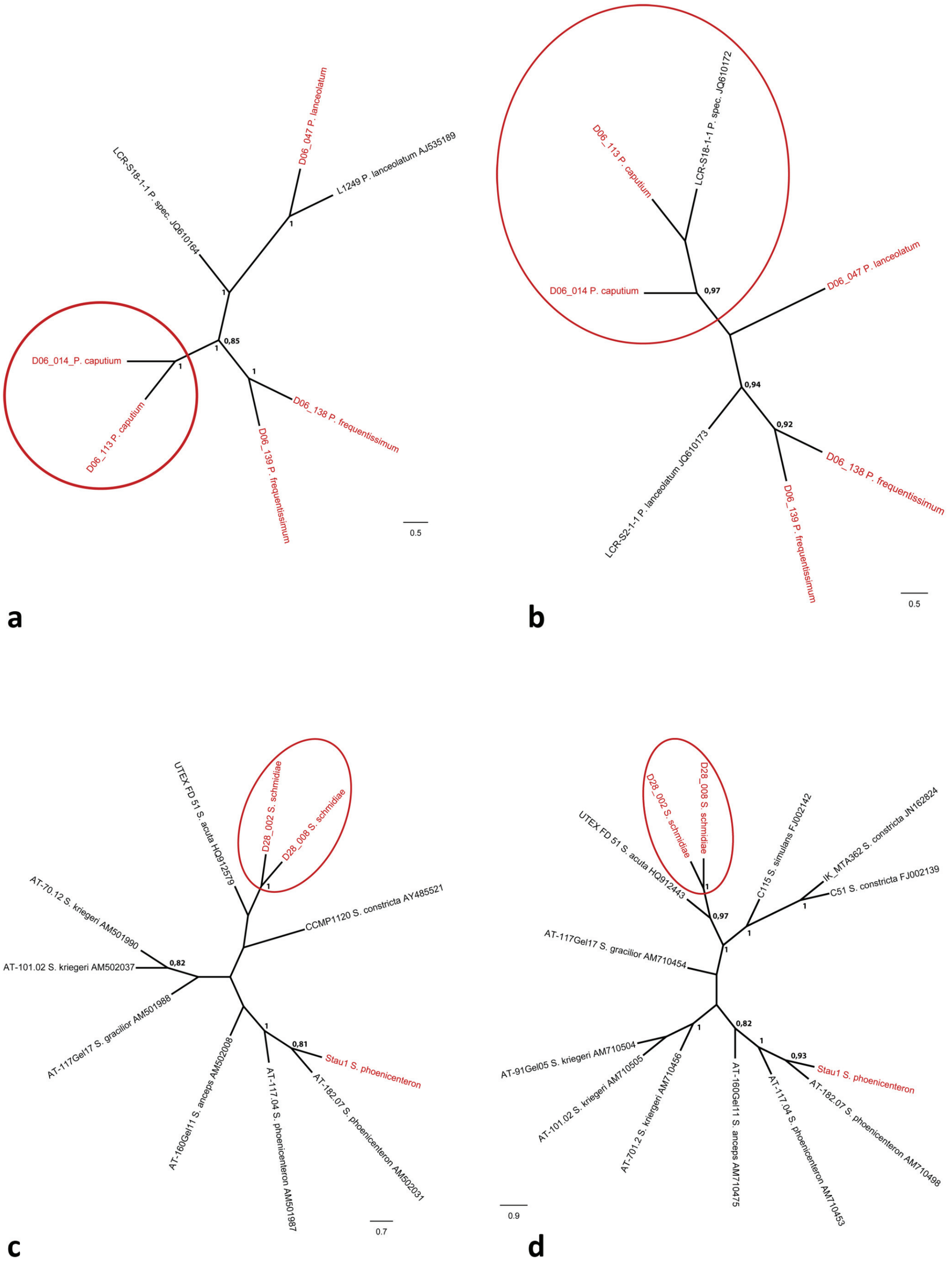

a
b
c
d

Figure 8. Neighbour Joining Tree (10 000 bootstrap replicates) including all sequences available from the INSDC databases for genus *Planothidium* (a) 18S V4, (b) *rbc*L as well as *Stauroneis* (c) 18S V4, (d) *rbc*L. Bootstrap support values >0.75 given at nodes. Red indicates data from new species.
doi:10.1371/journal.pone.0108793.g008

resolution at genus level. In diatoms, assignment to genus level is unproblematic [51,52]. Even identification to the species level is possible, but strongly depends on the quality of the reference database [52–58]. We here tested the taxonomic consistency of naviculoid diatom taxa at the species level by comparing our identified sequences with the published sequences in the repositories of the INSDC. We found that the taxonomic assignment in INSDC is currently unsatisfying, because it is often erroneous. In the data of the two commonly used DNA barcoding markers for diatoms 18S V4 and *rbc*L we analysed, we found that for *rbc*L 26% for the sequences listed under the same name as our strains more than 15 bp sequence difference were recorded (Fig. 5b); for 18S V4 this was 12% (Fig. 5a). For the 800 bp long *rbc*L fragment 15 bp difference amounts to roughly 2% sequence difference, in the shorter (400 bp) 18S V4 fragment 15 bp difference correlates to even 4%. The relatively high percentage of differences in these short DNA fragments suggests that the sequences belong to a different taxon. This implies morphology-related misidentification, mislabelling or cross-contamination. There are an additional 16% (*rbc*L, Fig. 5b) respectively 5% (18S V4, Fig. 5a) of the sequences where sequences with the same taxon name showed differences between 6 and 15 bp, here it is unclear whether these strains belong to a different taxon of a closely related cryptic species or whether they reflect natural intraspecific variation. Furthermore, we found that in 4% (*rbc*L, Fig. 5b) respectively 15% (18S V4, Fig. 5a) of the cases, identical sequences in the repositories of the INSDC were annotated with a different taxon name than the strains of this study. These sequences therefore provide an erroneous identification. In summary, the unevaluated use of information deposited in the INSDC leads to wrong identifications in at least 30% of the cases; in only about 20% of our cases, the identifications coincided unambiguously.

Unfortunately, in most cases it is not possible to trace the DNA sequence to the specimen from which it originated and, because of lacking voucher specimens, taxonomic evaluation is not possible; hence there are no means to verify whether a faulty taxon assignment had occurred or an interesting biological phenomenon. Therefore such sequences are of no future use and valuable information is lost to science. Assessment of diatom community composition through environmental DNA barcoding could greatly benefit from better documented reference libraries, especially because biodiversity in general should be evaluated at least on the species level [59].

Furthermore, the linkage between historically and morphologically described taxa and molecular sequences is not very strong. A possible threat is that two independent data clouds might develop [60]: one including large amounts of molecular data from environmental sequencing, the other species specific data (e.g. paleontological and recent distribution, ecology, phylogeny) linked to morphological descriptions. For organism groups where next to no morphology based data exist (e.g. many groups of bacteria), there is little harm if the information in the two clouds cannot be correlated. However, in groups like diatoms, where two centuries of data collection linked to morphologically described species exists, it would be a waste of painfully acquired data not to link these two groups of data. At the moment, this link would be a reference sequence that is connected to a morphological voucher (and DNA sample) deposited in a natural history collection and therefore available for multiple testing and verification of results as well as for long-term studies.

We here define a taxonomic reference library as an entity combining molecular data – in our case DNA sequence data of two markers – with morphological documentation of important features as well as a valid name. Also environmental information on the collecting site should be provided in a standardised format.

Documentation should also include the deposition of DNA in a curated repository. To ensure traceability of a name/sequence back to the specimen it originated from, morphological details important for identification should be provided in an online photographic documentation, this includes high-resolution photographs giving an overview of the cell as well as details produced by electron microscopy or comparable techniques. Another special aspect for diatoms (and some other microorganism groups) is that many sequences derive from cultured clonal strains, especially if they are linked to morphological entities. Therefore, the strain number and other strain specifications are valuable information that should be presented along with the sequence.

Ideally, all the necessary information for traceable taxonomic classification should be available in a single data portal; however, at the moment there are several technological limitations to deposit and/or respectively retrieve all the information in and from one location. The Consortium for the Barcode of Life (CBOL) aims at compiling DNA barcode records in a public library (Barcoding of Life Database BOLD) [53] and even designed a Barcode Submission Tool for submitting sequences to the INSDC databases. However, this tool is limited to one marker, namely the mitochondrial cytochrome oxidase subunit I (COI) e.g. [17,61–65]. For many groups, e.g. plants [66] but also diatoms, this barcoding marker is not routinely applicable [19,21–25], albeit there are BOLD supported activities to implement alternative solutions for some organism groups e.g. [28]. On the other hand, the Barcode Submission Tool provides possibilities to at least upload a pherogram (output of sanger sequencing), but no pictures of the organisms can be stored. Therefore, this tool does not require a link to a morphological voucher (digital and physical), which would allow for subsequent taxonomic validation. Also a link to a herbarium specimen is only indirectly possible if the accession number of the specimen collection is given and the respective collection has their specimen picture online available. Although, it seems generally possible to deposit pictures and other data along with the DNA sequence in BOLD [53], unfortunately, the data deposited within BOLD is often not open access, depending on the rights given by the administrator. Also, we heard reports that data is not released to the public even if requested by the author. In conclusion it would be preferable if INSDC would extend their service, as they are the most commonly used platform to deposit sequence data [58].

Here we present our strategy on how documentation can be performed to build a comprehensive reference database for diatoms even with inconvenient IT possibilities. The here presented materials and data have been documented as follows: The physical vouchers (microscopic slides and SEM stubs) have been deposited in the Berlin Herbarium (B), the DNA in the DNA bank network of the Botanic Garden and Botanical Museum Berlin-Dahlem [39]. The data for both items are made available through The Global Genome Biodiversity Network (GGBN [67]) and The Global Biodiversity Information Facility (GBIF [68]).

The sequences have been submitted to an INSDC database (EMBL) along with strain numbers, voucher number from the Berlin Herbarium (B) and DNA bank number. Also primer details and geo-references have been deposited there. Photographic documentation is online available from the AlgaTerra Information System [30], linked through INSDC accession number and accession number from the Berlin Herbarium. Morphological characters, cultivation details as well as sampling data of the collecting sites beyond the geo-references (e.g. ecological specifications) have also been deposited in the AlgaTerra Information System [30].

A carefully documented reference sequence could be considered as something similar to a molecular type of the name of a species. Biological taxon types should be documented with a maximum amount of data, which makes it possible for every researcher to determine whether a specific specimen belongs to the concept of the designated type. In the botanical [69] and zoological [70] codes of nomenclature the basis for species description is the deposition of physical specimen. A reference sequence or reference barcode should be similarly well documented.

Biological identification systems are in constant development, therefore a continuous process of confirmation, validation and updating in relation to alpha taxonomy is required to build a compressive and accurate reference library. Protocols for data curation and revision are indispensable for new species discovery as well as taxonomic revisions. Therefore, entries in a taxonomic reference library (e.g. in an extended INSDC like system) need to be curated and updated in order to be in line with current taxonomy. However, a huge impediment for data curation by the respective author – once it is submitted – is, that there is no reward system for researchers for curating their data [71]. It has been shown, that incentives for researchers for the publication of thoroughly documented datasets similar to the publication of the conclusions drawn from these could greatly increase the motivation to publish datasets [71]. Another approach would be that data curation would be carried out by professional personnel employed for this purpose or a combination of both approaches.

Not only DNA barcoding approaches would benefit from well documented and referenced molecular data but also taxonomic and phylogenetic studies of diatoms which could integrate published data more efficiently if better documentation linked to physical objects were available [72]. For example, the clusters found for the genus *Mayamaea*, based on available 18S V4 and *rbc*L sequences, show low taxonomic consistency (Fig. 7c, 7d). The INSDC data suggest that there are different groups of *Mayamaea* (*atomus* var.) *permitis*, and within the *Mayamaea atomus* (var. *atomus*) sequences is one sequence named *Mayamaea fossalis* var. *fossalis* (Fig. 7c, 7d, black and red). For two of the AT strains included in the *Mayamaea* analysis additional data is available from the AlgaTerra Information System [30] (Fig. 7c, 7d, green): (a) more taxonomic detail is given than deposited alongside the sequence in INSDC - strain AT-115Gel07 is identified as *Mayamaea atomus* var. *atomus* and AT-101Gel04 as *Mayamaea atomus* var. *permitis* - and (b) photographs with morphological details are provided. Therefore the identification of both strains could be checked and verified. Even though additional data for only two strains is available from the AlgaTerra Information System [30], this already aids in the interpretation of the trees given in Fig. 7c and 7d; especially for the tree based on 18S V4. There is a cluster of *Mayamaea permitis* (Syn. *Mayamaea atomus* var. *permitis*), incl. strain AT-101Gel04, and one strain (AT-115Gel07) belonging to *Mayamaea atomus* var. *atomus* (Fig. 7c, green). As *Mayamaea permitis* (Syn. *Mayamaea atomus* var. *permitis*) has been raised to species rank due to morphological reasons (see above), this allows the interpretation that *Mayamaea fossalis* could be an independent taxon (Fig. 7c, green). For the tree based on *rbc*L, however, only an informed guess can be made: for two strains, namely (Wes2)f and AT-199Gel01, no additional data is available to check the identification (Fig. 7d). If it could be assumed that (Wes2)f was misidentified and AT-199Gel01 belongs to *Mayamaea permitis*, again four independent taxa could be assumed: *Mayamaea atomus*, *Mayamaea fossalis*, *Mayamaea permitis* and *Mayamaea terrestris*. This example, particularly the different interpretation possibilities between 18S V4 and *rbc*L trees, clearly shows how valuable additional data can be for the interpretation of sequence based analyses.

Due to the fact that species descriptions in diatoms are based on morphology derived from microscopic pictures (of variable quality) of single, or a limited number, of valves from a presumed population in mixed samples, it is often difficult to unambiguously identify a strain. Even within a single clonal culture, morphological variation sometimes fits in parts to different species circumscriptions [45]. In addition, size wise clonal cultures are often at the lower end of the morphometrics of a taxon description; if cultured for too long and if no auxosporulation has taken place, diatom valves tend to lose their typical morphological features because they get smaller with each cell division. This leads to the problem how to link sequences derived from cultures to a type specimen or at least to a current species concept. If a type specimen is designated, this can be achieved e.g. through epitypification as has been done for *Cocconeis pediculus* and *C. placentula* [73,74]. But in most cases, this will be done in the context of a taxonomic revision of a species group as e.g. for *Gomphonema saprophilum* [45] and needs to be done for the two unidentified *Pinnularia* species of this study. For the purpose of a reference library, if no unambiguous identification seems possible, the sequence could either be designated as belonging to a certain "formenkreis" (taxon group) marked as *affine* (e.g. *Amphora* sp. aff. *atomoides*), as not exactly fitting the original descriptions marked as *confer* (e.g. *Amphora* cf. *pediculus*) [http://bionomenclature-glossary.gbif.org/], or a new taxon has to be described formally along with providing the reference sequence (e.g. *Amphora berolinense*). The first two options are a practical way to make re-users of the data aware of an "uncertainty level" concerning the taxonomic identification; this is better than providing no guidance to the species group by giving just the genus name such as *Amphora* sp. As we documented in this study, the marine or halophilic species of *Amphora* sensu lato have been recently moved into the genus *Halamphora*; for a freshwater reference library, this is important ecological data. In addition, this information might become valuable for the interpretation of taxonomic discrepancies

Conclusions

As here shown exemplarily for some naviculoid diatoms, taxonomic reference libraries could serve as an online accessible and algorithmically searchable equivalent to commonly used printed identification literature. They are needed to link molecular based identification technologies with correct organism references. However, up to now searchable data bases often include large percentages of wrongly annotated sequencesand provide no possibility to trace the identification back to the respective specimen, leaving molecular based techniques often with identifications only to family or genus level. While for some studies this level of taxonomic depth seems to suffice (e.g. large scale biodiversity assessments), there are many studies that could profit from well documented molecular data (e.g. species inventories, monitoring, taxonomy, phylogeny). Therefore, it would be worth the effort to provide all material needed for identification of an organism.

Supporting Information

Appendix S1 Table of INSDC accessions with strain numbers and references.
(XLSX)

Appendix S2 Table with uncorrected p distances given for individual genera. Intra-taxon variability is highlighted.
(XLSX)

Appendix S3 Neighbour Joining Trees (10 000 bootstrap replicates) derived from individual datasets 18S V4 including all sequences from this study. All bootstrap support values given above branches.
(TIFF)

Appendix S4 Neighbour Joining Trees (10 000 bootstrap replicates) derived from individual datasets *rbc*L including all sequences from this study. All bootstrap support values given above branches.
(TIFF)

Acknowledgments

The help of Jana Bansemer and Ilona Danßmann with cultivation and retrieval of molecular data and of Monika Lüchow at the SEM is gratefully acknowledged. We thank Gabriele Dröge for providing counsel concerning the DNA Bank Network and Birgit Gemeinholzer for fruitful discussions.

Author Contributions

Conceived and designed the experiments: JZ NE NA RJ WHK OS. Performed the experiments: JZ NA NE. Analyzed the data: JZ NE. Wrote the paper: JZ NA NE RJ. Conceived the idea: RJ. Cultivation: OS. Composition of the manuscript: NE JZ. Taxonomical and morphological analyses: NA RJ. Data curation: WHK RJ.

References

1. Falkowski PG, Barber RT, Smetacek V (1998) Biogeochemical Controls and Feedbacks on Ocean Primary Production. Science 281: 200–206.
2. Field CB, Behrenfeld MJ, Randerson JT, Falkowski P (1998) Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. Science 281: 237–240.
3. Smetacek V (1999) Diatoms and the Ocean Carbon Cycle. Protist 150: 25–32.
4. Mann DG (1999) The species concept in diatoms. Phycologia 38: 437–495.
5. Poulíčková A, Špačková J, Kelly M, Duchoslav M, Mann D (2008) Ecological variation within *Sellaphora* species complexes (Bacillariophyceae): specialists or generalists? Hydrobiologia 614: 373–386.
6. Vanelslander B, Créach V, Vanormelingen P, Ernst A, Chepurnov VA, et al. (2009) Ecological differantiation between sympatric pseudocryptic species in the estuarine benthic diatom *Navicula phyllepta* (Bacillariophyceae). Journal of Phycology 45: 1278–1289.
7. Jahn R, Zetzsche H, Reinhardt R, Gemeinholzer B (2007) Diatoms and DNA barcoding: A pilot study on an environmental sample. In: Kusber W, Jahn R, (ed.) Proceedings of the 1st Central European Diatom Meeting 2007 Berlin: Botanic Garden and Botanical Museum Berlin-Dahlem. 63–68.
8. Schaumburg J, Schranz C, Hofmann G, Stelzer D, Schneider S, et al. (2004) Macrophytes and phytobenthos as indicators of ecological status in German lakes — a contribution to the implementation of the water framework directive. Limnologica - Ecology and Management of Inland Waters 34: 302–314.
9. Schaumburg J, Schranz C, Stelzer D, Hofmann G, Gutowski A, et al. (2005) Bundesweiter Test: Bewertungsverfahren "Makrophyten & Phytobenthos" in Fließgewässern zur Umsetzung der WRRL. Bayerisches Landesamt für Umwelt, Endbericht im Auftrag der LAWA (Projekt Nr O204).
10. Watanabe T, Asai K, Houki A (1988) Numerical water quality monitoring of organic pollution using diatom assemblages. In: Round, FE (ed.), Proceedings of the 9th Diatom Symposium, Biopress Ltd,: 123–141.
11. Kelly MG, Cazaubon A, Coring E, Dell'Uomo A, Ector L, et al. (1998) Recommendations for the routine sampling of diatoms for water quality assessments in Europe. Journal of Applied Phycology 10: 215–224.
12. European-Committee-for-Standardization (2003) EuropeanStandard. EN 14407. Water Quality – Guidance Standardfor the Identification, Enumeration and Interpretation of Benthic Diatom Samples from Running Waters. CEN.
13. Kusber WH (2001) Mikroalgen und Naturschutz - Rote Listen, Bewertungsinstrumentarium und Auswertungsansätze. Ökologie & Umweltsiherung 21: 197–228.
14. Mann DG, Droop SJM (1996) 3. Biodiversity, biogeography and conservation of diatoms. Hydrobiologia 336: 19–32.
15. Mann D, Sato S, Trobajo R, Vanormelingen P, Souffreau C (2010) DNA barcoding for species identification and discovery in diatoms. Cryptogamie Algologie 31: 557–577.
16. Rach J, DeSalle R, Sarkar IN, Schierwater B, Hadrys H (2008) Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata. Proceedings of the Royal Society B: Biological Sciences 275: 237–247.
17. Blaxter ML (2004) The promise of a DNA taxonomy. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences 359: 669–679.
18. Hebert PDN, Gregory TR (2005) The Promise of DNA Barcoding for Taxonomy. Systematic Biology 54: 852–859.
19. 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 & Evolution 11: 173–192.
20. Hamsher SE, Evans KM, Mann DG, Poulíčková A, Saunders GW (2011) Barcoding Diatoms: Exploring Alternatives to COI-5P. Protist 162: 405–422.
21. MacGillivary ML, 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: 529–536.
22. Moniz MBJ, Kaczmarska I (2009) Barcoding diatoms: Is there a good marker? Molecular Ecology Resources 9: 65–74.
23. Moniz MBJ, Kaczmarska I (2010) Barcoding of Diatoms: Nuclear Encoded ITS Revisited. Protist 161: 7–34.
24. Evans KM, Wortley AH, Mann DG (2007) An assessment of potential diatom "barcode" genes (*cox*1, *rbc*L, 18S and ITS rDNA) and their effectiveness in determining relationships in *Sellaphora* (Bacillariophyta). Protist 158: 349–364.
25. Evans KM, Wortley AH, Simpson GE, Chepurnov VA, Mann DG (2008) A molecular systematic approach to explore diversity within the *Sellaphora pupula* species complex (Bacillariophyta). Journal of Phycology 44: 215–231.
26. Eiler A, Drakare S, Bertilsson S, Pernthaler J, Peura S, et al. (2013) Unveiling Distribution Patterns of Freshwater Phytoplankton by a Next Generation Sequencing Based Approach. PLoS ONE 8: e53516.
27. Shokralla S, Spall JL, Gibson JF, Hajibabaei M (2012) Next-generation sequencing technologies for environmental DNA research. Molecular Ecology 21: 1794–1805.
28. Pawlowski J, Audic S, Adl S, Bass D, Belbahri L, et al. (2012) CBOL Protist Working Group: Barcoding Eukaryotic Richness beyond the Animal, Plant, and Fungal Kingdoms. PLoS Biol 10: e1001419.
29. Kermarrec L, Franc A, Rimet F, Chaumeil P, Frigerio J-M, et al. A next-generation sequencing approach to river biomonitoring using benthic diatoms. Freshwater Science 33(1): 349–363.
30. Jahn R, Kusber WH (2005+) AlgaTerra Information System (online). Botanic Garden and Botanical Museum Berlin-Dahlem, Freie Universität Berlin. http://www.algaterra.org.
31. Geissler U, Kies L (2003) Artendiversität und Veränderungen in der Algenflora zweier städtischer Ballungsgebiete Deutschlands: Berlin und Hamburg. Nova Hedwigia, Beihefte. pp. 777.
32. Hofmann G, Werum M, Lange-Bertalot H (2011) Diatomeen im Süßwasser - Benthos von Mitteleuropa. Bestimmungsflora Kieselalgen für die ökologische Praxis. Über 700 der häufigsten Arten und ihre Ökologie. Lange-Bertalot H (ed.), Gantner, Ruggell, Liechtenstein, Gantner.
33. Guillard RRL, Lorenzen CJ (1972) Yellow-green algae with chlorophyllide C. Journal of Phycology 8: 10–14.
34. Bold H, Wynne M (1978) Cultivation of algae in the laboratory. Introduction to the algae: structure and reproduction. Englewood Cliffs, N. J.: Prentice-Hall INC. pp. 571–578.
35. Krammer K, Lange-Bertalot H (1997) Bacillariophyceae. 1. Teil: Naviculaceae; Ettl H, Gerloff J, Heynig H, Mollenhauer D, editors. Jena, Germany: Fischer.
36. Ettl H, Gärtner G (2013) Syllabus der Boden-, Luft- und Flechtenalgen. 2nd ed. Springer: 773.
37. Levkov Z (2009) Amphora sensu lato. In.: H. Lange-Bertalot (ed.), Diatoms of Europe: Diatoms of the European Inland Waters and Comparable Habitats. pp. 5–916.
38. Levkov Z, Metzeltin D, Pavlov A (2014) *Luticola* and *Luticolopsis*. In: Lange-Bertalot H (ed.), Diatoms of Europe: Diatoms of the European Inland Waters and Comparable Habitats 7. pp7–697.
39. Gemeinholzer B, Droege G, Zetzsche H, Knebelsberger T, Raupach M, et al. (2009+) DNA Bank Network Webportal.
40. Abarca N, Enke N, Zimmermann J, Jahn R (2014) Does the cosmopolitan diatom *Gomphonema parvulum* (Kützing) Kützing have a biogeography? PLOS One 9: e86885.

41. Messing J (1983) [2] New M13 vectors for cloning. In: Ray Wu LGKM, ed. Methods in Enzymology: Academic Press. pp. 20–78.
42. Ivanova NV, Zemlak TS, Hanner RH, Hebert PDN (2007) Universal primer cocktails for fish DNA barcoding. Molecular Ecology Notes 7: 544–548.
43. Müller J MK, Neinhuis C, Quandt D (2010) PhyDE–Phylogenetic Data Editor. Computer program.
44. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32: 1792–1797.
45. Swofford DL (2002) Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4 ed: Sinauer Associates, Sunderland, Massachusetts.
46. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, et al. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28: 2731–2739.
47. Lange-Bertalot H (ed.) (2001) Navicula sensu stricto. 10 genera separated from Navicula sensu lato. Frustulia. Diatoms of Europe 2. Gantner, Ruggell.
48. Bruder K, Medlin L (2007) Molecular assessment of phylogenetic relationships in selected species/genera in the Naviculoid diatoms (Bacillariophyta). Nova Hedwigia 85.
49. Straub F (1990) Compared variability of *Achnanthes lanceolata* (Bréb.) Grunow. 2. Biometrical approach of several races of the sub-species *frequentissima* Lange-Bertalot. In: Ricard M, editor. Ouvrage dedie à la memoire du Professeur Hanry Germain (1903–1989. Königstein, Germany; Champaign, Illinois, USA: Koeltz Scientific Books.
50. Van der Vijver B, Beyens L, Lange-Bertaltot H (ed.) (2004) The genus *Stauroneis* in the Arctic and (Sub-)Antarctic Regions. Bibliotheca Diatomologica, 51.
51. Kermarrec L, Franc A, Rimet F, Chaumeil P, Humbert JF, et al. (2013) Next-generation sequencing to inventory taxonomic diversity in eukaryotic communities: a test for freshwater diatoms. Molecular Ecology Resources 13: 607–619.
52. Zimmermann J, Glöckner G, Jahn R, Enke N, Gemeinholzer B (subm.) Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies. Molecular Ecology Resources.
53. Ratnasingham S, Hebert PDN (2007) BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). Molecular Ecology Notes 7: 355–364.
54. Becker S, Hanner R, Steinke D (2011) Five years of FISH-BOL: Brief status report. Mitochondrial DNA 22: 3–9.
55. Yoccoz NG (2012) The future of environmental DNA in ecology. Molecular Ecology 21: 2031–2038.
56. Bortolus A (2008) Error Cascades in the Biological Sciences: The Unwanted Consequences of Using Bad Taxonomy in Ecology. AMBIO: A Journal of the Human Environment 37: 114–118.
57. Collins RA, Cruickshank RH (2013) The seven deadly sins of DNA barcoding. Molecular Ecology Resources 13: 969–975.
58. Kvist S (2013) Barcoding in the dark?: A critical view of the sufficiency of zoological DNA barcoding databases and a plea for broader integration of taxonomic knowledge. Molecular Phylogenetics and Evolution 69: 39–45.
59. Carpenter SR, Mooney HA, Agard J, Capistrano D, DeFries RS, et al. (2009) Science for managing ecosystem services: Beyond the Millennium Ecosystem Assessment. Proceedings of the National Academy of Sciences 106: 1305–1312.
60. Marakeby H, Badr E, Torkey H, Song Y, Leman S, et al. (2014) A System to Automatically Classify and Name Any Individual Genome-Sequenced Organism Independently of Current Biological Classification and Nomenclature. PLoS ONE 9: e89142.
61. Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proceedings of the Royal Society of London Series B: Biological Sciences 270: 313–321.
62. Ward RD, Zemlak TS, Innes BH, Last PR, Hebert PDN (2005) DNA barcoding Australia's fish species. Philosophical Transactions of the Royal Society B: Biological Sciences 360: 1847–1857.
63. Saunders GW (2005) Phil Trans R Soc B 360: 1879.
64. Robba L, Russell SJ, Barker GL, Brodie J (2006) Assessing the use of the mitochondrial cox1 marker for use in DNA barcoding of red algae (Rhodophyta). American Journal of Botany 93: 1101–1108.
65. Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM (2004) Identification of Birds through DNA Barcodes. PLoS Biol 2: e312.
66. Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, et al. (2009) From the Cover: A DNA barcode for land plants. Proceedings of the National Academy of Sciences 106: 12794–12797.
67. Droege G, Barker K, Astrin JJ, Bartels P, Butler C, et al. (2013) The Global Genome Biodiversity Network (GGBN) Data Portal. Nucleic Acids Research.
68. GBIF (2014) Global Biodiversity Information Facility. Published on the Internet at http://www.gbif.org, accessed 2014 April 29.
69. McNeill J, Barrie FR, Buck WR, Demoulin V, Greuter W, et al. (2012) Regnum Vegetabile 154: 1.
70. (1999) Nomenclature. ICoZ, Ride WDL, Nomenclature. ITfZ, Sciences. IUoB, Museum NH International code of zoological nomenclature = Code international de nomenclature zoologique. London: International Trust for Zoological Nomenclature, c/o Natural History Museum.
71. Enke N, Thessen A, Bach K, Bendix J, Seeger B, et al. (2012) The user's view on biodiversity data sharing — Investigating facts of acceptance and requirements to realize a sustainable use of research data —. Ecological Informatics 11: 25–33.
72. Cocquyt C, Jüttner I, Kusber W-H (2013) Reinvestigation of West African Surirellaceae (Bacillariophyta) described by Woodhead and Tweed from Sierra Leone. Diatom Research 28: 121–129.
73. Jahn R, Kusber W-H, Romero OE (2009) *Cocconeis pediculus* Ehrenberg and *C. placentula* Ehrenberg: Typification and taxonomy. Fottea 9: 275–288.
74. Romero OE, Jahn R (2013) Typification of *Cocconeis lineata* and *Cocconeis euglypta* (Bacillariophyta). Diatom Research 28: 175–184.

METABARCODING VS. MORPHOLOGICAL IDENTIFICATION TO ASSESS DIATOM DIVERSITY IN ENVIRONMENTAL STUDIES

JONAS ZIMMERMANN

GERNOT GLÖCKNER

REGINE JAHN

NEELA ENKE

BIRGIT GEMEINHOLZER

2014

Molecular Ecology Resources. DOI: 10.1111/1755-0998.12336

MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2014) doi: 10.1111/1755-0998.12336

Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies

JONAS ZIMMERMANN,*† GERNOT GLÖCKNER,‡§ REGINE JAHN,† NEELA ENKE† and BIRGIT GEMEINHOLZER*
*Justus-Liebig-University Giessen, AG Spezielle Botanik, Heinrich-Buff-Ring 38, 35392 Giessen, Germany, †Botanic Garden and Botanical Museum Berlin-Dahlem, Freie Universität Berlin, Königin-Luise-Str. 6-8, 14195 Berlin, Germany, ‡Institute for Biochemistry I, Medical Faculty, University of Cologne, Joseph-Stelzmann-Str. 52, 50931 Cologne, Germany, §Molecular Biology Group, Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Müggelseedamm 310, 12587 Berlin, Germany

Abstract

Diatoms are frequently used for water quality assessments; however, identification to species level is difficult, time-consuming and needs in-depth knowledge of the organisms under investigation, as nonhomoplastic species-specific morphological characters are scarce. We here investigate how identification methods based on DNA (metabarcoding using NGS platforms) perform in comparison to morphological diatom identification and propose a workflow to optimize diatom fresh water quality assessments. Diatom diversity at seven different sites along the course of the river system Odra and Lusatian Neisse from the source to the mouth is analysed with DNA and morphological methods, which are compared. The NGS technology almost always leads to a higher number of identified taxa (270 via NGS vs. 103 by light microscopy LM), whose presence could subsequently be verified by LM. The sequence-based approach allows for a much more graduated insight into the taxonomic diversity of the environmental samples. Taxa retrieval varies considerably throughout the river system, depending on species occurrences and the taxonomic depth of the reference databases. Mostly rare taxa from oligotrophic parts of the river systems are less well represented in the reference database used. A workflow for DNA-based NGS diatom identification is presented. 28 000 diatom sequences were evaluated. Our findings provide evidence that metabarcoding of diatoms via NGS sequencing of the V4 region (18S) has a great potential for water quality assessments and could complement and maybe even improve the identification via light microscopy.

Keywords: environmental DNA barcoding, next-generation sequencing, diatoms, DNA-based specimen identification, water monitoring

Received 22 February 2014; revision received 20 September 2014; accepted 24 September 2014

Introduction

DNA barcoding (Hebert *et al*. 2003) introduced a DNA sequence based identification alternative. DNA barcodes link an individual to a taxon with a similar genetic sequence in a reference database (Ratnasingham & Hebert 2007). The DNA barcode consists of a marker fragment easily sequenced in one Sanger read that unambiguously identifies a given taxon (Hebert *et al*. 2003; Stoeckle 2003; Moritz & Cicero 2004; Zimmermann *et al*. 2011), independent of its life stage (Hajibabaei *et al*. 2005, 2012; Mann *et al*. 2010; Shokralla *et al*. 2012; Eiler *et al*. 2013). Following the concept of Collins & Cruickshank (2013), DNA barcoding can be used for specimen identification as well as species discovery.

Correspondence: Jonas Zimmermann, Fax: +393083850142; E-mail: j.zimmermann@bgbm.org

While DNA barcoding describes the linkage of a specimen to a unique DNA sequence, environmental DNA (eDNA) barcoding (Kermarrec *et al*. 2014) or metabarcoding refers to an approach that uses this linkage to identify taxa from mixed samples to evaluate the community composition in a target environment; an approach that relies on high-throughput techniques, such as next-generation sequencing (NGS). NGS-based eDNA barcoding allows for automated sample handling as well as standardized laboratory protocols. All these points could increase the possibility to compare data sets from different studies and laboratories. A standardized approach will also facilitate monitoring programs with large spatiotemporal dimensions (Hajibabaei *et al*. 2011, 2012). The progress of DNA sequencing technologies will also allow for sequencing multiple markers in parallel (multiplexing) (Dupuis *et al*. 2012; Fujita *et al*. 2012;

Collins & Cruickshank 2013) or even whole genomes of at least small organelles (Taberlet *et al.* 2012a,c; Eiler *et al.* 2013).

However, whereas the protocols for generating sequences from environmental samples are well established, currently the assignment of sequences to taxon names for protists (e.g. diatoms) is still in development.

Diatoms are photoautotrophic unicellular microalgae that occur in mostly aquatic habitats – freshwater as well as marine – but also in soil or as aeroplankton (Kawecka & Olech 1993; Spaulding & McKnight 1999). Life strategies include planktonic forms, benthic species adhering to many different kinds of substrate (e.g. mud, sand, stones, plants, algae, animals), but also appearing in interstitial and intertidal communities. An estimated number of over 200 000 diatom species worldwide exist (Smetacek 1999; Mann & Vanormelingen 2013), many of them yet undiscovered and not described. The species differ considerably in size and in the shape of their silica cell frustule (Mann & Droop 1996).

Diatoms are important elements of ecosystems as food source for many other organisms, due to their capability to fixate carbon dioxide and silica. Diatoms are responsible for at least 25% of global carbon dioxide fixation (Falkowski *et al.* 1998; Field *et al.* 1998; Mann 1999; Smetacek 1999). Furthermore, they contribute approximately 20% of the global net primary production (Mann 1999). Some diatom species have a broad ecological plasticity while others are adapted to specific environmental conditions. Even closely related species are often found under vastly different ecological conditions (Poulíčková *et al.* 2008; Vanelslander *et al.* 2009). Therefore, the specific composition of diatom communities is a sensitive indicator of changes in their environment, for example due to climatic change, altering nutrient influx and an increasing anthropogenic pressure (Eiler *et al.* 2013).

In many countries (Descy & Coste 1991; Kelly 1998; Stevenson & Pan 1999; Prygiel *et al.* 2002; Zalack *et al.* 2010), diatom indices based on microscopic approaches have been developed and are being used for water quality assessments, for example in the European Water Framework Directive (WFD) (Kelly *et al.* 1998; European-Committee-for-Standardization 2003, 2004). Quality control of water bodies is crucial as they account for over 70% of the earth's surface (not counting phreatic water systems and ice formations) and represent a large economic value for human society (Zinger *et al.* 2012). Diatom-based indices require unambiguous taxa identification to the species level. The characters of the frustules have been the basis for the applied morphological and mostly descriptive species concept in diatoms so far. It needs expert taxonomic knowledge and often expensive infrastructure to identify diatoms morphologically below the genus level as many of the morphological characters can only be detected by scanning electron microscopy or similar high-resolution technologies. Furthermore, morphological characters can vary substantially within populations or clonal cultures (Bailey-Watts 1976; Jahn 1984; Medlin *et al.* 1991; Zinger *et al.* 2012; Vijver *et al.* 2013). And even if all necessary characters for identification are available, trained taxonomists sometimes come to different conclusions, either due to the use of different criteria of species delimitations or because the descriptions available allow for different interpretations (Mann *et al.* 2010). For these reasons, many ecological assessments are misleading due to misidentifications (Archibald 1984; Morales *et al.* 2001).

Metabarcoding as applied here for the identification of specimens is based on the similarity between known and unknown sequences, estimated in terms of differences in the nucleotide sequences. Contrary to morphological characters, and given that the method of genetic distance calculation is provided, genetic distances estimations are not taxonomist dependent. Furthermore, our goal is only to identify specimens (which does not rely on any given species concept), not to newly delimit species, and the identified units can be transferred to any taxonomic concept (Rach *et al.* 2008).

Most of the published analyses up till now tested the feasibility of DNA barcoding of diatoms using classical Sanger sequencing on cultured strains, *in silico*, or in mock communities (artificially mixed samples). The latter showed that NGS-based DNA barcoding retrieves a high percentage of taxa (Kermarrec *et al.* 2013b), respectively, all the taxa (Zimmermann *et al.* 2010) that were included in the artificial mixed sample. First attempts at using eDNA barcoding for diatom specimen identification in environmental samples produced multiple sequences from one sample via cloning and Sanger sequencing (Jahn *et al.* 2007; Stoof-Leichsenring *et al.* 2012); however, Kermarrec *et al.* (2014) recently presented a NGS-based study for eDNA barcoding in diatoms with promising results.

Even though the applicability of DNA barcoding and eDNA barcoding for the identification of diatoms has generally been demonstrated, there is an ongoing discussion about the 'ideal' DNA barcoding marker. Mainly 18S SSU rDNA (V4 region) (Zimmermann *et al.* 2011; Luddington *et al.* 2012), 28S LSU rDNA (D2/D3 region) (Hamsher *et al.* 2011), *cox*1 (Evans *et al.* 2007, 2008) and different parts of *rbc*L (Hamsher *et al.* 2011; MacGillivary & Kaczmarska 2011; Kermarrec *et al.* 2013b) are being discussed as standard DNA barcoding marker. We decided to use the 18S V4 region as previous studies showed its feasibility (Zimmermann *et al.* 2011; Luddington *et al.* 2012), especially for eDNA barcoding. The V4 region is the most variable region in the approximately 1800 bp long 18S rDNA gene (Alverson *et al.* 2006) and has been declared the prebarcode for protists by the CBOL protist working group (ProWG) (Pawlowski *et al.* 2012).

The aim of this study is to investigate whether the identification of diatom taxa from environmental samples using metabarcoding is comparable to the identification using light microscopy on a level currently used for water quality assessments under the WFD. We present the results from a NGS eDNA barcoding approach on environmental samples from the Lusatian Neisse/Odra river system on the border between Poland and Germany. The direct comparison of taxa lists derived by each method showed that the NGS eDNA barcoding approach is qualitatively at least on the same level as light microscopy, but often provides a greater taxonomic depth than the classical morphology-based approach.

Materials and methods

Sampling

In May 2011, benthic water samples were taken from seven sampling sites along the River Lusatian Neisse and the River Odra, from the Lusatian Neisse spring in the Czech Republic (near Liberec/Jablonec), through Ratzdorf where the River Neisse joins the Odra, to the last distinct limnic part of the Odra/Lusatian Neisse river system in Schwedt at the German/Polish border, approximately 50 km away from the estuary of the Baltic Sea (Fig. 1, Table 1). Figure 2 summarizes the following procedural steps.

For each sampling site, benthic epilithic diatoms were harvested with a toothbrush (a new toothbrush for each sample site) from stones located in 10–30 cm depth, according to the sampling strategy of the European Water Framework Directive (WFD) (Kelly *et al.* 1998; European-Committee-for-Standardization 2003, 2004). The diatoms attached to the toothbrush were resuspended in a total amount of 60 ml filtered (0.45 μm) river water. The samples were homogenized and partitioned into three subsamples (a, b, c) of 20 ml each. Two samples (a, b) were immediately deep frozen (−24°C) in an Engel MB-15® (Roega/Engel, Freilassing, Germany) mobile freezer. Subsample c was stored by room temperature with regular air supply until further processing. The first subsample (i) was used for DNA extraction, the second (ii) for the morphological analysis and the third (iii) for the establishment of clone cultures from individual cells to broaden the reference database (Jahn & Kusber 2005+).

Molecular analysis via NGS

DNA extraction. For the extraction of the genomic DNA, the NucleoSpin ® Plant II Mini Kit (Machery and Nagel, Düren, Germany) was used. Subsamples a and the material from the clonal cultures established from subsample c were defrosted and centrifuged with 6797 *g* for 5 min. The supernatants were carefully discarded. The pellets were resuspended in the DNA extraction buffer PL1. From here on, DNA extraction was performed following the manufacturers protocol, expect for the two elution steps. Here, only 32 μl instead of 50 μl elution buffer PE was used. Until further use, the DNA was stored by −24°C.

Fig. 1 Map of the seven sampling sites RL1.1 (1), RL2.1 (2), RL3.1 (3), RL4.1 (4), RL5.1 (5), RL6.1 (6), RL7.1 (7) alongside the Lusatian Neisse/Odra river system.

PCR. For amplifying the 18S V4 locus (approximately 390–400 bp) using the primer pair M13F-D512for 18S and M13R-D978rev 18S (Zimmermann *et al.* 2011), we followed Zimmermann *et al.* (2011) with two ameliorations; the total reaction volume was increased to 50 μl, and the primer annealing temperature was 56°C.

The PCR products were checked, purified and measured for amount of DNA according to Zimmermann *et al.* (2011). For each site, PCR products of three independent reactions were pooled.

Primer specificity. To test the specificity of M13F-D512 for 18S and M13R-D978rev 18S, TestPrime (Klindworth *et al.* 2013) was used. A maximum of 1 mismatch was allowed with a 0-mismatch zone at 3′end of 5 nucleotides as this is a more realistic simulation of PCR behaviour according to the TestPrime tutorial. The primers were checked against EMBL/GenBank entries.

Table 1 Sampling site locations and georeferences along the Lusatian Neisse/Odra river system

Site	Location	River	Latitude	Longitude
RL1.1	Jablonec nad Nisou	Lusatian Neisse	50°43′56.90″N	15°13′34.08″E
RL2.1	Hagenwerder	Lusatian Neisse	51°04′12.10″N	14°58′03.60″E
RL3.1	Briesnig	Lusatian Neisse	51°48′29.20″N	14°36′05.20″E
RL4.1	Ratzdorf	Lusatian Neisse	52°04′08.70″N	14°45′21.50″E
RL5.1	Slubice	Odra	52°21′04.00″N	14°33′17.10″E
RL6.1	Hohenwutzen	Odra	52°50′44.30″N	14°07′22.80″E
RL7.1	Schwedt/Krajnik Dolny	Odra	53°02′07.60″N	14°18′40.20″E

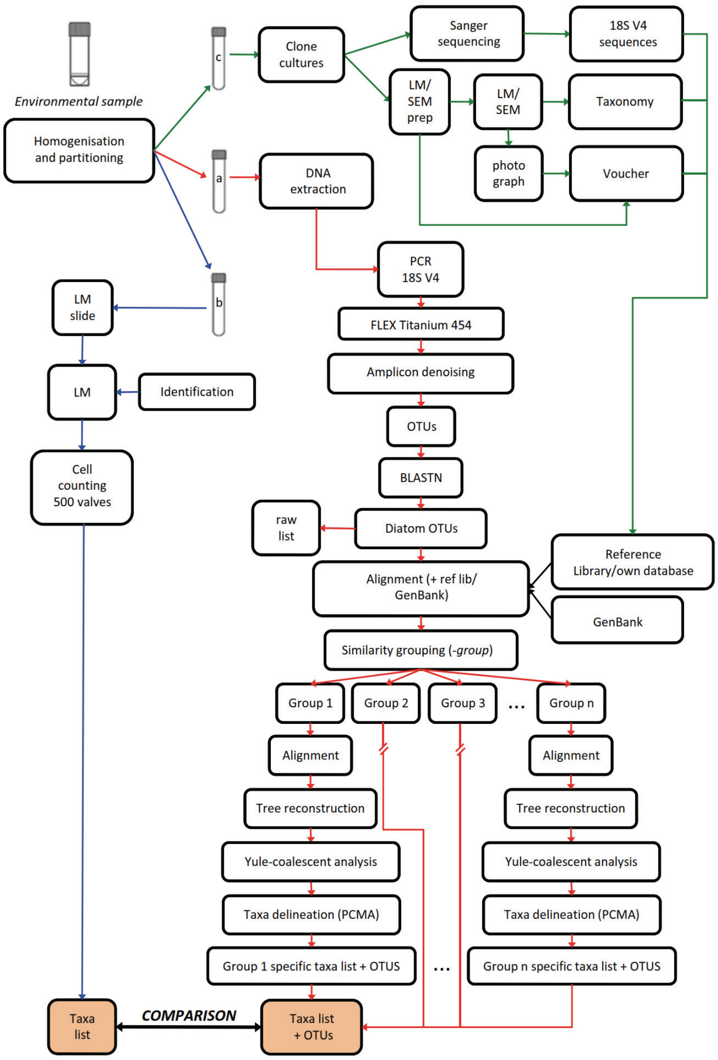


Fig. 2 General schema of the workflow for the handling of the environmental samples. Sample a treatment and evaluation strategy for the NGS approach. Sample b treatment and evaluation for the LM method. Sample c treatment to broaden the own reference library/database.

Roche®/454 sequencing. Adaptors containing multiplex identifier (MIDs) tags for sample tagging were ligated to the PCR amplicons. These tagged amplicons were pyrosequenced on a Roche®/454 GS FLX system (454 Life Sciences, Branford, CT, USA) using titanium chemistry according to the manufacturers' protocol.

Base calling and quality trimming were executed with the software integrated in the Roche® 454 system.

Data processing. Only sequences with a complete primer sequence and longer than 250 bp were retained for subsequent analysis. Further filtering was executed following Huse *et al.* (2007). The reads were sorted into bins according to the MID tag sequences unique for each of the seven sample sites. To remove PCR and 454 pyrosequencing noise AmpliconNoise Version 1.24 (Quince *et al.* 2011) with denoising, algorithms were performed.

The sequences were run with the BLASTn (Altschul *et al.* 1990) algorithm for a preliminary genera and subgeneric specimen identification and to get first insights into the approximate diversity. As the focus of this study is on diatoms in environmental samples, the none diatom sequences were filtered out and neglected. In a 2nd step, the received, edited data sets were evaluated according to sequence quality via the sequence E values. Additionally, a 90% identity value cut-off for phylogenetic reconstructions was applied.

Reference database. As reference we used a) our own reference database (Jahn & Kusber 2005+; Zimmermann *et al.* 2014) with extensions (J. Zimmermann, N. Abarca, W.H. Kusber, N. Enke, & R. Jahn, unpublished) of diatom sequences (it contains about 1500 strains of ca 600 taxa, alone 120 different species from Germany and the investigated area) and b) all diatom sequences available from GenBank. Our own reference database was mainly used for initial taxon clustering, the trees were constructed mainly on sequences publically available in GenBank. An exception the genus *Cocconeis*, which is underrepresented in GenBank.

Operational taxonomic unit (OTU) construction. For OTU construction, we used the program USEARCH (Edgar 2010) and its implemented tools. Here, the raw reads were aligned and identical or nearly identical sequences (threshold value 98% identity) were clustered to yield the OTUs. Our reference database (Jahn & Kusber 2005+; Zimmermann *et al.* 2014) was used as template to identify potential chimeric sequences and singletons in our data set. As this reference database cannot cover all sequence variability in nature and therefore our approach might yield falsely classified chimeras, we kept the OTUs assigned to be chimeras for subsequent analyses. The chimeras were evaluated manually as well as using the software KeyDNATools (Guillou *et al.* 2008) for an internal check against our reference database (J. Zimmermann, N. Abarca, W.H. Kusber, N. Enke, & R. Jahn, unpublished) and, for example (Zimmermann *et al.* 2014), which consists of curated sequences annotated to accessible and taxonomically curated slides deposited at the Herbarium Berolinense (B). The assignable OTUs were reintegrated in the data set and subsequently used as otus in the manuscript (always in minuscule). Distinct chimeras and singletons were removed following Behnke *et al.* (2011).

Specimen identification. For further analysis, sequences representing centric diatoms were removed from the data set as mainly benthic pennates are used for water quality assessments according to the WFD. Sample site-specific data sets for all seven subsamples were created including the newly generated and edited 454 sequences from this study, diatom sequences from GenBank and our own yet unpublished diatom reference database. These data sets were aligned using the MEGA software (Tamura *et al.* 2011) and the implemented MUSCLE (Edgar 2004) alignment algorithm. Instead of the default settings, the –group function for clustering the sequences after similarity within the alignment was used and the gap penalties were altered (gap open −350/gap extend −260) after tests to find the best fitting alignments.

The obtained clusters/clades were re-aligned in new cluster-/clade-specific optimized alignments. Different tree constructing programs and models were evaluated on random cluster alignments. The neighbour-joining algorithm (Saitou & Nei 1987) with the Kimura 2-parameter model (Kimura 1980) implemented in MEGA (Tamura *et al.* 2011) was used to generate phylogenetic trees with 10 000 bootstrap replicates for every cluster-specific alignment. A multiple threshold general mixed Yule-coalescent (GMYC) model approach (Pons *et al.* 2006; Fontaneto *et al.* 2007) was performed. The GMYC model approach attempts to identify taxon boundaries from the variation in branching rates of a given tree (Monaghan *et al.* 2009). The GMYC approach was executed with trees reconstructed with a relaxed lognormal coalescent according to the findings of Monaghan *et al.* (2009). The trees were imported into the statistics software R 2.15.1 (ProjectR-http://cran.r-project.org). This GMYC analysis required the R package SPLITS obtainable from the RForge website (Monaghan *et al.* 2009; Hoef-Emden 2012; Barraclough 2013). Due to the weaknesses ascribed to the multiple thresholds, GMYC model approach (Hoef-Emden 2012) and to refine the specimen identification an additional evaluation of the bootstrap values within the cluster-specific trees was carried out. A 60% bootstrap support was chosen according to values after Aquilino *et al.* (2011), Chen *et al.* (2011) and Göker *et al.* (2010) to represent a distinct taxon. This 60% bootstrap support was then used to re-evaluate the clusters on the phylogenetic tree. Supported clusters were considered separate taxa. This approach is subsequently referred to as phylogenetic-based coalescent model approach (PCMA).

All corresponding cluster alignments from the seven sampling sites were manually concatenated to 7 combined data sets. The Newick format of the bootstrapped phylogenetic subtrees was used to construct amalgamated trees in a 'supertree' approach, based on a schematic standard 'supertree' as no comprehensive limnic diatom classification has been published, yet (Medlin & Kaczmarska 2004; Bruder *et al.* 2008; Theriot *et al.* 2010; Ruck & Theriot 2011) (File S2: Figs S2–S7). This approach was necessary, as the V4 region is too diverse for correct phylogenetic cluster assignment throughout the diatoms to prevent clusters from collapsing. The 'supertrees' were built for visualization purposes only. The Newick data sets were transferred to MEGA (Tamura *et al.* 2011) for visualization.

Morphological analysis via LM

Microscopic identification and counting. For the morphological analysis of the diatom composition in each sample, the b subsamples containing the diatom frustules were used. Sampling, cleaning, identification and counting procedures were following standard procedures (Kelly *et al.* 1998; European-Committee-for-Standardization 2003, 2004), which correspond to the strategy of the WFD. The diatom frustules were purified in 30% H_2O_2, and in following steps, the peroxide residues were washed of by alternating repeats of centrifugation and rinsing. The cleaned subsamples were dispersed on cover glasses, dried and embedded in the high refractive medium Naphrax™ for further analysis.

Counts and identifications of the diatom valves were carried out at 1000× magnification using an Axio Imager 2® (Carl Zeiss, Oberkochen, Germany). Microphotographs were produced with the integrated camera using a differential interference contrast (DIC) filter. Identification is mainly based on Hofmann *et al.* (2011).

All taxa encountered in transects across the slides were identified and counted until a minimum of 500 valves was scored. Relative frequencies of all taxa in each sample were calculated.

Comparison of NGS and morphological analysis

To access the comparability of both methodologies, the taxa lists derived from both NGS and morphological approaches were compared in regard to taxa presence or absence and species composition.

Results

Molecular analysis via NGS

Sequence output. The number of V4 sequences per site after removal of incomplete reads and reads with insufficient quality was 9629 for site RL1.1, 28242 for site RL2.1, 7632 for site RL3.1, 9832 for site RL4.1, 10075 for site RL5.1, 13415 for site RL6.1 and 6942 for site RL7.1 (Table 1 and Table 2). The total number of diatom sequences (pennates and centrics) for each sampling site and in total is given in Table 2. For further evaluation, only the sequences of benthic pennate diatoms were processed as these are mainly employed in the WFD. An overview on centric diatom genera found in the samples (neglected in this study) is given in File S1: Table S1. The number of detected pennate diatom sequences per sample ranged from about 1000 to more than 3000 depending on the sampling site. The average number of qualitatively sufficient sequences per site was 1630. The total number of pennate sequences across all seven sampling sites was 11 411 (Table 2).

Primer specificity. The test of primer specificity showed that 81% of sequences deposited as Bacillariophyceae in EMBL/GenBank and identified as long enough were recovered by the here used primers (a total of 880 sequences or 22%). 11% of all eukaryotic sequences in EMBL/GenBank were recovered, but these also included the diatom sequences. In total, 3020 sequences (~ 78%) of nondiatom eukaryotes (approximately 1600 sequences from other Stramenopiles) were recovered in comparison to 880 diatom sequences. Therefore the ratio of diatom sequences to other eukaryote sequences is about 1:3.4.

Table 2 Number of sequences (and OTUs/otus) from 454 sequencing for sampling sites RL1.1-RL7.1 for main diatom groups pennates and centrics. otus indicate false chimeras that were reintegrated into data set after automatic and manual re-evaluation against our own reference database.

	Sampling Site							
Taxon	RL1.1	RL2.1	RL3.1	RL4.1	RL5.1	RL6.1	RL7.1	RL1.1–7.1
Pennates *no. of sequences*	1060	3037	1066	1015	1760	2295	1178	11 411
Pennates *no. of OTUs*	34	61	48	40	42	54	37	316
Pennates *no. of otus*	59	118	45	60	97	95	72	546
Centrics *no. of sequences*	0	680	847	1577	5523	4579	3121	16 327

Specimen identification. The MUSCLE (Edgar 2004) – group function in MEGA (Tamura *et al.* 2011) produced 9 clusters/clades per sampling site, which were split and re-aligned to create cluster-/clade-specific alignments.

The test of several tree reconstruction programs, for example MEGA (Tamura *et al.* 2011), PAUP (Swofford 2002) and MrBayes (Huelsenbeck & Ronquist 2001) and the different implemented algorithms on randomly chosen alignments showed no significant differences. So the cluster-/clade-specific trees with 10 000 bootstrap replicates were constructed in MEGA (Tamura *et al.* 2011) with the implemented neighbour-Joining algorithm with the Kimura 2-parameter model as it was the best trade-off between accuracy and velocity. The cluster-/clade-specific trees per sampling site were assembled in a 'supertree' approach into one general diatom tree per sampling site in accordance to the presumed phylogenetic diatom relationships (File S2: Figs S2–S7).

The multiple threshold general mixed Yule-coalescent (GMYC) model approach (Pons *et al.* 2006; Fontaneto *et al.* 2007) after Monaghan *et al.* (2009) did not recover the pre-assumed taxa discrimination; hence, it was combined with the evaluation of the bootstrap supports which led to satisfactory results.

Per sampling site 7–9 group-specific neighbour-joining trees of here generated sequences and additional sequences retrieved via BLAST searches from GenBank were combined in one 'supertree' each. Without the 'supertree' approach, clustering would not reflect phylogenetic relationships. The tree of sampling site RL6.1 is shown in Fig. 3 as example for specimen identification via PCMA. The trees for the other sampling sites are available in the File S2: Figs S2–S7. Clades supported by >60% bootstrap values were treated as distinct. Clades were unambiguously assigned to a specific taxon if a database hit as well as a morphological correspondence in the LM analysis was found (e.g. *Reimeria sinuata*, *Encyonema silesacum*, *Mayamaea atomus* var. *permitis* Fig. 3 and File S3: Table S8), sequences associated in distinct but less supported clades were treated as unspecified members of a genus (e.g. *Amphora* spec., *Fragilaria* spec. Fig. 3 and File S3: Table S8), respectively, as cf. if a morphological or a molecular correlation was possible (e.g. *Gomphonema* cf. *bourbonense*, *Navicula* cf. *cryptocephala* Fig. 3 and File S3: Table S8).

The number of genera recovered by name assignment through BLAST and, respectively, through PCMA was similar for sites RL1.1, RL2.1 and RL6.1. At the other sampling sites, PCMA recovered from 3 to 9 more genera (Table 3). Over all, with PCMA more genera could be assigned than with BLAST, with the exception of sample site RL6.1. This trend is even more distinct and increasing when looking at subgeneric taxa: PCMA recovered at least twice as many taxa (Table 3). Furthermore, all undefined units by the BLAST only approach – for example uncultured diatoms or Bacillariophyta – could be assigned using the PCMA approach (Table 3). With NGS sequencing using the PCMA approach from 11 to 20 genera per site in Table 3 and from 12 to 19 genera per site in Table 4 could be unambiguously assigned, leading to 28 identified genera in total and the arithmetic mean of 16 genera per sampling site. Subgeneric specimen identification via NGS and PCMA varied from 50 to 83, with 263 assigned taxa in total and 70 per site as arithmetic mean (Table 4).

The taxa lists derived from these approaches are given in File S3: Table S8 and serve as basis for the comparison with the morphological approach.

Morphological analysis via light microscopy (LM)

For the compilation of the taxa lists, 500 valves were counted for each sampling site using LM recording also the quantity of the occurrence of each taxon. 17–21 genera per site in Table 4 and 17–20 genera per site in Table 4 were identified with 30 different genera in total and an arithmetic mean of 18 genera per site (Table 4). The diatom valves could be assigned to 31–41 different subgeneric taxa, leading to 102 taxa in total and 37 as arithmetic mean per sampling site. A taxa list for all seven sampling sites is given in File S3: Table S2, the quantities are available from the first author upon request.

Comparison of NGS and LM analyses

A total of 34 different genera could be detected; 28 genera were retrieved by the NGS approach and 30 genera by the LM approach (Table 4). On the subgeneric level, the NGS approach allowed for the identification of 263 taxa versus 102 taxa detected via the LM approach.

In general, the NGS approach recovered more taxa than the classical approach using LM (e.g. sample site RL2.1: LM 34 taxa, NGS 83 taxa, Table 4). Also, more taxa recovered with NGS were missing using the LM, than vice versa (e.g. sample site RL2.1: 9 LM taxa were missing in NGS analysis, whereas 69 NGS taxa were missing in the LM approach, Table 4). On the genus level, the numbers of recovered taxa by both methods were more similar than on the subgenus level (Table 4), the number of taxa recovered by NGS was either the same or lower than with LM (e.g. sample site RL2.1: 18 genera with LM, 16 with NGS, Table 4). The highest difference on the genus level was found on sample site RL1.1, where 17 genera could be identified via LM and 12 genera via NGS (Table 4).

Figures 4 and 5 summarize the differences between the LM and the NGS approach individually for each genus and each sampling site. Seven genera (*Achnanthidium*,

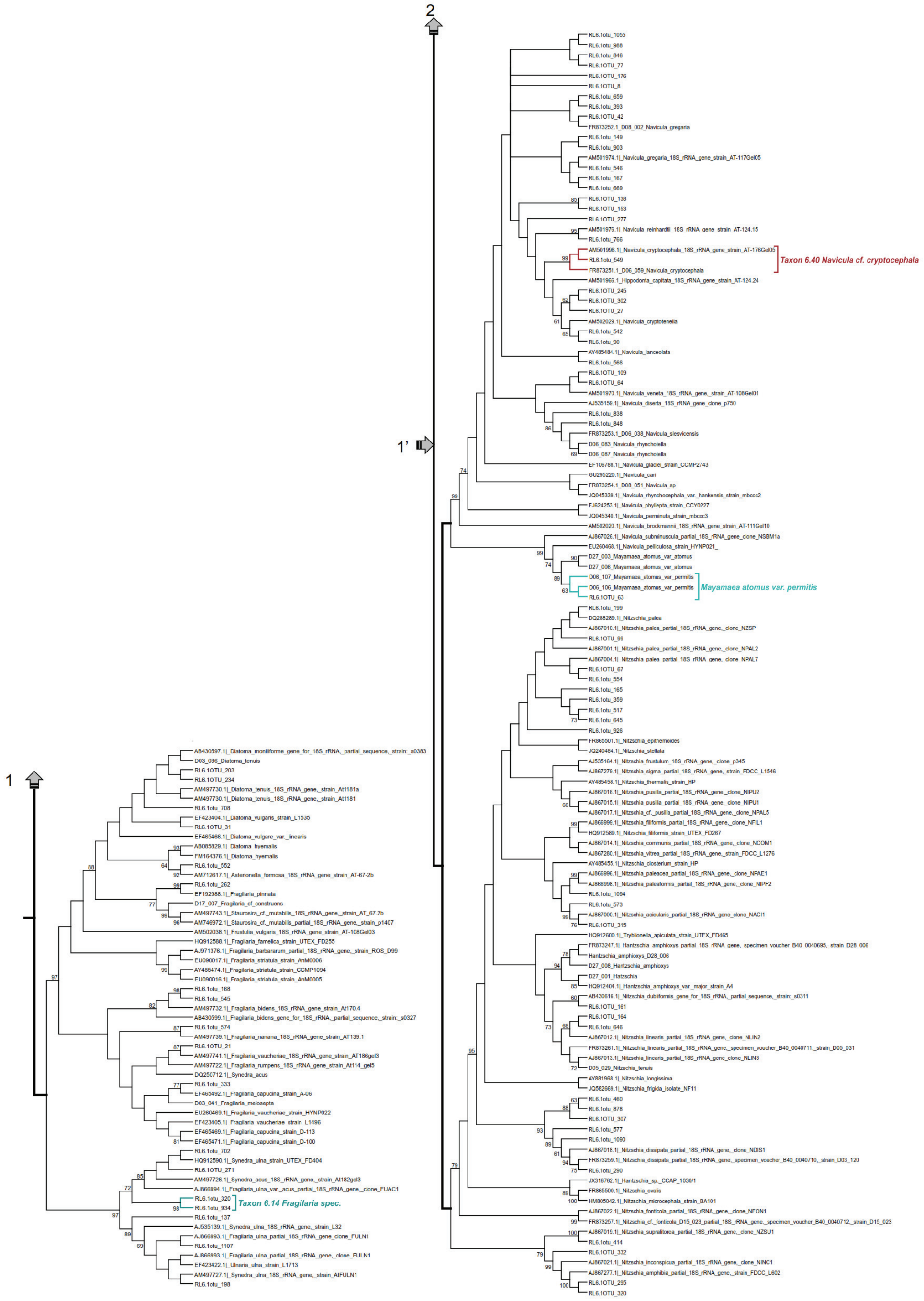


Fig. 3 General diatom tree derived from sampling site RL6.1 and additional GenBank entries, generated from 8 independent cluster-/clade-specific neighbour-joining trees, in accordance to the presumed phylogenetic diatom relationships. 'Supertree' branches are depicted in bold black lines while nonbold lines refer to phylogenetic reconstructions. Bootstrap values >50 are shown above the branches. Brackets refer to examples in the text.

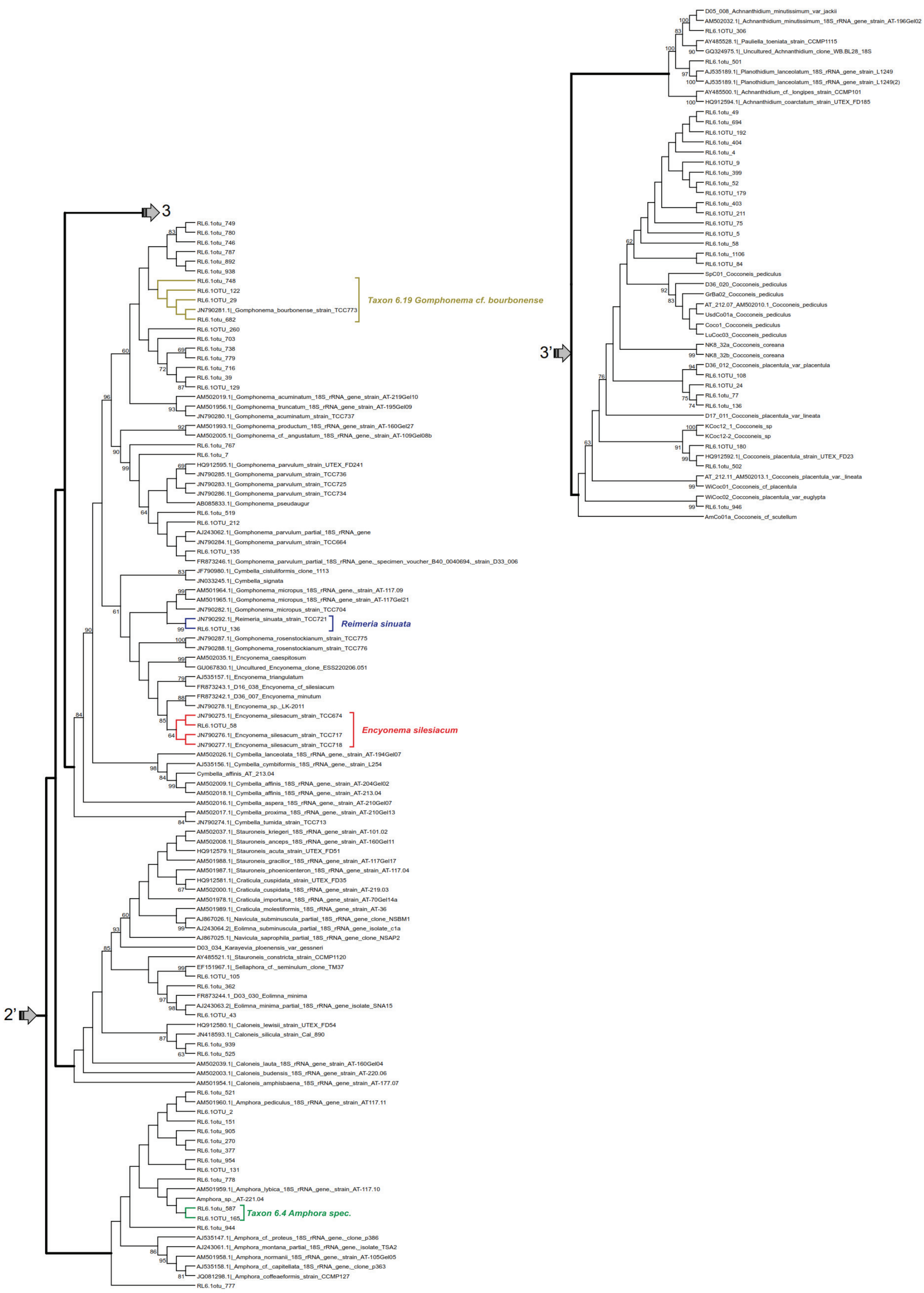


Fig. 3 Continued.

Table 3 Comparison of specimen identification between BLASTn and the phylogenetic-based coalescent model approach (PCMA) based on the data from 454 sequencing

	Approach				
	BLASTn			PCMA	
Sample site	Genera	Subgenus taxa	Unassigned units	Genera	Subgenus taxa
RL1.1	10	20	28	11	50
RL2.1	16	38	30	16	83
RL3.1	11	25	21	20	63
RL4.1	9	24	25	16	67
RL5.1	16	34	33	19	83
RL6.1	17	52	54	16	78
RL7.1	13	30	27	16	69

Table 4 Number of retrieved taxa on different taxonomic levels compared between LM (light microscopy) and NGS (next-generation sequencing) approaches given for all seven sampling sites

	Genera		Subgenus taxa			
Sample site	LM	NGS	LM	NGS	Missing in NGS	Missing in LM
RL1.1	17	12	36	50	18	31
RL2.1	18	16	34	83	9	69
RL3.1	19	19	38	63	15	40
RL4.1	18	16	36	67	13	45
RL5.1	19	19	40	83	10	56
RL6.1	16	16	31	78	7	55
RL7.1	20	16	41	69	16	44
Arithmetic mean	18	16	37	70	13	49
RL1.1–RL7.1	30	28	102	263		

Diatoma, Fragilaria, Gomphonema, Navicula, Nitzschia, Planothidium) were identified with both methods at all sites, while *Encyonema* and *Eolimna* are presumably also present on all sites, however, were missing either by LM (for *Encyonema* RL6.1, Fig. 5) or NGS (for *Encyonema* and *Eolimna* RL1.1, Fig. 5) identification. Taxa composition at RL1.1 varied distinctly in comparison to all other sites, which was consistent in both methods. Four genera, namely *Asterionella, Craticula, Cymbella* and *Tryblionella* were only found by NGS, whereas LM identification recovered five genera not found with NGS (*Geissleria, Luticola, Meridion, Parlibellus* and *Rhoicosphenia*; Figs 4, 5). Via the NGS approach, the largest amount of taxa was found in the genera *Nitzschia* (49); *Navicula* (40) and *Fragilaria* (29); using the LM approach the highest taxa number was found in the genera *Gomphonema* (37), *Fragilaria* (18) and *Nitzschia* (17) (Figs 4, 5). Except for very few genera (e.g. *Karayevia, Meridion, Stauroneis*; Figs 4, 5), NGS provided the higher number of taxa.

The comparison of abundances of the counted cells via LM and the received number of sequences from the NGS approach for taxa which were detected by both methods, showed no significant correlation.

Discussion

Molecular methodology

Metabarcoding of environmental samples can be applied on different organismic levels: either to access the biodiversity over a wide taxonomic range or of a specific group of organisms. But even with NGS approaches, only a limited number of sequences can be accurately generated and evaluated in a time- and cost-efficient manner. Retrieval of all organismal groups (e.g. animals, bacteria, fungi, protists) present in an environmental sample by representative sequences would result in a low number of sequences per organism group and would likely not recover some of the diversity found in each of the groups (Eiler *et al.* 2013). If the target organism group is more limited, the same number of sequences will likely recover all the diversity found in this specific group, in our case the diatoms or Bacillariophyta.

Group-specific primers and stringent PCR conditions can be used as presequencing filters. Preselection of organismic sequences during PCR amplification (e.g. through primer binding preferences) significantly influences the taxa representation in the postsequencing data and should be taken into account (Hajibabaei *et al.* 2011, 2012; Zinger *et al.* 2012) prior to any investigation. Here, the number of amplified nondiatom eukaryotic sequences is about 58 000 (approximately 67%) compared with about 28 000 diatom sequences (approximately 33% of total sequences). Therefore, the percentage of diatom sequences retrieved is higher *in vivo* than *in silico* (33% vs. 22%). In a metabarcoding study on fungal community composition, the number of nonfungal sequences was around 60% (Schmidt *et al.* 2013), a similar range to our 67%.

The here used primer pair (M13F-D512for 18S/M13R-D978rev 18S) is not strictly diatom specific, but the best coverage in all eukaryotic groups is found in the Stramenopiles which include the diatoms. Furthermore, it allows for coverage of all important diatom groups (Zimmermann *et al.* 2011), which reflects the trade-off between specificity and universality within diatoms; a trade-off commonly recognized in metabarcoding approaches (Toju *et al.* 2012; Schmidt *et al.* 2013). The here used primers show the most promising features of all 18S primers so far used for DNA barcoding and metabarcoding in diatoms (Zimmermann *et al.* 2011).

	no. taxa/site+method															
	RL1.1		RL2.1		RL3.1		RL4.1		RL5.1		RL6.1		RL7.1		RL1.1-RL7.1	
Genus	LM	NGS	LM	NGS	LM	NGS	LM	NGS	LM	NGS	LM	NGS	LM	NGS	LM	NGS
Achnanthidium	2	8	1	9	1	1	1	1	2	3	1	1	1	2	4	14
Amphora	-	-	1	3	1	1	-	-	1	2	2	6	2	3	2	8
Asterionella	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	1
Caloneis	-	-	-	-	-	-	-	-	-	-	1	1	-	-	1	1
Cocconeis	-	-	1	2	1	1	1	3	1	3	1	7	2	2	2	15
Craticula	-	-	-	-	-	1	-	1	-	-	-	-	-	-	-	2
Cymatopleura	-	-	-	-	-	-	-	-	-	1	-	-	1	5	1	5
Cymbella	-	-	-	-	-	1	-	-	-	1	-	-	-	-	-	2
Diadesmis	1	1	-	-	-	-	-	-	-	-	-	-	-	-	1	1
Diatoma	1	1	1	1	1	1	1	1	4	3	2	3	3	4	6	11
Encyonema	1	-	3	8	3	3	2	3	2	4	-	1	1	2	3	7
Eolimna	1	-	2	7	1	1	1	1	1	1	1	2	1	1	2	7
Eunotia	1	6	-	-	-	-	-	-	-	-	-	-	-	-	1	5
Fragilaria	5	7	3	6	5	3	5	4	3	4	4	9	5	5	18	29
Frustulia	-	-	1	1	1	-	-	-	-	-	-	-	-	-	2	1
Geissleria	1	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-
Gomphonema	2	1	4	27	2	4	2	6	6	14	2	12	3	6	10	28
Hantzschia	-	-	1	1	-	1	-	-	-	-	-	-	-	-	1	2
Hippodonta	-	-	1	1	1	1	1	2	1	1	-	-	1	-	2	2
Karayevia	-	-	-	-	1	1	-	-	-	-	1	-	1	-	2	1
Luticola	-	-	-	-	1	-	-	-	-	-	-	-	-	-	1	-
Mayamaea	-	-	-	-	-	-	-	-	1	1	1	1	-	-	1	2
Meridion	2	-	1	-	-	-	1	-	1	-	-	-	2	-	2	-
Navicula	6	19	4	11	5	12	6	17	6	24	5	14	5	8	9	40
Nitzschia	3	5	6	8	7	24	6	18	5	13	5	17	5	14	17	49
Parlibellus	-	-	-	-	-	-	-	-	1	-	-	-	1	-	1	-
Pinnularia	1	1	-	-	-	-	2	4	-	-	-	-	-	-	3	5
Planothidium	1	2	2	3	2	2	1	1	1	2	1	1	1	2	2	9
Reimeria	-	-	1	2	1	1	1	1	1	1	1	1	1	1	1	7
Rhoicosphenia	-	-	1	-	1	-	1	-	1	-	1	-	1	-	1	-
Sellaphora	1	1	-	-	1	-	1	1	1	1	1	1	1	2	2	6
Stauroneis	1	-	-	-	-	-	1	-	-	-	-	-	-	-	1	-
Surirella	2	2	1	3	1	2	2	2	2	3	-	-	2	3	3	8
Tryblionella	-	-	-	-	-	1	-	-	-	1	-	-	-	-	-	2

Fig. 4 'Heat map' and numbers of taxa detected in the specific genera for both methods and sampling sites. Colour code for the heat map: from lighter shades for 1 to darker shades for the highest number of taxa for the single sample sites, from lighter shades for lower taxa numbers to darker shades for higher taxa numbers for the summary of the seven sampling sites.

Taxa with many copies per cell could be overrepresented while rare taxa with a low-copy number could be missed or only be represented as singleton in the NGS output (Behnke *et al.* 2011). Therefore, singletons should be evaluated critically and not automatically be discarded from the data set.

Preselection by PCR can lead to over- or underestimation or loss of taxa diversity, whereas postsequencing reduction of the data set to the targeted taxonomic group is a way to set an artificial focus for a specific purpose in a study. Data neglected in one study can be used in future studies with a different focus and analytical methodology.

Specimen identification

Up to now, most biodiversity assessment of micro-organisms via NGS technology assigned sequences mainly to families or genera; partly due to the vast number of sequence reads, and partly because the studies are focused on a broad range of micro-organismic groups (Eiler *et al.* 2011, 2013; Pawlowski *et al.* 2011). For fresh water quality assessments using diatoms, however, genus delimitation is insufficient and most indices demand species level discovery. Furthermore, biodiversity in general is evaluated on the species level (Carpenter *et al.* 2009). For NGS biodiversity investigations on higher taxonomic levels (e.g. genus, family), the BLASTn algorithm is sufficient for specimen identification and provides a fast and powerful approach. For the identification on the species level, the BLASTn is not suitable as it uses shortcuts to save computation time and therefore compares sequences only imperfectly (Edgar 2010). Especially, small genetic differences able to discriminate between taxa are not detected. As we wanted to compare sequences to the commonly used LM approach, an assignment to subgenus or even subspecific level is necessary.

The PCMA approach, based on phylogenetic trees, proves to be more precise for lower taxonomic levels

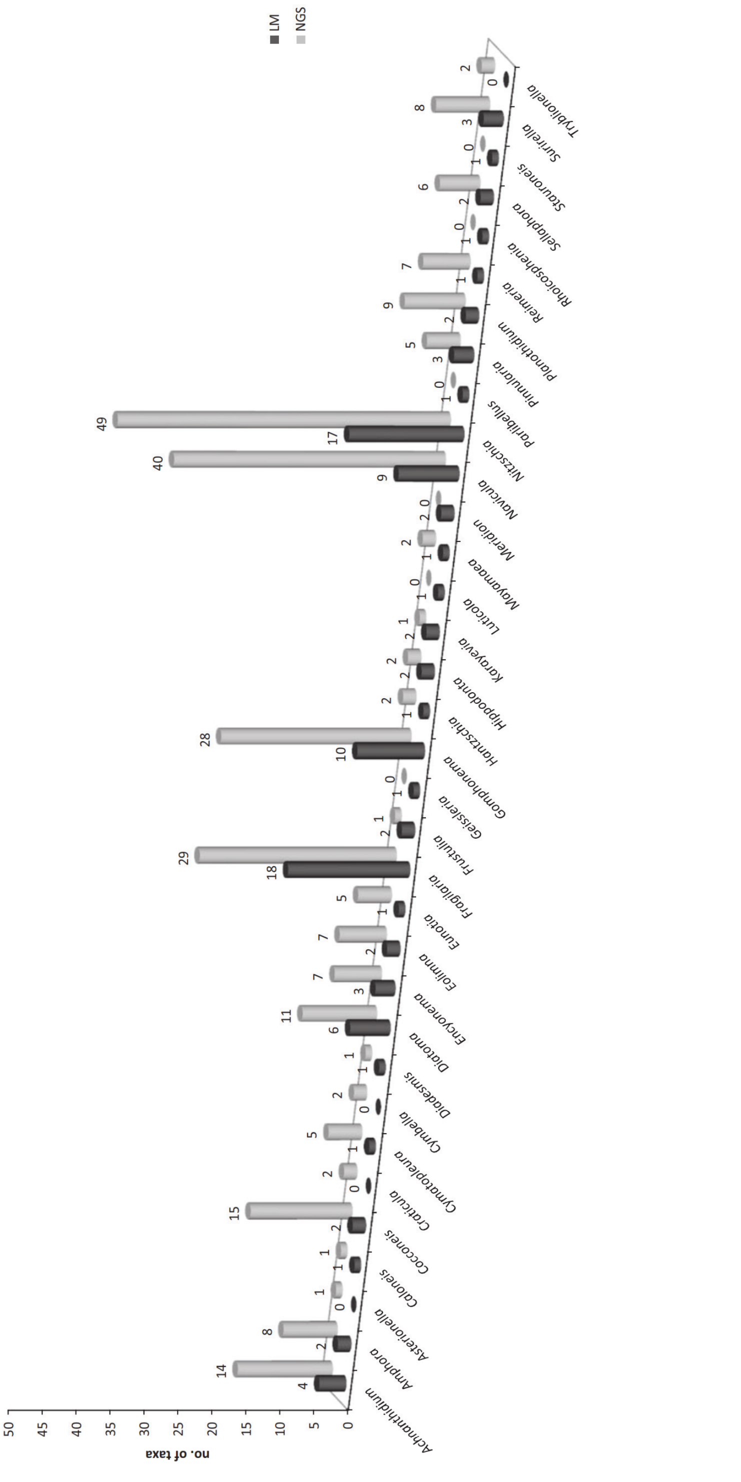


Fig. 5 Summary of taxa numbers per genus given for both methods [next-generation sequencing (NGS), (light microscopy) LM] and all sampling sites (RL1.1–RL7.1).

(Aquilino *et al.* 2011; Hajibabaei *et al.* 2011; Wilson *et al.* 2011; Bik *et al.* 2012). While genera assignment was roughly comparable between BLASTn and PCMA, the latter performs better in many cases (Table 3). BLASTn also produces many undefined units, which can later be assigned using PCMA (Table 3). The undefined units binned by BLASTn are partly due to undefined GenBank entries (e.g. uncultured eukaryote, uncultured diatom and uncultured stramenopile). Especially, these undefined units can be assigned by the PCMA as the sequence appears in the tree in a defined phylogenetic context (Monaghan *et al.* 2009; Hoef-Emden 2012). However, the bootstrap cut-off value for the PCMA approach is empirical and needs to be tested on further data sets and simulations.

One key issue of taxon/specimen identification via DNA sequences is the reference database used (e.g. GenBank, Silva, BOLD) (Ratnasingham & Hebert 2007). If the database is incomplete, that is. sequences for some taxa are missing or sequences are erroneously linked to a taxon name, specimen identification can be flawed (Becker *et al.* 2011; Yoccoz 2012). In this study, some taxa that could be identified morphologically had no entry in the reference database and thus could not be detected by sequence comparisons. It is possible that PCMA retrieved a corresponding OTU to the missing taxon in the NGS data, but no authoritative relation between the OTU and the taxon could be determined.

Regardless of the method used it is important to use the same taxonomic concept in all approaches to compare performance: for example *Fragilaria*, *Synedra* and *Ulnaria* can be counted as individual genera or all as *Fragilaria sensu lato*. If one is not aware of such taxonomic details, the assignment can differ between BLASTn (giving always the genus name an entry is listed under in GenBank or other databases) and the PCMA approach, that is based on the taxonomic concept of the individual scientist interpreting the tree. Taxonomic background knowledge on the investigated group is crucial for data interpretation and can lead to wrong conclusions if not taken into account (Bortolus 2008).

Comparison of LM and NGS

In general, it can be said that nearly all taxa that were found by LM were also recovered by NGS; those taxa that were missing in NGS were not present in the reference databases, but for most of the missing taxa in NGS often there are significant hits on the genus level. Due to the missing reference sequences also no PCMA clusters could be assigned to the missing taxa. This situation is best demonstrated by the fact that at sample location RL1.1 the highest discrepancy between LM and NGS is found, because the oligotrophic conditions at RL1.1 are represented by rare taxa that are more often missing from the reference database than common taxa. This problem will diminish in future, as taxonomic gaps in the reference databases will be filled. Broader integration of taxonomic knowledge and enhanced further development of reliable databases (incl. online accessible vouchers) is crucial for high-quality DNA barcoding and metabarcoding (Bortolus 2008; Collins & Cruickshank 2013; Kvist 2013; Zimmermann *et al.* 2014). Vice versa, taxonomy and species description can benefit from the integration of the information contained in data sets generated by DNA barcoding and metabarcoding: the incongruence between molecular variation and morphological perceptions can trigger further taxonomic investigations.

In addition, NGS recovered conspicuously more taxa than LM. One reason for this is the deeper taxon sampling due to the higher number of retrieved sequences compared with the 500 valves that have been counted for the LM analyses (Stoeck *et al.* 2009). Although during the LM analysis process rare taxa that have been found only in the initial screening of the slides were also included in the species list, the thus increased taxonomic completeness of the morphologically derived species lists could not match the NGS list.

One reason for the higher number of taxa received from the NGS approach is the so-called concealed diversity: morphologically identical taxa can have considerable genetic variation (e.g. *Gomphonema parvulum* (Abarca *et al.* 2014; Kermarrec *et al.* 2013a), *Nitzschia palea* complex (Mann *et al.* 2010; Trobajo *et al.* 2010), *Sellaphora* (Evans *et al.* 2007, 2008). In our study, the genus *Nitzschia* is a good example for incongruence between morphological and molecular variation: LM detected 17 taxa versus 49 taxa via NGS (Figs 4, 5), which would be 50% of the nearly 100 *Nitzschia* taxa described for entire Central Europe by Hofmann *et al.* (2011). The discussion whether this genetic variation indicates real cryptic species or rather genetic diversity in populations is ongoing. In some cases, refined morphological follow-up examinations induced by genetic variation lead to the description of new species e.g. (Abarca *et al.* 2014). The taxonomic concepts of diatoms currently in use will be improved as more and more studies are carried out combining morphological and DNA sequence data, but NGS already allows the direct comparison of genotypes and specimen between different sampling sites. But whether these sequences belong to ecodemes or morphodemes and how these are related and linked to environmental conditions remains a taxonomical and ecological issue. The linkage of sequence diversity to known species is also important for the classification of new species as a large proportion of them are still undiscovered and/or not described (Wilson 2003; Monaghan *et al.* 2009); for diatoms, the estimated number of species not yet discovered lies between 30 000 and 100 000 (Mann & Vanormelingen 2013).

Some of the higher diversity found in NGS could also be attributed to PCR bias and pyrosequencing artefacts. While automated denoising and manually sequence evaluation reduces these error prone sequences, but not all problematic sequences can be eliminated this way (Behnke *et al.* 2011; Pawlowski *et al.* 2011; Bik *et al.* 2012). However, the PCMA OTUs provide a cushioning effect as they cluster sequences with only minor differences.

In conclusion, the comparison of LM and NGS can be constrained by differences in the morphological and genetic classification concepts (Eiler *et al.* 2013).

If the reference databases and the bioinformatic pipelines for the analysis of sequence data sets are further standardized and automated, NGS has the potential to increase the comparability between samples from different sites and produced by different research teams. Automated refined bioinformatic pipelines could help computationally untrained biologists to circumvent data misinterpretation (Bik *et al.* 2012) and also decrease the number of false negatives and positives during specimen identification (Yoccoz 2012; Kermarrec *et al.* 2013b). Also, bioinformatic software tools for NGS-based eDNA barcoding purposes need to be implemented or linked to the respective sequence reference databases in the future to ease specimen identification (Collins & Cruickshank 2013).

The number of NGS sequences for the mostly planktonic living centric diatoms is higher than the number of benthic pennates in the lowland sampling sites RL4.1-7.1 (Table 2, File S1) even though these samples were scratched from substrate at the littoral. As the flow velocity decreases while the water body volume increases downstream from the spring region to river sections with lowland characteristics, it is expectable that the probability rises that the mostly planktonic centric diatoms accumulate more often in these river sections. In the LM analysis, the number of centric frustules was comparable to the pennate valves. The reason for the overrepresentation of centric diatom sequences in the lower river regions may be found in the points discussed above or in not yet explored processes.

One further advantage of the NGS approach is the following: samples prepared for LM are treated with H_2O_2 leaving only the diatom silica valves intact; all other cells including those of other organism groups are destroyed. The DNA of the environmental sample used for NGS, however, remains complete and can be used for later studies on, for example different organism groups or the reproduction of diatom studies with advanced NGS methodology (Shokralla *et al.* 2012).

Conclusion

Our findings provide evidence that eDNA barcoding of diatoms via NGS sequencing of the V4 region (18S) can be suitable for standardly used water quality assessments in the future and in general could complement or even improve the traditional identification technique via LM. The NGS technology almost always leads to a higher number of identified taxa, which presence could subsequently being verified by LM. Taxa retrieval varies considerably throughout the river system, largely depending on taxa coverage in the reference database. Taxonomically validated databases (incl. online accessible vouchers) are crucial for reliable taxa identification via NGS.

More knowledge about copy number and genome size and correlations of these two factors to cell size of different diatoms is needed to tackle the problem of quantification via NGS. Another challenge lies in the bioinformatic data handling: the scientist to tackle these challenges in future will neither be a biologist with some insight into informatics nor a computer scientist with biological basics but a true interdisciplinary scientist, who will at least support the communication between the two fields (Smith & Staetsky 2007; Yoccoz 2012; Collins & Cruickshank 2013).

Next-generation sequencing based eDNA barcoding is not a swiss army knife, but provides a more comprehensive insight into diatom diversity or other protist communities and therefore could be the basis for the ecological projection of global diversity (Stoeck *et al.* 2009). If thoroughly conducted, the here presented approach not only bears the potential to supplement and improve the old identification system, but beyond that opens up many new opportunities and challenges: diversity data from NGS eDNA barcoding of environmental samples can easily be compared and combined on different spatial (α-, β-, γ-diversity) (Taberlet *et al.* 2012b), temporal and taxonomical levels. Therefore, it is applicable for large scale biomonitoring (Hajibabaei *et al.* 2011) and the quality management of water bodies, for example under governmental frameworks. The method has the potential to be extended to analyse not only taxon abundances but also functional and ecological interrelations within communities, for example by including real metagenomics and the investigation of metabolic and catabolic products.

This study demonstrates that environmental NGS eDNA barcoding of diatoms is a competitive approach to describe communities and furthermore holds the potential for a large application spectrum.

Acknowledgements

The authors wish to thank the diatom research group at the BGBM for fruitful discussions, Jana Bansemer for the support in laboratory and Michael Rodewald for the help with the maps. 454 sequencing was executed at the BeGenDiv (Berlin Consortium for Genomics in Biodiversity Research). This study was funded by the DFG (Deutsche Forschungsgemeinschaft) project GE 12 42/11-1.

References

Abarca N, Enke N, Zimmermann J, Jahn R (2014) Does the cosmopolitan diatom *Gomphonema parvulum* (Kützing) Kützing have a biogeography? *PLoS ONE*, **9**, e86885.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Alverson AJ, Cannone JJ, Gutell RR, Theriot EC (2006) The evolution of elongate shape in diatoms. *Journal of Phycology*, **42**, 655–668.

Aquilino SVL, Tango JM, Fontanilla IKC *et al.* (2011) DNA barcoding of the ichthyofauna of Taal Lake, Philippines. *Molecular Ecology Resources*, **11**, 612–619.

Archibald REM (1984) Diatom illustrations—an appeal. *Bacillaria*, **7**, 173–178.

Bailey-Watts AE (1976) Planktonic diatoms and some diatom-silica relations in a shallow eutrophic Scottish loch. *Freshwater Biology*, **6**, 69–80.

Barraclough T (2013) The Barralab Website: Description of how to install Splits as a library in R. Available: http://barralab.bio.ic.ac.uk/downloads.html. Accesed 2013.

Becker S, Hanner R, Steinke D (2011) Five years of FISH-BOL: brief status report. *Mitochondrial DNA*, **22**, 3–9.

Behnke A, Engel M, Christen R *et al.* (2011) Depicting more accurate pictures of protistan community complexity using pyrosequencing of hypervariable SSU rRNA gene regions. *Environmental Microbiology*, **13**, 340–349.

Bik HM, Porazinska DL, Creer S *et al.* (2012) Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology & Evolution*, **27**, 233–243.

Bortolus A (2008) Error cascades in the biological sciences: the unwanted consequences of using bad taxonomy in ecology. *AMBIO: A Journal of the Human Environment*, **37**, 114–118.

Bruder K, Sato S, Medlin LK (2008) Molecular assessment of phylogenetic relationships in selected species/genera in the Naviculoid diatoms (Bacillariophyta). *Diatom*, **24**, 8–24.

Carpenter SR, Mooney HA, Agard J *et al.* (2009) Science for managing ecosystem services: Beyond the Millennium Ecosystem Assessment. *Proceedings of the National Academy of Sciences*, **106**, 1305–1312.

Chen J, Li Q, Kong L, Yu H (2011) How DNA barcodes complement taxonomy and explore species diversity: the case study of a poorly understood marine fauna. *PLoS ONE*, **6**, e21326.

Collins RA, Cruickshank RH (2013) The seven deadly sins of DNA barcoding. *Molecular Ecology Resources*, **13**, 969–975.

Descy J, Coste M (1991) A test of methods for assessing water quality based on diatoms. *Negotiations of the International Society of Limnology*, **24**, 2112–2116.

Dupuis JR, Roe AD, Sperling FAH (2012) Multi-locus species delimitation in closely related animals and fungi: one marker is not enough. *Molecular Ecology*, **21**, 4422–4436.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792–1797.

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Eiler A, Heinrich F, Bertilsson S (2011) Coherent dynamics and association networks among lake bacterioplankton taxa. *The ISME Journal*, **6**, 330–342.

Eiler A, Drakare S, Bertilsson S *et al.* (2013) Unveiling distribution patterns of freshwater phytoplankton by a next generation sequencing based approach. *PLoS ONE*, **8**, e53516.

European-Committee-for-Standardization (2003) EuropeanStandard. EN 14407. Water Quality – Guidance Standardfor the Identi?cation, Enumeration and Interpretation of Benthic Diatom Samples from Running Waters. *CEN*, 14 pp.

European-Committee-for-Standardization (2004) EuropeanStandard. EN 14407. Water Quality – Guidance Standardfor the Identification, Enumeration and Interpretation of Benthic Diatom Samples from Running Waters. *CEN* Brussels, 12 pp.

Evans KM, Wortley AH, Mann DG (2007) An assessment of potential diatom "barcode" genes (*cox*1, *rbc*L, 18S and ITS rDNA) and their effectiveness in determining relationships in *Sellaphora* (Bacillariophyta). *Protist*, **158**, 349–364.

Evans KM, Wortley AH, Simpson GE, Chepurnov VA, Mann DG (2008) A molecular systematic approach to explore diversity within the *Sellaphora pupula* species complex (Bacillariophyta). *Journal of Phycology*, **44**, 215–231.

Falkowski PG, Barber RT, Smetacek V (1998) Biogeochemical controls and feedbacks on ocean primary production. *Science*, **281**, 200–206.

Field CB, Behrenfeld MJ, Randerson JT, Falkowski P (1998) Primary production of the biosphere: integrating terrestrial and oceanic components. *Science*, **281**, 237–240.

Fontaneto D, Herniou EA, Boschetti C *et al.* (2007) Independently evolving species in asexual bdelloid rotifers. *PLoS Biology*, **5**, e87.

Fujita MK, Leaché AD, Burbrink FT, McGuire JA, Moritz C (2012) Coalescent-based species delimitation in an integrative taxonomy. *Trends in Ecology & Evolution*, **27**, 480–488.

Göker M, Grimm GW, Auch AF, Aurahs R, Kučera M (2010) A clustering optimization strategy for molecular taxonomy applied to planktonic foraminifera SSU rDNA. *Evolutionary Bioinformatics*, **6**, 97–112.

Guillou L, Viprey M, Chambouvet A *et al.* (2008) Widespread occurrence and genetic diversity of marine parasitoids belonging to Syndiniales (Alveolata). *Environmental Microbiology*, **10**, 3349–3365.

Hajibabaei M, deWaard JR, Ivanova NV *et al.* (2005) Critical factors for assembling a high volume of DNA barcodes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **360**, 1959–1967.

Hajibabaei M, Shokralla S, Zhou X, Singer G, Baird D (2011) Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS ONE*, **6**, e17497.

Hajibabaei M, Spall J, Shokralla S, van Konynenburg S (2012) Assessing biodiversity of a freshwater benthic macroinvertebrate community through non-destructive environmental barcoding of DNA from preservative ethanol. *BMC Ecology*, **12**, 1–10.

Hamsher SE, Evans KM, Mann DG, Poulíčková A, Saunders GW (2011) Barcoding diatoms: exploring alternatives to COI-5P. *Protist*, **162**, 405–422.

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **270**, 313–321.

Hoef-Emden K (2012) Pitfalls of establishing DNA barcoding systems in protists: the cryptophyceae as a test case. *PLoS ONE*, **7**, e43652.

Hofmann G, Werum M, Lange-Bertalot H (2011) *Diatomeen im Süßwasser-Benthos von Mitteleuropa*. Gantner Verlag, Ruggell.

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754–755.

Huse SM, Huber JA, Morrison HG, Sogin ML, Welch D (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology*, **8**, 143–152.

Jahn R (1984) The structure of the frustule and its variability in clones and populations. *Proceedings of the 8th International Diatom Symposium 1984*. 191–204.

Jahn R, Kusber WH (2005+) AlgaTerra Information System (online). Botanic Garden and Botanical Museum Berlin-Dahlem, Freie Universität Berlin. http://www.algaterra.org.

Jahn R, Zetzsche H, Reinhardt R, Gemeinholzer B (2007) Diatoms and DNA barcoding: a pilot study on an environmental sample. In: *Proceedings of the 1st Central European Diatom Meeting 2007* (eds Kusber W. & Jahn R.), pp. 63–68. Botanic Garden and Botanical Museum Berlin-Dahlem, Berlin.

Kawecka B, Olech M (1993) Diatom communities in the Vanishing and Ornithologist Creek, King George Island, South Shetlands, Antarctica. *Hydrobiologia*, **269–270**, 327–333.

Kelly MG (1998) Use of the trophic diatom index to monitor eutrophication in rivers. *Water Research*, **32**, 236–242.

Kelly MG, Cazaubon A, Coring E *et al.* (1998) Recommendations for the routine sampling of diatoms for water quality assessments in Europe. *Journal of Applied Phycology*, **10**, 215–224.

Kermarrec L, Bouchez A, Rimet F, Humbert J-F (2013a) First evidence of the existence of semi-cryptic species and of a phylogeographic structure in the gomphonema parvulum (Kützing) Kützing complex (Bacillariophyta). *Protist*, **164**, 686–705.

Kermarrec L, Franc A, Rimet F *et al.* (2013b) Next-generation sequencing to inventory taxonomic diversity in eukaryotic communities: a test for freshwater diatoms. *Molecular Ecology Resources*, **13**, 607–619.

Kermarrec L, Franc A, Rimet F *et al.* (2014) A next-generation sequencing approach to river biomonitoring using benthic diatoms. *Freshwater Science*, **33**, 349–363.

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, **16**, 111–120.

Klindworth A, Pruesse E, Schweer T *et al.* (2013) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Research*, **41**, e1.

Kvist S (2013) Barcoding in the dark? A critical view of the sufficiency of zoological DNA barcoding databases and a plea for broader integration of taxonomic knowledge. *Molecular Phylogenetics and Evolution*, **69**, 39–45.

Luddington IA, 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**, e45664.

MacGillivary ML, 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**, 529–536.

Mann DG (1999) The species concept in diatoms. *Phycologia*, **38**, 437–495.

Mann DG, Droop SJM (1996) Biodiversity, biogeography and conservation of diatoms. *Hydrobiologia*, **336**, 19–32.

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

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

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

Medlin LK, Elwood HJ, Stickel S, Sogin ML (1991) Morphological and genetic variation within the diatom *Skeletonema costatum* (Bacillariophyta): evidence for a new species *Skeletonema pseudocostatum*. *Journal of Phycology*, **27**, 514–524.

Monaghan MT, Wild R, Elliot M *et al.* (2009) Accelerated species inventory on madagascar using coalescent-based models of species delineation. *Systematic Biology*, **58**, 298–311.

Morales EA, Siver PA, Trainor FR (2001) Identification of diatoms (Bacillariophyceae) during ecological assessments: comparison between Light Microscopy and Scanning Electron Microscopy techniques. *Proceedings of the Academy of Natural Sciences of Philadelphia*, **151**, 95–103.

Moritz C, Cicero C (2004) DNA barcoding: promise and pitfalls. *PLoS Biology*, **2**, e354.

Pawlowski J, Christen R, Lecroq B *et al.* (2011) Eukaryotic richness in the abyss: insights from pyrotag sequencing. *PLoS ONE*, **6**, e18169.

Pawlowski J, Audic S, Adl S *et al.* (2012) CBOL Protist Working Group: barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLoS Biology*, **10**, e1001419.

Pons J, Barraclough TG, Gomez-Zurita J *et al.* (2006) Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology*, **55**, 595–609.

Poulíčková A, Špačková J, Kelly M, Duchoslav M, Mann D (2008) Ecological variation within *Sellaphora* species complexes (Bacillariophyceae): specialists or generalists? *Hydrobiologia*, **614**, 373–386.

ProjectR- The Comprehensive R Archive Network. Available: http://cran.r-project.org. Accessed 2013.

Prygiel J, Carpentier P, Almeida S *et al.* (2002) Determination of the biological diatom index (IBD NF T 90–354): results of an intercomparison exercise. *Journal of Applied Phycology*, **14**, 27–39.

Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ (2011) Removing noise from pyrosequenced amplicons. *BMC Bioinformatics*, **12**, 38.

Rach J, DeSalle R, Sarkar IN, Schierwater B, Hadrys H (2008) Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata. *Proceedings of the Royal Society B: Biological Sciences*, **275**, 237–247.

Ratnasingham S, Hebert PDN (2007) BOLD: the Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes* **7**, 355–364.

Ruck EC, Theriot EC (2011) Origin and evolution of the canal raphe system in diatoms. *Protist*, **162**, 723–737.

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**, 406–425.

Schmidt P-A, Bálint M, Greshake B *et al.* (2013) Illumina metabarcoding of a soil fungal community. *Soil Biology and Biochemistry*, **65**, 128–132.

Shokralla S, Spall J, Gibson J, Hajibabaei M (2012) Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology*, **21**, 1794–1805.

Smetacek V (1999) Diatoms and the ocean carbon cycle. *Protist*, **150**, 25–32.

Smith TMF, Staetsky L (2007) The teaching of statistics in UK universities. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **170**, 581–622.

Spaulding SA, McKnight DM (1999) Assessing ecological conditions in rivers and streams with diatoms. In: *The diatoms: Applications to the environmental and earth sciences*(eds Stoermer E. P. & Smol J. P.), pp. 245–260. Cambridge University Press, Cambridge.

Stevenson RJ, Pan Y (1999) Assessing ecological conditions in rivers and streams with diatoms. In: *The diatoms: Applications to the environmental and earth sciences*(eds Stoermer E. P. & Smol J. P.), pp. 11–40. Cambridge University Press, Cambridge.

Stoeck T, Behnke A, Christen R *et al.* (2009) Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities. *BMC Biology*, **7**, 72.

Stoeckle M (2003) Taxonomy, DNA and the barcode of life. *BioScience*, **53**, 2–3.

Stoof-Leichsenring KR, Epp LS, Trauth MH, Tiedemann R (2012) Hidden diversity in diatoms of Kenyan Lake Naivasha: a genetic approach detects temporal variation. *Molecular Ecology*, **21**, 1918–1930.

Swofford DL (2002) *Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Sinauer Associates, Sunderland, Massachusetts.

Taberlet P, Coissac E, Hajibabaei M, Rieseberg L (2012a) Environmental DNA. *Molecular Ecology*, **21**, 1789–1793.

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

Taberlet P, Prud'homme S, Campione E *et al.* (2012c) Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies. *Molecular Ecology*, **21**, 1816–1820.

Tamura K, Peterson D, Peterson N *et al.* (2011) MEGA5: Molecular Evolutionary Genetics Analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, **28**, 2731–2739.

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

Toju H, Tanabe AS, Yamamoto S, Sato H (2012) High-coverage ITS primers for the DNA-based identification of ascomycetes and basidiomycetes in environmental samples. *PLoS ONE*, **7**, e40863.

Trobajo R, Mann DG, Clavero E *et al.* (2010) The use of partial *cox*1, *rbc*L and LSU rDNA sequences for phylogenetics and species identification within the *Nitzschia palea* species complex (Bacillariophyceae). *European Journal of Phycology*, **45**, 413–425.

Vanelslander B, Créach V, Vanormelingen P *et al.* (2009) Ecological differantiation between sympatric pseudocryptic species in the estuarine benthic diatom *Navicula phyllepta* (Bacillariophyceae). *Journal of Phycology*, **45**, 1278–1289.

Vijver BVD, Jarlman A, Haan MD, Compere P (2013) The ultrastructure of *Gomphonema augur* and *Gomphonema gautieri* (Bacillariophyta). *Cryptogamie, Algologie*, **34**, 103–116.

Wilson EO (2003) The encyclopedia of life. *Trends in Ecology & Evolution*, **18**, 77–80.

Wilson J, Rougerie R, Schonfeld J *et al.* (2011) When species matches are unavailable are DNA barcodes correctly assigned to higher taxa? An assessment using sphingid moths. *BMC Ecology*, **11**, 18.

Yoccoz NG (2012) The future of environmental DNA in ecology. *Molecular Ecology*, **21**, 2031–2038.

Zalack JT, Smucker NJ, Vis ML (2010) Development of a diatom index of biotic integrity for acid mine drainage impacted streams. *Ecological Indicators*, **10**, 287–295.

Zimmermann J, Jahn R, Gemeinholzer B (2010) Diatom barcoding: water monitoring using molecular tools. 2nd Conference of the European Consortium for the Barcode of Life, 38.

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 & Evolution*, **11**, 173–192.

Zimmermann J, Abarca N, Enke N *et al.* (2014) Taxonomic reference libraries for environmental barcoding: a best practice example from diatom research. *PLoS ONE*, **9**, e108793.

Zinger L, Gobet A, Pommier T (2012) Two decades of describing the unseen majority of aquatic microbial diversity. *Molecular Ecology*, **21**, 1878–1896.

J.Z. carried out data collection, morphological and NGS data analyses and drafted the manuscript. B.G. conceived of the general idea, contributed the 'Supertrees'. R.J. supported JZ in questions arising from the morphological analysis. G.G. supported the NGS data analysis and participated in the design of the project. J.Z. and G.G. developed the NGS analysis approach. J.Z. and NE did the sampling trips. All authors critically revised the manuscript.

Data Accessibility

Parts of our internal reference database are already submitted to GenBank and available under accession no. FR873231–FR873265, KM084866–KM084994, HG529999–HG530017, KM350053–KM350154. Newly submitted SSU rDNA 18S V4 sequences used in trees: GenBank accession no. KM592926–KM592938. Input files (read libraries) and phylogenetic tree files are available at Dryad: http://doi.org/10.5061/dryad.kb88d.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

File S1. Table S1. Overview of the detected centric diatom genera and number of sequences for all seven sampling sites.

File S2. Figs S2–S7 'Supertrees' compiled from phylogenetic subtrees based on 18S V4 region gene sequences (~390 bp) of the 18S rDNA for all diatoms occurring on sampling sites RL1.1, RL2.1, RL3.1, RL4.1, RL5.1, RL7.1.

File S3. Table S8. List of detected specimens/taxa via LM and NGS (with OTUs, no. of OTUs and sequences) for the sampling sites RL1.1-RL7.1.

Part III

Appendix

Although this may seem a paradox, all exact science is dominated by the idea of approximation. When a man tells you that he knows the exact truth about anything, you are safe in inferring that he is an inexact man.

Bertrand Russell

Der Lebenslauf wurde aus der elektronischen Version der Arbeit entfernt.

The curriculum vitae was removed from the electronic version of the paper.

LIST OF PUBLICATIONS

PEER REVIEWED PUBLICATIONS

Enke, N., Kunze, R., Pustahija, F., Glöckner, G., **Zimmermann, J.**, Oberländer, J., Kamari, G. & Siljak-Yakovlev, S. (subm.). Genome Size Shifts: Karyotype Evolution in *Crepis* section *Neglectoides* (Asteraceae). Plant Biology

Enke, N., Kulikovskiy, M.,Abarca, N., **Zimmermann, J.**, Gusev, E., Skibbe, O., Kusber, W-H., Witkowsi, A. & Jahn, R. (subm.). Settling an Old Dispute? A Molecular based Approach to Resolve the Phylogenetic Relationship of *Pinnularia* Ehrenb. and *Caloneis* P.T. Cleve (Bacillariophyceae). PLoS ONE.

Zimmermann, J., Glöckner, G., Jahn, R., Enke, N. & Gemeinholzer, B. (2014). Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies. Molecular Ecology Resources. DOI: 10.1111/1755-0998.12336.

Zimmermann, J., Abarca, N., Skibbe, O., Kusber, W.-H., Enke, N. & Jahn, R. (2014). Taxonomic Reference Libraries for Environmental Barcoding: a Best Practice Example from Diatom Research. PLoS ONE 9(9): e108793.

Kistenich, S., Dressler, M., **Zimmermann, J.**, Hübener, T., Bastrop, R. & Jahn, R. (2014). An investigation into the morphology and genetics of *Cyclotella comensis* and closely related taxa. Diatom Research 29 (4): 423-440.

Abarca, N., Jahn, R., **Zimmermann, J.** & Enke, N. (2014). Does the Cosmopolitan Diatom *Gomphonema parvulum* (Kützing) Kützing have a Biogeography? PLoS ONE 9(1): e86885.

Pawlowski, J., Audic, S., Adl, S., Bass, D., Belbahri, L., Berney, C., Bowser, S.S., Cepicka, I., Decelle, J., Dunthorn, M., Fiore-Donno, A.M., Gile, G.H., Holzmann, M., Jahn, R., Jirků, M., Keeling, P.J., Kostka, M., Kudryavtsev, A., Lara, E., Lukeš, J., Mann, D.G., Mitchell, E.A., Nitsche, F., Romeralo, M., Saunders, G.W., Simpson, A.G., Smirnov, A.V., Spouge, J.L., Stern, R.F., Stoeck, T., **Zimmermann, J.**, Schindel, D., de Vargas, C. (2012). CBOL Protist Working Group: Barcoding Eurakyotic Richness beyond the Animal, Plant, and Fungal Kingdoms. PLoS Biol.;10(11):e1001419.

Zimmermann, J., Jahn, R. & Gemeinholzer, B. (2011). Barcoding diatoms: evaluation of the V4 subregion on the 18S rRNA gene, including new primers and protocols. Organism Diversity & Evolution 11 (4): 173-192.

TALKS

Zimmermann, J., Glöckner, G., Jahn, R., Enke, N. & Gemeinholzer, B. (2014). Metabarcoding of freshwater diatoms. [invited speaker] Berlin Center for Genomics in Biodiversity Research (BeGenDiv) – Autumn Symposium 2014, Berlin, Germany.

Abarca, N., Enke, N., Kusber, W.-H., **Zimmermann, J.**, Jahn, R. (2014). Reconciling morphological taxonomic concepts with molecular data - a case study on several clone cultures. 23rd International Diatom Symposium, Nanjing, China.

Enke, N., **Zimmermann, J.**, Abarca, N., Kusber, W.-H., Skibbe, O., Acs, E., Bouchez, A., Dvorak, P., Kelly, M., Mann, D., Poulickova, A., Rimet, F., Sabbe, K., Trobajo, R., Vyverman, W., Jahn, R. (2014). Standardising Information for Environmental DNA barcoding of Diatoms in Taxonomic Reference Libraries - Examples from Berlin Waters. 23rd International Diatom Symposium, Nanjing, China.

Zimmermann, J., Glöckner, G., Jahn, R., Enke, N., Gemeinholzer, B. (2014), NGS barcoding vs. morphological identification to assess diatom composition in environmental studies. 23rd International Diatom Symposium, Nanjing, China.

Jahn, R., Abarca, N., Enke, N., Kusber, W.-H. & **Zimmermann, J.** (2014). DNA barcoding and reference libraries – impact on taxonomy. 8th Central European Diatom Meeting, Zagreb, Croatia.

Enke, N., **Zimmermann, J.**, Abarca, N., Kusber, W.H., & Jahn, R. 2014. Calibration of new identification tools for diatoms: A case study from Berlin waters. 15. Wissenschaftliche Tagung der Sektion Phykologie der Deutschen Botanischen Gesellschaft, Stralsund, Germany.

Enke, N., Acs, E., Bouchez, A., Dvorak, P., Jahn, R., Kelly, M., Kusber, W. H., Mann, D., Poulickova, A., Rimet, F., Sabbe, K., Trobajo, R., Vyverman, W. & **Zimmermann, J.** (2013). The EU Diatom Barcoding Initiative: standardisation of sampling, taxonomic validation and data storage. 20th International Conference on Environmental Indicators, Trier, Germany.

Enke N., Abarca N., **Zimmermann, J.**, Skibbe O., & Jahn R. 2013: Gomphonema parvulum revisited: A critical assessment of the taxonomic concept. BioSyst.EU 2013 Global systematics, Vienna, Austria.

Kusber, W.-H., Abarca, N., **Zimmermann, J.**, von Mering, S. & Jahn, R. (2013). Open access to protist biodiversity information: from file cards to GBIF . BioSyst.EU 2013 Global systematics, Vienna, Austria.

Zimmermann, J., Glöckner, G., Jahn, R., Pawlowski, J., de Vargas, C. & Gemeinholzer, B. 2013: Barcoding diatoms with the 18S V4 region: setting an example for protists? BioSyst.EU 2013 Global systematics, Vienna, Austria.

Kusber, W.-H., Abarca, N., Skibbe, O., **Zimmermann, J.**, Jahn, R. (2012). Reference library of DNA-barcoced diatoms - A use case for publishing data via the GBIF database AlgaTerra. Institute Ruder Boskovic, Center for Marine Research, Rovinj, Croatia.

Zimmermann, J. (2012). Next-Generation sequencing (NGS) methods in environmental barcoding – a progress report on biomonitoring of benthic river diatoms". Institute Ruder Boskovic, Center for Marine Research, Rovinj, Croatia.

Jahn, R., **Zimmermann, J.**, Romero, O. R. (2012). A molecular approach to the taxonomy of Cocconeis placentula sensu lato. 22nd International Diatom Symposium Gent, Belgium.

Kusber, W.-H., Abarca, N., Skibbe, O., **Zimmermann, J.**, Jahn, R. (2012). Reference library of DNA-barcoded diatoms - A use case for publishing data via the GBIF database AlgaTerra. 22nd International Diatom Symposium Gent, Belgium.

Romero, O., Jahn, R., **Zimmermann, J.** (2012). A culture based study of Cocconeis lineate Ehrenberg and Cocconeis euglypta Ehrenberg (Bacillariophyta): morphology, typification and barcoding. 22nd International Diatom Symposium Gent, Belgium.

Zimmermann, J., Glöckner, G., Jahn, R., Gemeinholzer, B. (2012). Next-generation sequencing (NGS) methods in environmental barcoding - a progress report on biomonitoring of benthic river diatoms. 22nd International Diatom Symposium Gent, Belgium.

Jahn, R., **Zimmermann, J.**, Abarca, N., von Mering, S., Kusber, W.-H. (2012). Diatom research: towards cybertaxonomy. 26. Treffen Deutschsprachiger Diatomologen, 6th Central European Diatom Meeting, Innsbruck, Austria.

Zimmermann, J., Jahn, R., Gemeinholzer, B. (2012). Diatom DNA barcoding revisited. [invited speaker] 14. Wissenschaftliche Tagung der Sektion Phykologie in der Deutschen Botanischen Gesellschaft, Wissenschaftliche Tagung der Deutschen Gesellschaft für Protozoologie Wuppertal, Germany.

Abarca, N. **Zimmermann, J.** Jahn, R. (2011). Gomphonema parvulum Kützing: Evaluation of some morphological and molecular characters. 5th Central European Diatom Meeting, Szczecin, Poland.

Abarca, N., **Zimmermann, J.** & Jahn, R. (2010). Taxonomy of some species of the diatom genus Gomphonema using morphological and morphological data. 21st International Diatom Symposium, St. Paul, USA.

Zimmermann, J., Gemeinholzer, B. & Jahn, R. (2010). Barcoding diatoms: water monitoring via molecular methods. 21st International Diatom Symposium, St. Paul, USA.

Zimmermann, J., Gemeinholzer, B., Jahn, R. (2010). Diatom barcoding: water monitoring using molecular tools. 2nd Conference of the European Consortium for the Barcode of Life, Braga, Portugal.

Zimmermann J., Gemeinholzer B. & Jahn R. (2010). Evaluation of *cox*1 and 18S as DNA barcoding loci for diatoms. 13. Wissenschaftliche Tagung der Sektion Phykologie, Reichenau, Germany.

Zimmermann J., Gemeinholzer B. & Jahn R. (2010). Barcoding diatoms: Common marker, new primers! 4th Central European Diatom Meeting 2010, Reichenau/Bodensee, Germany.

Pfannkuchen, M., Marić, D., Godrijan, J., Precali, R., Jahn, R., Gemeinholzer, B., **Zimmermann, J.** & Batel, R. (2010). Regularly occurring domination of Skeletonema sp. in the winter phytoplankton of the northern Adriatic Sea. How to resolve the species? 4th Central European Diatom Meeting 2010, Reichenau/Bodensee, Germany.

Gemeinholzer, B., **Zimmermann, J.** & Jahn, R. (2008). Establishing DNA-Barcoding Methods in Diatoms for Diversity Assessments. Botany 2008, Vancouver, Canada

Zimmermann, J., Gemeinholzer, B., Kube, M. & Jahn, R. (2008). Diatoms as bioindicators: watermonitoring via DNA barcoding. Systematics 2008, Göttingen, Germany.

POSTER

Pfannkuchen, M., Godrijan, J., Marić Pfannkuchen, D., Kusber, W.-H., **Zimmermann, J.**, Dröge, G., Enke, N., Abarca, N. & Jahn, R. (2014). Diatom species from the northern Adriatic: Life cell cultures, voucher material and DNA-banks. 8th Central European Diatom Meeting, Zagreb, Croatia.

Enke, N., Acs, E., Bouchez, A., Dvorak, P., Jahn, R., Kelly, M., Kusber, W.-H., Mann, D., Poulickova, A., Rimet, F., Sabbe, K., Trobajo, R., Vyverman, W., **Zimmermann, J.** (2013). DiaDem EU Implementing DNA Barcoding into the Water Framework Directive: standardisaton of sampling, taxonomic validation and data storage. 7th Central European Diatom Meeting, Thonon, France.

Pfannkuchen, M., Godrijan, J., Marić Pfannkuchen, D., Dröge, G., Kusber, W.-H., **Zimmermann, J.**, Enke, N., Abarca, N. & Jahn, R. (2013). Access to the planktonic biodiversity of the northern Adriatic. BioSyst.EU 2013 Global systematics, Vienna, Austria.

Pfannkuchen, M., Godrijan, J., Marić Pfannkuchen, D., Kusber, W. H., **Zimmermann, J.**, Dröge, G., Enke, N., Abarca, N. & Jahn, R. 2013. Access to the planktonic biodiversity of the northern Adriatic. Biosyst.EU 2013, Vienna, Austria.

Jahn, R., **Zimmermann, J.**, Abarca, N., von Mering, S., Kusber, W.-H. (2012). Diatom research: towards cybertaxonomy. 13. Jahrestagung der Gesellschaft für Biologische Systematik (GfBS), Bonn, Germany.

Zimmermann, J., Jahn, R., Glöckner, G., Gemeinholzer, B. 2012: "DNA barcoding of diatoms: establishment of routine laboratory. 13. Jahrestagung der Gesellschaft für Biologische Systematik (GfBS), Bonn, Germany.

Zimmermann, J., Gemeinholzer, B., Kube, M., Reinhardt, R. & Jahn, R. (2009). Watermonitoring via DNA barcoding: diatoms as bioindicators. Systematics, Leiden, Netherlands.

ERKLÄRUNG

Ich erkläre: Ich habe die vorgelegte Dissertation selbständig und ohne unerlaubte fremde Hilfe und nur mit den Hilfen angefertigt, die ich in der Dissertation angegeben habe. Alle Textstellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen sind, und alle Angaben, die auf mündlichen Auskünften beruhen, sind als solche kenntlich gemacht. Bei den von mir durchgeführten und in der Dissertation erwähnten Untersuchungen habe ich die Grundsätze guter wissenschaftlicher Praxis, wie sie in der 'Satzung der Justus-Liebig-Universität Gießen zur Sicherung guter wissenschaftlicher Praxis' niedergelegt sind, eingehalten.

Gießen, November 2014

JONAS ZIMMERMANN

ACKNOWLEDGEMENTS

I express my deepest gratitude to

- PD Dr B. Gemeinholzer, Dr R. Jahn and PD Dr G. Glöckner for the initiation of the project, the encouragement to follow my ideas, the hard time they gave me to defend them and the excellent supervision during the last years.
- the dissertation committee including PD Dr B. Gemeinholzer and PD Dr G. Glöckner for mentoring the thesis and Prof Dr V. Wissemann and Prof Dr T. Wilke for participating in the defence of the thesis.
- The Deutsche Forschungsgemeinschaft for funding.
- the Systematic Botany Group at Justus-Liebig-University Giessen for taking me under their wing and giving me the freedom to do most of my work at the Botanic Garden and Botanical Museum in Berlin; A. Mehl for always being so supportive with the administrative peculiarities.
- the Research Group Diatoms at the Botanic Garden and Botanical Museum, ZE Freie Universität Berlin for hosting me, providing me with a fruitful research environment.
- my colleagues Dr N. Abarca, Dr N. Enke, W.-H. Kusber, Dr. D. Lauterbach, D. Mora, Prof Dr K. Sabbe, Dr K. Stachura-Suchoples and Dr H. Zetzsche for the help and the productive discussions and cooperation on all aspects of diatom research as well as their moral support.
- J. Bansemer, M. Lüchow, I. Danßmann for their support with the lab work.
- Dr O. Skibbe (Larger Than Life) for his brown-greenish thumbs in diatom cultivation
- the Team of the Berlin Center for Genomics in Biodiversity, especially S. Mbedi, Dr M.T. Monaghan and K. Richter for the help with theoretical and practical aspects of next generation sequencing.
- my colleagues J. Godrijan, Dr D. Marić-Pfannkuchen and Dr M. Pfannkuchen for giving me insights in the biodiversity of the Adriatic diatoms.
- Dr H. Enke, M. Simon and C. Breuer for extensive advice on computer eccentricities.
- M. Rodewald for the discussion on all matters of the general aesthetics.
- my extended “Patchwork family”, NE for almost everything as well as to my friends, especially EEN, FOW, IR, CR, LK, S, MKN, FG for enduring my bad temper, running through the world, keeping me sane and healthy; they never doubted me.