Reference: *Biol. Bull.* **227:** 93–106. (October 2014)

Next-Generation Environmental Diversity Surveys of Foraminifera: Preparing the Future

J. PAWLOWSKI[1,*], F. LEJZEROWICZ[1], AND P. ESLING[2]

[1]*Department of Genetics and Evolution, University of Geneva, Switzerland; and* [2]*IRCAM, UMR 9912, Université Pierre et Marie Curie, Paris, France*

Abstract. Foraminifera are commonly defined as marine testate protists, and their diversity is mainly assessed on the basis of the morphology of their agglutinated or mineralized tests. Diversity surveys based on environmental DNA (eDNA) have dramatically changed this view by revealing an unexpected diversity of naked and organic-walled lineages as well as detecting foraminiferal lineages in soil and freshwater environments. Moreover, single-cell analyses have allowed discrimination among genetically distinctive types within almost every described morphospecies. In view of these studies, the foraminiferal diversity appeared to be largely underestimated, but its accurate estimation was impeded by the low speed and coverage of a cloning-based eDNA approach. With the advent of high-throughput sequencing (HTS) technologies, these limitations disappeared in favor of exhaustive descriptions of foraminiferal diversity in numerous samples. Yet, the biases and errors identified in early HTS studies raised some questions about the accuracy of HTS data and their biological interpretation. Among the most controversial issues affecting the reliability of HTS diversity estimates are (1) the impact of technical and biological biases, (2) the sensitivity and specificity of taxonomic sequence assignment, (3) the ability to distinguish rare species, and (4) the quantitative interpretation of HTS data. Here, we document the lessons learned from previous HTS surveys and present the current advances and applications focusing on foraminiferal eDNA. We discuss the problems associated with HTS approaches and predict the future trends and avenues that hold promises for surveying foraminiferal diversity accurately and efficiently.

Received 11 April 2014; accepted 8 July 2014.

* To whom correspondence should be addressed. E-mail: jan.pawlowski@unige.ch

Abbreviations: aDNA, ancient DNA; eDNA, environmental DNA; HTS, high-throughput sequencing; NGS, next-generation sequencing; OTU, operational taxonomic unit; SSU, small subunit.

Introduction

During the last two decades there have been tremendous changes in our understanding of the diversity and evolutionary history of microbial eukaryotes. In the early stage, the development of molecular systematics led to the discovery of cryptic diversity in most protist phyla and completely altered their classification. More recently, the advances in next-generation sequencing (NGS) technologies applied to environmental surveys revealed a large number of novel micro-eukaryotic lineages known only by their DNA sequences. Here, we trace the changes in diversity assessment from morphological to genetic and environmental genomic perspectives, using the example of foraminifera, the group of protists best known for their fossil record, which has been shown to comprise an unexpected richness of non-fossilized species.

Morphological diversity

The morphology of foraminiferal tests (shells) is the basic feature used to assess foraminiferal diversity. The test can be soft or hard, composed of a single or multiple chambers, bearing one or multiple apertures. The test wall can be formed of organic lining, sometimes covered with agglutinated particles gathered from the surrounding sediments, or it can be composed of secreted calcium carbonate crystals of different mineral compositions and orientation. The characteristics of the wall represent key diagnostic features for the identification of foraminiferal orders, while the form of the test, its internal structure, and the position of the apertures are commonly used to distinguish lower level taxa. Species are usually characterized by the size of the test, the number

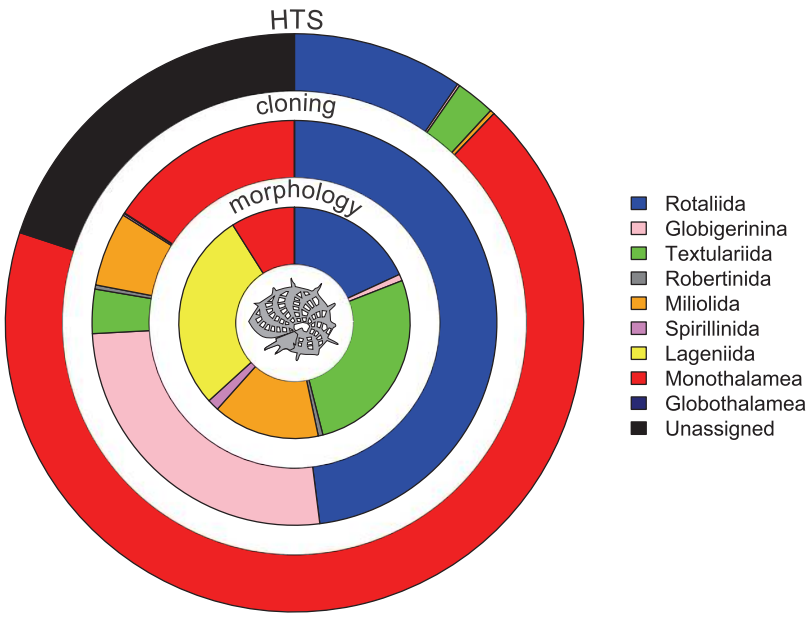

Figure 1. Comparison of the relative proportions of the major foraminiferal taxa found using morphospecies number (inner circle), cloning and Sanger-sequencing (middle circle), and high-throughput sequencing (outer circle). The morphospecies counts are based on the foraminiferal entries to the World Register of Marine Species (WoRMS) database. The Sanger data are based on single-cell sequences available in the GenBank and our database. The HTS data correspond to a concatenation of 8 Illumina sequencing runs gathering 26,135,707 sequences.

and form of their chambers, or their ornamentation design (Sen Gupta, 1999).

There are 6705 morphospecies of extant foraminifera, according to the latest edition of the World Register of Marine Species (WoRMS Editorial Board, 2014). In addition, the micropaleontologists recognize about 40,000 fossil species preserved in the geological record (Buzas and Culver, 1991; Murray, 2007). Traditionally, the foraminifera are divided into 8–12 orders (Loeblich and Tappan, 1988), which have been recently placed into 3 or 4 major classes (Pawlowski *et al*., 2013; Mikhalevich, 2013). The most diversified of these orders are the calcareous and agglutinated multi-chambered benthic Rotaliida, Miliolida, Lagenida, and Textulariida, each comprising more than 1000 species. The planktonic order Globigerinida and the benthic orders Robertinida and Spirillinida, as well as the paraphyletic assemblage of monothalamous (single-chambered) taxa are represented by relatively few species (Fig. 1).

Although the morphological diversity of foraminifera is widely used to characterize past and present environments (Murray, 2006), the reliability of morphological features alone for species distinction is questionable. The morphological definition of species differs between "splitters" and "lumpers." The majority of conflicts stem from the morphological plasticity of foraminiferal tests and the lack of clear distinction between ecophenotypes and morphospecies. Experimental studies of eco-morphological variations are rare due to the difficulties in cultivation of foraminiferal species (Schnitker, 1974). The lack of easily cultivable species also considerably limits our knowledge about the reproduction, life cycle, and cell biology of foraminifera. Unsurprisingly, most of the foraminiferal diversity studies are based on analyses of dead assemblages using morphology as the unique feature for species distinction (Buzas and Culver, 1991; Hayward *et al*., 1999, 2010; Murray, 2006; Debenay, 2012).

Molecular revolution

Since the nineties, the traditional view of foraminiferal diversity has been challenged by molecular studies that consistently demonstrated the pitfalls and limitations of morphotaxonomy. As for most protistan phyla, the molecular data revealed that practically every foraminiferal morphospecies could be subdivided into several genetically distinctive phylotypes. The coherence of this subdivision was supported by further detailed studies showing that some of these cryptic species may have restricted geographic distribution, occupying separate ecological niches, and/or possessing distinctive morphological features (de Vargas *et al*., 1999; Hayward *et al*., 2004; Darling *et al*., 2007; Morard *et al*., 2009). The diversity of some taxonomic groups, such as the extensively studied planktonic foraminifera, composed of 50 morphospecies, has been multiplied by 4 or

5 (Darling and Wade, 2008). Similarly, high genetic diversity was observed in common shallow-water benthic foraminifera—for example, in genera *Ammonia* (Hayward *et al*., 2004) and *Elphidium* (Pillet *et al*., 2013). By contrast, the abyssal benthic species showed an astonishingly low level of genetic variation across the world oceans (Pawlowski *et al*., 2007).

The most spectacular finding of early molecular studies was the high diversity of the single-chambered species belonging to the paraphyletic class "Monothalamea" (Pawlowski *et al*., 2013). Because their organic or agglutinated tests are not well preserved, the monothalamids were largely ignored by micropaleontologically oriented foraminiferologists. New monothalamid species have traditionally been described by protistologists and marine biologists, but their diversity remained unexplored due to the paucity of distinctive morphological characters and the limited interest of micropaleontologists. The evolutionary importance of the group has been highlighted by molecular evidence that the large radiation of monothalamids by far predates the first appearance of testate multi-chambered foraminifera in Cambrian sediments (Pawlowski *et al*., 2003). As could be expected, every monothalamid genus also comprises several genetically distinctive species that often demonstrate geographically restricted ranges (Pawlowski *et al*., 2002, 2008). It has rapidly become evident that the diversity of this group is much higher that suggested by morphological studies (Fig. 1).

Molecular analyses of foraminiferal diversity are mainly based on nuclear ribosomal genes, because their evolution rates are faster than in other eukaryotes (Pawlowski *et al*., 1997) and because their recovery from single-cell extracts is relatively easy (Pawlowski *et al*., 2000). The fragment located at the 3′end of the small subunit rDNA (SSU/18S), is commonly accepted as the standard foraminiferal DNA barcode (Pawlowski *et al*., 2012; Pawlowski and Holzmann, 2014). This fragment is interspersed by six hypervariable regions entailing a phylogenetic signal able to resolve relationships within foraminiferal clades and down to the species level (Pawlowski and Lecroq, 2010). One of these hypervariable regions, the helix 37f, has been shown to be particularly variable and was chosen as the ideal mini-barcode (length 30–60 nucleotides) for the high-throughput sequencing (HTS) environmental DNA (eDNA) studies of foraminiferal diversity (Lecroq *et al*., 2011; Lejzerowicz *et al*., 2013a). Other regions of the rRNA gene cluster, such as internal transcribed spacers (ITS1 and ITS2) or the 5′ end of the long subunit, are used sporadically (Holzmann *et al*., 1996; Tsuchyia *et al*., 2003). A manually curated and constantly updated database contains the sequences of the SSU rDNA barcoding 3′ region, as well as the illustrations and description of sequenced species.

Environmental DNA Surveys

The most important contribution to revolutionizing the view of foraminiferal diversity was made by eDNA surveys. The high specificity of the foraminiferal primers designed for amplifying rRNA genes from single cells allowed the detection of foraminiferal SSU rDNA copies in various environments. The presence of foraminiferal eDNA was confirmed not only in marine environments ranging from coastal (Habura *et al*., 2004, 2008; Bernhard *et al*., 2013; Edgcomb *et al*., 2014) to deep-sea sediments (Pawlowski *et al*., 2011b; Lejzerowicz *et al*., 2013b), but also in freshwater lakes (Holzmann *et al*., 2003) and rivers (L. Apotheloz-Gentil-Perret, University of Geneva; unpubl. data), and most astonishingly, in diverse soil samples (Lejzerowicz *et al*., 2010).

Although relatively long SSU rDNA sequences (*ca.* 1 kb) resulted from early foraminiferal eDNA surveys based on cloning and Sanger-sequencing, only a few sequences could be related to the reference database. Many remained unassigned at the species or genus levels and populated clades branching at the base of the foraminifera phylogenetic trees. Eight of these environmental clades related to early, monothalamous lineages (ENFOR1 to ENFOR8) have been identified in the deep Southern Ocean sediment (Pawlowski *et al*., 2011a). They are mainly composed of eDNA sequences, but some of them also comprise sequences obtained from isolated specimens. As none of the later sequences could be univocally assigned to any well-defined morphotype, they are assumed to represent foraminiferal squatters inhabiting the inside or outside of other foraminiferal tests (Moodley, 1990; Grimm *et al*., 2007).

Among the four environmental clades that gather soil and freshwater foraminiferal eDNA sequences (Lejzerowicz *et al*., 2010), only one could be characterized morphologically. This clade comprises the cultured freshwater species *Reticulomyxa filosa* (Pawlowski *et al*., 1999), whose genome has been recently sequenced (Glöckner *et al*., 2014), as well as another recently described species *Haplomyxa saranae* (Dellinger *et al*., 2014). Although other eDNA clades are much more species rich, and several freshwater amoeboid protists have been considered as related to foraminifera (Holzmann and Pawlowski, 2002), none of them could be isolated and morphologically characterized despite intensive efforts (L. Apotheloz-Perret-Gentil, University of Geneva; unpubl.). Basic biology, morphology, and ecology of these clades, therefore, remain one of the major questions in foraminiferal research.

The advent of the high-throughput sequencing era

The main limitation of early eDNA surveys was the low number of amplicon copies that could be obtained by the cloning and Sanger-sequencing approach. The millions of sequences generated by HTS technologies removed this

constraint, opening new perspectives for the development of eDNA surveys of foraminifera.

The first foraminiferal HTS eDNA study explored the diversity of deep-sea benthic foraminifera from Arctic, Antarctic, and Atlantic oceans (Lecroq *et al*., 2011). In that study, we sequenced the 37f hypervariable region of the SSU rDNA exclusively found among foraminifera (Pawlowski and Lecroq, 2010). At that time, the early version of the Illumina/Solexa technology limited the reads length to 36 nt only, encompassing roughly half of the 37f hypervariable region. However, despite such a short size, it was still possible to assign half of the sequences to different taxonomic levels on the basis of the phylogenetic signal present at the beginning of the 37f helix (Lecroq *et al*., 2011). Although the length of Illumina sequences recently expanded to up to 600 nt (longer sequences can be obtained using Roche 454 technologies), the helix 37f remains a favorite foraminiferal minibarcode for HTS studies.

So far we have generated 8 Illumina sequencing runs to address the foraminiferal diversity in six different projects, including ancient DNA (Lejzerowicz *et al*., 2013b; Pawlowska *et al*., 2014), biomonitoring (Pawlowski *et al*., 2014; X. Pochon *et al*., Cawthron Institute, New Zealand; unpubl.), and biogeographical survey (Lecroq *et al*., 2011; Lejzerowicz *et al*., 2014; F. Gschwend *et al*., University of Geneva; unpubl.). Here, we reanalyze the 26,135,707 reads obtained for all these projects, except the 36-nt reads (Lecroq *et al*., 2011) and those generated for the ancient DNA (aDNA) studies (Fig. 1). The results of this analysis confirmed the previous eDNA surveys based on the cloning approach. In particular, we brought compelling evidence that deep-sea foraminiferal communities are dominated by early-evolved monothalamous lineages. The operational taxonomic units (OTUs) assigned to Monothalamea accounted for up to 50% of the total diversity in some samples, whereas the hard-shelled taxa (Rotaliida, Textulariida, Miliolida) that are familiar to micropaleontologists represented less than 30% of OTUs. The remaining OTUs have been assigned to the environmental clades (ENFOR) or remain undetermined. In the survey of the Southern Ocean deep-sea sediment samples (Lejzerowicz *et al*., 2014), at each station, the second most-sequenced taxon corresponded to ENFOR2 clade, so far detected only in this area. Interestingly, this study also showed that the unassigned OTUs are often rare and can be eliminated if only the diversity found across sample replicates is analyzed. Some of them could represent rare species known to be abundant at the deep-sea bottom (Gooday and Jorissen, 2012), but their taxonomic status is disputable (see the following discussion).

The monothalamids, environmental clades, and undetermined lineages also dominated in the Roche 454 survey of eukaryotic diversity in the European coastal waters (Logares *et al*., 2014). This study revealed a high number of unknown foraminiferal lineages in both water and sediment samples, but interestingly, the diversity of monothalamids in water samples was clearly lower compared to rotaliids.

Applications of foraminiferal eDNA surveys

In addition to the exploration of recent foraminifera diversity, we also applied the HTS tool to environmental and industrial projects, for which foraminifera represent promising indicators of change in past and present ecosystems.

We conducted a series of studies of foraminiferal aDNA preserved in subsurface marine sediments. In the first study, we recovered extremely short DNA fragments corresponding to foraminiferal and radiolarian rDNA templates buried in abyssal sediment of the South Atlantic dated to about 32.5 thousand years (Lejzerowicz *et al*., 2013b). In the second study, we used a HTS eDNA approach to analyze foraminiferal communities in a sediment core from a Svalbard fjord encompassing one millennium of history (Pawlowska *et al*., 2014). In both studies, we used the taxonomic resolution of the short 37f hypervariable region to compare past foraminiferal OTU diversity with microfossil assemblages. We found that about half of the species archived in the fossil record could also be recovered from aDNA data. However, there was a limited match in terms of stratigraphic occurrence for some fossil species with respect to their aDNA sequences, especially in the case of rare taxa.

Beyond the investigation of foraminiferal community changes at geological time scales, we also used the HTS eDNA for environmental biomonitoring. In a recent study, we assessed the ability of HTS to reveal the response of benthic foraminiferal communities to the variation of environmental gradients associated with fish farming in the coastal environment (Pawlowski *et al*., 2014). Using ribosomal minibarcodes amplified from DNA and RNA sediment extracts, we showed that foraminiferal richness decreases in highly enriched sites. We also detected foraminiferal species that could be used as potential candidate bioindicators of environmental impact induced by fish farming. On the basis of this proof-of-concept study, we conclude that environmental barcoding using foraminifera and other protists has considerable potential to become a powerful tool for surveying the impact of aquaculture and other industrial activities in the marine environment.

Challenges of Next-Generation eDNA Surveys

The routine eDNA surveys of foraminiferal diversity for bioassessment and paleogenomic studies require standardized and robust HTS methods to overcome various technical and biological biases. Here, we present some of these biases (Table 1, Fig. 2), and we discuss their impact on eDNA diversity studies with emphasis on foraminiferal projects, although the problems we highlight are broadly applicable

Table 1

Main next-generation sequencing biases stemming from in vivo, in vitro, *and* in silico *aspects of the molecular studies with ease of detection, effect on diversity analysis, and recommendations to alleviate its effect*

Name of Bias	Detect	Recommendations
In situ		
Micro-patchiness	+	Beta-diversity and sampling pattern
Extracellular DNA	−	Systematic RNA co-sequencing
Intra-genomic polymorphism	+	Studying small subunit secondary structure
In vitro		
Contaminations	+	Stringent procedures and replicates
PCR-related biases	+	Lowering number of PCR cycles
Chimeras	++	PCR cycles + software detection
Mistagging	+	Latin Square experimental design
Cross-talks	−	Lower sequencing cluster density
Primers dimers	++	Trim reads on sequence length
In silico		
Reads quality	++	Stringent filtering parameters
Low abundance	−	Abundance-based filtering
Per-base errors	++	Clustering and pairwise global alignments

to other protists. We also introduce the three major challenges related to eDNA studies concerning taxonomic assignment, origin of unassigned sequences, and quantification of HTS data.

Principal biases of eDNA studies

In situ. The magnitude of biological biases affecting the composition and accuracy of HTS datasets is difficult to distinguish from technical biases. Beginning with the sampling process itself, it is not always clear to what extent the eDNA data are representative of the investigated communities. Although the sampling design and effort required for characterizing morphological diversity is well documented (see Gray and Elliott, 2009, for sampling marine sediments), there is yet no standards for eDNA sampling. For example, the patchiness characterizing the distribution of deep-sea benthic foraminifera (Gooday and Jorissen, 2012) consequently affects their eDNA diversity, which differs considerably in sediments collected a few centimeters apart (Lejzerowicz *et al*., 2014). Hence, such *micro-patchiness* invalidates studies based on point samples and supports the collection of numerous sample replicates across a broad range of spatial scales.

Another factor that biases the eDNA studies of marine sediments is the presence of *extracellular DNA* actively extruded by living cells or released after cell death (Vlassov *et al*., 2007). Such "dead" DNA is particularly well preserved when adsorbed on mineral surfaces (Naviaux *et al*., 2005), and it has been estimated that its abundance may reach up to 0.45 Gt at the surface of deep-sea sediments (Dell'Anno and Danovaro, 2005). Although this extracellular DNA can be easily detected when it corresponds to planktonic organisms sequenced from deep-sea floor (Pawlowski *et al*., 2011b), the presence of DNA molecules preserved across large areas may artificially increase the homogeneity of beta-diversity measures and the permeability of biogeographical and ecological barriers. Given the exceptional potential for DNA molecules to be preserved in downcore sediments (Lejzerowicz *et al*., 2013b; Pawlowska *et al*., 2014), the possibility that most of the eDNA material sequenced from deep-sea sediments does not belong to active species cannot be ruled out. While this is a minor problem for diversity exploration studies, the distinction of "dead" DNA is crucial in eDNA surveys focusing on the detection of ecologically active species. In this case, the environmental RNA (eRNA) may provide a better proxy.

In vitro. Extraneous DNA contamination as well as DNA cross-contamination threaten every step of the data acquisition process, including DNA/RNA extraction, PCR amplification, and library preparation. It is therefore mandatory to establish extremely careful laboratory procedures. Ideally, each step of the workflow should be associated with dedicated material and facilities that are regularly cleaned with DNA and RNA removal solutions. Typical contamination sources can also be detected by taxonomic analyses when specific taxa are targeted (Orsi *et al*., 2013).

The most pervasive source of errors and artifacts in eDNA amplicon sequencing studies is the PCR amplification (Berney *et al*., 2004; Acinas *et al*., 2005; Aird *et al*., 2011). The major types of PCR errors are insertions of bases and chimeras, which are readily formed in samples characterized by high DNA template diversity (Fonseca *et al*., 2012). The formation of chimeras is facilitated by the mosaic structure of rRNA genes intermingling the conserved and variable regions. Optimizing the PCR conditions can help reduce the prevalence of chimeras in sequencing libraries (Stevens *et al*., 2013). Furthermore, computational solutions are also available to screen chimeric sequences from HTS datasets (Edgar *et al*., 2011).

Another type of error results from the multiplexing of PCR amplicon samples, referred to as mistagging (Carlsen *et al*., 2012; Kircher *et al*., 2012; Carew *et al*., 2013). Mistagging consists of a shuffling of the barcode sequences appended either during the PCR amplification with tagged primers or during the preparation of the sequencing libraries. Hence, a large part of the reads are identified with the wrong tag combination, resulting in intractable cross-contamination events. To detect and remove these reads, systematic paired-end tagging and careful experimental planning must be applied. Our recent experiments demonstrate the importance of de-saturating the number of tagged primer combinations employed relative to the total number of possible combinations, and balancing the tag usage frequencies (Esling *et al*., unpubl.). Finally, during the Illumina sequencing process, the fragments could be sequenced in

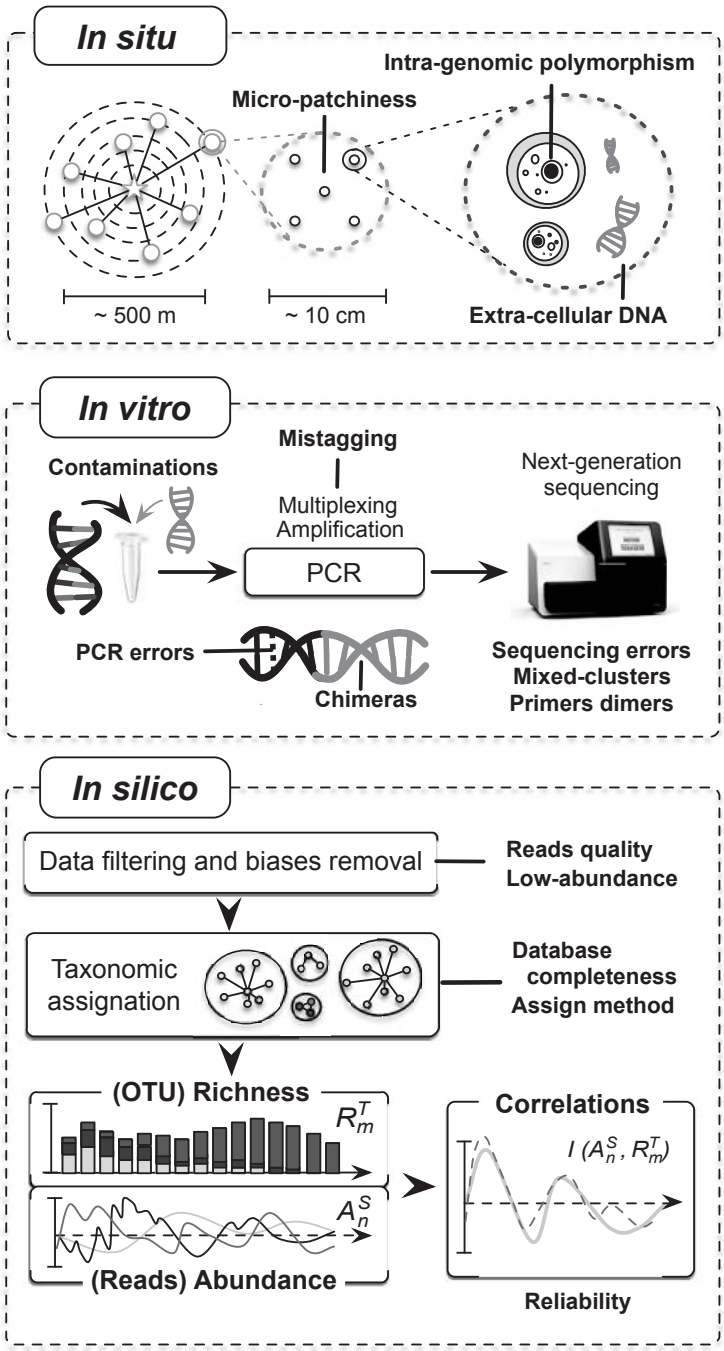

Figure 2. The schematic workflow of high-throughput sequencing (HTS) environmental DNA studies with various biological and technical biases indicated in bold.

mixed clusters, where overlapping base calls would result in additional substitution errors. It has been shown that reducing the cluster density on the flow cell increases the overall quality of a sequencing run (Kozich *et al*., 2013).

In silico. During the raw data processing using bioinformatics, it is recommended to remove spurious reads associated with low-quality scores or containing errors that can be recognized in expected sequence regions (Minoche *et al*., 2011). For example, in foraminifera, a conserved region 80 bases long adjacent to the hypervariable region is used to assess the magnitude of sequence errors. Slight variations can be assumed negligible according to the knowledge acquired from analyses of reference sequences and can be removed using supervised clustering techniques. Typically, the majority of HTS dataset sequences are rare. Discriminating the sequences that truly represent genuine species of

the rare biosphere from spurious sequences remains a major challenge in sequence analysis (Kunin *et al*., 2010). Some authors suggest that using objective thresholds based on sequence quality or low abundance is efficient at removing artifacts (Bokulich *et al*., 2012; Caporaso *et al*., 2011).

Taxonomic assignments

Current methods of HTS reads assignment rely on different procedures, algorithms, and parameters that often have in common fixed thresholds. Their accuracy can be high when very similar sequences are present in the reference database. In most of eukaryotic diversity surveys, however, the fixed thresholds are inappropriate given the large variability of the rates of evolution and diversification across eukaryotic taxa (Caron *et al*., 2009; Pernice *et al*., 2013). Assignments can be realized on the basis of phylogenetic signals, sequence similarity measures, or diagnostic signatures, but the suitability of the method varies depending on the marker and the evolutionary history of the species (van Velzen *et al*., 2012). Assignments based on BLAST searches are not always reliable, and the results require careful examination (Koski and Golding, 2001), especially in the case of ribosomal DNA (rDNA) sequences where conserved regions are often longer than the taxonomically informative, hypervariable regions. The reliability of species-level assignment depends on the amount of taxonomic knowledge incorporated in the analyses (Hoef-Emden, 2012). There is yet no established system incorporating patterns of marker sequence variability to supervise species delineations for all micro-eukaryotes (Boenigk *et al*., 2012). Such systems are only starting to emerge for the stramenopiles (Massana *et al*., 2014) and for the ciliates (Dunthorn *et al*., 2012, 2014).

In foraminiferal HTS studies, the taxonomic assignment is based on the compound diagnostic signatures that are present in the 5′ end of the foraminiferal 37f hypervariable region and which allow unambiguous pre-assignment to family (or clade) level (Lecroq *et al*., 2011; Lejzerowicz *et al*., 2013b). Then, the assignment to genus or species level is based on distances calculated from Needleman-Wunsch alignments between complete 37f sequences and subsets of reference sequences belonging to assigned higher-level taxa (Pawlowski *et al*., 2014). On the basis of the reanalyzed dataset, it appears that the database completeness and the taxonomic coverage within pre-assigned families or clade determine the identification depth of foraminiferal environmental sequences (Fig. 3). Unsurprisingly, the clades well documented in the reference database, both in terms of number of species entries and number of isolates sequenced per species, are also more sequenced in the environment. This could simply reflect the taxonomic coverage offered by a continuous effort of sequencing SSU rRNA gene copies from isolated specimens. However, when sequences are assigned to clades that enclose numerous reference sequences (*e.g.,* Clade 3 or Clade E), they are more readily assigned down to the species level. This is displayed on Figure 3 both in terms of number of OTUs (inner heatmap) and number of reads (outer heatmap), mostly assigned to the species level (outer block) when the database coverage is high within the family. The pre-assignment is necessary since the minimum distance between clades ($K^{\min}$) is in general smaller than the maximum distance among the sequences of a given clade ($K^{\max}$) (Fig. 3). The $K^{\max}$ distance represents a conservative threshold for well-defined clades. For poorly described clades that gather very different genera, a species-level assignment will be more difficult as a result of increased taxonomic conflicts. A challenging task is to refine the foraminiferal phylogenies to better delineate the clades and thus improve species-level assignment.

Taxonomic assignment can be additionally complicated by intraspecific and intra-genomic polymorphisms. It is also possible that several species hide behind a single sequence because of a lack of resolution within taxa, as shown for dinoflagellates (Stern *et al*., 2010), or in the worst case, as a result of convergence between short sequences due to mutational saturation in fast-evolving sites. In foraminifera, intra-genomic polymorphisms remain a major challenge for closely related species, creating overlapping conflicting assignments even at low distances. As shown by our recent study, this type of polymorphism is observed in SSU rRNA genes of almost all foraminiferal species (Weber and Pawlowski, 2014). In particular cases, the sequence divergence of SSU rDNA can reach up to 8%, and more than five distinct ribotypes can be found within a single specimen (Pillet *et al*., 2012). In this case, even a stringent interpretation of HTS data may still lead to an overestimation of foraminiferal richness, and only extensive database coverage with many SSU copies sequenced for each specimen could help mitigate its effect.

Characterizing the unknowns

In all of our HTS studies (Lecroq *et al*., 2011; Lejzerowicz *et al*., 2013a; Pawlowski *et al*., 2014), a large part of the sequences would pertain either to environmental clades or to unknown sequences. The latter could not be assigned to any database reference; but because they were amplified using specific primers and do contain the conserved region upstream from the diagnostic 37f region, their genuine foraminiferal origin is unquestionable. These sequences should be handled with extreme caution as they may include both artifacts and new lineages never sequenced before.

As shown by the compiled analyses of all our HTS studies (Fig. 4), the proportion of unassigned reads, averaging 20%–25%, is equivalent across various geographical regions and projects. This observation casts doubt on the

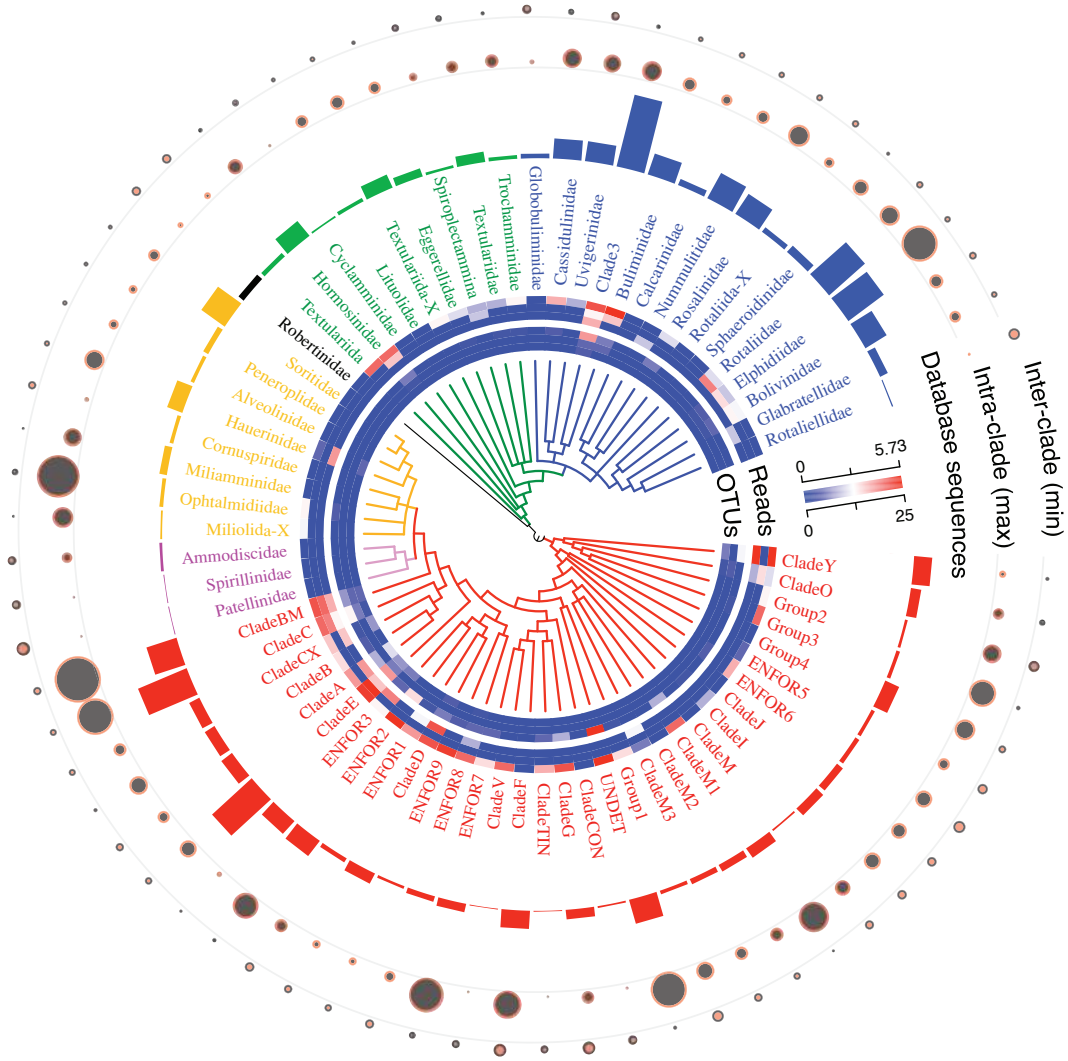

Figure 3. Reference sequence database variability and high-throughput sequencing (HTS) reads assignment depth of benthic foraminiferal families. For each foraminiferal family (or clade when no monophyletic family could be defined) represented at the tips of the central phylogenetic tree are indicated the assignment depth of environmental OTUs (inner circular heatmap) and reads (outer circular heatmap), as well as the number and variability of the related database reference sequences (barplots and dots). In the heatmaps, the assignment depths are indicated for the family (inner blocks), the genus (middle blocks), and the species level (outer blocks), according to the color legend ranging from blue to red. For each family the intra-clade (K^{max}, inner dots) and the inter-clade (K^{min}, inner dots) variability is represented, as well as the number of unique reference sequences ranging from 77 (Clade3) to 1 (Rotaliellidae). The families are colored according to the order-level classification: monothalamiids (red), Spirillinida (violet), Miliolida (yellow), Textulariida (green), and Rotaliida (blue).

validity of these sequences. Indeed, the reproducibility of this distribution points to the existence of a linear HTS bias appearing independently of the region or molecule sequenced. This is further supported by the results of our recent study performed on synthetic communities, where similar proportions of unassigned sequences were found among the expected sequences (Esling *et al.*, unpubl.). The artifactual character of unknowns is also reinforced by the fact that most of them generate an incredibly high number of low-abundance OTUs found in only a few samples (Fig. 4).

Nevertheless, not all unassigned sequences should be considered as technical or biological errors. Some of them might be derived from truly unknown species, reflecting a part of biological diversity that is yet to be discovered. As shown in Figure 4, a small portion of the unassigned reads are actually grouped in highly abundant OTUs, represented by a large number of reads, and found in a large number of samples. The discovery of totally novel lineages is not uncommon in protists (Kim *et al.*, 2011; Massana *et al.*, 2014). In the case of foraminifera, we can expect that many

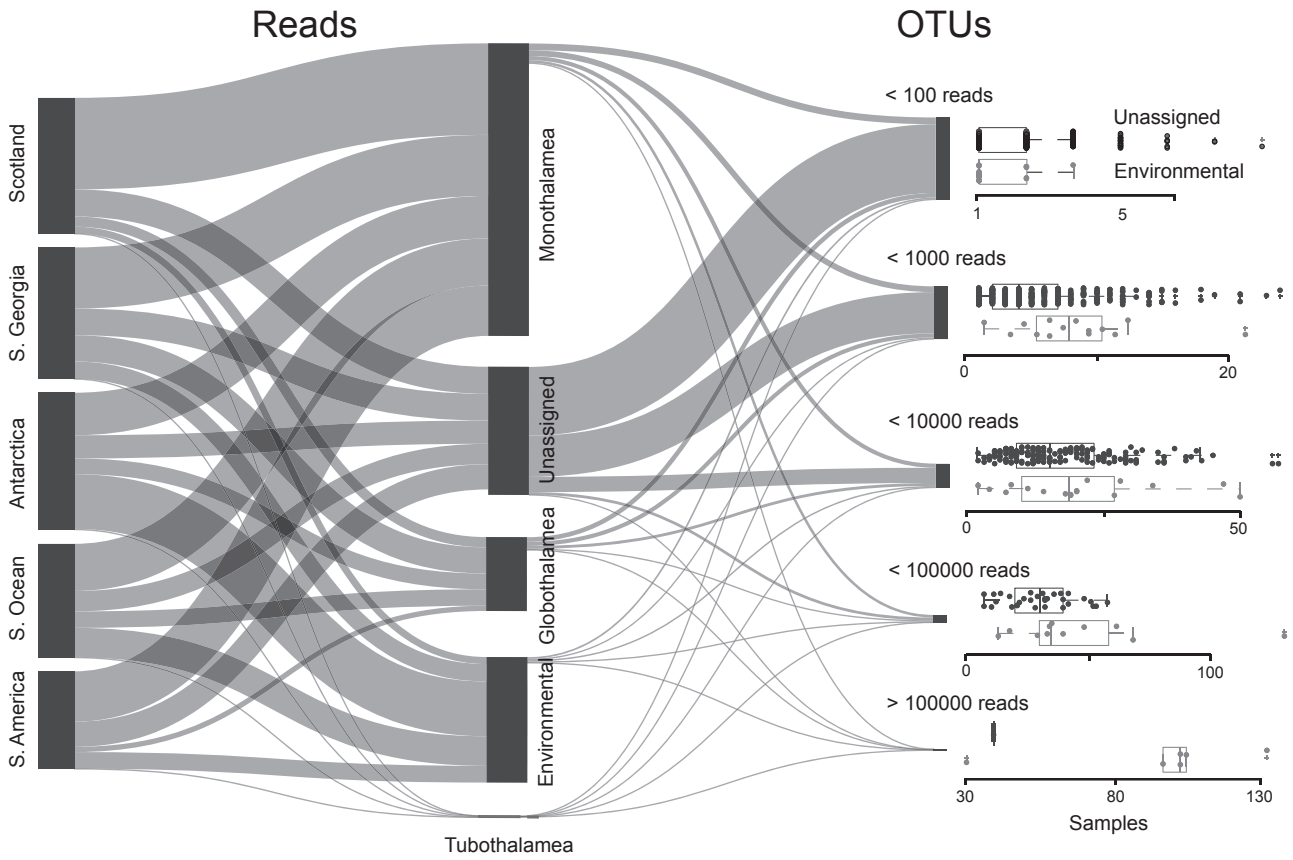

Figure 4. The flow of high-throughput sequencing (HTS) reads sequenced from different regions (left) and assigned to different foraminiferal groups (middle). The reads are grouped into OTUs, each of which has a different abundance category (right). The box and scatterplots represent the number of samples that each of the OTUs formed by unassigned or environmental reads can be found in for each abundance category.

enigmatic lineages occurring in the marine environment correspond to pico-sized or parasitic species that necessitate special efforts to be isolated and described.

The proportion of unknown reads would be much higher if we also included the sequences assigned to environmental clades. Indeed, none of the OTUs belonging to these clades have ever been identified. However, the chances that these OTUs are artefactual are very low. As shown in Figure 4, the OTUs formed by environmental sequences are more equally distributed among abundance categories and found in large number of samples. Hence, there is a higher probability that they stem from yet-undescribed lineages.

Quantitative inference from HTS data

Certainly the most controversial challenge related to HTS diversity surveys is to infer the species abundance from eDNA samples. Traditionally, the determination of protist abundance relied mainly on direct microscopic counting, quantitative PCR (Zhu *et al*., 2005; Zhang and Fang, 2006), and flow cytometry (Shi *et al*., 2011). Although quantitative HTS data are often used as a basis for estimators of species richness (Chao and Bunge, 2002; Bunge *et al.,* 2012) and may even preserve relative biomass information (Andersen *et al*., 2012), it remains an analytical challenge to infer the relative abundance of organisms (Deagle *et al*., 2013) and especially of microbial eukaryote cells. Comparison of pyrosequencing data and morphological counts to infer seasonal protist species turnover led to very divergent results (Medinger *et al*., 2010). In experimental testing of fungal mock communities, a difference of one order of magnitude was found between the most and least abundant species, mixed in the same initial concentration (Amend *et al*., 2010).

The main factor that could explain these discrepancies is the variation of the rDNA copy number across species. Different protist taxa harbor variable numbers of nuclear rDNA copies as a result of different genome size (Prokopowich *et al*., 2003), biovolume (Godhe *et al*., 2008), or number of nuclei (Heyse *et al*., 2010). Accounting for these sources of variation is particularly challenging for foraminifera, given the complexity of their life cycle and the

multiplicity of factors influencing genomic dynamics (Parfrey *et al*., 2008; Parfrey and Katz, 2010). According to the qPCR assays, the number of rDNA copies in foraminifera varies between 40,000 in monothalamous genus *Allogromia* and 5,000–10,000 in rotaliid genera *Bolivina* and *Rosalina* (Weber and Pawlowski, 2013). The authors of this study show that under controlled conditions in which species diversity is known, it is possible to correctly determine abundance of species on the basis of proportions of sequences by using normalization factors that take into account the variations of rDNA copy number or rRNA expression level.

Various technical biases can also influence the abundance of HTS reads. Among them, the most important is the relative efficiency of the PCR amplification step. The number of reads obtained for each sequence depends on its properties and on the overall sample complexity. Indeed, the PCR efficiency is higher for shorter fragments (Huber *et al*., 2009) and varies for each DNA template according to the global composition of the sample (Gonzalez *et al*., 2012). This effect of PCR competition is further complicated by a primer bias if long constructs including trailing sequences are used to bypass the step of preparing the sequencing library (Berry *et al*., 2011). Finally, the number of reads available for each of the samples multiplexed in a sequencing run can vary greatly owing to the accuracy of the quantification and pooling of the samples, as well as the composition and quality of the libraries.

In foraminifera, some of these PCR-related abundance biases are of less importance because the PCR conditions have been optimized and the primer specificity has been tested for all taxonomic groups. However, another technical problem specific to this group is the selective efficiency of nucleic acid extraction protocols. As many foraminiferal cells are protected by hard shells, it is clear that their extraction may be much more difficult than in the case of organic-walled species. We cannot exclude the possibility that the abundance of organic-walled monothalamids in all our eDNA studies (Fig. 1) is partially due to this bias. This factor has also been given as an explanation for the lack of some testate calcareous species in HTS data in spite of their abundance in morphological counts (Pawlowski *et al*., 2014).

An alternative for resolving the abundance issue is to analyze metatranscriptomic rather than metagenomic data. The main advantage of using RNA molecules rather than DNA is that it more accurately depicts the diversity of species alive at the time of sampling (Stoeck *et al*., 2007). This has been supported for foraminifera in our studies of deep-sea (Lejzerowicz *et al*., 2013b) and anoxic (Langlet *et al*., 2013) sediments. Although not generally accepted (Orsi *et al*., 2013), it is assumed that the RNA molecules have a much shorter life span compared to DNA and therefore are less subjected to preservation in the form of extracellular molecules. Moreover, RNA sequencing data have the potential to be used to infer relative levels of species abundance (Logares *et al*., 2012). In our assessment of aquaculture impact (Pawlowski *et al*., 2014), we show that by retaining the OTUs that are present in both DNA and RNA, the accuracy of HTS data increases considerably. Moreover, the species richness inferred from RNA sequences was strongly correlated to environmental gradients and morphological counts, suggesting that RNA could possibly surpass DNA as an indicator of the abundance of ecologically relevant species.

Preparing the Future

The biases and unresolved issues that have been cited above are so numerous that one could conclude that they prohibit the use of HTS for diversity surveys. Here, we argue that this is not the case and that most of these biases can be mitigated and even mostly removed through careful procedures and stringent examination of the data. Below we delineate the main avenues of research in HTS eDNA studies that should be explored in upcoming years.

High-throughput accuracy

To mitigate the effects of various artifacts, most studies use filtering techniques that vary in their functionalities, parameters, and adequacy to remove spurious reads. These filters are usually parameterized empirically—that is, based on one or several control samples aimed at extracting the baseline values of abundance or sequence divergence. However, it is obvious that all these choices are biased by the composition of the control sample itself. Several authors (Degnan and Ochmann, 2011; Egge *et al*., 2013; Zhan *et al*., 2013) have shown that it is impossible to remove all biased reads without collaterally discarding genuine OTUs. Hence, we advocate a new generation of internal controls and mathematical rationale to improve HTS accuracy through replicate analyses and techniques specifically tailored for bias removal. To avoid diversity over-estimation, alternative enrichment methods based on highly parallelized microreactors (Leung *et al*., 2012) or PCR-free hybridization methods capturing a large taxonomic spectrum (Mason *et al*., 2011) are being developed and should be more widely used in HTS eDNA studies.

Database and taxonomic assignment

To achieve robust assignments, we need to continuously expand the taxonomic coverage of the reference database by keeping the pace of single-cell sequencing efforts. However, it is important to increment the database not only vertically (*i.e.,* to add new entries corresponding to taxa not sequenced yet), but also horizontally (*i.e.,* to sequence all the variable SSU rDNA copies that exist in the genome of

each single species). Then, the resolution of such a bi-dimensional database could be assessed exhaustively through various analyses accounting for rRNA secondary structures, coordinates of sequence motifs, and signature, as well as intra-genomic polymorphisms. As a result, each foraminiferal species could be characterized by refined genetic information and ascribed a specific method ensuring optimal species-level assignment of the environmental sequences. Finally, it is noteworthy that the development of well-curated reference databases for taxon-specific DNA barcodes holds great promise for delineating species that would be indistinguishable using SSU rDNA sequences only (Pawlowski *et al*., 2012).

Distinguishing unknowns from artifacts

In the same line of thought, the question of the rare biosphere is still one of the most controversial issues. As the majority of rare sequences are unassigned, this question is often eluded by withdrawing low-abundance taxa to remove potential errors. However, many of these so-called errors might truly be derived from real species. Indeed, there are no biological reasons that rarity necessarily implies falseness. To solve this problem, we recommend conducting an in-depth characterization of each single unassigned sequence by searching for unexpected chimera recombination, primer motifs, and other characteristics. If such a sequence passes the stringent filtering conditions, it should be considered as corresponding to a putatively new lineage.

Chasing down the abundance

The use of abundance information in HTS data is impeded by a combination of biological and technical biases that are very difficult to overcome. The simplest solution would be to estimate the number of copies of rRNA genes for all concerned species. Another possibility would be to sequence the same communities repeatedly with slightly varying amounts of each species, to understand the relation between the number of PCR amplicons and the abundance of specimens. However, such solutions would be time-consuming and rather impractical, especially in the case of uncultivable protists, such as foraminifera. Therefore, it seems more reasonable to abandon attempts to infer absolute abundance with reference to effective number of specimens and replace it by using relative OTU abundances. Such a system, in conjunction with RNA rather than DNA sequencing, may lead to more efficient measurement of species activity in relation to environmental changes.

Concluding Remarks

These few remarks provide a glimpse of what could be done to improve the accuracy and interpretation of HTS-based surveys of environmental diversity. By focusing on a single taxonomic group, our reflection may be biased by certain aspects of the HTS approach specific to foraminifera. Nevertheless, many of our comments also apply to other groups of protists and can help in analyzing the diversity of complex microbial eukaryotic communities.

Acknowledgments

We thank Bruce Hayward for help accessing the WoRMS database. This work was supported by the Swiss National Science Foundation grant 31003A_140766, and G. & A. Claraz Donation. Some of the unpublished data reported here were obtained with help of the Biodiversity of Marine euKaryotes (BioMarKs, http://www.biomarks.eu) consortium, which was funded by the European Union ERANet program BiodivERsA (2008–6530).

Literature Cited

Acinas, S. G., R. Sarma-Rupavtarm, V. Klepac-Ceraj, and M. F. Polz. 2005. PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Appl. Environ. Microbiol*. **71:** 8966–8969.

Aird, D., M. G. Ross, W. S. Chen, M. Danielsson, T. Fennell, C. Russ, D. B. Jaff, C. Nusbaum, and A. Gnirke. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol*. **12:** R18.

Amend, A. S., K. A. Seifert, and T. D. Bruns. 2010. Quantifying microbial communities with 454 pyrosequencing: does read abundance count? *Mol. Ecol*. **19:** 5555–5565.

Andersen, K., K. L. Bird, M. Rasmussen, J. Haile, H. Breuning-Madsen, K. H. Kjaer, L. Orlando, M. T. P. Gilbert, and E. Willerslev. 2012. Meta-barcoding of 'dirt'DNA from soil reflects vertebrate biodiversity. *Mol. Ecol*. **21:** 1966–1979.

Berney, C., J. Fahrni, and J. Pawlowski. 2004. How many novel eukaryotic "kingdoms"? Pitfalls and limitations of environmental DNA surveys. *BMC Biol*. **2:** 13.

Bernhard, J. M., V. P. Edgcomb, P. T. Visscher, A. McIntyre-Wressnig, R. E. Summons, M. L. Bouxsein, L. Louis, and M. Jeglinski. 2013. Insights into foraminiferal influences on microfabrics of microbialites at Highborne Cay, Bahamas. *Proc. Natl. Acad. Sci. USA* **110:** 9830–9834.

Berry, D., K. B. Mahfoudh, M. Wagner, and A. Loy. 2011. Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. *Appl. Environ. Microbiol*. **77:** 7846–7849.

Boenigk, J., M. Ereshefsky, K. Hoef-Emden, J. Mallet, and D. Bass. 2012. Concepts in protistology: species definitions and boundaries. *Eur. J. Protistol*. **48:** 96–102.

Bokulich, N. A. , S. Subramanian , J. J. Faith , D. Gevers , J. I. Gordon, R. Knight, D. A. Mills, and J. G. Caporaso. 2012. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat. Methods* **10:** 57–59.

Bunge, J., L. Woodard, D. Böhning, J. A. Foster, S. Connolly, and H. K. Allen. 2012. Estimating population diversity with CatchAll. *Bioinformatics* **28:** 1045–1047.

Buzas, M. A., and S. J. Culver. 1991. Species diversity and dispersal of benthic foraminifera. *Bioscience* **41:** 483–489.

Caporaso, J. G., C. L. Lauber, W. A. Walters, D. Berg-Lyons, C. A. Lozupone, P. J. Turnbaugh, N. Fierer, and R. Knight. 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. USA* **108:** 4516–4522.

Carew, M. E., V. J. Pettigrove, L. Metzeling, and A. A. Hoffmann.

2013. Environmental monitoring using next generation sequencing: rapid identification of macroinvertebrate bioindicator species. *Front. Zool.* **10:** 45.

Carlsen, T., A. B. Aas, D. L. Lindner, T. Vråstad, T. Schumacher, and H. Kauserud. 2012. Don't make a mista(g)ke: is tag switching an overlooked source of error in amplicon pyrosequencing studies? *Fungal Ecol.* **5:** 747–749.

Caron, D. A., P. D. Countway, P. Savai, R. J. Gast, A. Schnetzer, S. D. Moorthi, M. R. Dennett, D. M. Moran, and A. C. Jones. 2009. Defining DNA-based operational taxonomic units for microbial-eukaryote ecology. *Appl. Environ. Microbiol.* **75:** 5797–5808.

Chao, A., and J. Bunge. 2002. Estimating the number of species in a stochastic abundance model. *Biometrics* **58:** 531–539.

Darling, K. F., and C. M. Wade. 2008. The genetic diversity of planktic foraminifera and the global distribution of ribosomal RNA genotypes. *Mar. Micropaleontol.* **67:** 216–238.

Darling, K. F., M. Kucera, and C. M. Wade. 2007. Global molecular phylogeography reveals persistent Arctic circumpolar isolation in a marine planktonic protist. *Proc. Natl. Acad. Sci. USA* **104:** 5002–5007.

Deagle, B. E., A. C. Thomas, A. K. Shaffer, A. W. Trites, and S. N. Jarman. 2013. Quantifying sequence proportions in a DNA-based diet study using Ion Torrent amplicon sequencing: which counts count? *Mol. Ecol. Resour.* **13:** 620–633.

Debenay, J.-P. 2012. *A Guide to 1,000 Foraminifera from Southwestern Pacific, New Caledonia*. IRD Editions Marseille/Publication Scientifiques du Muséum, Paris. p.378.

Degnan, P. H., and H. Ochman. 2011. Illumina-based analysis of microbial community diversity. *ISME J.* **6:** 183–194.

Dell'Anno, A., and R. Danovaro. 2005. Extracellular DNA plays a key role in deep-sea ecosystem functioning. *Science* **309:** 2179.

Dellinger, M., A. Labat, L. Perrouault, and P. Grellier. 2014. *Haplomyxa saranae* gen. nov. et sp. nov., a new naked freshwater foraminifer. *Protist* **165:** 317–329.

de Vargas, C., R. Norris, L. Zaninetti, S. W. Gibb, and J. Pawlowski. 1999. Molecular evidence of cryptic speciation in planktonic foraminifers and their relation to oceanic provinces. *Proc. Natl. Acad. Sci. USA* **96:** 2864–2868.

Dunthorn, M., J. Klier, J. Bunge, and T. Stoeck. 2012. Comparing the hyper-variable V4 and V9 regions of the small subunit rDNA for assessment of ciliate environmental diversity. *J. Eukaryot. Microbiol.* **59:** 185–187.

Dunthorn, M., J. Otto, S. A. Berger, A. Stamatakis, F. Mahé, S. Romac, C. de Vargas, S. Audic, BioMarKs Consortium, A. Stock, F. Kauff, and T. Stoeck. 2014. Placing environmental next generation sequencing amplicons from microbial eukaryotes into a phylogenetic context. *Mol. Biol. Evol.* **31:** 993–1009.

Edgar, R. C., B. J. Haas, J. C. Clemente, C. Quince, and R. Knight. 2011. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27:** 2194–2200.

Edgcomb, V. P., J. M. Bernhard, R. E. Summons, W. Orsi, D. Beaudoin, and P. T. Visscher. 2014. Active eukaryotes in microbialites from Highborne Cay, Bahamas, and Hamelin Pool (Shark Bay), Australia. *ISME J.* **8:** 418–429.

Egge, E., L. Bittner, T. Andersen, S. Audic, C. de Vargas, and B. Edvardsen. 2013. 454 pyrosequencing to describe microbial eukaryotic community composition, diversity and relative abundance: a test for marine haptophytes. *PLoS One* **8:** e74371.

Fonseca, V. G., B. Nichols, D. Lallias, C. Quince, G. R. Carvalho, D. M. Power, and S. Creer. 2012. Sample richness and genetic diversity as drivers of chimera formation in nSSU metagenetic analyses. *Nucleic Acids Res.* **40:** e66–e66.

Glöckner, G., N. Hülsmann, M. Schleicher, A. A. Noegel, L. Eichinger, C. Gallinger, J. Pawlowski, R. Sierra, U. Euteneuer, L. Pillet *et al.* 2014. The genome of the foraminiferan *Reticulomyxa filosa*. *Curr. Biol.* **24:** 11–18.

Godhe, A., M. E. Asplund, K. Höm, V. Saravanan, A. Tyagi, and I. Karunasagar. 2008. Quantification of diatom and dinoflagellate biomasses in coastal marine seawater samples by real-time PCR. *Appl. Environ. Microbiol.* **74:** 7174–7182.

Gonzalez, J. M., M. C. Portillo, P. Belda-Ferre, and A. Mira. 2012. Amplification by PCR artificially reduces the proportion of the rare biosphere in microbial communities. *PLoS One* **7:** e29973.

Gooday, A. J., and F. J. Jorissen. 2012. Benthic foraminiferal biogeography: controls on global distribution patterns in deep-water settings. *Annu. Rev. Marine Sci.* **4:** 237–262.

Gray, J. S., and M. Elliott. 2009. *Ecology of Marine Sediments: From Science to Management.* Oxford University Press, Oxford.

Grimm, G. W., K. Stögerer, K. T. Ertan, H. Kitazato, M. Kucera, V. Hemleben, and C. Hemleben. 2007. Diversity of rDNA in *Chilostomella*: Molecular differentiation patterns and putative hermit types. *Mar. Micropaleontol.* **62:** 75–90.

Habura, A., J. Pawlowski, S. D. Hanes, and S. S. Bowser. 2004. Unexpected foraminiferal diversity revealed by small-subunit rRNA analysis of Antarctic sediment. *J. Eukaryot. Microbiol.* **51:** 173–179.

Habura, A., S. T. Goldstein, S. Broderick, and S. S. Bowser. 2008. A bush, not a tree: the extraordinary diversity of cold-water basal foraminiferans extends to warm-water environments. *Limnol. Oceanogr.* **53:** 1339–1351.

Hayward, B. W., M. Holzmann, H. R. Grenfell, and J. Pawlowski. 2004. Morphological distinction of molecular types in *Ammonia*—towards a taxonomic revision of the world's most common and misidentified foraminiferal genus. *Mar. Micropaleontol.* **50:** 237–271.

Hayward, B. W., H. R. Grenfell, C. M. Reid, and K. A. Hayward. 1999. *Recent New Zealand Shallow-Water Benthic Foraminifera: Taxonomy, Ecologic Distribution, Biogeography, and Use in Paleoenvironmental Assessment*, Institute of Geological & Nuclear Sciences Monograph, Vol. 21. Institute of Geological and Nuclear Sciences, Wellington, New Zealand. 258 pp.

Hayward, B. W., M. Holzmann, H. R. Grenfell, and J. Pawlowski. 2004. Morphological distinction of molecular types in *Ammonia*—towards a taxonomic revision of the world's most common and misidentified foraminiferal genus. *Mar. Micropaleontol.* **50:** 237–271.

Hayward, B. W., H. R. Grenfell, A. T. Sabaa, H. L. Neil, and M. A. Buzas. 2010. *Recent New Zealand Deep-Water Benthic Foraminifera: Taxonomy, Ecologic Distribution, Biogeography, and Use in Paleoenvironmental Assessment*, Institute of Geological & Nuclear Sciences Monograph, Vol. 26. Institute of Geological and Nuclear Sciences, Wellington, New Zealand. 363 pp.

Heyse, G., F. Jönsson, W.-J. Chang, and H. J. Lipps. 2010. RNA-dependent control of gene amplification. *Proc. Natl. Acad. Sci. USA* **107:** 22134–22139.

Hoef-Emden, K. 2012. Pitfalls of establishing DNA barcoding systems in protists: the *Cryptophyceae* as a test case. *PloS One* **7:** e43652.

Holzmann, M., and J. Pawlowski. 2002. Freshwater foraminiferans from Lake Geneva: past and present. *J. Foraminiferal Res.* **32:** 344–350.

Holzmann, M., W. Piller, and J. Pawlowski. 1996. Sequence variations in large-subunit ribosomal RNA gene of *Ammonia* (Foraminifera, Protozoa) and their evolutionary implications. *J. Mol. Evol.* **43:** 145–151.

Holzmann, M., A. Habura, H. Giles, S. S. Bowser, and J. Pawlowski. 2003. Freshwater foraminiferans revealed by analysis of environmental DNA samples. *J. Eukaryot. Microbiol.* **50:** 135–139.

Huber, J. A., H. G. Morrison, S. M. Huse, P. R. Neal, M. L. Sogin, and D. B. Mark Welch. 2009. Effect of PCR amplicon size on assessments of clone library microbial diversity and community structure. *Environ. Microbiol.* **11:** 1292–1302.

Kim, E., J. W. Harrison, S. Sudek, M. D. Jones, H. M. Wilcox, T. A. Richards, A. Z. Worden, and J. M. Archibald. 2011. Newly identified and diverse plastid-bearing branch on the eukaryotic tree of life. *Proc. Natl. Acad. Sci. USA* **108:** 1496–1500.

Kircher, M., S. Sawyer, and M. Meyer. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* **40:** e3.

Koski, L. B., and G. B. Golding. 2001. The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol.* **52:** 540–542.

Kozich, J. J., S. L. Westcott, N. T. Baxter, S. K. Highlander, and P. D. Schloss. 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl. Environ. Microbiol.* **79:** 5112–5120.

Kunin, V., A. Engelbrektson, H. Ochman, and P. Hugenholtz. 2010. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ. Microbiol.* **12:** 118–123.

Langlet, D., E. Geslin, C. Baal, E. Metzger, F. Lejzerowicz, B. Riedel, M. Zuschin, J. Pawlowski, M. Stachowitsch, and F. J. Jorissen. 2013. Foraminiferal survival after long-term in situ experimentally induced anoxia. *Biogeoscience* **10:** 7463–7480.

Lecroq, B., F. Lejzerowicz, R. Christen, P. Esling, L. Baerlocher, M. Osteras, L. Farinelli, and J. Pawlowski. 2011. Ultra-deep sequencing of foraminiferal microbarcodes unveils hidden richness of early monothalamous lineages in deep-sea sediments. *Proc. Natl. Acad. Sci. USA* **108:** 13177–13182.

Lejzerowicz, F., J. Pawlowski, L. Fraissinet-Tachet, and R. Marmeisse. 2010. Molecular evidence for widespread occurrence of Foraminifera in soils. *Environ. Microbiol.* **12:** 2518–2526.

Lejzerowicz, F., I. Voltsky, and J. Pawlowski. 2013a. Identifying active foraminifera in the Sea of Japan using metatranscriptomic approach. *Deep-Sea Res. II* **86–87:** 214–220.

Lejzerowicz, F., M. Majewski, W. Szczuciński, J. Decelle, C. Obadia, P. Martinez Arbizu, and J. Pawlowski. 2013b. Ancient DNA complements microfossil record in deep-sea subsurface sediments. *Biol. Lett.* **9:** 20130238.

Lejzerowicz, F., P. Esling, and J. Pawlowski. 2014. Patchiness of deep-sea benthic Foraminifera across the Southern Ocean: insights from high-throughput DNA sequencing. *Deep-Sea Res.* II, doi:/10.1016/j.dsr2.2014.07.018.

Leung, K., H. Zahn, T. Leaver, K. M. Konwar, N. W. Hanson, A. P. Pagé, C.-C. Lo, P. S. Chain, S. J. Hallam, and C. L. Hansen. 2012. A programmable droplet-based microfluidic device applied to multiparameter analysis of single microbes and microbial communities. *Proc. Natl. Acad. Sci. USA* **109:** 7665–7670.

Loeblich, A. R., and H. Tappan. 1988. *Foraminiferal Genera and Their Classification,* 2 Vols. Van Nostrand Reinhold, New York.

Logares, R., S. Audic, S. Santini, M. C. Pernice, C. de Vargas, and R. Massana. 2012. Diversity patterns and activity of uncultured marine heterotrophic flagellates unveiled with pyrosequencing. *ISME J.* **6:** 1823–1833.

Logares, R., S. Audic, D. Bass, L. Bittner, C. Boutte, R. Christen, J.-M. Claverie, J. Decelle, J. R. Dolan, M. Dunthorn *et al.* 2014. Patterns of rare and abundant marine microbial eukaryotes. *Curr. Biol.* **24:** 1–9.

Mason, V. C., G. Li, K. M. Helgen, and W. J. Murphy. 2011. Efficient cross-species capture hybridization and next-generation sequencing of mitochondrial genomes from noninvasively sampled museum specimens. *Genome Res.* **21:** 1695–1704.

Massana, R., J. del Campo, M. E. Sieracki, S. Audic, and R. Logares. 2014. Exploring the uncultured microeukaryote majority in the oceans: reevaluation of ribogroups within stramenopiles. *ISME J.* **8:** 854–66.

Medinger, R., V. Nolte, R. V. Pandey, S. Jost, B. Ottenwälder, C. Schlötterer, and J. Boenigk. 2010. Diversity in a hidden world: potential and limitation of next-generation sequencing for surveys of molecular diversity of eukaryotic microorganisms. *Mol Ecol*. **19 Suppl. 1:** 32–40.

Mikhalevich, V. I. 2013. New insight into the systematics and evolution of the foraminifera. *Micropaleontology* **59:** 493–527.

Minoche, A. E., J. C. Dohm, and H. Himmelbauer. 2011. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol.* **12:** R112.

Moodley, L. 1990. "Squatter" behaviour of soft-shelled foraminifera. *Mar. Micropaleontol.* **16:** 149–153.

Morard, R., F. Quillévéré, G. Escarguel, Y. Ujiie, T. de Garidel-Thoron, R. D. Norris, and C. de Vargas. 2009. Morphological recognition of cryptic species in the planktonic foraminifera *Orbulina universa*. *Mar. Micropaleontol*. **71:** 148–165.

Murray, J. W. 2006. *Ecology and Applications of Benthic Foraminifera.* Cambridge University Press, Cambridge.

Murray, J. W. 2007. Biodiversity of living benthic foraminifera: how many species are there? *Mar. Micropaleontol.* **63:** 163–176.

Naviaux, R. K., B. Good, J. D. McPherson, D. L. Steffen, D. Markusic, B. Ransom, and J. Corbeil. 2005. Sand DNA: a genetic library of life at the water's edge. *Mar. Ecol. Progr. Ser.* **301:** 9–22.

Orsi, W., J. F. Biddle, and V. Edgcomb. 2013. Deep sequencing of subseafloor eukaryotic rRNA reveals active fungi across marine subsurface provinces. *PloS One* **8:** e56335.

Parfrey, L. W., and L. A. Katz. 2010. Genome dynamics are influenced by food source in *Allogromia laticollaris* strain CSH (Foraminifera). *Genome Biol. Evol*. **2:** 678–685.

Parfrey, L. W., D. J. G. Lahr, and L. A. Katz. 2008. The dynamic nature of eukaryotic genomes. *Mol. Biol. Evol.* **25:** 787–794.

Pawłowska, J., F. Lejzerowicz, P. Esling, W. Szczuciński, M. Zajączkowski, and J. Pawłowski. 2014. Ancient DNA sheds new light on the Svalbard foraminiferal fossil record of the last millennium. *Geobiology* **12:** 277–288.

Pawlowski, J. 2000. Introduction to the molecular systematics of foraminifera: *Micropaleontology* **46 Suppl. 1:** 1–12.

Pawlowski, J., and M. Holzmann. 2014. A plea for DNA barcoding of Foraminifera. *J. Foraminiferal Res.* **44:** 62–67.

Pawlowski, J., and B. Lecroq. 2010. Short rDNA barcodes for species identification in foraminifera. *J. Eukaryot. Microbiol.* **57:** 197–205.

Pawlowski, J., I. Bolivar, J. F. Fahrni, C. de Vargas, M. Gouy, and L. Zaninetti. 1997. Extreme differences in rates of molecular evolution of foraminifera revealed by comparison of ribosomal DNA sequences and the fossil record, *Mol. Biol. Evol.* **14:** 498–505.

Pawlowski, J., I. Bolivar, J. Fahrni, C. de Vargas, and S. S. Bowser. 1999. Molecular evidence that *Reticulomyxa filosa* is a freshwater naked foraminifer. *J. Eukaryot. Microbiol.* **46:** 612–617.

Pawlowski, J., J. F. Fahrni, U. Brykczynska, A. Habura, and S. S. Bowser. 2002. Molecular data reveal high taxonomic diversity of allogromiid Foraminifera in Explorers Cove (McMurdo Sound, Antarctica). *Polar Biol*. **25:** 96–105.

Pawlowski, J., M. Holzmann, C. Berney, J. Fahrni, A. J. Gooday, T. Cedhagen, A. Habura, and S. S. Bowser. 2003. The evolution of early Foraminifera, *Proc. Natl. Acad. Sci.* **100:** 11494–11498.

Pawlowski J., J. Fahrni, B. Lecroq, D. Longet, N. Cornelius, L. Excoffier, T. Cedhagen, and A. J. Gooday. 2007. Bipolar gene flow in deep-sea benthic foraminifera. *Mol. Ecol.* **16:** 4089–4096.

Pawlowski, J., W. Majewski, D. Longet, J. Guiard, T. Cedhagen, A. J. Gooday, S. Korsun, A. Habura, and S. S. Bowser. 2008. Genetic differentiation between Arctic and Antarctic monothalamous foraminifera. *Polar Biol*. **31:** 1205–1216.

Pawlowski, J., D. Fontaine, A. Aranda da Silva, and J. Guiard. 2011a. Novel lineages of Southern Ocean deep-sea foraminifera revealed by

environmental DNA sequencing. *Deep-Sea Research II* **58:** 1996–2003.

Pawlowski, J., R. Christen, B. Lecroq, D. Bachar, H. R. Shahbazkia, L. Amaral-Zettler, and L. Guillou. 2011b. Eukaryotic richness in the abyss: insights from pyrotag sequencing. *PLoS One* **6:** e18169.

Pawlowski, J., S. Audic, S. Adl, D. Bass, L. Belbahri, C. Berney, S. S. Bowser, I. Cepicka, J. Decelle, M. Dunthorn *et al.* 2012. CBOL Protist Working Group: barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLoS Biol.* **10:** e1001419.

Pawlowski, J., M. Holzmann, and J. Tyszka. 2013. New supraordinal classification of Foraminifera: molecules meet morphology. *Mar. Micropaleontol.* **100:** 1–10.

Pawlowski, J., P. Esling, F. Lejzerowicz, T. Cedhagen, and T. A. Wilding. 2014. Environmental monitoring through protist NGS metabarcoding: assessing the impact of fish farming on benthic foraminifera communities. *Mol. Ecol. Res.* **14:** 1129–1140.

Pernice, M. C., R. Logares, L. Guillou, and R. Massana. 2013. General patterns of diversity in major marine microeukaryote lineages. *PLoS One* **8:** e57170.

Pillet, L., D. Fontaine, and J. Pawlowski. 2012. Intra-genomic rRNA polymorphism and morphological variation in *Elphidium macellum* suggest inter-specific hybridization in foraminifera. *PLoS One* **7:** e32373.

Pillet, L., I. Voltski, and J. Pawlowski. 2013. Molecular phylogeny of Elphidiidae (Foraminifera). *Mar. Micropaleontol.* **103:** 1–14.

Prokopowich, C. D., T. R. Gregory, and T. J. Crease. 2003. The correlation between rDNA copy number and genome size in eukaryotes. *Genome* **46:** 48–50.

Schnitker, D. 1974. Ecotypic variation in *Ammonia beccarii* (Linné): *J. Foraminiferal Res*. **4:** 217–223.

Sen Gupta, B. K. 1999. Systematics of modern Foraminifera. Pp. 7–36 in *Modern Foraminifera*, B. K. Sen Gupta, ed. Kluwer Academic Publishers, Dordrecht.

Shi, X. L., C. Lepère, D. J. Scanlan, and D. Vaulot. 2011. Plastid 16S rRNA gene diversity among eukaryotic picophytoplankton sorted by flow cytometry from the South Pacific Ocean. *PLoS One* **6:** e18979.

Stern, R. F., A. Horak, R. L. Andrew, M.-A. Coffroth, R. A. Andersen, F. C. Küpper, I. Jameson, M. Hoppenrath, B. Véron, F. Kasai *et al.* 2010. Environmental barcoding reveals massive dinoflagellate diversity in marine environments. *PLoS One* **5:** e13991.

Stevens, J. L., R. L. Jackson, and J. B. Olson. 2013. Slowing PCR ramp speed reduces chimera formation from environmental samples. *J. Microbiol. Methods* **93:** 203–205.

Stoeck, T., A. Zuendorf, H. W. Breiner, and A. Behnke. 2007. A molecular approach to identify active microbes in environmental eukaryote clone libraries. *Microb. Ecol.* **53:** 328–339.

Tsuchiya, M., H. Kitazato, and J. Pawlowski, 2003. Analysis of Internal Transcribed Spacer of ribosomal DNA reveals cryptic speciation in *Planoglabratella opercularis*: *J. Foraminiferal Res*. **33:** 285–293.

van Velzen, R., E. Weitschek, G. Felici, and F. T. Bakker. 2012. DNA barcoding of recently diverged species: relative performance of matching methods. *PLoS One* **7:** e30490.

Vlassov, V. V., P. P. Laktionov, and E. Y. Rykova. 2007. Extracellular nucleic acids. *Bioessays* **29:** 654–667.

Weber, A. A.-T., and J. Pawlowski. 2013. Can abundance of protists be inferred from sequence data: a case study of Foraminifera. *PLoS One* **8:** e56739.

Weber, A. A.-T., and J. Pawlowski. 2014. Wide occurrence of SSU rDNA intragenomic polymorphism in Foraminifera and its implications for molecular species identification. *Protist* **165:** 645–661.

Zhan, A., M. Hulak, F. Sylvester, X. Huang, A. A. Adebayo, C. L. Abbott, S. L. Adamowicz, D. D. Heath, M. E. Cristescu, and H. J. MacIsaac. 2013. High sensitivity of 454 pyrosequencing for detection of rare species in aquatic communities. *Methods Ecol. Evol*. **4:** 558–565.

Zhang, T., and H. H. P. Fang. 2006. Applications of real-time polymerase chain reaction for quantification of microorganisms in environmental samples. *Appl. Microbiol. Biotechnol.* **70:** 281–28.

Zhu, F., R. Massana, F. Not, D. Marie, and D. Vaulot. 2005. Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiol. Ecol.* **52:** 79–92.