Current practice in the approach to species

Kirsten Hoef-Emden & David Bass

Until the 'molecular revolution' of the 1990s, morphological differences between protist lineages were detected by microscopy, and species definitions made on this basis. Inconsistencies in morphological species delimitations usually originate from a lack of awareness of evolutionary histories behind phenotypic characters, in the case of protists often in combination with a lack of resolution at the morphological level. The majority of species descriptions – especially from the era of light microscopy – have been based on individual cells observed in field material. Under such conditions, it is not possible to evaluate characters with respect to natural variability or stability in a clone of asexually reproducing protists or in a protistan biological species. Also unique (= synapomorphic) traits suited as diagnostic characters cannot be safely identified from among the noise of misleading homoplasious characters. As a consequence, important morphological characters have been ignored/overlooked, while others have been weighted differently depending in an author's interpretation of their value. A patchwork of text and figures of quite variable quality to deal with adds to the problems in assigning species names to protists.

The introduction of molecular sequence data (originally and still principally SSU rRNA gene sequences) in combination with morphological data quickly showed two things: 1) that genetic diversity within an apparently tightly-defined morphotype was much higher than expected (cryptic diversification), and 2) that some similar morphotypes actually belonged to a different part of the eukaryotic tree, sometimes whole phyla or supergroups away (convergent evolution). The latter case is quite simple to deal with, particularly as the previously apparently similar morphotypes usually turn out to be less similar when examined more carefully (and/or by investigating their ultrastructure) in the light of the molecular data. However, from a taxonomic point of view, cryptic diversification – especially at species level – can be more difficult to resolve.

Two cases in point are glissomonad and cercomonad flagellates (respectively orders Glissomonadida and Cercomonadida; phylum Cercozoa). Glissomonads are generally very small (~ 3-8 μm in length) and have few cellular characters that distinguish them from each other using standard light microscopy. However, by isolating many strains from a wide range of habitats, Howe et al. (2009, 2011) found sufficient differences in morphology, mode of movement, and SSU rDNA divergence to propose a total of twelve glissomonad genera, almost all of which can be identified on the basis of the first two sets of characters alone. Only four of these genera were previously recognized but at least two (*Heteromita* and *Bodomorpha*) were used very inconsistently and arbitrarily. A similarly large-scale study of cercomonads (Bass et al. 2009) also prompted a radical taxonomic revision. In some respects, cercomonads are generally easier to identify morphologically than glissomonads because as a whole they are more variable in size and types of pseudopodia and movement. However, many strains are so amoeboid and plastic that morphological boundaries between more closely related SSU rDNA types are blurred, even though the sequence divergence is comparable with species or even genus level differences in other groups. This is less of a problem when strains can be compared under standardized conditions over a relatively long period of time, but this cannot always be the case. In cercomonads, therefore, SSU rDNA sequences are particularly useful in distinguishing species. However, other experiments (Bass et al. 2007) showed that the SSU rDNA does not have sufficient evolutionary resolution to separate ecologically/ behaviourally (and to some extent morphologically) distinct lineages that deserve species-level distinction. In this study, robust differences between strains with identical 18S but clearly distinct ITS1 rDNA sequences were found in characters such as salinity tolerance, propensity to form cysts, growth rate, and behavioural characteristics such as movement, cell clustering, etc. This indicates that ITS1/2 rDNA might be better markers for species-level differences in this group.

In cryptophytes, intensive work has been carried out to overcome problems arising from the use of phenotypic characters (morphology, physiology, behaviour, pigments, etc.) outside of a robust evolutionary context. Such an approach requires examining a larger number of clonal cultures by a combination of traditional morphology-based methods with molecular phylogenetic analyses. Clonal cultures provide important

advantages over field material. Sufficient numbers of cells are available to investigate phenotypic characters at a larger scale (DIC/fluorescence light microscopy of fixed and/or live cells, transmission/scanning electron microscopy, physiology) and to extract DNA several times for detailed molecular characterisation. Quantitative characters can be repeatedly measured to compute mean values, standard abbreviations, maxima and minima of cell dimensions. Clonal cultures, thus, allow for identification of variable, potentially uninformative, as well as of stable, reliable, characters in an exponentially growing culture. They may provide the only means to identify and study heteromorphic life histories, and the influence of environmental factors on phenotypic characters can be tested under defined laboratory conditions (nutrient depletion, different culture media or food organisms, temperature, light intensity, wave lengths of light, mating experiments, presence/absence of predators) (Komárek 1964; Luo et al. 2006; Pringsheim 1968; Van Donk et al. 2011, and references therein).

These experiments, however, will not yield information about the suitability of identified stable traits as diagnostic characters (Pringsheim 1968). Identification of potentially species-specific morphological synapomorphies is possible only by mapping the characters onto a highly resolved phylogenetic tree so that they can be understood in an evolutionary context (Hoef-Emden 2008, 2007; Hoef-Emden & Melkonian 2003).

Prospects of ribosomal gene phylogenies

Despite of the increasing use of multi-protein or phylogenomic data sets for inferring deep phylogenetic trees, the eukaryotic ribosomal operon still serves as a versatile tool for phylogenies of lower rank taxa (e.g. Amato et al. 2007; Edvardsen et al. 2003; Scorzetti et al. 2002). In contrast to protein genes, no evidence for lateral gene transfer of ribosomal rRNA genes has been reported to date. Highly resolved phylogenetic trees and species delimitation, however, require different molecular tool sets. The ribosomal operon offers solutions to both analytic approaches, with three conserved rRNA genes (SSU, 5.8S and LSU rDNA) that can be concatenated to a multi-gene alignment to increase resolution, and with highly variable regions (internal transcribed spacers 1 and 2 [ITS1, ITS2] and 5' terminus of LSU rDNA) potentially suited for species delimitation.

Whereas the conserved rRNA genes can be aligned across all eukaryotes, ITS regions are often too variable to be aligned across genera in a group.

By using the inferred molecular phylogeny as a guide to trace evolutionary pathways, morphological traits of the respective cultures can be evaluated. In a completely resolved phylogeny without severe artifacts (i.e. using likelihood-based methods with appropriate evolutionary models), synapomorphic morphological features will be grouped monophyletically, plesiomorphic, i.e. ancestral characters will be para- or polyphyletic and characters originating from convergent evolution will be unveiled by polyphyly.

Mating experiments in green algae of the order Volvocales (Chlorophyceae) demonstrated a correlation between reproductive barriers and compensatory base changes (CBCs) in conserved positions of the secondary structure of the internal transcribed spacer 2 (ITS2) (Coleman 2000, and references therein). In several genera belonging to diverse evolutionary lineages (predominantly diatoms and brown algae, but also some embryophytes, animals, fungi) evidence for such a correlation could be found (summarised in Coleman 2009). The ITS2 RNA shares a common secondary structure across eukaryotic lineages (Schultz et al. 2005). Prior to excision from the primary transcript, it folds up to a hand-like structure with – in most cases – four helical domains. The 3' terminus of the 5.8S rRNA and the 5' terminus of the LSU rRNA pair to form a stem (Côté et al. 2002).Most conserved parts in terms of primary sequence are the proximal part of helix II and helix III, the latter containing the important C2 cleavage site (Côté et al. 2002; Henras et al. 2008). The crossing experiments have shown that a CBC in one of these conserved domains correlated with loss of sexual compatibility between evolutionary lineages.

Therefore, it has been proposed to establish a consistent species concept by defining clades containing no CBCs in the conserved parts of ITS2 secondary structure as species (Coleman 2000). Terminal clades separated by CBCs from each other would then belong to two different species. Considering the currently available data, ITS2 apparently evolves at a higher rate than mating genes in the tested lineages and, thus, seemed to be a safe predictor of interbreeding incompatibility. One has to be aware of, though, that a)

ITS2 is not a mating gene, the correlation between reproductive barriers and ITS2 secondary structure is most likely by chance and b) CBCs – thus the term CBC "clade" is inappropriate – do not necessarily occur in terminal clades only. It cannot be ruled out that a clade with even identical ITS2 sequences may encompass more than one biological species and in some lineages with extremely low evolutionary rates, CBCs in the ITS2 may resolve only at higher classification ranks, e.g. genera or families (Vrålstad 2011). Nevertheless, in sexually reproducing lineages with known inductors, a correlation of ITS2 secondary structure with reproductive barriers can be tested (Amato et al. 2007; Behnke et al. 2004). In evolutionary lineages with suspected sexual reproduction, the CBC "clade" concept can be used to approximate biological species and in asexually reproducing lineages, molecular approaches are the only option (Hoef-Emden 2007; Luo et al. 2006). In all three cases, applying the CBC concept to a morphologically and phylogenetically well-studied group of organisms will result in a consistent and reproducible species concept congruent with the phylogeny of organisms. If future mating experiments prove that ITS2 did not resolve down to the level of biological species in an organismic group, this finding probably will not again stir up the established systematics, but will only result in a refinement of the existing system.

Barcoding protists: the COI region

Another means proposed for an identification of species are DNA barcodes, short highly variable DNA regions (Hebert et al. 2003). DNA barcoding became a common practise for fast identification of animal species and has been adopted for other eukaryotic lineages (Hajibabaei et al. 2007). Hebert et al. (2003) hypothesized that the COI region comprising around 500-800 nt of the 5' terminus of the mitochondrial gene *cox*1 would act as a useful DNA barcode marker. *Cox*1 codes for the small subunit 1 of cytochrome c oxidase, the enzyme constituting complex IV of the mitochondrial respiratory chain. Studies addressing the identification of animal species by COI sequences have often been exclusively based on comparisons of morphological species with distance-based clustering trees inferred from COI sequences with one individual representing a species. Hebert et al. (2003) proposed COI as a DNA barcode marker for several reasons. They pointed out multiple possibilities to be misled by heteromorphic life cycles, sexual dimorphisms or plasticity of morphological characters and emphasised the possibility to

(a) identify known species rapidly without the need of expertise from a "dwindling pool of taxonomists" and (b) to find new species.

Both potential tools for species delimitation, ITS2 and COI, have pitfalls to address. Apart from potentially too low mutation rates to resolve biological species in some evolutionary lineages, ITS sequences (as well as the flanking rRNA genes) are notorious for indels and also intragenomic variation has been reported frequently (Stage & Eickbush 2007; Thornhill et al. 2007). Alignment of ITS2 sequences and prediction of their secondary structure can be a challenging task in groups with high evolutionary rates and/or long ITS2 sequences Gillespie 2004; Letsch & Kjer 2011). ITS2 seems to be a good tool to approximate biological species, but it definitely cannot be used as a molecular "Swiss army knife". Gapped and/or non-alignable regions prohibit the computation of genetic distances as currently done in DNA barcoding with COI and in biodiversity surveys based on sequencing of environmental DNA without secondary structure prediction, intragenomic variation will result in erroneously increased species counts (Thornhill et al. 2007).

Some points of criticism of DNA barcoding with COI address current practice

In DNA barcoding, species are supposed to be identified by setting a threshold of genetic divergence in percent as a species delimiter. This approach assumes the presence of a barcode gap between frequency distributions of intra- and interspecific genetic distances (Meyer and Paulay 2005). Studies have shown, however, that COI performs poorly, if applied in suboptimally sampled and systematically insufficiently characterised groups (Meyer and Paulay 2005). Given the heterogeneous mutation rates, almost always present across ingroup taxa in a phylogenetic analysis, and also present in mitochondrial genomes, it can be predicted, that the notion of a reliable barcode gap is an illusion (Hendrich et al. 2010; Rubinoff et al. 2006). DNA barcoding based on genetic distances will likely fail also in well-characterised and well-sampled groups.

Accuracy of DNA barcoding with COI may also be impaired by defective copies of the *cox*1 gene (pseudogenes) and by *cox*1 fragments exported to the nuclear genome (numts) (Rubinoff et al. 2006). In both cases, species counts in environmental DNA projects will be biased (Thornhill et al. 2007). More severe points of critics have been

raised against the choice of COI in general as a tool for species delimitation (Rubinoff et al. 2006). Mitochondrial genomes are subject to different selective pressures than nuclear genomes and show some specifics that may result in misleading branching patterns, such as maternal inheritance, heteroplasmy, introgression, absence of recombination or genetic bottlenecks.

The potential of COI as a tool to delimit species can only be assessed if a highly resolving phylogenetic tree inferred from other DNA sequences is used as a guidance tree to detect anomalies and if in addition, several clonal cultures/individuals are sequenced as representatives of a putative species. In some protist groups, COI may turn out to be a suitable tool to identify species, provided species delimitation uses **character-based** rather than distance-based methods and appropriate evolutionary models (Evans et al. 2007). Due to heterogeneous mutation rates, however, a safe identification of new species using short DNA barcodes alone without any other data at hand is unlikely. In addition, genes from organellar genomes cannot be applied to all evolutionary lineages (e.g. amitochondriate or heterotrophic taxa).

Many arguments against the use of either ITS sequences or COI apply more or less to all molecular markers, no matter whether chosen from nuclear, micro- or macronuclear, nucleomorph, mitochondrial or plastid genomes: intragenomic variation, orthologous and paralogous copies of genes, pseudogenes, unequal mutation rates, lateral gene transfer, introgression. Heterogeneous evolutionary rates, e.g., have been found in most genes/DNA regions from different genomes (Gillman et al. 2010; Hendrich et al. 2010; Hoef-Emden et al. 2005; Lopez et al. 2002). This implies that a safe identification of new species using a genetic threshold as a delimiter likely will fail not only in COI, but also with other potential barcode markers. A constant threshold across lineages, however, is a prerequisite for the use of distance-based automated identification procedures as proposed e.g. for ITS2 (Leliaert et al. 2009; Müller et al. 2007). We will not find the perfect molecular marker to distinguish species and definitely not across all lineages of the eukaryotic tree of life.

Environmental molecular data and its relevance to taxonomy

As molecular surveys have revealed very high levels of cryptic diversity in the vast majority of known protist taxa, a fundamental question arises: does this (micro)variation have any taxonomy interest, or does it represent (to some extent) neutral genetic variation within larger, phenotypically defined taxonomic units. We argue that an untested assumption that microvariation is biologically irrelevant or taxonomically uninformative is insufficient; the discovery of this diversity should be an incentive to study the organismal differences more closely to tease apart what phenotypic differences there may be and the ecological preferences that define their core niche. In any case, because of ancestral polymorphism, hybridization, introgression, and horizontal transfer, molecular characters are not inherently better than morphological characters; it is more a question of degree. A more scientifically transparent and informative approach would be to adopt a testable, but skeptical stance that the smaller differences between molecular markers revealed by cell-independent environmental studies is redundant - the null hypothesis - which could be rejected after biological studies and/or studies involving other loci. Under the neutral theory, the diversity expected to be theta = 4Nu, where u is the mutation rate, and N the effective size. For single-celled organisms, the population size can often be very large, and we therefore expect a huge diversity of effectively neutral variation. Furthermore, coalescence theory for clonal genomes indicates that this diversity will often be clustered into deep and shallow branches.

We should also bear in mind a psychological influence on our previous understanding of these 'lumped' lineages: if there is only one identity (name) associated with them, variations within the definition of the taxon will be seen as transitory differences, or intra-taxon unimportant variations. While we to no extent deny the reality of such phenomena, it may be that some of these differences represent biologically distinct entities within a pool of others. To avoid overlooking these elements of microbial biodiversity we believe the scientifically responsible approach is to hypothesise that the genetic differences correlate with phenotypic and/or ecological differences until shown otherwise. The glissomonad work described above illustrates this point well. For over 150 years, the literature has referred to a very small number of inconsistently

misidentified glissomonads, making most of this work useless for understanding the ecology of these organisms based on that literature alone.

Awareness of high intra-morphospecies genetic diversity has come not only from sequencing cultured or cell isolates, but also from environmental clone libraries, and more recently 454 and Illumina library sequencing. Because this environmental sequencing approach a) lacks the observer and culturing bias, b) can screen more environmental material in less time, and c) is much less labour-intensive per sequence generated than the cell-based approach, in general much more diversity has been revealed. This presents a further taxonomic challenge as sequences are obtained without any phenotypic data at all. Where environmental sequences cluster with known sequence types, we can infer something about the diversity of that known group, but an environmental sequence does not have to be very divergent from a characterized sequence before it becomes taxonomically uninformative. Another interesting aspect of clone library studies is that, in some cases where they have been paired with cell-based (cultures isolation, cell screening, etc.) the sets of lineages recovered from each approach have shown little or no overlap. This highlights the extent to which laboratory isolation conditions can be very highly selective. The most commonly isolated strains are not necessarily the most common in the sites sampled. The number of possible causes of this is large, and a source of important ecological information: differences in tolerances to nutrient levels, temperature, other aspects of the abiotic environment, and the biotic environment, including the possibility that the lineages missed by the cell-based approaches may require a particular interaction with other organisms in the original sample that is disrupted by the isolation process, or is intrinsically difficult to separate, e.g. a symbiotic or parasitic relationship.

In some respects, the recent advent of massively highly parallel sequencing (454, Illumina, and others) here referred to collectively (if unsatisfactorily) as next generation sequencing (NGS) has made this problem even worse. Millions of marker gene sequences from multiple libraries can now be generated extremely easily and quickly. It is also increasingly clear that these sequences can carry high error loads, for example from mis-readings of homopolymer regions, chimaera formation, and amplification of levels of PCR errors and intragenomic variation, all of which will artefactually inflate diversity

estimates (e.g. Pandey et al. 2011; Medinger et al. 2010; Quince et al. 2009; Stoeck et al. 2010). Some NGS technologies are currently more prone to sequencing errors than others, but some sources of error, such as PCR chimera formation and intragenomic sequence variation are more difficult to control. On the other hand, this huge amount of data, allied to the fact that individual samples will be sequenced much closer to saturation than clone libraries could ever manage, means that larger-scale algorithms can be applied to the data to statistically test the robustness of different lineages. Furthermore, high levels of sequencing coverage can help eliminate individual sequence errors, where reliable references sequences are available. Where environmental variables are available for the environmental samples being analysed, these can be incorporated into multi-variate analyses along with the sequence data, and sets of sequences with significantly different clustering in environmental parameter space can be tested for. If found, these could be used as a basis for defining genetic ecotypes, which might turn out to be the most valid and useful taxonomic units for understanding otherwise intractable protist diversity. As ecotypes are almost always going to be more highly resolved in evolutionary terms than morphospecies, and have the advantage that their initial detection depends on measurable and informative biological differences, NGS may offer an unexpectedly useful tool for taxonomists from the perspective of cell-independent experiments. One way in which an indication of potential functionally relevant taxonomic boundaries could be indicated using NGS data is where deeply sequenced PCR amplicon datasets derive from multiple samples with co-measured environmental variables available for each sampling point. Assuming sufficient phylogenetic signal is present in the sequenced fragment, the environmental variables relating to each data point (sequence) could be used to cluster sequences into potential ecotypes. Thus, an adaptation of the approach pioneered by Fonataneto et al. (2007) for detecting independently evolving entities and adaptive divergence in long-term asexual bdelloid rotifers could be used, substituting morphological measurements with environmental data. This could be particularly useful for (the assumed majority of) asexual protist taxa, but would also be valuable for sexuals. It is worth noting that for both clone library and NGS work so far, most studies have focused on SSU rDNA sequences. This has been a sensible strategy for many reasons, including good database representation of SSU sequences, the potential for

building reasonable phylogenies based on that gene, and a good choice and availability of PCR primer sites, among others. However, if as we suggest above, the ITS regions offer better species-level markers then the SSU approach may only be providing a relatively crude overview (perhaps with self-cancelling internal inconsistencies among taxa) of protist diversity distribution. Nonetheless, for most groups (notably excepting fungi) there are too many technical/conceptual difficulties involved in large-scale ITS environmental sequencing studies (and here is not the place to discuss them) to make them generally feasible at the moment. However, for some better characterized, well-studied, and tightly phylogenetically defined group, ITS environmental sequencing may provide bases for much more realistic ecological models than those currently undertaken.

As sequencing technologies continue to advance, making large-scale analyses (environmental, genomic, transcriptomic) faster and more affordable, the ways in which adaptively relevant differences between lineages can be investigated will become more numerous and diverse. On the one hand, the high levels of sampling made possible by NGS technologies open up much potential for multivariate environmental correlation analyses, as suggested above for large-scale, multi-sample amplicon datasets. Another approach is to use NGS to investigate differences between putative species at the level of whole genomes and/or transcriptomes. This can be done by genome-wide sequencing of individual isolates and cells, or, as NGS read lengths increase in length, but obtaining contiguous multilocus sequences from environmental samples. Logares et al. (2009) showed the potential value of such approaches with a genomic fingerprinting (AFLP) analysis of strains within recognized dinoflagellate species, revealing a large diversity genotypically distinct, sub-species level lineages that showed by phylogenetic and ecological distinctiveness, even in sympatry. Sequence-based approaches at this level have so far been mostly used for prokaryotes (e.g. Holt et al. 2008), but the idea that functionally-linked, integrated metagenomic approaches are/will be important for meaningful biodiversity assessment, including species recognition, is now much more realistic than it was even a few years ago, and is cogently argued by Bittner et al. (2010).

References

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.

Bass, D., Richards, T.A., Matthai, L., Marsh, V., Cavalier-Smith, T., 2007. DNA evidence for global dispersal and probable endemicity of protozoa. *BMC Evolutionary Biology* 7, 162.

Bass, D., Howe, A.T., Mylnikov, A.P., Vickerman, K., Chao, E.E., Edwards Smallbone, J., Snell, J., Cabral, Jr. C., Cavalier-Smith, T., 2009. Phylogeny and classification of Cercomonadida: *Cercomonas*, *Eocercomonas*, *Paracercomonas*, and *Cavernomonas* gen. n. *Protist* 160, 483-521.

Behnke. A., Friedl, T., Chepurnov, V.A., Mann, D.G., 2004. Reproductive compatibility and rDNA sequence analyses in the *Sellaphora pupula* species complex (Bacillariophyta). J. Phycol. 40, 193-208.

Bittner, L., Halary, S., Payri, C., Cruaud, C., de Reviers, B., Lopez, P., Bapteste, E., 2010. Some considerations for analysing biodiversity using integrative metagenomics and gene networks. Biol. Dir. 5, 47.

Coleman, A.W., 2000. The significance of a coincidence between evolutionary landmarks found in mating affinity and a DNA sequence. Protist 151, 1-9.

Coleman, A.W., 2009 Is there a molecular key to the level of "biological species" in eukaryotes? A DNA guide. Mol. Phylogenet. Evol. 50, 197-203.

Côté, C.A., Greer, C.L., Peculis, B.A., 2002 Dynamic conformational model for the role of ITS2 in pre-RNA processing in yeast. RNA 8, 786-797.

Edvardsen, B., Shalchian-Tabrizi, K., Jakobsen, K.S., Medlin, L.K., Dahl, E., Brubak, S., Paasche, E., 2003. Genetic variability and molecular phylogeny of *Dinophysis* species (Dinophyceae) from Norwegian waters inferred from single cell analyses of rDNA. J. Phycol. 39, 395-408.

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

Fontaneto, D., Herniou, E.A., Boschetti, C., Caprioli, M., Melone, G., Ricci, C., Barraclough, T.G., 2007. Independently evolving species in asexual bdelloid rotifers. PLoS Biology 5, e87.

Gillespie, J.J., 2004. Characterizing regions of ambiguous alignment caused by the expansion and contraction of hairpin-stem loops in ribosomal RNA molecules. Mol. Phylogenet. Evol. 33, 936-943.

Gillman, L.N., Keeling, D.J., Gardner, R.C., Wright, S.D., 2010. Faster evolution of highly conserved DNA in tropical plants. J. Evol. Biol. 23, 1327-1330.

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 Genet. 23, 167-172.

Hebert, P.D.N., Cywinska, A., Ball, S.L., deWaard, J., 2003. Biological identifications through DNA barcodes. Proc. R. Soc. Lond. B 270, 313-321

Hendrich, L., Pons, J., Ribera, I., Balke, M., 2010. Mitochondrial *cox*1 data reliably uncover patterns of insect diversity but suffer from high lineage-specific idiosynctratic error rates. PloS One 5, e14448.

Henras, A.K., Soudet, J., Gérus, M., Lebaron, S., Caizergues-Ferrer, M., Mougin, A., Henry, Y., 2008. The post-transcriptional steps of eukaryotic ribosomal biogenesis. Cell. Mol. Life Sci. 65, 2334-2359.

Hoef-Emden, K., 2007. Revision of the genus *Cryptomonas* (Cryptophyceae) II: Incongruences between the classical morphospecies concept and molecular phylogeny in smaller pyrenoid-less cells. Phycologia 46, 402-428.

Hoef-Emden, K., 2008. Molecular phylogeny of the phycocyanin-containing cryptophytes: Evolution of biliproteins and geographical distribution. J. Phycol. 44, 985-993.

Hoef-Emden, K., Melkonian, M., 2003. Revision of the genus *Cryptomonas* (Cryptophyceae): a combination of molecular phylogeny and morphology provides insights into a long-hidden dimorphism. Protist 154, 371-409. Corrigendum: Hoef-Emden, K., Melkonian, M., 2008. Protist 159, 507.

Hoef-Emden, K., Tran, H-D., Melkonian, M., 2005. Lineage-specific variations of congruent evolution among DNA sequences from three genomes, and relaxed selective constraints on *rbc*L in *Cryptomonas* (Cryptophyceae). BMC Evol. Biol. 5, 56.

Holt, K.E., Parkhill, J., Mazzoni, C.J., Roumagnac, P., Weill, F-X., Goodhead, I., Rance, R., Baker, S., Maskell, D.J., Wain, J., Dolecek, C., Achtmann, M., Dougan, G., 2008. High-throughput sequencing provides insights into genome variation and evolution in *Salmonella typhi*. Nat. Gen. 40, 987-993.

Howe, A.T., Bass, D., Vickerman, K., Chao, E.E., Cavalier-Smith, T., 2009. Phylogeny, taxonomy, and astounding genetic diversity of Glissomonadida ord. nov., the dominant gliding zooflagellates in soil (Protozoa: Cercozoa). Protist 160, 159-189.

Howe, A.H., Bass, D., Chao, E.E., Cavalier-Smith, T., 2011. New genera, species, and improved phylogeny of Glissomonadida (Cercozoa). Protist, in press.

Komárek, J., 1964. Utility of synchronized algal cultures for experimental taxonomy. Plant Cell. Physiol. 5, 385-391.

Leliaert, F., Verbruggen, H., Wysor, B., de Clerck, O., 2009. DNA taxonomy in morphologically plastic taxa: Algorithmic species delimitation in the *Boodlea* complex (Chlorophyta: Cladophorales). Mol. Phylogenet. Evol. 53, 122-133.

Letsch, H.O., Kjer, K.M., 2011. Potential pitfalls of modelling ribosomal RNA data in phylogenetic tree reconstruction: Evidence from case studies in the Metazoa. BMC Evol. Biol. 11, 146.

Logares, R., Boltovskoy, A., Bensch, S., Laybourn-Parry, J., Rengefors, K., 2009. Genetic diversity patterns in five protist species occurring in lakes. Protist 160, 301-317.López, P., Casane, D., Philippe, H., 2002. Heterotachy, an important process of protein evolution. Mol. Biol. Evol. 19, 1-7.

Luo, W., Pflugmacher, S., Pröschold, T., Walz, N., Krienitz, L., 2006. Genotype versus phenotype variability in *Chlorella* and *Micractinium* (Chlorophyta, Trebouxiophyceae). Protist 157, 315-333.

Medinger, R., Nolte, V., Vinay Pandey, R., 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. Mol. Ecol. 19, 32-40.

Meyer, C.P., Paulay, G., 2005. DNA barcoding: error rates based on comprehensive sampling. PloS Biol. 3, e422.

Müller, T., Philippi, N., Dandekar, T., Schultz, J., Wolf, M., 2007. Distinguishing species. RNA 13, 1469-1472.

Pandey, R.V., Nolte, V., Boenigk, J., Schlötterer, C., 2011. CANGS DB: a stand-alone web-based database tool for processing, managing and analyzinf 454 data in biodiversity studies. BMC Research Notes 4, 227.

Pringsheim, E.G., 1968. Zur Kenntnis der Cryptomonaden des Süßwassers. Nova Hedwigia 16, 367-401.

Quince, C., Lanzen, A., Curtis, T.P., Davenport, R.J., Hall, N., Head, I.M., Read, L.F., Sloan, T., 2009. Accurate determination of microbial diversity from 454 pyrosequencing data. Nature Met. 6, 639-641.

Rubinoff, D., Cameron, S., Kipling, W., 2006. A genomic perspective on the shortcomings of mitochondrial DNA for „barcoding“ identification. J. Hered. 97, 581-594.

Schultz, J., Maisel, S., Gerlach, D., Müller, T., Wolf, M., 2005. A common core of secondary structure of the internal transcribed spacer 2 (ITS2) throughout the Eukaryota. RNA 11, 361-364.

Scorzetti, G., Fell, J.W., Fonseca, A., Statzell-Tallman, A., 2002. Systematics of basidiomycetous yeasts: a comparison of large subunit D1/D2 and internal transcribed spacer rDNA regions. FEMS Yeast Research 2, 495-517.

Stage, D.E., Eickbush, T.H., 2007. Sequence variation within the rRNA gene loci of 12 *Drosophila* species. Genome Res. 17, 1888-1897.

Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M.D.M., Breiner, H.-W., Richards, T.A., 2010. Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. Mol. Ecol. 19, 21-31.

Thornhill, D.J., Lajeunesse, T.C., Santos, S.R., 2007. Measuring rDNA diversity in eukaryotic microbial systems: how intragenomic variation, pseudogenes, and PCR artifacts confound biodiversity estimates. Mol. Ecol. 16, 5326-5340.

Van Donk, E., Ianora, A., Vos, M., 2011. Induced defenses in marine and freshwater phytoplankton: a review. Hydrobiologia 668, 3-19.

Vrålstad, T., 2011. ITS, OTUs and beyond – fungal hyperdiversity calls for supplementary solutions. Mol. Ecol. 20, 2873-287.