Cryptogamie, Algologie, 2010, 31 (4): 557-577

DNA barcoding for species identification and discovery in diatoms

David G. MANN [a], Shinya SATO [a], Rosa TROBAJO [b], Pieter VANORMELINGEN [c] & Caroline SOUFFREAU [c]*

[a] Royal Botanic Garden Edinburgh, Inverleith Row, Edinburgh EH3 5LR, Scotland, UK

[b] Aquatic Ecosystems, Institute for Food and Agricultural Research and Technology (IRTA), Ctra Poble Nou Km 5.5, Sant Carles de la Ràpita, Catalunya, E-43540, Spain

[c] Laboratory of Protistology and Aquatic Ecology, Department of Biology, Ghent University, Krijgslaan 281-S8, B-9000 Gent, Belgium

Abstract – Diatoms are the largest group of microalgae, play an enormous role in the biosphere, and have major significance as bioindicators. Traditional identification requires *inter alia* long training, considerable microscopical skill, and use of a vast and scattered literature. During the life cycle, diatom cells change in size and pattern, often also shape, but the full cycle is known in <1% of described species. Recent evidence shows that there are many pseudocryptic and cryptic species of diatoms, requiring molecular methods for discovery and recognition. These and other factors argue that DNA barcoding would be highly beneficial. It could be 'strong', resolving nearly all species, or 'weak', resolving mostly species already recognized from light microscopy. Attempts have already been made to identify suitable genes and we evaluate these on the basis of universality and practicality, and ability to discriminate between species in the very few 'model' systems offering likely examples of sister-species-pairs. No candidate marker is ideal but LSU rDNA and *rbc*L may be acceptable, though their discriminatory power is lower than that of some other markers. We discuss the next steps in developing a full barcode system.

COI / barcoding / diatoms / DNA barcodes / identification / ITS rDNA / molecular systematics / *rbc*L / LSU rDNA / species discovery / SSU rDNA / taxonomy

Résumé – Les diatomées représentent le plus grand groupe d'algues unicellulaires, jouent un rôle fondamental dans la biosphère, et sont importantes en tant qu'indicateurs biologiques. L'identification sur des bases morphologique exige entre autres un long entraînement, des compétences considérables en microscopie, et l'accès à une vaste littérature publiée pour l'essentiel sous forme d'articles. Durant leur cycle de vie, les cellules de diatomées modifient leur taille et leurs motifs, souvent aussi leur forme, mais le cycle complet n'est connu que pour moins de 1 % des espèces décrites. De récentes études ont démontré l'existence d'espèces pseudocryptiques et cryptiques parmi les diatomées, nécessitant des méthodes moléculaires pour leur découverte et reconnaissance. Ces facteurs entre autres soutiennent le développement d'un système de taxinomie moléculaire tel que le code barre ADN. Ceci pourrait être un barcoding « puissant », pouvant résoudre presque toutes les espèces, ou « faible », ne pouvant résoudre que des espèces déjà reconnues en

* Correspondence and reprints: d.mann@rbge.org.uk

microscopie optique. Plusieurs essais d'identification de gènes convenables sont achevées et nous évaluons l'universalité, la facilité d'emploi et le pouvoir discriminatif interspécifique de ces gènes sur les quelques espèces modèles offrant des paires d'espèces-sœurs probables. A cet instant, aucun marqueur génétique n'est prouvé idéal, mais le LSU rDNA et le *rbc*L peuvent être acceptables, bien que leurs capacités discriminatoires soient plus faibles que ceux de quelques autres marqueurs. Nous discutons les prochaines étapes pour développer un système complet de code-barres.

COI / code-barres /diatomées / code-barres ADN / identification / ITS 2 ADN / systématique moléculaire / *rbc*L / LSU rADN / découverte d'espèces / taxonomie

INTRODUCTION

In angiosperms, vertebrates and the larger invertebrates, species-level taxonomy is fairly stable. This happy state of affairs is demonstrated, for example, by the continued usefulness of many floras, monographs and field guides decades after they were produced, even if the names of some genera have had to be changed to conform with the rules of nomenclature. Stability has been achieved because of many years of work by taxonomists, but also because of the abundance of genetically based variation that can be used for comparisons, requiring no special equipment beyond a hand lens. However, identification can still be problematic, e.g. because species differences are subtle or difficult to describe, or because specimens lack the particular structures required for identification. For example, diagnostic keys for angiosperms usually depend on characteristics of the flowers or fruits, which are rarely available throughout the year, and only a few experts may be able to identify species from vegetative material. The development of cheap, fast methods of sequencing DNA has provided a new approach to identification, 'DNA barcoding', in which short sections of one or a few genes or spacer regions from nuclear or organelle genomes are used to identify species. Since the identifiers are parts of the genome itself, rather than the phenotype, they are available all the time and from every part and life-cycle stage of an organism.

The diatoms are an enormous group, probably containing c. 200,000 species (Mann & Droop, 1996). They are also highly significant ecologically, contributing c. 20% of net global primary production (Mann, 1999). The classification and identification of diatoms is therefore neither a trivial task nor an unimportant one. Traditionally, diatom taxonomy has depended on microscopical features of the cell wall and a huge amount of work has been done to document diversity and relationships within the group. As a result, species-level taxonomy has become sufficiently good to support extensive and successful use of diatoms in biomonitoring and environmental change research (Smol & Stoermer, 2010). However, the taxonomy is not stable, most species have not yet been described and, even after training and long experience, diatomists often find it difficult to agree about identifications. Consequently, DNA barcoding deserves careful consideration as a means of improving the reliability of identifications and discovering species, and also (though coincidentally) increasing the quality and quantity of other taxonomic information, including morphology.

In principle, the application of DNA barcoding to diatoms is no different from its application to higher plants or animals or macroalgae. The aim is the same: to provide unambiguous identification whatever the phenotype of the

specimen, using one or more short stretches of coding or noncoding DNA. However, the special characteristics of diatoms and diatom taxonomy make barcoding both particularly advantageous and challenging in these organisms. We will review the theoretical pros and cons of barcoding in diatoms, before discussing the practicalities of developing a full barcode system.

THE ADVANTAGES OF BARCODING DIATOMS

Reducing costs and ambiguities

To evaluate the potential of DNA barcoding in diatoms, we must first examine how diatoms have previously been classified and identified. Except in some early 19th century works, in which colony morphology was used as a taxonomic character (as in species of '*Homoeocladia*' and '*Schizonema*', e.g. Kützing, 1844), diatom species have generally been described on the basis of the morphology of their frustules (the silicified part of the cell wall composed of two valves and a number of thinner girdle bands). To study frustules in detail, diatomists usually clean away the cell contents (with acids or strong oxidizing agents) and mount material in a suitable resin that provides good contrast with the glassy silica of the valves and girdle bands. Correct identification can depend on very subtle differences, such as whether the density of striation on the valves is 25 or 30 in 10 μm, tiny inflexions of the valve outline, or the occurrence and distribution of special pores or spines. To assess such characteristics and determine species accurately requires initial training in the basics of diatom biology and biodiversity, a high quality light microscope handled expertly, good powers of observation, access to scanning electron microscopy to check light microscope detail and examine structures beyond the resolution of the light microscope, an extensive library (because descriptions of species are scattered across many journals and books, some old and rare), and the advice of colleagues (because probably no diatomist since the 19th century, even Friedrich Hustedt, has had a comprehensive knowledge of all genera).

This is a formidable list of requirements and it is difficult and expensive to supply them all. Some aspects, such as the availability of literature and the ease of communication, have become lesser constraints on diatom research than they used to be. However, the cost of training biologists to identify diatoms (generally requiring masters' – or doctoral-level qualifications) and the capital costs of good quality microscopes (which may exceed US $50,000) are high, and specialists cannot identify more than a few hundred specimens per day unless material is very clean and species-poor. Preparation of cleaned material for study also represents a significant expenditure in technician time. Not surprisingly, therefore, considerable efforts have already been made to develop identification systems that have less need for human intervention, and proof of concept was achieved for several morphology-based systems, reviewed by Mann *et al.* (2006), involving optical diffraction or automated feature extraction. However, no working system has been developed.

Furthermore, even after training and long experience, diatom experts frequently disagree about identifications, because (1) there are different available classifications, or (2) the taxonomic descriptions or the specimens themselves can be interpreted differently, or (3) the specimens being examined lack the

characteristics required for diagnosis. Thus, in relation to (1) some taxonomists are splitters, while others are lumpers, e.g. contrast the treatments of *Amphora* by Levkov (2009) and Krammer & Lange-Bertalot (1986-91). Also, the flora now in general use for identifying freshwater diatoms (Krammer & Lange-Bertalot, 1986-91) differs considerably from those (e.g. Hustedt, 1930: Patrick & Reimer, 1966, 1975) used to train a previous generation of diatomists. As an illustration of (2), it is sometimes difficult to decide in *Nitzschia* whether or not the central fibulae are more widely spaced or not, which is a critical character for correct identification according to modern taxonomy, e.g. Krammer & Lange-Bertalot (1986-91). An example of (3) is that, as diatoms become smaller during the life cycle, the shape and structure of their valves often become simpler and may converge with the morphology of other species (e.g. Hustedt, 1937). Several examples have been documented in which cultured diatoms become extremely small and lose the normal morphology of the species (e.g. Round, 1993). Hence, identification may become impossible unless the small valves can be plausibly linked to larger valves in the same sample or culture that do possess the features characteristic of the species. Furthermore, the full ranges of size and shape are known for very few diatom species because the whole of the life cycle has not been observed, so that valves are often found that in most respects correspond to the species description, but are smaller or larger than those previously recorded leading to uncertainty in identification. Gathering information about the total variation occurring during the whole life cycle is tedious, because of the length of the cycle in nature and laboratory culture, and complications of the mating system (Mann, 1988; Chepurnov *et al.*, 2004). And there may be undetected polymorphism, involving gradual or abrupt changes in morphology.

DNA barcoding could offer a solution to many of the difficulties of identifying diatoms. It would not in itself solve taxonomic disputes (point 1 above), although like any source of data about genotypic differentiation, barcodes would provide additional evidence about whether or not speciation has occurred (de Queiroz, 2007). However, in other respects barcoding would offer major advantages relative to microscope-based identification. For example, diatoms could be identified from material at any stage of the life cycle, including stages never encountered before, because barcoding directly characterizes the genome. Polymorphisms (such as in *Phaeodactylum*) and the formation of modified resting stages (as in many centric and a few pennate diatoms), and cells whose phenotype has been altered by environmental conditions, would also pose no problem. Communicating about sequences and their variation would be unproblematic, compared to the description and illustration of morphological variation. The resulting standardization of taxonomy and identification would greatly facilitate the combination and integration of datasets in community ecology and biogeography. Furthermore, because of the drive to sequence genomes for medical research, sequencing is inexpensive and likely to become even more so, and new protocols and bioinformatic methods have been developed to allow characterization of mixtures of genomes or genome fragments extracted from natural material (environmental DNA samples). Hence there is great potential for eventual replacement of many of the laborious microscopical counts made during ecological and applied research on contemporary diatom communities. Furthermore, if a barcode system were established for diatoms, in which barcode sequences were linked to correctly named specimens, material could be accurately identified by anyone with a basic training in DNA technology – someone who might not know what a diatom looks like.

Cataloguing, taxonomic revision and benefits for microscope-based identification

Since they represent new information about the genotypes of organisms, barcodes will directly aid taxonomic revision, supplementing the morphological data already available. Arguably more important, however, are the new collections that the development of barcoding will require and produce. In order to produce the reference barcodes, many new samples will have to be collected and clonal cultures isolated from them (which could be used for many purposes besides barcoding). The cultures will then need to be vouchered via slides, unmounted material and SEM stubs, documented via images, morphometrics and descriptions, and their DNA extracted and preserved. Barcoding will therefore create, as a by-product, enormous new resources for diatom systematics and microscope-based identification of diatoms. Information about these resources will be freely and easily accessible via the Barcode of Life Data Systems website (e.g. choose Heterokontophyta at http://www.boldsystems.org/views/taxbrowser_root.php) and linked sites giving further morphological detail, as well as taxonomic and ecological data. Hence barcoding will bring huge benefits even for disciplines that cannot directly employ barcodes for identification, such as palaeoecology or stratigraphy.

Barcodes also have great potential for stabilizing the nomenclature of extant diatoms. At present, the types of diatoms are physical specimens, which are almost always cleaned of all organic material (including DNA) and mounted in resin. They are held in many different herbaria, some without adequate microscope facilities, some unwilling or unable to lend specimens: poor access to type material is a major source of instability and confusion in diatom taxonomy. Furthermore, the fact that most types can only be examined by light microscopy because they are permanently mounted in resin on slides, means that they often do not provide enough information to control the use of the name they define. Even where unmounted material is available for examination with SEM, it can sometimes be impossible to be sure how a name should be applied (e.g. Trobajo *et al.*, submitted). Introduction of barcoding will not immediately solve these problems, but once barcodes have been linked to type specimens and made effective nomenclaturally (e.g. by the expedient of designating an epitype whose barcode is known: Jahn *et al.*, 2008; Evans & Mann, 2009), the need to refer to physical type specimens will be greatly reduced. A barcode sequence — essentially a molecular type — is unambiguous and can be communicated easily, unlike the morphology of a physical specimen.

Species discovery

DNA barcoding was initially developed for groups of organisms, such as birds (e.g. Hebert *et al.*, 2004) and fish (e.g. Ward *et al.*, 2005) where the alpha taxonomy (the initial recognition, description and naming of species, principally using morphological data) is already more or less complete. In such organisms, and in some others where barcoding has lagged behind because of methodological problems (e.g. angiosperms), the purpose of barcoding is clearly (and almost solely) identification. In diatoms, on the other hand, there is still a huge amount of alpha taxonomy to be done. Many regions, especially in the tropics, have not been explored, and the benthic diatoms of the marine sublittoral are almost unknown.

Thus, even if we had a database of barcodes representing all the diatom species described so far, many barcode sequences obtained from natural populations or cultures would correspond to nothing in the database. Some might be

very similar to the barcode for a known species and could represent hitherto unrecognized intraspecific variation. Many others would represent undescribed species, requiring further characterization, description, and allocation to the appropriate supraspecific category (section, genus, family, etc). For this, it is desirable that the barcode can be used not only for identification, its primary purpose, but also for a very preliminary phylogenetic analysis. For this, it should be easily aligned and contain relatively conserved sites or regions (giving higher-level information), as well as rapidly evolving parts (for the barcoding itself).

A valid criticism of species discovery through barcodes (or other single molecular markers) is that it assumes that, below some threshold level of divergence, speciation has not occurred. This is unjustified because molecular divergence is not causally linked to speciation. Neutral genetic differences between sister species accumulate with time in a more or less clock-like manner but there may be a long delay after speciation has occurred before sister species become reciprocally monophyletic with respect to a barcode marker (see e.g. de Queiroz, 2007). The faster the barcode marker evolves, the less likely it is that recently evolved species will go unnoticed. Moniz & Kaczmarska (2010) suggested that ITS-2 rDNA is intrinsically superior to other candidate barcode regions, because of an apparent relationship between the presence of compensatory base changes (CBCs) in the helices of the molecule and reproductive isolation. However, even if the two are correlated, this is only a refinement of the idea of a molecular threshold for detecting species, because again there is no causal relationship between compensatory base-change and speciation. When a CBC is present, speciation has most likely occurred (e.g. Coleman, 2007) but the absence of CBCs doesn't imply the absence of speciation.

Despite these caveats, the extent to which a newly determined barcode diverges from the barcodes of all known species, or the discovery of a new barcode clade, is a useful first indication of the existence of undescribed species, which can then be confirmed using other data. For example, Vanormelingen *et al.* (unpublished data) investigated Australian and European *Sellaphora* using the COI barcode developed by Evans *et al.* (2007) and discovered several new COI clades among clones initially identified from morphology as *S. auldreekie* D.G. Mann & S.M. McDonald. On further examination, using other markers of speciation (i.e. applying the multidisciplinary approach to species discovery advocated e.g. by de Queiroz, 2007 and Mann, 1999), the COI clades were shown to correspond to reproductively isolated species, some of which could also be separated by very subtle differences in valve shape and striation.

There is mounting evidence that many diatom 'species' are like *S. auldreekie sensu lato*, containing cryptic, semicryptic or pseudocryptic species (semicryptic species-pairs are species whose ranges of morphological variation partially overlap so that only a proportion of individual cells can be identified; in pseudocryptic species, all individuals can be told apart morphologically, but only with great difficulty). The existence of cryptic and semicryptic species in protists was originally detected by the observation that sets of morphologically very similar or identical populations were reproductively isolated from each other, e.g. in *Paramecium* and *Pandorina* (Sonneborn, 1975; Coleman, 1959). However, testing compatibility with known 'test strains' is impractical for routine identification and this was the principal reason why Sonneborn did not initially give formal taxonomic recognition to the 'syngens' he discovered in *Paramecium*. The first known cases of cryptic and pseudocryptic diatom species have been reviewed by Mann (1999) and Mann & Evans (2007), while more recent examples include Evans *et al.* (2007, 2008), Vanormelingen *et al.* (2008b), Mann *et al.* (2008), Quijano-Scheggia

et al. (2009), Vanelslander *et al.* (2009), Trobajo *et al.* (2009 and in press), Poulíčková *et al.* (2010) and Sorhannus *et al.* (2010). The only practical way to discover and identify cryptic and semicryptic species is through molecular methods.

Pseudocryptic species, on the other hand, can by definition be detected and characterized morphologically, without recourse to molecular methods or mating experiments. However, the work necessary (application of morphometric and statistical methods of various kinds, or LM and SEM studies of many individuals from different populations) is enormous and experience suggests that pseudocryptic species are often unrecognized during microscopical examination, remaining confused with each other within an undifferentiated nominal species until information from other sources shows that taxonomic reinterpretation is necessary. For example, differentiation of 'varieties' within *Cocconeis placentula* Ehrenb. was aided by information about the mating system and reproductive isolation (Geitler, 1982: see Mann, 1999, p. 462) and reinterpretation of *Sellaphora pupula* (Kütz.) Mereschk. as a complex of many pseudocryptic species, rather than a single highly variable species, was prompted by observations of mating barriers between sympatric populations (Mann, 1984), with later confirmations from molecular data (e.g. Evans *et al.*, 2007, 2008). It is highly unlikely that the pseudocryptic species here, or in *Skeletonema* (Sarno *et al.* 2005) or *Pseudo-nitzschia* (e.g. Amato *et al.*, 2007), would have been discovered so readily if morphological data alone had been available. DNA barcoding therefore represents an unequalled opportunity to discover the full extent of species-level biodiversity in diatoms, rather than the fraction that can be detected microscopically.

New possibilities for studying diatom biogeography and the ecology of living diatoms

Because of the need to observe the fine detail of frustule ornamentation and structure for accurate identification, diatoms are usually dead when they are identified and in many cases it is not clear whether they were killed by the cleaning itself or were already dead when sampled. Thus it is often difficult to determine whether particular individuals were contemporary, living alongside each other in nature, or had allochthonous or allochronic origins. The true composition of communities may therefore be unclear and the ecological requirements of individual species masked (cf. Round, 1971, p. 102).

At first sight, the solution seems simple: identify diatoms while they are still alive. However, using microscopical methods, this is difficult, for the following reasons. Firstly, the difference in refractive index between water (RI = c. 1.3: International Association for the Properties of Water and Steam, 1997) and diatom silica (RI = 1.43: Lewin, 1962) is much less than between diatom silica and mountants such as Naphrax (RI = c. 1.7: Fleming, 1954), so that frustule details are more difficult to see in living material. Secondly, the chloroplasts and other cell contents obscure cell wall patterning, although improvements in detection can sometimes be made by using interference contrast optics (with a short effective depth of focus) and applying high-pass filters to digital images (an example is given at http://rbg-web2.rbge.org.uk/algae/sellaphora/sellaphora_valve_formation.html). Thirdly, although chloroplast morphology provides additional diagnostic information (and should therefore be recorded during barcoding programmes), this gain does not outweigh the loss of frustule detail, because chloroplast morphology varies little within most genera and scarcely at all among centric diatoms (most of which contain many small, undistinctive chloroplasts). Thus, if

based on living diatoms, microscopical identification will always be less certain than when using cleaned frustules. Furthermore, there are currently no identification guides for living diatoms, apart from the ground-breaking key produced by Cox (1996), which covers the most common freshwater diatoms.

An extra advantage of DNA barcodes over microscopical identification, therefore, could be to provide information on the composition of living diatom communities, especially where it is difficult or impossible to separate live cells physically from dead cells or from their environment. Characterization of a community might be only qualitative, in the form of a list of species present, but could also be quantitative. For example Créach *et al.* (2006) used quantitative PCR to examine temporal and spatial patterns of abundance in *Navicula phyllepta* Kütz. The community studied — estuarine epipelon — is especially difficult to study directly because the sediment particles are not dissimilar to diatoms in size, and all methods for separating epipelic diatoms from the sediment (e.g. Eaton & Moss, 1966) are inefficient. Créach *et al.* made indirect estimates of the abundance of living *N. phyllepta* by determining the relative abundance of *N. phyllepta* from counts of cleaned diatoms and then using the mean chlorophyll a content per cell of *N. phyllepta* to apportion the total chlorophyll a content per square metre. The results showed good agreement between indirect microscope-based counts and PCR estimates, but the PCR estimates had the advantage of speed and in addition could differentiate unambiguously and easily between two pseudocryptic species within the *N. phyllepta* complex, which have subsequently been shown to differ in their physiologies and ecological requirements (Vanelslander *et al.*, 2009). Barcodes could be used to design specific probes for the detection of species that have useful properties (e.g. that they indicate particular environmental conditions) or undesirable characteristics (e.g. production of domoic acid) (e.g. Diercks *et al.*, 2008). So far, however, 'molecular ecology' of diatoms has not been undertaken using explicitly designated barcodes and the gene regions used (Créach *et al.* used ITS1 and Diercks *et al.* mostly SSU rDNA) are, in our opinion, suboptimal for general species barcoding in diatoms, though they may be highly effective in particular cases.

The examples given above show how barcodes might be used for autecological studies and the detection of individual species that are important in some special way, e.g. as bioindicators or producers of toxins. In addition, whole communities could be characterized via 'environmental DNA' (metagenomic) analyses, in which the identities of organisms would be determined from the barcodes. Next-generation sequencing of whole-community SSU and ITS rDNA has already been used to demonstrate seasonal turn-over in fungal and protist communities (Jumpponen *et al.*, 2010; Medinger *et al.*, 2010; Nolte *et al.*, 2010), and to fully sample a lake's protist community diversity (Nolte *et al.*, 2010). It should be noted however that serious drawbacks for identification and quantification exist, including PCR biases, sequencing errors, a restricted sequence fragment size, and incomplete reference databases (Kunin *et al.*, 2010; Tedersoo *et al.*, 2010). In parallel, functional aspects of the community could be monitored via genes involved in specific activities such as photosynthesis, nitrogen fixation, etc (e.g. Wu *et al.*, 2001). However, the use of barcodes in ecology is not a magic solution to the problems of distinguishing living from dead, allochthonous from autochthonous, or synchronous from asynchronous. Short DNA sequences derived either from dead organisms or from the excretions and secretions of living ones, may persist in the environment for long periods. For higher animals, this offers the advantage that sites can be surveyed for the presence of species when the small numbers of individuals present would make detection very difficult by

normal survey methods (Ficetola *et al.*, 2008). Persistence of degrading DNA could also be an advantage for biogeographical studies of diatoms, because species abundance can vary greatly between seasons and it may be more effective to search for molecular traces of a species' presence than to look for visible remains. However, biogeographical studies require not only that we establish the *presence* of species in an area but also that we show their *absence* elsewhere. Finlay *et al.* (2002) argued that proving absence by microscopical examination is impossible for diatoms and the same argument applies (though with slightly less force) to molecular surveys, unless there are inherent limits to rarity, e.g. through a requirement for opposite sexes to be sufficiently common to mate and generate offspring. When there is a need to be sure that molecular methods are detecting living cells, rather the degrading DNA from dead ones, special techniques can be applied such as those discussed by Fittipaldi *et al.* (in press).

Use of DNA barcodes could also allow new examination of the dispersal of diatoms and the role of humans in altering distributions. It appears that many diatoms must move around rapidly without human aid, relative to the rate of change in niche requirements, because otherwise diatom biomonitoring and palaeoecology would not work so well (e.g. Bennett *et al.,* 2010). On the other hand, statistical analyses of large geographical datasets suggest that dispersal is indeed a constraint on the distribution of diatoms (Telford *et al.*, 2006; Vyverman *et al.*, 2007; Verleyen *et al.*, 2009), as do population genetic data (Evans *et al.*, 2009; Casteleyn *et al.*, 2010). Furthermore, circumstantial evidence, such as the apparent lack of *Asterionella* in pre-European lake sediments in New Zealand (Harper, 1994) and perhaps other recent floristic changes (see Vanormelingen *et al.*, 2008a; Spaulding & Elwell, 2007), suggests that humans may have accidentally extended the ranges of some diatom species by overcoming dispersal limitation (rather than by creating new areas of suitable habitat). If the natural spread of diatom species is indeed significantly constrained by dispersal over periods of centuries to millions of years (Vanormelingen *et al.*, 2008a), then new measures may be necessary to avoid unintentional transfers by humans, to avoid problems like those caused by alien plants and animals (e.g. Lowe *et al.*, 2000). However, there is almost no information about how diatoms actually disperse nor about the effectiveness of different kinds of barrier (e.g. fronts between different water masses for oceanic diatoms, open sea for benthic marine diatoms, salinity gradients, land, etc) in preventing migration. Data on introductions of fish (e.g. Rahel, 2007), including the remarkable observation that 95% of c. 16 000 high-altitude lakes in the western USA were originally fishless whereas 9500 of these lakes (c. 60%) now contain introduced fish (Bahls, 1992), suggest how huge the effect of unintentional human activity could be in diatom biogeography, creating artificially widespread or cosmopolitan species.

In order to make progress in understanding dispersal and geographical distribution, we must first be sure that the geographical ranges recorded are not artifacts of poor taxonomy and DNA barcodes could provide for this. Barcoding could also be used to test for the presence of live propagules on or in potential vectors. For example, it has long been known (e.g. Darwin, 1846) that diatom remains can be transported very long distances in the air, but most or all of the cells are dead; furthermore, most diatoms that have been tested exhibit low tolerance to desiccation (e.g. Souffreau *et al.*, 2010). As with ecological studies, possible dispersal agents need to be tested for their capacity to transport live cells. This can be done by microscopical examination (e.g. as done for ballast water by Klein *et al.*, 2010), but this may be like 'looking for a needle in a haystack'. Another is to culture material, but culturing is highly selective and live cells may

be accompanied by huge amounts of detritus that makes it impractical to isolate them. An alternative may be to use fluorescence-based cell sorting (e.g. using chlorophyll autofluorescence) to find and isolate live diatoms from washings off water-birds, the guts of aquatic insects, etc., followed by barcode identification of single cells or cultures.

Barcodes could be useful for tracking the spread of putative aliens and to determine their source, as in the study by Brawley *et al.* (2009) of the invasion of NE North America by the brown alga *Fucus serratus* L. and the mollusc *Littorina littorea* (L.). In this work, it was possible to identify the likely agents of dispersal, by matching the distributions of individual genotypes to records of shipping between the British Isles and America. Similar work would be done in diatoms if the barcode marker exhibits enough infraspecific variation.

LIMITATIONS OF BARCODING

Barcoding depends on the assumption that speciation (whatever the species concept in use) is generally accompanied by divergence in the sequence of the barcode gene(s). However, sequence divergence is stochastic rather than steady and so some closely related species will not be resolved by barcoding, even if parts of the barcode evolve rapidly. To identify such species, other evidence (from further molecular markers or morphology) will be needed. There is an extra problem, worse with the 'weak' barcodes that we define and discuss later ('weak' barcodes are those that reflect use of a coarse-grained taxonomy: see "The taxonomic basis of barcoding"), or with recently evolved species such as the North American *Stephanodiscus* studied by Zechman *et al.* (1994), that it may be impossible to barcode some species simply (i.e. via a single reference sequence, with identification based on divergence only), because they are paraphyletic. This can be illustrated in *Sellaphora*, where two of the most commonly recorded nominal species are *S. pupula* and *S. bacillum* (Ehrenb.) D.G. Mann (Hustedt, 1930; Krammer & Lange-Bertalot, 1986-91, as *Navicula pupula* and *N. bacillum*). According to available molecular data (Evans *et al.*, 2007, 2008), *S. bacillum* evolved as one lineage among many within *S. pupula sensu lato*, which makes the latter paraphyletic. Hence, any barcode marker that separates *S. bacillum* from *S. pupula sensu lato* (to reflect the taxonomy of Krammer & Lange-Bertalot 1986-91) is unlikely to provide a single barcode sequence for *S. pupula* (this difficulty could be by-passed by using character-based comparisons, rather than simple divergence). Barcoding therefore has limitations and it is inherently incapable of identifying all diatoms. However, since speciation is a process in which different species characteristics are reached in a different order and at different times in different lineages (de Queiroz, 2007), all characteristics, including morphology and reproductive isolation, will sometimes fail if used singly for species differentiation.

There will probably always be some species or groups in which the chosen barcode cannot be used. For example, if *rbc*L were to be chosen as the preferred barcode marker for diatoms (or any other plastid marker whose function is to code for proteins involved in photosynthesis), barcoding would not work for the few species that lack a functional plastid and are obligately heterotrophic, e.g. *Nitzschia alba* Lewin & Lewin; an alternative marker would have to be developed for these. Even DNA extraction itself appears to be

problematic in some diatoms that produce large amounts of mucilage, e.g. *Mastogloia* species (D.G. Mann & G.E. Simpson, unpublished data).

It should be emphasized too that barcoding does not make microscopy redundant. Many aspects of community structure and function — such as the three-dimensional arrangement of cells, motility, or cell-size spectra (reflecting the proportions of cells at different stages of size reduction) — cannot be determined without use of some kind of microscopical or optical technique, and much would therefore be lost if barcoding was regarded as an alternative to microscopy, rather than complementary to it.

CHALLENGES IN DEVELOPING BARCODING FOR DIATOMS

The principal challenges are (1) to choose the taxonomic basis for barcoding, (2) development and testing of candidate barcodes, and (3) generation of a sufficiently comprehensive set of barcodes to make barcode identification practical.

The taxonomic basis of barcoding

We mentioned earlier that one source of disagreement about the identification of diatoms is that people adopt different taxonomies. This is especially evident in the ecological and palaeoecological literature, where it is common to find continued successful use of classifications that have been superseded. For example, the 2010 issues of *Journal of Paleolimnology* include several papers (e.g. Ampel *et al.*, 2010; Norberg *et al.*, 2010; Shinneman *et al.*, 2010; Wiklund *et al.*, 2010) that rely for identification primarily on Krammer & Lange-Bertalot (1986-91) or older literature. These papers successfully investigate, analyse and interpret historical changes in the environment without needing to incorporate the many taxonomic changes that have been made since 1991 (e.g. Krammer, 2000, 2002; Lange-Bertalot, 2001; Levkov, 2009; and volumes of *Iconographia Diatomologica*). There is therefore a conflict between (1) the taxonomists, who are clearly unsatisfied with what they have done so far and wish to continue to develop their classifications until they represent biodiversity, comprehensive and faithfully, to species level and beyond, and (2) some or many of those who use classifications for ecology or palaeoecology, who find the scope and detail of late 20th century taxonomy adequate, even if they may have problems in applying the taxonomy (because of access to literature, lack of illustrations, poor descriptions, etc).

This conflict is a problem for the development of DNA barcoding in diatoms. Most taxonomists would probably wish to design a barcode system that is sensitive enough to discriminate between all the new species they are describing and expect to describe, including cryptic and pseudocryptic forms. Taxonomists might therefore choose a molecular marker that evolves comparatively rapidly, such as ITS-1 or -2 (Moniz & Kaczmarska, 2009, 2010), or COI (Evans *et al.*, 2007). We will call this a ‘strong’ barcode. In contrast, those who already use diatoms successfully for biomonitoring and find that a crude taxonomy is sufficient for their purposes, might be content with a ‘weak’ barcode, with lower discriminatory power. It might be possible, for example, to design a barcode

system that largely emulates the widely used freshwater flora of Krammer & Lange-Bertalot (1986-91), replacing microscopical features by molecular ones. This seems to have been the philosophy of Jahn *et al.* (2007) during their investigation of SSU rDNA as a possible barcode marker. It should be noted, however, that for some diatoms, such as the marine benthos, there are *no* floras or monographs that can be used as a basis for making a list of the nominal 'species' that would need to be differentiated by 'weak' barcoding.

In making a choice between 'weak' and 'strong' barcodes, it would be helpful to know whether speciation in diatoms is usually or always accompanied by differentiation in physical or chemical requirements, or sensitivity to biotic factors, or whether instead clades of closely related species share the same niche. To discover the answer will require detailed studies of speciation and accompanying ecological studies in contrasting environments and different kinds of diatom, e.g. planktonic and benthic; freshwater, brackish and marine; motile and nonmotile; oogamous and isogamous, heterothallic, homothallic, automictic and apomictic; and different lineages of centric and pennate diatoms. These studies are in their infancy but the results so far tend to indicate that speciation (including speciation at cryptic and pseudocryptic levels) *is* associated with niche differentiation (e.g. Poulíčková *et al.*, 2008; Špačkova *et al.*, 2009; Vanelslander *et al.*, 2009; Quijano-Scheggia *et al.*, 2009), implying that the resolution of ecological monitoring could be enhanced by applying a 'strong' barcode. Therefore, choice of a 'weak' barcode is likely to inhibit use of these organisms for biotechnology and biomonitoring, and would hinder studies of population genetics, biogeography and ecology.

Thus the kind of barcode has important repercussions. The extra resolution provided by a 'strong' barcode system will probably allow future refinements of biomonitoring and ecological research. It will also allow recognition and study of cryptic species. A 'weak' barcode system, on the other hand, will tend to stabilize taxonomy in the state reached in the mid to late 20th century. It would translate a classification developed largely by light microscopy, and already known to be insufficient for some research fields, into a molecular identification system. The precautionary principle suggests that, if there is any doubt about which kind of barcode is needed, a 'strong' one should be chosen, because data obtained using this system can always be combined later to provide information at a coarser level (i.e. with less taxonomic resolution). However, the cost of producing a 'strong' barcode system will certainly be higher than that of a 'weak' system.

Development and testing of candidate barcodes

Several papers have been published that seek to evaluate potential barcode regions in diatoms. The regions assessed so far are SSU rDNA, ITS rDNA, LSU rDNA, the universal plastid amplicon (UPA), *rbc*L and COI. As in any other groups of organisms, the criteria for assessing barcodes are: (A) universality (can the barcode protocol be applied in the whole of the group of organisms for which it is designed?), (B) practicality (is the protocol cheap and easy to use, and are the results easily interpretable?), and (C) discrimination (can the barcode distinguish among the species or other taxa that we wish to identify?). The ability of a marker to discriminate is also referred to as its 'power' (e.g. Hamsher *et al.*, in press).

Universality can be assessed by testing the barcode on a variety of taxa from across the diatom phylogeny. There is as yet no consensus about the branching order of the major lineages of diatoms (Mann & Evans, 2007; Theriot *et al.*, 2009), but enough is known about relationships to show that a universality test should include representatives of each of the major lineages of 'radial centric diatoms' (e.g. the informally named classes of 'Coscinodiscophytina' listed by Adl *et al.*, 2005), several 'multipolar centric diatoms', and a diversity of pennate diatoms. Practicality requires that the barcode is sufficiently short to allow bidirectional reads using a single pair of primers, and that subsequent analysis is straightforward, not requiring complex algorithms to obtain the necessary alignment. Practicality changes with time and generally becomes less of a constraint, as equipment and bioinformatic protocols improve. Thus, for example, the difficulties caused now by a marker that is over-long (e.g. > 700 bp) may become trivial within a few years. Practicality might be enhanced if there were already many sequences that could be re-used as barcodes (e.g. SSU rDNA in diatoms: Jahn *et al.*, 2007). However, the quality of identifications and vouchering of diatom sequences in GenBank is very uneven (Mann & Evans, 2007, Table 13.1) and the proportion of diatom species for which there is even one gene sequence is tiny.

The most disappointing region with respect to universality has been the partial COI region (close to the 5' end) used extensively for barcoding metazoa, which was tested in one of the first diatom barcode papers by Evans *et al.* (2007). Evans *et al.* showed that COI sequences were highly variable and provided excellent discrimination between closely related species in *Sellaphora*, a genus in which there was already abundant information about speciation (*sensu* de Queiroz, 2007) from morphology, various molecular markers, the mating system, and reproductive compatibility (e.g. Evans *et al.,* 2008; Mann *et al.*, 2009). The COI barcode region was also easy to align and compare, and generally required only a single pair of primers for amplification and sequencing. Subsequent papers (e.g. Jahn *et al.*, 2008; Evans & Mann, 2009) explored how barcodes might be made nomenclaturally effective in diatoms, through use of a barcoded epitype. In *Sellaphora*, therefore, and also in some other cases (e.g. *Gomphonema parvulum* Kütz.: Kermarrec *et al.*, 2010), COI seemed to be ideal. However, outside these few cases, amplification and sequencing COI has often proved to be problematic (e.g. Moniz & Kaczmarska, 2009; Hamsher *et al.*, in press), despite the availability of full sequences of COI from a diversity of diatoms from the work of Ehara *et al.* (2000a, b) and trials of many primer combinations. Trobajo *et al.* (in press) tested COI in *Nitzschia palea* (Kütz.) W. Smith and showed that, where sequences could be obtained (> 80% of isolates), COI outperformed *rbc*L and LSU rDNA for discrimination, but it failed the universality test, with closely related diatoms differing in whether any diatom COI product could be amplified and sequenced. The authors have had the same experience in *Pinnularia* (C. Souffreau, unpublished data) and *Eunotia* (P. Vanormelingen, unpublished data): COI generally works well, but fails in particular lineages, e.g. *P. borealis* Ehrenb. and *E. mucophila* (Lange-Bert. *et* Nörpel-Schempp) Lange-Bert.. Hence, it seems that COI may be valuable only in special cases, such as *Sellaphora*, though it is conceivable that a technical solution to the universality problem may yet be found if the reasons for COI failures can be determined.

SSU rDNA has been studied extensively in diatoms and there is abundant information to guide choice of a barcode region within it and design of appropriate primers. The other proposed markers have not been so well tested for universality, the 'radial centrics' being poorly represented in tests and the raphid

diatoms overrepresented (although the raphids are the most species-rich group within the diatoms, they are of phylogenetically recent origin and may therefore exhibit fewer changes at conserved sites). Nevertheless, although more tests are desirable, it seems likely that partial LSU rDNA, partial ITS-1–5.8S–ITS-2, *rbc*L or partial *rbc*L and UPA all have the necessary universality (Moniz & Kaczmarska, 2009, 2010; Hamsher *et al.*, in press). So, are any or all of these rDNA regions practical as barcodes, and if so, do they show sufficient discrimination to act as 'strong' or 'weak' barcodes?

The problem with any of the rDNA regions is **practicality**: there is often intragenomic variation (reviewed by Alverson, 2008), as a result of the presence of multiple non-identical copies of the rDNA cistron, which may be distributed among one, two or several loci (e.g. Álvarez & Wendel, 2003). Although there may be one predominant copy, others may be sufficiently abundant to prevent unambiguous reads during direct sequencing, especially because there is often variation in length because of insertions and deletions. Direct sequencing of ITS was impossible in species of several genera, including *Nitzschia* (Trobajo *et al.*, 2009), *Achnanthes* (E. Clavero, unpublished data), *Eunotia* (Vanormelingen *et al.*, 2007), *Navicula* (A. Poulíčková, personal communication), *Pseudo-nitzschia* (D'Alelio *et al.*, 2009) and *Sellaphora* (Behnke *et al.*, 2004). Furthermore, alignment of rDNA sequences is not straightforward, becoming more difficult with phylogenetic separation, at a rate determined by functional constraints on the evolution of the molecule (so that on average divergences are in the order ITS-1 > ITS-2 > SSU > 5.8S). Alignment is not a prerequisite for identification, since sequences can be compared using algorithms such as BLAST. However, if barcode regions are to be used for species discovery as well as identification, which is desirable given the current state of diatom taxonomy, then ease of alignment becomes an important practical issue.

For these reasons, rDNA is suboptimal for barcoding and Evans *et al.* (2007), Trobajo *et al.* (in press) and Hamsher *et al.* (in press) have argued that practical issues should rule out ITS rDNA for barcoding diatoms (a view not accepted by Moniz & Kaczmarska, 2010). However, intragenomic variation may be sufficiently low and interspecific variation sufficiently high with some rDNA regions, e.g. the D1-D3 segment of LSU rDNA, to make them workable as barcodes (Hamsher *et al.*, in press, used D2-D3; Trobajo *et al.*, in press, used D1-D2).

Protein-encoding genes such as COI and *rbc*L pose fewer practical problems than rDNA, once they have been obtained (i.e. if the universality criterion is satisfied): there is rarely any intragenomic variation (almost none in length, while the few intragenomic substitutions can be detected as resolvable ambiguities at one or a few sites), and they are very easily aligned and compared. There is an extra advantage too, that sequencing errors can often be detected by frame shifts and unlikely amino acid changes, e.g. exchange of one type of amino acid by a different one (e.g. polar by non-polar, basic by acidic, etc). Furthermore, COI and *rbc*L are organellar genes and many copies are present in each plastid or mitochondrion, so that amplification is as straightforward as with rDNA, once suitable primers have been developed.

Once the universality and practicality criteria have been satisfied, barcode markers need to be tested for **discrimination**. Of the three criteria, this is the most difficult to assess in diatoms, because of the state of current taxonomy and the choice that has to be made between 'weak' and 'strong' barcodes. Testing discrimination requires that we have already discovered enough about relationships to know which species are sisters and therefore likely to be minimally

different genetically. It is pointless to examine the performance of a barcode in discriminating between randomly selected species whose relationships are unknown or known only from the barcode itself. Unfortunately, there are very few groups of species that are understood well enough to be used as a basis for testing. The best are *Pseudo-nitzschia*, because of the extensive work on the genus in connection with toxin production, and the *Sellaphora pupula-bacillum* clade, although there are almost certainly species still to be discovered in both, especially *Sellaphora*. Also useful are *Eunotia* (e.g. Vanormelingen *et al.*, 2008b) and *Navicula* (Poulíčková *et al.*, 2010). Among centric diatoms, the best are probably *Aulacoseira* (Edgar & Theriot, 2004), *Cyclotella meneghiniana* Kütz. (Beszteri *et al.*, 2007) and *Skeletonema* (e.g. Sarno *et al.*, 2005; Godhe *et al.*, 2006), although these lack the extra dimension provided by data on reproductive isolation and mating systems. Other 'model systems' will undoubtedly become available as taxonomists continue to use a battery of molecular, morphological and other markers to investigate species complexes in detail.

Of the model systems listed above, the only one used explicitly for testing barcodes has been *Sellaphora* (Evans *et al.*, 2007; Hamsher *et al.*, in press). Using this system, Hamsher *et al.* found that a combination of partial LSU rDNA and partial *rbc*L could provide (just) enough discrimination to be used for barcoding *Sellaphora* species. These two markers also perform well in *Pseudo-nitzschia*, judging by previous phylogenetic studies (e.g. Amato *et al.*, 2007), and apparently also in *Nitzschia palea* (Trobajo *et al.*, in press). COI, favoured by Evans *et al.* (2007) on the basis of its performance in *Sellaphora*, would be preferable to either LSU rDNA or *rbc*L in *Nitzschia palea* because of its high discriminatory power (see also Trobajo *et al.*, in press), but it lacks universality, as already noted. UPA, which was selected as a promising marker for barcoding on the basis of its performance in other groups of organisms (Sherwood & Presting, 2007), does not discriminate sufficiently between closely related species of diatoms (Hamsher *et al.*, in press).

Have we reached the end of the search for a barcode marker for diatoms, with the selection of *rbc*L-3P and the D2/D3 region of LSU as a non-ideal but workable solution ('borderline strong'), as recommended by Hamsher *et al.* (in press)? It is hard to believe that better markers could not be found, and with increasing numbers of whole nuclear and organelle genome sequences available, there are now new opportunities to search objectively for regions that can be reliably and easily amplified from all diatoms but show better discrimination than LSU rDNA and *rbc*L. However, barcoding is not just a scientific project but a sociological phenomenon. The idea of barcoding all life attracted attention because of its apparent simplicity and novelty, and its grand scale — biodiversity's counterpart of the human genome project. But the novelty is already almost gone and there is a real danger that diatom barcoding will not receive a fair share of funding and support unless the barcode region is decided soon.

Generation of diatom barcodes and the establishment of an identification system

The prospect of barcoding 200,000 extant diatom species is daunting, but the task can be tackled in stages, for example by habitat or genus or bioindicator value. In multicellular organisms, obtaining material for producing the reference barcode is rarely a problem: a leaf or small piece of tissue cut from a plant or animal represents a single genotype and provides sufficient DNA for analysis; to provide equivalent numbers of cells and amounts of DNA in diatoms requires

culturing. However, many diatoms have never been cultured successfully (Mann & Chepurnov, 2004, p. 258) and others cannot be maintained permanently in culture because of their mating system (Chepurnov *et al.*, 2004). Consequently, diatoms are poorly and unevenly represented in culture collections and to barcode diatoms will require a major new effort to isolate and grow strains. These will need to be vouchered, via slides and unmounted material of cleaned frustules, documented via images and textual data on clone sources and characteristics, and identified as far as possible using microscopical criteria (including comparisons with type material). Our experience (P. Vanormelingen) suggests that it is the documentation and identification of strains, rather than routine isolation and culturing, that will be the bottleneck in barcoding programmes. The sequencing itself is probably the least challenging stage.

It is imperative to preserve DNA (either extracted or in frozen pellets of the cultures), in case the barcode needs to be checked and to provide material for subsequent research. This is especially important given that the perfect diatom barcode marker has not been found (see above) but may yet be developed. Future sequencing of a new barcode marker for all specimens in the barcode database is feasible given the present low cost of sequencing, *provided that DNA is available*. However, redoing the huge culturing and vouchering effort that will be needed (and is already underway) to build the barcode database would be near-impossible. For taxa that resist attempts to isolate and culture them, there is now the possibility of amplifying the whole genome from one or a few cells (e.g. Medlin *et al.*, 2008), though this is problematic for barcoding because of the difficulty of documenting and vouchering the morphology of the cells extracted.

The steps to be taken are therefore:

(1) Agreement on one or two barcode regions (based on their universality, practicality and discriminatory power) that must be sequenced for all species (additional regions may be useful for particular groups). Currently, 3'-*rbc*L and partial (e.g. D1-D3) LSU rDNA are the strongest candidates.

(2) Establishment of common protocols, including culturing, vouchering, characterization, DNA preservation, primers used, and most importantly the addition of all information to the central database at www.boldsystems.org.

(3) Further testing of potential barcode markers, and continued efforts to pinpoint species limits in model groups. Development and improvement of existing markers and protocols.

(4) Scaling up the culturing, vouchering, characterization and identification effort. Culturing, vouchering and DNA harvesting can be delegated to trained technicians and are aspects where major new funding is needed. Initially, there should be a focus on particular habitats and model groups, but in parallel also a wider coverage. Detailed studies of selected diatom communities and genera, leading to ± complete, usable identification systems would help to convince diatomists that the project is worthwhile. Meanwhile, a continuous background of sampling from a wide selection of habitats and diatom groups would make the database useful to non-diatomists seeking to identify sequences from environmental sequencing of complete microbial communities, but also encourage the maximum involvement of diatomists, whatever their specialization. The contributions of every available diatom taxonomist will be essential to ensure that barcode identifiers are correctly linked to the existing microscope-based taxonomy and ways should be sought to achieve maximum integration between the barcode effort and the large body of alpha-taxonomic work that is continuously being delivered in diatoms (e.g. via *Iconographia Diatomologica*).

Acknowledgements. We are grateful to many people for discussions about speciation and molecular species identification in diatoms: in the context of DNA barcoding, particularly Katharine Evans, who has done so much to pioneer barcoding in diatoms, Sarah Hamsher, Regine Jahn, Irena Kaczmarska, Lenaïg Kermarrec, Linda Medlin, Mónica Moniz, Aloisie Poulíčková, Frédéric Rimet, Koen Sabbe and Gary Saunders; we also thank three anonymous reviewers for their careful reading of the manuscript and constructive comments; and Line Le Gall and Gary Saunders for their kind invitation to submit this contribution to a special issue of *Cryptogamie-Algologie* on algal DNA-barcoding. This paper was informed by studies supported by EU FP6 SYNTHESYS awards to Rosa Trobajo and Pieter Vanormelingen, and we also acknowledge a Fund for Scientific Research – Flanders (FWO-Flanders, Belgium) award to Caroline Souffreau.

REFERENCES

ADL S.M., SIMPSON A.G.B., FARMER M.A., ANDERSEN R.A., ANDERSON O.R., BARTA J.R., BOWSER S.S., BRUGEROLLE G., FENSOME R.A., FREDERICQ S., JAMES T.Y., KARPOV S., KUGRENS P., KRUG J., LANE C.E., LEWIS L.A., LODGE J., LYNN D.H., MANN D.G., MCCOURT R.M., MENDOZA L., MOESTRUP Ø., MOZELEY-STANDRIDGE S.E., NERAD T.A., SHEARER C.A., SMIRNOV A.V., SPIEGEL F.W. & TAYLOR M.F.J.R., 2005 — The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *Journal of eukaryotic microbiology* 52: 399-451.

ÁLVAREZ I. & WENDEL J.F., 2003 — Ribosomal ITS sequences and plant phylogenetic inference. *Molecular phylogenetics and evolution* 29: 417-434.

ALVERSON A.J., 2008 — Molecular systematics and the diatom species. *Protist* 159: 339-353.

AMATO A. KOOISTRA W.H.C.F., LEVIALDI GHIRON J.H., MANN D.G., PRÖSCHOLD T. & MONTRESOR M., 2007 — Reproductive isolation among sympatric cryptic species in marine diatoms. *Protist* 158: 193-207.

AMPEL L., WOHLFARTH B., RISBERG J., VERES D., LENG M.J. & TILLMAN P.K., 2010 — Diatom assemblage dynamics during abrupt climate change: the response of lacustrine diatoms to Dansgaard-Oeschger cycles during the last glacial period. *Journal of paleolimnology* 44: 397-404.

BAHLS, P., 1992 — The status of fish populations and management of high mountain lakes in the western United States. *Northwest science* 66: 189-193.

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.

BENNETT J.R., CUMMING B.F., GINN B.K. & SMOL J.P., 2010 — Broad-scale environmental response and niche conservatism in lacustrine diatom communities. *Global ecology and biogeography* 19: 724-732.

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 polymorphisms. *European journal of phycology* 42: 47-60.

BRAWLEY S.H., COYER J.A., BLAKESLEE A.M.H., HOARAU G., JOHNSON L.E., BYERS J.E., STAM W.T. & OLSEN J.L., 2009 — Historical invasions of the intertidal zone of Atlantic North America associated with distinctive patterns of trade and emigration. *Proceedings of the National Academy of Sciences of the United States of America* 106: 8239-8244.

CASTELEYN G., LELIAERT F., BACKELJAU T., DEBEER A.-E., KOTAKI Y., RHODES L., LUNDHOLM N., SABBE K. & VYVERMAN W., 2010 — Limits to gene flow in a cosmopolitan marine planktonic diatom. *Proceedings of the National Academy of Sciences of the United States of America* 107: 12952-12957.

CHEPURNOV V.A., MANN D.G., SABBE K. & VYVERMAN W., 2004 — Experimental studies on sexual reproduction in diatoms. *International review of cytology* 237: 91-154.

COLEMAN A.W., 1959 — Sexual isolation in *Pandorina morum*. *Journal of protozoology* 6: 249-264.

COLEMAN A.W., 2007 — Pan-eukaryote ITS2 homologies revealed by RNA secondary structure. *Nucleic acids research* 35: 3322-3329.

COX E.J., 1996 — *Identification of freshwater diatoms from live material*. London, Chapman & Hall, 158 pp.

CRÉACH V., ERNST A., SABBE K., VANELSLANDER B., VYVERMAN W. & STAL L.J., 2006 — Using quantitative PCR to determine the distribution of a semicryptic benthic diatom, *Navicula phyllepta* (Bacillariophyceae). *Journal of phycology* 42: 1142-54.

DE QUEIROZ K., 2007 — Species concepts and species delimitation. *Systematic biology* 56: 879-886.

D'ALELIO D., AMATO A., KOOISTRA W.H.C.F., PROCACCINI G., CASOTTI R. & MONTRESOR M., 2009 — Internal Transcribed Spacer polymorphism in *Pseudo-nitzschia multistriata* (Bacillariophyceae) in the Gulf of Naples: recent divergence or intraspecific hybridization? *Protist* 160: 9-20.

DARWIN C., 1846 — An account of the fine dust which often falls on vessels in the Atlantic Ocean. *Quarterly journal of the geological society of London* 2: 26-30.

DIERCKS S., METFIES K., & MEDLIN L.K., 2008 — Molecular probe sets for the detection of toxic algae for use in sandwich hybridization formats. *Journal of plankton research* 30: 439-448.

EATON J.W. & MOSS B., 1966 — The estimation of numbers and pigment content in epipelic algal populations. *Limnology and oceanography* 11: 584-595.

EDGAR S.M. & THERIOT E.C., 2004 — Phylogeny of *Aulacoseira* (Bacillariophyta) based on molecules and morphology. *Journal of phycology* 40: 771-788.

EHARA M., INAGAKI Y., WATANABE K.I. & OHAMA T., 2000a — Phylogenetic analysis of diatom cox1 genes and implications of a fluctuating GC content on mitochondrial genetic code evolution. *Current genetics* 37: 29-33.

EHARA M., WATANABE K.I. & OHAMA T., 2000b — Distribution of cognates of group II introns detected in mitochondrial cox1 genes of a diatom and haptophyte. *Gene* 256: 157-167.

EVANS K.M., WORTLEY A.H. & MANN D.G., 2007 — An assessment of potential diatom "barcode" genes (cox1, rbcL, 18S and ITS rDNA) and their effectiveness in determining relationships in *Sellaphora* (Bacillariophyta). *Protist* 158: 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.

EVANS K.M. & MANN D.G., 2009 — A proposed protocol for nomenclaturally effective DNA barcoding of microalgae. *Phycologia* 48: 70-74.

EVANS K.M., CHEPURNOV V.A., SLUIMAN H.J., THOMAS S.J., SPEARS B.M. & MANN D.G., 2009 — Highly differentiated populations of the freshwater diatom *Sellaphora capitata* suggest limited dispersal and opportunities for allopatric speciation. *Protist* 160: 386-396.

FICETOLA G.F., MIAUD C., POMPANON F. & TABERLET P., 2008 — Species detection using environmental DNA from water samples. *Biology letters* 4: 423-425.

FITTIPALDI M., CODONY F., ADRADOS B., CAMPER A.K. & MORATÓ J., In Press — Viable real-time PCR in environmental samples: can all data be interpreted directly? *Microbial ecology* (published online 15 July 2010: DOI 10.1007/s00248-010-9719-1).

FLEMING W.D., 1954 — Naphrax: a synthetic mounting medium of high refractive index. New and improved methods of preparation. *Journal of the royal microscopical Society* 74: 42-44.

GEITLER L., 1982 — Die infraspezifischer Sippen von *Cocconeis placentula* des Lunzer Seebachs. *Archiv für Hydrobiologie, Supplement* 63 (*Algological studies* 30): 1-11.

GODHE A., MCQUOID M.R., KARUNASAGAR I., KARUNASAGAR I., & REHNSTAMHOLM A.S., 2006 — Comparison of three common molecular tools for distinguishing among geographically separated clones of the diatom *Skeletonema marinoi* Sarno et Zingone (Bacillariophyceae). *Journal of phycology* 42: 280-291.

HAMSHER S.E., EVANS K.M., MANN D.G., POULÍČKOVÁ A. & SAUNDERS G.W., In Press — Barcoding diatoms: exploring alternatives to COI-5P. *Protist.*

HARPER M.A., 1994 — Did Europeans introduce *Asterionella formosa* Hassall to New Zealand? *In*: Kociolek J.P. (ed) *Proceedings of the 11th International Diatom Symposium*. San Francisco, California Academy of Sciences, pp. 479-484.

HEBERT P.D.N., STOECKLE M.Y., ZEMLAK T.S. & FRANCIS C.M., 2004 — Identification of birds through DNA Barcodes. *PLoS Biology* 2(10): e312.

HUSTEDT F., 1930 — Bacillariophyta (Diatomeae). *In*: Pascher A. (ed.) *Die Süsswasserflora Mitteleuropas*, 2nd ed., vol. 10, Jena, G. Fischer, 466 p.

HUSTEDT F., 1937 — Zur Systematik der Diatomeen. II. Der Begriff des "Typus" bei den Diatomeen und der Umfang der Diagnosen. III Konvergenzerscheinungen und Kummerformen. *Berichte der Deutschen botanischen Gesellschaft* 55: 465-472.

INTERNATIONAL ASSOCIATION FOR THE PROPERTIES OF WATER AND STEAM, 1997 — *Release on the refractive index of ordinary water substance as a function of wavelength, temperature and pressure*. Erlangen, 7 p.

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

JAHN R., MANN D.G., EVANS K.M. & POULÍČKOVÁ A., 2008 — The identity of *Sellaphora bacillum* (Ehrenberg) D.G. Mann. *Fottea* 8: 121-124.

JUMPPONEN A., JONES K.L., MATTOX J.D. & YAEGE C., 2010 — Massively parallel 454-sequencing of fungal communities in *Quercus* spp. ectomycorrhizas indicates seasonal dynamics in urban and rural sites. *Molecular ecology* 19 (Suppl. 1): 41-53.

KERMARREC L., RIMET F. & BOUCHEZ A. 2010 — Genetic diversity within the *Gomphonema parvulum* complex. *In: Abstracts, 21st International Diatom Symposium, St Paul, Minnesota, United States, 29 August – 3 September 2010,* p. 30.

KLEIN G., MACINTOSH K., KACZMARSKA I. & EHRMAN J.M., 2010 — Diatom survivorship in ballast water during trans-Pacific crossings. *Biological invasions* 12: 1031-1044.

KRAMMER K. & LANGE-BERTALOT H., 1986-91 — Bacillariophyceae. *In:* Ettl H., Gerloff J., Heynig H. & Mollenhauer D. (eds), *Süsswasserflora von Mitteleuropa*, 4 vols. Stuttgart & New York, G. Fischer.

KRAMMER K., 2000 — The genus *Pinnularia*. *In:* Lange-Bertalot H. (ed.), *Diatoms of Europe: diatoms of the European inland waters and comparable habitats*, vol. 1. Ruggell, A.R.G. Gantner, 703 p.

KRAMMER K., 2002 — *Cymbella. In:* Lange-Bertalot H. (ed.), *Diatoms of Europe: diatoms of the European inland waters and comparable habitats*, vol. 3. Ruggell, A.R.G. Gantner, 584 p.

KUNIN V., ENGELBREKTSON A., OCHMAN H. & HUGENHOLTZ P., 2010 — Wrinkles in the rare biosphere: pyrosequencing errors can lead to the artificial inflation of diversity estimates. *Environmental microbiology* 12: 118-123.

KÜTZING F.T., 1844 — *Die kieselschaligen Bacillarien oder Diatomeen*. Nordhausen, W. Köhne, 152 p.

LANGE-BERTALOT H., 2001 — *Navicula sensu stricto*, 10 genera separated from *Navicula sensu lato*, *Frustulia*. *In:* Lange-Bertalot H. (ed.), *Diatoms of Europe: diatoms of the European inland waters and comparable habitats*, vol. 2. Ruggell, A.R.G. Gantner, 526 p.

LEVKOV Z., 2009 — *Amphora* sensu lato. *In:* Lange-Bertalot H. (ed.), *Diatoms of the European inland waters and comparable habitats*, vol. 5. Ruggell, A.R.G. Gantner, 916 p.

LEWIN J.C., 1962 — Silicification. *In:* Lewin R.A. (ed.), *Physiology and biochemistry of algae.* New York and London, Academic Press, pp. 445-453.

LOWE S., BROWNE M., BOUDJELAS S. & DE POORTER M. (2000) — *100 of the world's worst invasive alien species a selection from the global invasive species database.* Auckland, New Zealand, The Invasive Species Specialist Group (ISSG), World Conservation Union (IUCN), 12 p.

MANN D.G., 1984 — Observations on copulation in *Navicula pupula* and *Amphora ovalis* in relation to the nature of diatom species. *Annals of botany* 54: 429-438.

MANN D.G., 1988 — Why didn't Lund see sex in Asterionella? A discussion of the diatom life cycle in nature. In: Round F.E. (ed.), *Algae and the aquatic environment*. Bristol, Biopress, pp. 383-412.

MANN D.G. & DROOP S.J.M., 1996 — Biodiversity, biogeography and conservation of diatoms. *Hydrobiologia* 336: 19-32.

MANN D.G., 1999 — The species concept in diatoms. *Phycologia* 38: 437-495.

MANN D.G. & CHEPURNOV V.A. 2004 — What have the Romans ever done for us? The past and future contribution of culture studies to diatom systematics. *Nova Hedwigia* 79: 237-291.

MANN D.G., DROOP S.J.M., HICKS Y.A., MARSHALL A.D., MARTIN R.R. & ROSIN P.L., 2006 — Alleviating the taxonomic impediment in diatoms: prospects for automatic and web- based identification systems. *In:* Witkowski A. (ed.), *Proceedings of the 18th International Diatom Symposium*. Bristol, Biopress, pp. 265-285.

MANN D.G. & EVANS K.M., 2007 — Molecular genetics and the neglected art of diatomics. *In:* Brodie J. & Lewis J. (eds), *Unravelling the algae – the past, present and future of algal systematics*. Boca Raton, Florida, CRC Press, pp. 231-265.

MANN D.G., THOMAS S.J. & EVANS K.M., 2008 — Revision of the diatom genus *Sellaphora*: a first account of the larger species in the British Isles. *Fottea* 8: 15-78.

MANN D.G., EVANS K.M., CHEPURNOV V.A. & NAGAI S., 2009 — Morphology and formal description of *Sellaphora bisexualis*, sp. nov. (Bacillariophyta). *Fottea* 9: 199-209.

MEDINGER R., NOLTE N., PANDEY R.V., JOST S., OTTENWÄLDER B., SCHLÖTTERER C. & BOENIGK J., 2010 — Diversity in a hidden world: potential and limitation of next-generation sequencing for surveys of molecular diversity of eukaryotic microorganisms. *Molecular ecology* 19 (Suppl. 1): 32-40.

MEDLIN L.K., SATO S., MANN D.G. & KOOISTRA W.H.C.F., 2008 — Molecular evidence confirms sister relationship of *Ardissonea*, *Climacosphenia* and *Toxarium* within the bipolar centric diatoms (Bacillariophyta, Mediophyceae) and cladistic analyses confirms that extremely elongated shape has arisen twice in the diatoms. *Journal of phycology* 44: 1340-1348.

MONIZ M.B.J. & KACZMARSKA I., 2009 — Barcoding diatoms: is there a good marker? *Molecular ecology resources* 9 (Suppl. 1): 65-74.

MONIZ M.B.J. & KACZMARSKA I., 2010 — Barcoding of diatoms: nuclear encoded ITS revisited. *Protist* 161: 7-34.

NOLTE V., PANDEY R.M., JOST S., MEDINGER R., OTTENWÄLDER B., BOENIGK J. & SCHLÖTTERER C., 2010 — Contrasting seasonal niche separation between rare and abundant taxa conceals the extent of protist diversity. *Molecular ecology* 19: 2908-2915.

NORBERG M., BIGLER C. & RENBERG I., 2010 — Comparing pre-industrial and post-limed diatom communities in Swedish lakes, with implications for defining realistic management targets. *Journal of paleolimnology* 44: 233-242.

PATRICK R. & REIMER C.W., 1966 — The diatoms of the United States, exclusive of Alaska and Hawaii, Volume 1. Fragilariaceae, Eunotiaceae, Achnanthaceae, Naviculaceae. *Academy of natural sciences of Philadelphia, Monograph* 13: 1-688.

PATRICK R. & REIMER C.W., 1975 — The diatoms of the United States, exclusive of Alaska and Hawaii, Volume 2, Part 1. Entomoneidaceae, Cymbellaceae, Gomphonemaceae, Epithemaceae. *Academy of natural sciences of Philadelphia, Monograph* 13: 1-213.

POULÍČKOVÁ A., ŠPAČKOVA J., KELLY M.G., DUCHOSLAV M. & MANN D.G., 2008 — Ecological variation within *Sellaphora* species complexes (Bacillariophyceae): specialists or generalists? *Hydrobiologia* 614: 373-386.

POULÍČKOVÁ A., VESELÁ J., NEUSTUPA J. & SKALOUD P., 2010 — Pseudocryptic diversity versus cosmopolitanism in diatoms: a case study on *Navicula cryptocephala* Kütz. (Bacillariophyceae) and morphologically similar taxa. *Protist* 161: 353-369.

QUIJANO-SCHEGGIA S.I., GARCÉS E., LUNDHOLM N., MOESTRUP Ø., ANDREE K. & CAMP J., 2009 — Morphology, physiology, molecular phylogeny and sexual compatibility of the cryptic *Pseudo-nitzschia delicatissima* complex (Bacillariophyta), including the description of *P. arenysensis* sp. nov. *Phycologia* 48: 492-509.

RAHEL F.J., 2007 — Biogeographic barriers, connectivity and homogenization of freshwater faunas: it's a small world after all. *Freshwater biology* 52: 696-710.

ROUND F.E., 1971 — Benthic marine diatoms. *Oceanography and marine biology annual review* 9: 83-139.

ROUND F.E., 1993 — A *Synedra* (Bacillariophyta) clone after several years in culture. *Nova Hedwigia, Beiheft* 106: 353-359.

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.

SHERWOOD A.R. & PRESTING G.G., 2007 — Universal primers amplify a 23S rDNA plastid marker in eukaryotic algae and Cyanobacteria. *Journal of phycology* 43: 605-608.

SHINNEMAN A.L.C., BENNETT D.M., FRITZ S.C., SCHMIEDER J., ENGSTROM D.R., EFTING A. & HOLZ J., 2010 — Inferring lake depth using diatom assemblages in the shallow, seasonally variable lakes of the Nebraska Sand Hills (USA): calibration, validation, and application of a 69-lake training set. *Journal of paleolimnology* 44: 443-464.

SMOL J.P. & STOERMER E.F., 2010 — *The diatoms: applications for the environmental and earth sciences*, Ed. 2. Cambridge, Cambridge University Press, 688 p.

SONNEBORN T.M., 1975 — The *Paramecium aurelia* complex of fourteen sibling species. *Transactions of the American microscopical society* 94: 155-178.

SORHANNUS U., ORTIZ J.D., WOLF M. & FOX M.G., 2010 — Microevolution and speciation in *Thalassiosira weissflogii* (Bacillariophyta). *Protist* 161: 237-249.

SOUFFREAU C., VANORMELINGEN P., VERLEYEN E., SABBE K. & VYVERMAN W., 2010 — Tolerance of benthic diatoms from temperate aquatic and terrestrial habitats to experimental desiccation and temperature stress. *Phycologia* 49: 309-324.

ŠPAČKOVA J., HAŠLER P., ŠTĚPÁNKOVÁ J. & POULÍČKOVÁ A., 2009 — Seasonal succession of epipelic algae: a case study on a mesotrophic pond in a temperate climate. *Fottea* 9: 121-133.

SPAULDING S.A. & ELWELL L., 2007 — *Increase in nuisance blooms and geographic expansion of the freshwater diatom Didymosphenia geminata.* U.S. Geological Survey Open-File Report 2007-1425, 38 p.

TEDERSOO L., NILSSON R.H., ABARENKOV K., JAIRUS T., SADAM A., SAAR I., BAHRAM M., BECHEM E., CHUYONG G. & KÕLJALG U., 2010 — 454 Pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. *New phytologist* 188: 291-301.

TELFORD R.J., VANDVIK V. & BIRKS H.J.B., 2006 — Dispersal limitations matter for microbial morphospecies. *Science* 312: 1015.

THERIOT E.C., CANNONE J.J., GUTELL R.R. & ALVERSON A.J., 2009 — The limits of nuclear encoded SSU rDNA for resolving the diatom phylogeny. *European Journal of phycology* 44: 277-290.

TROBAJO R., CLAVERO E., CHEPURNOV V.A., SABBE K., MANN D.G., ISHIHARA S. & COX E.J., 2009 — Morphological, genetic, and mating diversity within the widespread bioindicator *Nitzschia palea* (Bacillariophyta). *Phycologia* 48: 443-459.

TROBAJO R., CLAVERO E., EVANS K.M., VANORMELINGEN P., MCGREGOR R.C. & MANN D.G., In Press — The use of partial cox1, *rbc*L and LSU rDNA sequences for phylogenetics and species identification within the *Nitzschia palea* complex (Bacillariophyceae). *European journal of phycology.*

TROBAJO R., MANN D.G. & COX E.J., submitted — Studies of the type material of *Nitzschia abbreviata* (Bacillariophyta). *Nova Hedwigia, Beih.*

VANELSLANDER B., CRÉACH V., VANORMELINGEN P., ERNST A., CHEPURNOV V.A., SAHAN E., MUYZER G., STAL L.J., VYVERMAN W. & SABBE K., 2009 — Ecological differentiation between sympatric pseudocryptic species in the estuarine benthic diatom *Navicula phyllepta* (Bacillariophyceae). *Journal of phycology* 45: 1278-1289.

VANORMELINGEN P., CHEPURNOV V.A., MANN D.G., COUSIN S. & VYVERMAN W., 2007 — Congruence of morphological, reproductive and ITS rDNA sequence data in some Australasian *Eunotia bilunaris* (Bacillariophyta). *European Journal of Phycology* 42: 61-79.

VANORMELINGEN P., VERLEYEN E. & VYVERMAN W., 2008a — The diversity and distribution of diatoms: from cosmopolitanism to narrow endemism. *Biodiversity and Conservation* 17: 393-405.

VANORMELINGEN P., CHEPURNOV V.A., MANN D.G., SABBE K. & VYVERMAN W., 2008b — Genetic divergence and reproductive barriers among sympatric but morphologically heterogeneous sympatric clones of *Eunotia bilunaris sensu lato* (Bacillariophyta). *Protist* 159: 73-90.

VERLEYEN E., VYVERMAN W., STERKEN M., HODGSON D.A., DE WEVER A., JUGGINS S., VAN DE VIJVER B., JONES V.J., VANORMELINGEN P., ROBERTS D., FLOWER R., KILROY C., SOUFFREAU C. & SABBE K., 2009 — The importance of dispersal related and local factors in shaping the taxonomic structure of diatom metacommunities. *Oikos* 118: 1239-1249.

VYVERMAN W., VERLEYEN E., SABBE K., VANHOUTTE K., STERKEN M., HODGSON D.A., MANN D.G., JUGGINS S., VAN DE VIJVER B., JONES V., FLOWER R., ROBERTS D., CHEPURNOV V.A., KILROY C., VANORMELINGEN P. & DE WEVER A., 2007 — Historical processes constrain patterns in global diatom diversity. *Ecology* 88: 1924-1931.

WARD R.D., ZEMLAK 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*, B 360: 1847-1857.

WIKLUND J.A., BOZINOVSKI, N., HALL R.I. & WOLFE B.B., 2010 — Epiphytic diatoms as flood indicators. *Journal of paleolimnology* 44: 25-42.

WU L. THOMPSON D.K., LI G., HURT R.A., TIEDJE J.M. & ZHOU J. 2001 — Development and evaluation of functional gene arrays for detection of selected genes in the environment. *Applied and environmental microbiology* 67: 5780-5790.

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: 507512.