PFR^2: a curated database of planktonic Foraminifera18S ribosomal DNA as a resource for studies of plankton ecology, biogeography, and evolution

Raphaël Morard, Kate F Darling, Frédéric Mahé, Stéphane Audic, Yurika Ujiié, Agnes K.F. Weiner, Aurore André, Heidi Seears, Chris M Wade, Frédéric Quillévéré, et al.

PFR²: a curated database of planktonic foraminifera18S ribosomal DNA as a resource for studies of plankton ecology, biogeography, and evolution.

Raphaël Morard[1,2,3], Kate F. Darling[4, 5], Frédéric Mahé[6], Stéphane Audic[1,2], Yurika Ujiié[7], Agnes K. M. Weiner[3], Aurore André[8,9], Heidi A. Seears[10,11], Chris M. Wade[10], Frédéric Quillévéré[8], Christophe J. Douady[12, 13],Gilles Escarguel[8], Thibault de Garidel-Thoron[14], Michael Siccha[3], Michal Kucera[3] and Colomban de Vargas[1,2]

[1]*Centre National de la Recherche Scientifique, UMR 7144, EPEP, Station Biologique de Roscoff, 29680 Roscoff, France*

[2]*Sorbonne Universités, UPMC Univ Paris 06, UMR 7144, Station Biologique de Roscoff, France*

[3]*MARUM Center for Marine Environmental Sciences, University of Bremen, Leobener Strasse, 28359 Bremen, Germany*

[4]*School of GeoSciences, University of Edinburgh, Edinburgh EH9 3JW, UK*

[5]*School of Geography and GeoSciences, University of St Andrews, Fife KY16 9AL, UK*

[6]*Department of Ecology, Technische Universität Kaiserslautern, 67663 Kaiserslautern, Germany*

[7]*Department of Biology, Shinshu University, Asahi3-1-1, Matsumoto, Nagano 390-8621, Japan*

[8]*CNRS UMR 5276, Laboratoire de Géologie de Lyon: Terre, Planètes, Environnement, Université Claude Bernard Lyon 1, 69622 Villeurbanne, France*

[9]*Université de Reims-Champagne-Ardenne, UFR Sciences Exactes et Naturelles, Campus Moulin de la Housse, Batiment 18, 51100 REIMS, France*

[10]*School of Life Sciences, University of Nottingham, University Park, Nottingham, NG7 2RD, UK*

[11]*Department of Biological Sciences, Lehigh University, Iacocca Hall, 111 Research Drive, Bethlehem, Pennsylvania, 18105, USA*

[12]*Université de Lyon ; UMR5023 Ecologie des Hydrosystèmes Naturels et Anthropisés ; Université Lyon 1 ; ENTPE ; CNRS ; 6 rue Raphaël Dubois, 69622 Villeurbanne, France.*

[13]*Institut Universitaire de France, 103 Boulevard Saint-Michel, 75005 Paris, France*

[14]*Aix-Marseille Université, CNRS, CEREGE UM34, Technopôle de l'Arbois, 13545 Aix-en-Provence, France*

Keywords: Planktonic foraminifera, 18S ribosomal DNA, molecular ecology, genetic diversity, molecular taxonomy, sequence database.

Corresponding Author: Raphaël Morard, MARUM Center for Marine Environmental Sciences, University of Bremen, Leobener Strasse, 28359 Bremen, Germany, Fax: +49 (0) 421 218 - 9865974, rmorard@marum.de.

Abstract

Planktonic foraminifera (Rhizaria) are ubiquitous marine pelagic protists producing calcareous shells with conspicuous morphology. They play an important role in the marine carbon cycle and their exceptional fossil record serves as the basis for biochronostratigraphy and past climate reconstructions. A major worldwide sampling effort over the last two decades has resulted in the establishment of multiple large collections of cryopreserved individual planktonic

foraminifera samples. Thousands of 18S rDNA partial sequences have been generated, representing all major known morphological taxa across their worldwide oceanic range. This comprehensive data coverage provides an opportunity to assess patterns of molecular ecology and evolution in a holistic way for an entire group of planktonic protists. We combined all available published and unpublished genetic data to build PFR², the *Planktonic Foraminifera Ribosomal Reference* database. The first version of the database includes 3,322 reference 18S rDNA sequences belonging to 32 out of the 47 known morphospecies of extant planktonic Foraminifera, collected from 460 oceanic stations. All sequences have been rigorously taxonomically curated using a six-rank annotation system fully resolved to the morphological species level and linked to a series of metadata. The PFR² website, available at http://pfr2.sb-roscoff.fr, allows downloading the entire database or specific sections, as well as the identification of new planktonic foraminiferal sequences. Its novel, fully documented curation process integrates advances in morphological and molecular taxonomy. It allows for an increase in its taxonomic resolution and assures that integrity is maintained by including a complete contingency tracking of annotations and assuring that the annotations remain internally consistent.

Introduction

Despite their ubiquity and the critical role they play in global biogeochemical cycles, unicellular eukaryotes (protists) remain the most poorly known domain of life (e. g., Pawlowski et al., 2012). Because of their extreme morphological and behavioral diversity, the study of even relatively narrow lineages requires a high degree of taxonomic expertise (e. g., Guillou et al., 2012, Pawlowski and Holzmann, 2014). As a result, the knowledge of protistan ecology and evolution is limited by the small number of taxonomists, resulting in scarcity of taxonomically

well-resolved ecological data. As an alternative approach, numerous studies have demonstrated the potential of identification of protists by means of short DNA sequences or barcodes (e. g., Saunders, 2005; Sherwood et al., 2007; Hollingsworth et al., 2009; Nossonova et al., 2010; Pawlowski and Lecroq, 2010; Hamsher et al., 2011; Stern et al., 2010; Schoch et al., 2012), both at the single-cell and metacommunity levels (e. g., Sogin et al., 2006; Logares et al., 2014). Such barcoding/metabarcoding approaches critically rely on the fidelity of the marker gene with respect to specificity (avoiding ambiguity in identification), comprehensiveness (assuring all taxa in the studied group are represented in the reference barcode database) and accuracy (assuring that barcode assignments are consistent with a coherent, phenotypic taxonomic framework; e. g., Zimmermann et al., 2014)). These three pre-requisites are rarely found in protists, where classical morphological taxonomy is often challenging, DNA extraction and sequencing from a single cell is prone to contamination, and a large portion of the diversity in many groups remains unknown (e. g., Mora et al., 2011). In this respect, planktonic foraminifera represent a rare exception.

Planktonic foraminifera are ubiquitous pelagic marine protists with reticulated pseudopods, clustering within the Rhizaria (Nikolaev et al., 2004). The group is marked by a rather low number of extant morphospecies (47; Hemleben et al., 1989), which can be distinguished using structural characteristics of their calcite shells. Their global geographic distribution, seasonal dynamics, vertical habitats and trophic behavior have been thoroughly documented by analyses of plankton hauls (e.g., Bé and Hudson, 1977), sediment trap series (e.g., Zaric et al., 2005) and thousands of surface sediment samples across the world oceans (e.g., Kucera et al., 2005). Their outstanding preservation in marine sediments resulted in arguably the most complete fossil record, allowing comprehensive reconstruction of the evolutionary history

of the group (Aze et al., 2011). Over the last two decades, the morpho-taxonomy and phylogeny of the group have been largely confirmed by molecular genetic analyses (e.g., Aurahs et al., 2009a) based on the highly informative, ~1,000 bp fragment at the 3'end of the 18S rDNA gene. These analyses confirmed that the morphological characters used to differentiate planktonic foraminifera taxa are phylogenetically valid both at the level of morphological species and at the level of higher taxa. The studied gene fragment contains six hypervariable expansion segments, some unique to foraminifera, providing excellent taxonomic resolution (Pawlowski and Lecroq, 2010). Analyses of this fragment revealed the existence of genetically distinct lineages within most of the morphospecies, which likely represent reproductively isolated units (Darling et al., 1996, 1997, 1999, 2000, 2003, 2004, 2006, 2007, 2009; Darling and Wade, 2008; Wade et al., 1996; de Vargas et al., 1997, 1999, 2001, 2002, de Vargas and Pawlowski, 1998; Stewart et al., 2001; Aurahs et al., 2009b, 2011; Ujiié and Lipps, 2009; Ujiié et al., 2008, 2012; Morard et al., 2009, 2011, 2013; Seears et al., 2012; Quillévéré et al., 2013; Weiner et al., 2012, 2014; André et al., 2014). In order to assess the ecology and biogeography of such cryptic species, large numbers of rDNA sequences from single-cell extractions collected across the world oceans have been generated for most morphospecies (Figure 1). Due to this extensive single-cell rDNA sequencing, the genetic and morphological diversity of planktonic foraminifera have been linked together to a degree that now allows for transfer of taxonomic expertise. The knowledge of the genetic and morphological taxonomy of the group allows the establishment of an exceptionally comprehensive reference genetic database that can be further used to interpret complex data from plankton metagenomic studies with a high level of taxonomic resolution. Because planktonic foraminifera are subject to the same ecological forcing as other microplankton, including the dominance of passive transport in a relatively unstructured environment, huge population sizes,

and basin-scale distribution of species, they can potentially serve as a model for the study of global ecological patterns in other groups of pelagic protists, whose diversity remains largely undiscovered (Mora et al., 2011).

By early 2014, 1,787 partial 18S rDNA sequences from single-cell extractions of planktonic foraminifera were available in public databases. However, their NCBI taxonomy is often inconsistent, lacking standardization. It includes (and retains) obvious identification errors, as discussed by Aurahs et al. (2009a) and André et al. (2014), and their annotation lacks critical metadata. In addition, an equivalent number of rDNA sequences not deposited in public databases have been generated by the co-authors of the present study. Collectively, the existing rDNA sequences from single cells collected throughout the world oceans cover the entire geographic and taxonomic range of planktonic foraminifera. This collection unites the current morphological, genetic, ecological, and biogeographic knowledge of the group and may serve as a *Rosetta Stone/Philae Obelisk* for interpreting metabarcoding data (Pawlowski et al., 2014). To pave the way for future exploitation of this resource, we combined all published and unpublished planktonic foraminifera rDNA sequence data and curated the resulting database with a semi-automated bioinformatics pipeline. The resulting *Planktonic Foraminifera Ribosomal Reference* database (PFR^2) is a highly resolved, fully annotated and internally entirely consistent collection of 18S rDNA sequences of planktonic foraminifera, aligned and evaluated in a way that facilitates, among others, direct assessment of barcoding markers.

Material and Methods

Primary database assembly

A total of 1,787 18S rDNA sequences of planktonic foraminifera were downloaded from the GenBank query portal (http://www.ncbi.nlm.nih.gov/; release 201) on the 14th of May 2014. The taxonomic path and metadata for these sequences were extracted from NCBI and supplemented by information in original papers when available. The metadata associated to each sequence consisted of: (i) their organismal origin (specimen voucher, taxonomic path, infra specific genetic type assignment), (ii) their methodological origin (direct sequencing or cloning), and (iii) their spatio-temporal origin (geographic coordinates, depth, and time of collection). Metadata were described using standard vocabularies and data formats. For 47 sequences, the coordinates of the collection site could not be recovered, in which case the locality was described in words (Supplementary Material 1).

We next compiled all unpublished 18S rDNA sequences generated by the co-authors of this paper and linked them with the same suite of metadata. These sequences originate from single-cell extractions of planktonic foraminifera collected by stratified or non-stratified plankton net hauls, in-situ water pumping, as well as SCUBA diving. After collection, the specimens were individually picked under a stereomicroscope, cleaned, taxonomically identified and transferred into DNA extraction buffer or air-dried on cardboard slides and stored at -20°C or -80°C. DNA extractions were performed following the DOC (Holzmann & Pawlowski, 1996), the GITC* (Morard et al., 2009), or the Urea (Weiner et al., 2014) protocols. Sequences located at the 3' end of the 18S rDNA were obtained following the methodology described in de Darling et al. (1996, 1997), de Vargas et al. (1997), Aurahs et al. (2009b), Morard et al. (2011) and Weiner et al., (2014). A total of 820 new planktonic foraminiferal sequences were analyzed and annotated for this study. In addition, 925 unpublished sequences analyzed in Darling et al. (2000, 2003, 2004, 2006, 2007), Darling and Wade (2008), Seears et al. (2012), and Weiner et al. (2014) were also

included. All unpublished sequences, except 177 sequences shorter than 200 bp, were deposited in GenBank under the accession numbers KM19301 to KM194582. Overall, PFR^2 contains data from 460 sites sampled during 54 oceanographic cruises and 15 near shore collection campaigns between 1993 and 2013. It covers all oceanic basins, all seasons, and water depths ranging between the surface and 700 meters (Figure 1; Supplementary Material 1).

Taxonomy

Morphological taxonomy

As the first step in the curation process, the primary taxonomic annotations of all 3,532 18S rDNA sequences gathered from NCBI and our internal databases were harmonized. The identification of planktonic foraminifera is challenging especially for juvenile individuals, which often lack diagnostic characters (Brummer et al., 1986). Thus, many of the published and unpublished 18S rDNA sequences were mislabeled or left in open nomenclature. In some cases the same taxon has been recorded under different names, reflecting inconsistent use of generic names, synonyms and misspelling. To harmonize the taxonomy, we first carried out a manual curation of the original annotations to remove the most obvious taxonomic conflicts in the primary database. To this end, the sequence annotations were aligned with a catalog of 47 species names based on the taxonomy used in Hemleben et al. (1989), but adding *Globigerinoides elongatus* following Aurahs et al. (2011) and treating *Neogloboquadrina incompta* following Darling et al. (2006). Thus, the 109 sequences labelled as *Globigerinoides ruber* (pink) and the 63 labelled as *Globigerinoides ruber* (white) were renamed as *Globigerinoides ruber*. The 113 sequences of *Globigerinoides ruber* and *Globigerinoides ruber* (white) attributed to the genotype II were renamed *Globigerinoides elongatus* following Aurahs

et al. (2011). The 12 sequences labelled *Globigerinella aequilateralis* were renamed *Globigerinella siphonifera* following Hemleben et al. (1989). The 7 sequences corresponding to the right-coiled morphotype of *Neogloboquadrina pachyderma* were renamed *Neogloboquadrina incompta* following Darling et al. (2006). All taxonomic reassignments were checked by sequence similarity analyses to the members of the new group. Next, we attempted to resolve the attribution of sequences with unresolved taxonomy and searched manually for obviously misattributed sequences. This refers to sequences that are highly divergent from other members of their group but identical to sequences of other well-resolved taxa. Overall, these first steps of manual curation led to the taxonomic reassignment of 124 sequences. All corrections and their justification are documented in the Supplementary Material 1.

Annotation of genetic types

In order to preserve the information on the attribution of 18S rDNA sequences to genetic types (potential cryptic species), we harmonized the existing attributions at this level for species where extensive surveys have been carried out and published. A total of 1,356 sequences downloaded from NCBI were associated with a genetic type label, which was always retained. In addition, 19 sequences labelled as *Globigerinoides ruber*, 15 as *Globigerinoides sacculifer*, 36 as *Globigerinita glutinata*, 6 as *Globigerinita uvula*, 9 as *Globorotalia inflata*, 10 as *Neogloboquadrina incompta*, 6 as *Neogloboquadrina pachyderma*, 5 as *Orbulina universa*, 5 as *Pulleniatina obliquiloculata*, 30 as *Hastigerina pelagica*, and 32 as *Globigerinella siphonifera* have been analyzed after their first release in the public domain by Aurahs et al. (2009), Ujiié et al. (2012), Weiner et al. (2012, 2014), and André et al. (2013, 2014), and were attributed to a genetic type by these authors. These attributions differ from those in the NCBI label, but were retained in the PFR² database. In case of multiple attributions of the same sequence to different

genetic types by several authors, we retained the molecular taxonomy that was based on the study presenting the most resolved and comprehensive attribution. In addition, 877 unpublished sequences belonging to *Orbulina universa*, *Globigerina bulloides*, *Neogloboquadrina incompta*, *Neogoboquadrina dutertrei*, *Neogloboquadrina pachyderma,* and *Turborotalita quinqueloba* received a genotypic attribution following de Vargas et al. (1999) and Darling et al. (2004, 2006, 2007, 2008). Most of these sequences have been produced and identified within earlier studies, but were not originally deposited on NCBI. Their PFR² genotypic assignment is therefore entirely consistent with the attribution of the representative sequences of the same genetic type that were deposited on NCBI.

PFR^2 final taxonomic framework

As a result of the first manual curation and annotation to the genetic type level, the original 3,532 18S rDNA sequences were re-assigned to 33 species names and 2,276 sequences were annotated to the level of genetic types (Supplementary Material 1). For all sequences, we established a ranked taxonomy with six levels: 1- Morphogroup, 2-Genus, 3-Species, 4-Genetic type level 1, 5-Genetic type level 2, 6-Genetic type 3. For the "Morphogroup" rank we used the taxonomical framework of Hemleben et al. (1989), dividing the extant planktonic foraminifera species into five clades based on the ultrastructure of the calcareous shell: Spinose, Non-spinose, Microperforate, Monolamellar and Non-spiral. The "Genus" and "Species" ranks follow the primary annotation as described above. For the "Genetic type level 1", "Genetic type level 2" and "Genetic type level 3" ranks, we used the hierarchical levels presented in the labels of the genetic types of *Globigerinoides ruber*, *Globigerinoides elongatus*, *Globigerinella siphonifera*, *Globigerinella calida*, *Hastigerina pelagica*, *Globigerina bulloides*, *Neogloboquadrina dutertrei*, *Pulleniatina obliquiloculata,* and *Turborotalita quinqueloba*. Genetic type attributions lacking

hierarchical structure were reported in the rank "Genetic type level 1". After this step, the Primary Reference Database (Figure 2) of 3,532 sequences contained 113 different taxonomic paths (Supplementary Material 1).

Sequences partitioning into conserved and variable regions

Because PFR² is a resource not only for taxonomic assignment but also for ecological and biogeographical studies, all planktonic foraminiferal 18S rDNA sequences were included irrespective of length, as long as they contained taxonomically relevant information. As a result, the length of the sequences included in the annotated primary database ranges between 33 and 3,412 bp. To evaluate their coverage and information content, all sequences were manually aligned using Seaview 4 (Gouy et al., 2010) to the borders of each variable region of the 18S rDNA fragment. The positions of the borders were determined according to the SSU rDNA secondary structure of the monothalamous foraminifera *Micrometula hyalostera* presented by Pawlowski and Lecroq (2010), except for the region 37/f where a strict homology was difficult to establish for all sequences. Instead, we defined the end of this region by the occurrence of a pattern homologous to the series of nucleotides "CUUUCACAUGA" located at the 3' end of Helix 37. We also noticed that the short conserved fragment located between the variable regions 45/e and 47/f was difficult to identify across all sequences. We thus merged the regions 45/e, 46 and 47/f into a single region that we named 45E-47F (Table1). As a result, the position and length of six conserved (32-37, 37-41, 39-43, 44-45, 47-49, 50) and five variable (37F, 41F, 43E, 45E-47F, 49E) regions were identified for all sequences (Figure 2). The remaining part of the 18S rDNA sequence, only present in sequences EU199447, EU199448 and EU199449 and located before the motive "AAGGGCACCACAAGA" has not been analyzed in this way. All regions fully covered in a sequence and containing sequence motives observed at least twice in

247 the whole dataset were labelled as “complete”. Regions fully covered but containing a sequence
248 motive that was observed only once in the whole dataset were labelled as “poor”. This is because
249 we consider sequencing/PCR errors as the most likely cause for the occurrence of such unique
250 sequence motives. We realize that using this procedure, even genuine unique sequences may be
251 discarded from the analysis, but this would be the case only if such sequences deviated in all
252 regions. In all other cases, the regions were labelled as “partial” when only a part of the region
253 was present or “not available” if they did not contain any fragment of the sequence. As a result
254 we obtain the Partitioned Primary Reference Database (Figure 2). The coverage of each
255 individual region in the Partitioned Primary Reference Database is given in Supplementary
256 Material 1, and all sequence partitions are given in Supplementary Material 2.

257 *Semi-automated iterative curation pipeline for optimal taxonomic assignment*

258 The consistency of taxonomic assignments within the annotated database of partitioned
259 sequences was assessed using a semi-automated process (Figures 2 and 3). All “complete”
260 regions of sequences with the same taxonomic assignment at the morphospecies level were
261 automatically aligned using global pairwise alignment (Needleman & Wunsch 1970), as
262 implemented in the software *needle* from the Emboss suite of bioinformatics tools (Rice et al.,
263 2000). To detect annotation inconsistencies, mean pairwise similarities were computed for each
264 “complete” region of each sequence against all other sequences with the same taxonomic
265 assignment from the finest annotation level “Genetic type level 3” up to the “Species level” rank.
266 Results are provided in Supplementary Material 1 and were visualized using R (R Development
267 Core Team, 2014) and the ggplot2 library (Wickham, 2009). The resulting plots are given in
268 Supplementary Material 3. If all annotations are consistent and there is no variation within taxa,
269 each sequence within the analyzed taxon should only find an exact match and the mean pairwise

similarity for that taxon should be 1. However, beyond sequencing/PCR errors introducing spurious sequence differences, there are several reasons why the mean pairwise similarity within a taxon may be lower. First, if a sequence has been assigned the wrong name, its similarity to all other sequences labelled with that name will be low, thus decreasing the resulting mean pairwise similarity. Second, if a sequence has been assigned to the correct taxon, but the taxon comprises multiple sequence motives, that sequence will find a perfect match within the taxon but the mean pairwise similarity will also be lower than 1.

In order to deconvolve the different sources of sequence variability within taxa, we followed a three-step iterative approach, which was repeated for each of the 11 ”complete” regions of the analyzed SSU rDNA fragment. First, we considered the distribution of mean pairwise similarities for all sequences within each region assigned to one taxon at the finest rank of “Genetic type level 3”. Assuming that misidentifications are rare and result in large pairwise distances, we manually searched for sequences whose mean pairwise similarity deviates substantially from the rest of the sequences within the taxon. Such sequences were initially “invalidated”, whereas all other sequences analyzed at this level were “validated”. We then repeated the same procedure for the higher ranks of “Genetic type level 2”, “Genetic type level 1” and finally “Species level”, always starting with the full database (Figures 2 and 3A). Thus, at each level, we expected a misidentified sequence to have a pairwise similarity markedly lower than the mean of pairwise similarities between correctly assigned sequences (Figure 3B). This procedure had to be repeated for every rank, because not all sequences in the database are assigned to all ranks. Nevertheless, once “validated”, a sequence cannot be “invalidated” during analyses of higher rank taxa, because it represents an accepted variability within that taxon. In taxa where all sequences within

a region show low mean pairwise similarities all attributions are initially invalidated (this would be typically the case for a “wastebasket taxa”; Figure 3C).

In the second step, all sequences invalidated during step 1 were reconsidered based on their pairwise similarities with ‘validated’ sequences from the same region. The main goal of the curated taxonomy being to achieve correct taxonomic assignment at the species level, the pairwise comparison was carried out at this rank. If the best match is a ”validated” sequence with the same initial species attribution as the invalidated sequence, this sequence is “validated” at the species level and its assignment at the “genetic type” level is then deleted. Such a situation can only occur when the sequence was initially assigned to the wrong genetic type within the correct species. If the pairwise comparisons of all regions analyzed match sequences with different (but consistent) species attributions than the invalidated sequence, the sequence is reattributed to that species. If the pairwise comparisons indicate that the analyzed sequence has no close relative in the validated part of the database, the initial attribution is retained, provided that the initial attribution is not yet in the validated dataset. This case occurs when all sequences of one species have been initially invalidated because the same species name was associated with highly divergent sequences. When the sequence has no close relative but its initial attribution is represented in the validated part of the dataset, the initial attribution is discarded and the sequence receives an artificial attribution derived from the nearest higher rank that matches the pairwise comparisons. In all cases, the erroneous attributions are replaced by the corrected ones in the database (Figure 2, Supplementary Material 1).

In the third step, sequences that received new attributions were reanalyzed as described in step 1. If inconsistencies in the distribution of mean pairwise similarities remain, steps 2 and 3 are repeated until no inconsistency is observed.

As a final diagnosis we performed leave-one-out analyses to evaluate the robustness and potential limitations of the curated taxonomy, as well as a monophyly validation by Neighbor-Joining using only sequences that are covering the 6 conserved and 5 variable regions of the 5' end fragment. First, each individual sequence included in the first version of PFR² was blasted against the remaining part of the database including n-1 sequences using SWIPE (Rognes, 2011). The sequences among the "n-1 PFR² database" returning the highest score were retrieved and their taxonomic attribution compared to the one of the blasted sequence (Supplementary Material 1). Second, we retrieved all sequences covering the 5 variable and 6 conserved regions and divided them according to their assignment to higher taxa (here simplified by the morphogroups Monolamellar, Non-Spinose, Spinose, and Microperforates + Benthic). Each subset was automatically aligned using MAFFT v.7 (Katoh and Standley., 2013) and the subsequent alignments were trimmed off on the edges to conserve only homologous position, finally leading to 41, 583, 271, and 100 analyzed sequences for the Monolamellar, Non-Spinose, Spinose, and Microperforates + Non-spiral morphogroups, respectively. For each alignment, a tree was inferred using a Neighbor-Joining approach with Juke and Cantor distance while taking into account gap sites as implemented in SEAVIEW 4 (Supplementary Material 4) with 100 pseudo-replicates. The scripts used to perform the different curation steps are available as Supplementary Material 5.

Results

Of the 3,532 planktonic foraminiferal 18S rDNA partial sequences analyzed, 3,347 (94.8%) contained at least one "complete" gene region making possible the curation process. The remaining 185 sequences included 33 singletons (rare motives or poor quality sequences) and 152 sequences that were too short to cover at least one region (Supplementary Material 1).

Amongst the 3,347 curated sequences, the taxonomic assignment of 84 was initially invalidated. Of these, 3 represent cases where the morphospecies attribution was correct, but the attribution to a genetic type was erroneous. In 46 cases, the invalidated sequences found a perfect match with a different taxon and thus their taxonomic assignment was changed. In all of these cases, the novel taxonomic assignment corresponded to a morphologically similar morphospecies, explaining the original misidentification of the sequenced specimen. In 14 cases, the original assignment was retained because the sequences did not find any match and their original attribution did not appear in the validated part of the dataset. All of these sequences were labelled as *Hastigerinella digitata*. This species name had been entirely invalidated in the first step because of inconsistent use of the homonymous species name *Beella digitata*. Finally, 17 sequences received an unresolved artificial assignment. These represent six different sequence motives diverging substantially from all sequences in the validated part of the database and also between each other. Because the original attribution upon collection was obviously wrong, we could not reassign these sequences to the species level. In two cases, we could identify the most likely generic attribution, but four sequences are left with an entirely unresolved path. Finally, our procedure captured one sequence with a spelling error in its path and three sequences that appear to have been attributed correctly but represent small variants within species. After resolution of the 84 conflicts described above, the re-annotated dataset was subjected to a second round of the curation process for verification. All sequences were validated.

Based on this internally consistent taxonomic annotation for all 3,347 18S rDNA sequences from individual planktonic foraminifera, we generated the *Planktonic Foraminiferal Ribosomal Reference* or PFR^2 database. Of the 3,347 sequences, 25 were shorter than 200 bp, and could not be deposited in NCBI (see Supplementary Material 1). The $PFR^2$1.0 database thus includes 3,322

reference sequences assigned to 32 morphospecies and 6 taxa with unresolved taxonomy (Figure 2), and contains 119 unique taxonomic paths when including all three levels of genetic types.

The leave-one-out BLAST evaluation applied on the first version of PFR² to assess its robustness returned an identical taxonomic path for 2,509 sequences. For 614 sequences, the BLAST-determined taxonomic paths were identical between the "morphogroup" and "species" rank but displayed a different resolution between the ranks "genetic type level 1" and "genetic type level 3". This reflects a situation where some sequences belonging to one species are annotated to the level of a genetic type, whereas others are not. Finally, 19 sequences were assigned to the correct species but to a different genetic type. This illustrates the case of genetic types represented by only one sequence in the database, which were logically assigned to the closest genetic type within the same species by the leave-one-out procedure. Thus, 94.5 % of the sequences in the PFR^2 database find a nearest neighbor with a correct taxonomic assignment at the species target level. For the remaining 180 sequences, the returned taxonomic path was inconsistent at the species level. In two cases, the sequences were assigned to a morphologically and phylogenetically close sister species (*Globorotalia ungulata* and *Globorotalia tumida*), reflecting insufficient coverage in the database for these species. Two cases involved singleton sequences with unresolved taxonomy, which find no obvious nearest neighbor. Finally, 176 cases of inconsistent identification refer to sequences of *Globigerinella calida* and *Globigerinella siphonifera*, whose species names have been used interchangeably in the literature (Weiner et al., 2014) and the clade has been shown to be in need of a taxonomic revision (Weiner et al., 2015). The leave-one-out evaluation thus reveals excellent coverage of PFR^2 and confirms that the curated taxonomy is internally entirely consistent.

To further confirm the validity of morphospecies level taxonomy, we constructed NJ trees for the five clades including only the long sequences (Supplementary Material 4). This analysis confirmed the monophyly of all morphospecies, except the *Globigerinella calida/Globigerinella siphonifera* plexus. All clades were strongly supported except for the sister species *Globorotalia tumida* and *Globorotalia ungulata* and the monolamellar species *Hastigerina pelagica* and *Hastigerinella digitata*. In the first case, the poor support reflects the lack of differentiation between these two species in the conserved region of the gene, thus decreasing the bootstrap score; in the second case the extreme divergence of two genetic lineages of *Hastigerina pelagica* renders the phylogenetic reconstruction difficult (Weiner et al., 2012).

An analysis of the taxonomic annotations retained in PFR^2 reveals that the database covers at least 70-80% of the traditionally recognized planktonic foraminiferal species in each clade. The species represented in PFR^2 constitute the dominant part of planktonic foraminifera assemblages in the world oceans. Compared with a global database of census counts from surface sediments (MARGO database, Kucera et al., 2005), the species covered by PFR^2 account for >90% of tests larger than 150 μm found in surface sediments (Figure 4). In cold and temperate provinces, PFR^2 species account for almost the entire assemblages, while in warmer subtropical and tropical waters, only up to 4% of the sedimentary assemblages are not represented in PFR^2. Evidently, PFR^2 reference sequences cover most of the ecologically relevant portion of the morphological diversity and the taxa that are not yet represented in PFR^2 are small, rare or taxonomically obscure. It is possible that some of these taxa may correspond to the six sequences with still unresolved taxonomy. If so, PFR^2 may be considered to cover up to 38 of the 47 recognized species.

405 Finally, for each species present in PFR^2, we evaluated the ecological coverage of the global
406 sampling effort (Figure 4). Morphospecies of planktonic foraminifera are known to be
407 distributed zonally across the world oceans, reflecting the latitudinal distribution of sea surface
408 temperature (e. g., Bé and Tolderlund, 1971). A comparison between the temperature range of
409 each species as indicated by their relative abundance in surface sediment samples (Kucera et al.,
410 2005) and the temperatures measured at sampling localities shows that a large portion of the
411 ecological range of the species is covered by the reference sequences in PFR^2 (Figure 4).

412 *The PFR^2 web interface*

413 To facilitate data download and comparative sequence analyses, PFR² has been implemented into
414 a dedicated web interface, available at http://pfr2.sb-roscoff.fr. The website provides:

415 (1) a search/browse module, which allows the user to download parts of the database either by
416 taxonomic rank (morphogroup name, genus name, species name), geographic region (e. g.,
417 North Atlantic, Mediterranean Sea, Indian Ocean) or collection (cruise name) ;
418 (2) a classical BLAST/Similarity module that facilitates identification of unknown sequences;
419 (3) a map module displaying the localities for all sequences present in PFR² and facilitating
420 download of all data from each single locality;
421 (4) a download section with direct access to all data included in PFR². All sequences and
422 sequence partitions are available in FASTA format and the metadata are available in a
423 tabulated file.

424 **Discussion**

425 Comprehensive databases of ribosomal RNA sequences with curated taxonomy are available for
426 Protists (Protist ribosomal reference database, PR²; Guillou et al., 2013) and for the major

427 domains of life (SILVA; Yilmaz et al., 2013). These databases include sequences of planktonic
428 foraminifera. However, they are used mainly as benchmarks to annotate complex environmental
429 datasets (e.g., Logares et al., 2014) at the morphological species level. In contrast, PFR² has been
430 designed and implemented in a way that facilitates other applications.

431 First, because of structural limitations PR² contains "only" 402 sequences of planktonic
432 foraminifera (based on Released 203 of GenBank, October 2014), compared to PFR², which
433 contains for now 3,322 SSU rDNA sequences. Second, 2276 of the sequences present in PFR²
434 have an assignation to the genetic type level and as far as possible, the sequences are associated
435 with metadata related to the origin of each specimen and the conditions where it was collected,
436 thus forming a basis for ecological modelling. Third, most importantly, using planktonic
437 foraminifera as a case study, we propose and implement an annotation scheme with unmatched
438 accuracy and full tracking of changes. This is only possible because of the narrower focus of
439 PFR² combined with high-level expert knowledge of their taxonomy. The fidelity of the
440 annotations will facilitate a qualitatively entirely different level of analysis of eDNA libraries.

441 For example, the design of PFR² allows to incorporate advances in classical and molecular
442 taxonomy, particularly at the level of genetic types (e.g., André et al., 2014), which can be re-
443 evaluated depending of the criteria used to delineate molecular OTUs. Further, by retaining
444 information on clone attribution to specimens (vouchers), PFR² allows to evaluate intra-genomic
445 polymorphism, which offers excellent opportunity to identify the taxonomically relevant level of
446 variability (Weber and Pawlowski, 2014). Finally, the modular structure of PFR^2 (i.e., its
447 partitioning into variable and conserved regions) is particularly suitable for the evaluation of
448 existing barcodes or the design of new barcoding systems needed to capture total or partial
449 planktonic foraminiferal diversity within complex plankton assemblages. Indeed, an examination

of the length polymorphism in the 11 regions of the 18S rDNA fragment that have been aligned for all PFR^2 sequences reveals that next to the variable 37/f region identified as a barcode for benthic foraminifera (Pawlowski and Lecroq, 2010), several other regions may be suitable as targets for barcoding of planktonic foraminifera (Figure 5).

The main difference between PFR² and classical databases is in the association of sequence data with environmental and collection data. Such level of annotation is not feasible in large databases, which have to rely on the completeness and level of metadata details provided in GenBank. The association of metadata to PFR² sequences facilitates an assessment of biogeography and ecology of genetic types (potential cryptic species). This is significant for studies of evolutionary processes in the open ocean such as speciation and gene flow at basin scale, but also for paleoceanography, which exploits ecological preferences of planktonic Foraminferal species to reconstruct climate history of the Earth (e.g., Kucera et al., 2005). Modeling studies showed that the integration of cryptic diversity into paleoceanographic studies will improve their accuracy (Kucera and Darling, 2002; Morard et al., 2013). Together with the MARGO database (Kucera et al., 2005), which records the occurrence of morphospecies of planktonic foraminifera in surface sediments and the CHRONOS/NEPTUNE database (Spencer-Cervato et al., 1994; http://www.chronos.org/), which records their occurrence through geological time, PFR² represents the cornerstone to connect genetic diversity to the fossil record in an entire group of pelagic protists.

Conclusion and perspectives

The PFR^2 database represents the first geographically and taxonomically comprehensive reference barcoding system for an entire group of pelagic protists. It constitutes a pivotal tool to

investigate the diversity, ecology, biogeography, and evolution in planktonic foraminifera as a model system for pelagic protists. In addition, the database constitutes an important resource allowing reinterpretation and refinement of the use of foraminifera as markers for stratigraphy and paleoceanography. In particular, PFR^2 can be used to: (i) annotate and classify newly generated 18S rDNA sequences from single individuals; (ii) study the biogeography of cryptic genetic types; (iii) design rank-specific primers and probes to target any group of planktonic foraminifera in natural communities; and (iv) assign accurate taxonomy to environmental sequences from metabarcoding or metagenomic datasets. This last point is particularly worth noting. Indeed, future global metabarcoding of planktonic foraminifera covering comprehensive spatio-temporal scales will likely reveal the full extent and complexity of species diversity and ecology in this group, serving as a model system for studies of the evolutionary dynamics of the plankton and its interaction with the Earth system.

Acknowledgments:

We thank all crew members and scientist for their help in the collection of planktonic foraminifera that were used to generate the database. We thank Erica de Leau for her help in compiling the data and Dominique Boeuf and ABIMS for their help in designing and hosting the PFR² website. This work was supported by grants from ANR-11-BTBR-0008 OCEANOMICS, ANR-09-BLAN-0348 POSEIDON, ANR-JCJC06-0142-PALEO-CTD, from Natural Environment Research Council of the United Kingdom (NER/J/S2000/00860 and NE/D009707/1), the Leverhulme Trust and the Carnegie Trust for the Universities of Scotland, from DFG-Research Center/Cluster of Excellence “The Ocean in the Earth System” and from the Deutsche Forschungsgemeinschaft KU2259/19 and DU1319/1-1. This study is a contribution to

494 the effort of the SCOR/IGBP Working Group 138 “Modern planktonic Foraminifera and ocean
495 changes”.

496 **References**
497

498 André A, Weiner A, Quillévéré F *et al.* (2013) The cryptic and the apparent reversed : lack of
499 genetic differentiation within the morphologically diverse plexus of the planktonic
500 foraminifer *Globigerinoides sacculifer*. *Paleobiology*, **39**, 21–39.
501 André A, Quillévéré F, Morard R *et al.* (2014) SSU rDNA Divergence in Planktonic
502 Foraminifera: Molecular Taxonomy and Biogeographic Implications (V Ketmaier, Ed,).
503 *PLoS ONE*, **9**, 1–19.
504 Aurahs R, Göker M, Grimm GW *et al.* (2009a) Using the Multiple Analysis Approach to
505 Reconstruct Phylogenetic Relationships among Planktonic Foraminifera from Highly
506 Divergent and Length-polymorphic SSU rDNA Sequences. *Bioinformatics and biology*
507 *insights*, **3**, 155–177.
508 Aurahs R, Grimm GW, Hemleben V, Hemleben C, Kucera M (2009b) Geographical distribution
509 of cryptic genetic types in the planktonic foraminifer *Globigerinoides ruber*. *Molecular*
510 *ecology*, **18**, 1692–1706.
511 Aurahs R, Treis Y, Darling K, Kucera M (2011) A revised taxonomic and phylogenetic concept
512 for the planktonic foraminifer species *Globigerinoides ruber* based on molecular and
513 morphometric evidence. *Marine Micropaleontology*, **79**, 1–14.
514 Aze T, Ezard THG, Purvis A *et al.* (2011) A phylogeny of Cenozoic macroperforate planktonic
515 foraminifera from fossil data. *Biological reviews of the Cambridge Philosophical Society*,
516 **86**, 900–27.
517 Bé A.W.H., Tolderlund, D., (1971) Distribution and ecology of living planktonic foraminifera in
518 surface waters of the Atlantic and Indian Oceans. In: Funnell, B. M., and Riedel, W. R..
519 Eds., The micropalaeontology of oceans. London: Cambridge Univ. Press, pp. 105-149,
520 text-figs. 1-27.
521 Bé, A.W.H, Hudson WH (1977) Ecology of planktonic foraminifera and biogeographic patterns
522 of life and fossil assemblages in the Indian Ocean. *Micropaleontology*, **23**, 369–414.
523 Brummer GA, Hemleben C, Michael S (1986) Planktonic foraminiferal ontogeny and new
524 perspectives for micropalaeontology. *Nature*, **319**, 50–52.
525 Darling KF, Kroon D, Wade CM, Leigh Brown AJ (1996) Molecular Phylogeny of the planktic
526 foraminifera. *Journal of foraminiferal research*, **26**, 324–330.
527 Darling KF, Wade CM, Kroon D, Leigh Brown AJ (1997) Planktic foraminiferal molecular
528 evolution and their polyphyletic origins from benthic taxa. *Marine Micropaleontology*, **30**,
529 251–266.
530 Darling KF, Wade CM, Kroon D, Leigh Brown AJ, Bijma J (1999) The Diversity and
531 Distribution of Modern Planktic Foraminiferal Small Subunit Ribosomal RNA Genotypes
532 and their Potential as Tracers of Present and Past Ocean Circulations. *Paleoceanography*,
533 **14**, 3–12.
534 Darling KF, Wade CM, Stewart I a *et al.* (2000) Molecular evidence for genetic mixing of Arctic
535 and Antarctic subpolar populations of planktonic foraminifers. *Nature*, **405**, 43–7.

536 Darling KF, Kucera M, Wade CM, von Langen PJ, Pak DK (2003) Seasonal distribution of
537 genetic types of planktonic foraminifer morphospecies in the Santa Barbara Channel and its
538 paleoceanographic implications. *Paleoceanography*, **18**, 1–10.
539 Darling KF, Kucera M, Pudsey CJ, Wade CM (2004) Molecular evidence links cryptic
540 diversification in polar planktonic protists to Quaternary climate dynamics. *Proceedings of*
541 *the National Academy of Sciences of the United States of America*, **101**, 7657–62.
542 Darling KF, Kucera M, Kroon D, Wade CM (2006) A resolution for the coiling direction
543 paradox in *Neogloboquadrina pachyderma*. *Paleoceanography*, **21**, PA2011.
544 Darling KF, Kucera M, Wade CM (2007) Global molecular phylogeography reveals persistent
545 Arctic circumpolar isolation in a marine planktonic protist. *Proceedings of the National*
546 *Academy of Sciences of the United States of America*, **104**, 5002–5007.
547 Darling KF, Wade CM (2008) The genetic diversity of planktic foraminifera and the global
548 distribution of ribosomal RNA genotypes. *Marine Micropaleontology*, **67**, 216–238.
549 Darling KF, Thomas E, Kasemann SA *et al.* (2009) Surviving mass extinction by bridging the
550 benthic/planktic divide. *Proceedings of the National Academy of Sciences of the United*
551 *States of America*, **106**, 12629–33.
552 de Vargas C, Zaninetti L, Hilbrecht H, Pawlowski J (1997) Phylogeny and rates of molecular
553 evolution of planktonic foraminifera: SSU rDNA sequences compared to the fossil record.
554 *Journal of molecular evolution*, **45**, 285–294.
555 de Vargas C, Pawlowski J (1998) Molecular versus taxonomic rates of evolution in planktonic
556 foraminifera. *Molecular phylogenetics and evolution*, **9**, 463–469.
557 de Vargas C, Norris R, Zaninetti L, Gibb SW, Pawlowski J (1999) Molecular evidence of cryptic
558 speciation in planktonic foraminifers and their relation to oceanic provinces. *Proceedings of*
559 *the National Academy of Sciences of the United States of America*, **96**, 2864–2868.
560 de Vargas C, Renaud S, Hilbrecht H, Pawlowski J (2001) Pleistocene adaptive radiation in
561 *Globorotalia truncatulinoides*: genetic, morphologic, and environmental evidence.
562 *Paleobiology*, **27**, 104–125.
563 de Vargas C, Bonzon M, Rees NW, Pawlowski J, Zaninetti L (2002) A molecular approach to
564 biodiversity and biogeography in the planktonic foraminifer *Globigerinella siphonifera*
565 (d'Orbigny). *Marine Micropaleontology*, **45**, 101–116.
566 Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: A multiplatform graphical user
567 interface for sequence alignment and phylogenetic tree building. *Molecular biology and*
568 *evolution*, **27**, 221–4.
569 Guillou L, Bachar D, Audic S *et al.* (2013) The Protist Ribosomal Reference database (PR2): a
570 catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy.
571 *Nucleic acids research*, **41**, D597–604.
572 Hamsher SE, Evans KM, Mann DG, Poulíčková A, Saunders GW (2011) Barcoding diatoms:
573 exploring alternatives to COI-5P. *Protist*, **162**, 405–22.
574 Hemleben C, Spindler M, & Anderson OR (1989) Modern Planktonic Foraminifera. Springer-
575 Verlag New York Inc. pp. 363.
576 Hollingsworth, PM, Forrest, LL, Spouge JL, et al. (2009) A DNA barcode for land plants.
577 *Proceedings of the National Academy of Sciences of the USA*, **106**, 12,794-12,797.
578 Holzmann M, Pawlowski J (1996) Preservation of foraminifera for DNA extraction and PCR
579 amplification. *journal of foraminiferal research*, **26**, 264–267.
580 Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7:
581 improvements in performance and usability. *Molecular biology and evolution*, **30**, 772–80.

582 Kucera M, Darling KF (2002) Cryptic species of planktonic foraminifera: their effect on
583 palaeoceanographic reconstructions. *Philosophical transactions. Series A, Mathematical,*
584 *physical, and engineering sciences*, **360**, 695–718.
585 Kucera M, Weinelt M, Kiefer T *et al.* (2005) Reconstruction of sea-surface temperatures from
586 assemblages of planktonic foraminifera: multi-technique approach based on geographically
587 constrained calibration data sets and its application to glacial Atlantic and Pacific Oceans.
588 *Quaternary Science Reviews*, **24**, 951–998.
589 Logares R, Audic S, Bass D *et al.* (2014) Patterns of rare and abundant marine microbial
590 eukaryotes. *Current biology : CB*, **24**, 813–21.
591 Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B (2011) How many species are there on
592 Earth and in the ocean? *PLoS biology*, **9**, e1001127.
593 Morard R, Quillévéré F, Escarguel G *et al.* (2009) Morphological recognition of cryptic species
594 in the planktonic foraminifer *Orbulina universa*. *Marine Micropaleontology*, **71**, 148–165.
595 Morard R, Quillévéré F, Douady CJ *et al.* (2011) Worldwide genotyping in the planktonic
596 foraminifer *Globoconella inflata*: implications for life history and paleoceanography. *PLoS*
597 *ONE*, **6**, 1–12.
598 Morard R, Quillévéré F, Escarguel G, Garidel-thoron T de (2013) Ecological modeling of the
599 temperature dependence of cryptic species of planktonic foraminifera in the Southern
600 Hemisphere. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **391**, 13–33.
601 R Development Core Team (2014) R: a language and environment for statistical computing. R
602 Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.
603 Nassonova E, Smirnov A, Fahrni J, Pawlowski J (2010) Barcoding amoebae: comparison of
604 SSU, ITS and COI genes as tools for molecular identification of naked lobose amoebae.
605 *Protist*, **161**, 102–15.
606 Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in
607 the amino acid sequence of two proteins. *Journal of molecular biology*, **48**, 443–53.
608 Nikolaev SI, Berney C, Fahrni JF *et al.* (2004) The twilight of Heliozoa and rise of Rhizaria, an
609 emerging supergroup of amoeboid eukaryotes. *Proceedings of the National Academy of*
610 *Sciences of the United States of America*, **101**, 8066–71.
611 Pawlowski J, Lecroq B (2010) Short rDNA barcodes for species identification in foraminifera.
612 *The Journal of eukaryotic microbiology*, **57**, 197–205.
613 Pawlowski J, Audic S, Adl S *et al.* (2012) CBOL protist working group: barcoding eukaryotic
614 richness beyond the animal, plant, and fungal kingdoms. *PLoS biology*, **10**, e1001419.
615 Pawlowski J, Holzmann M (2014) A plea for DNA barcoding of foraminifera. *Journal of*
616 *foraminiferal research*, **44**, 62–67.
617 Pawlowski J, Lejzerowicz F, Esling P (2014) Next-Generation Environmental Diversity Surveys
618 of Foraminifera : Preparing the Future. *Biol. Bull.*, **227**, 93–106.
619 Quillévéré F, Morard R, Escarguel G *et al.* (2013) Global scale same-specimen morpho-genetic
620 analysis of *Truncorotalia truncatulinoides*: A perspective on the morphological species
621 concept in planktonic foraminifera. *Palaeogeography, Palaeoclimatology, Palaeoecology*,
622 **391**, 2–12.
623 Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open
624 Software Suite. *Trends in Genetics*, **16**, 2–3.
625 Rognes T (2011) Faster Smith-Waterman database searches with inter-sequence SIMD
626 parallelisation. *BMC bioinformatics*, **12**, 221.

627 Saunders GW (2005) Applying DNA barcoding to red macroalgae: a preliminary appraisal holds
628 promise for future applications. *Philosophical transactions of the Royal Society of London.*
629 *Series B, Biological sciences*, **360**, 1879–88.
630 Schoch CL, Seifert K a, Huhndorf S *et al.* (2012) Nuclear ribosomal internal transcribed spacer
631 (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National*
632 *Academy of Sciences of the United States of America*, **109**, 6241–6.
633 Seears HA, Darling KF, Wade CM (2012) Ecological partitioning and diversity in tropical
634 planktonic foraminifera. *BMC Evolutionary Biology*, **12**, 54.
635 Sherwood AR, Presting GG (2007) Universal primers amplify a 23S rDNA plastid marker in
636 eukaryotic algae and cyanobacteria. *Journal of Phycology*, **43**, 605–608.
637 Spencer-Cervato C, Thierstein HR, Lazarus DB, Beckmann J-P (1994) How synchronous are
638 neogene marine plankton events? *Paleoceanography*, **9**, 739.
639 Stern RF, Horak A, Andrew RL *et al.* (2010) Environmental barcoding reveals massive
640 dinoflagellate diversity in marine environments. *PloS one*, **5**, e13991.
641 Stewart IA, Darling KF, Kroon D, Wade CM, Troelstra SR (2001) Genotypic variability in
642 subarctic Atlantic planktic foraminifera. *Marine Micropaleontology*, **43**, 143–153.
643 Sogin ML, Morrison HG, Huber J a *et al.* (2006) Microbial diversity in the deep sea and the
644 underexplored "rare biosphere". *Proceedings of the National Academy of Sciences of the*
645 *United States of America*, **103**, 12115–20.
646 Ujiié Y, Kimoto K, Pawlowski J (2008) Molecular evidence for an independent origin of modern
647 triserial planktonic foraminifera from benthic ancestors. *Marine Micropaleontology*, **69**,
648 334–340.
649 Ujiié Y, Lipps JH (2009) Cryptic diversity in planktonic foraminifera in the northwest Pacific
650 Ocean. *Journal of foraminiferal research*, **39**, 145–154.
651 Ujiié Y, Asami T, de Garidel-Thoron T *et al.* (2012) Longitudinal differentiation among pelagic
652 populations in a planktic foraminifer. *Ecology and evolution*, **2**, 1725–37.
653 Wickham, H. (2009). ggplot2: elegant graphics for data analysis. Springer New York.
654 Wade CM, Darling KF, Kroon D, Brown AJL (1996) Early Evolutionary Origin of the Planktic
655 Foraminifera Inferred from Small Subunit rDNA Sequence Comparisons. *Journal of*
656 *molecular evolution*, **43**, 672–677.
657 Weber AA-T, Pawlowski J (2014) Wide occurrence of SSU rDNA intragenomic polymorphism
658 in foraminifera and its implications for molecular species identification. *Protist*, **165**, 645–
659 61.
660 Weiner A, Aurahs R, Kurasawa A, Kitazato H, Kucera M (2012) Vertical niche partitioning
661 between cryptic sibling species of a cosmopolitan marine planktonic protist. *Molecular*
662 *ecology*, **21**, 4063–73.
663 Weiner AKM, Weinkauf MFG, Kurasawa A *et al.* (2014) Phylogeography of the tropical
664 planktonic foraminifera lineage *Globigerinella* reveals isolation inconsistent with passive
665 dispersal by ocean currents. *PloS one*, **9**, e92148.
666 Weiner AKM, Weinkauf MFG, Kurasawa A, Darling KF, Kucera M (2015) Genetic and
667 morphometric evidence for parallel evolution of the *Globigerinella calida* morphotype.
668 *Marine Micropaleontology*, **114**, 19–35.
669 Yilmaz P, Parfrey LW, Yarza P *et al.* (2013) The SILVA and "All-species Living Tree Project
670 (LTP)" taxonomic frameworks. *Nucleic acids research*, **42**, D643–8.

Žarić S, Donner B, Fischer G, Mulitza S, Wefer G (2005) Sensitivity of planktic foraminifera to sea surface temperature and export production as derived from sediment trap data. *Marine Micropaleontology*, **55**, 75–105.

Zimmermann J, Abarca N, Enk N et al. (2014) Taxonomic reference libraries for environmental barcoding: a best practice example from diatom research. PloS one, 9, e108793.

Author contribution

KFD, CdV, YU, RM, TdG, AKMW, HAS, MK, AA, MS participated in sample collection, CdV, MK, KFD, CMW, CJD, FQ, GE, TdG provided laboratory infrastructure, KFD, YU, RM, AKMW, AA, HAS participated in laboratory work. FM and RM conceived and designed the bioinformatics pipeline, FM performed the computational work, SA built the website. RM wrote the manuscript with help from MK and CdV. All authors read, edited and approved the final manuscript.

Data Accessibility

Sequences, NCBI accession numbers and metadata are available in Supplementary Material 1 and 2 and on the PFR² website at http://pfr2.sb-roscoff.fr. The custom scripts used to perform the curation procedure are available in Supplementary Material 5; the results of the curation process are given in Supplementary Material 1 and 2.

Figures

Figure 1

Sampling Map. Location of the 460 oceanic stations sampled over 20 years for single-cell genetic studies of planktonic foraminifera. Each symbol corresponds to a scientific cruise or near shore collection site. Cruise names and dates of the collection expeditions are indicated in the legend. Grey shading shows ocean bathymetry.

Figure 2

Workflow to constitute PFR². In step I the sequences, metadata and taxonomic information are retrieved from public databases and literature or from the internal databases of the co-authors to constitute the Primary Reference Database. In step II, the coverage of each sequence is evaluated by alignment with structural regions of the 18S RNA secondary structure derived for the species *Micrometula hyalostera* (Pawlowski and Lecroq, 2010). In step III, the consistency of the annotation is checked from the most exclusive level of annotation "genetic type 3" up to the species level (Phase 1) to detect annotation inconsistencies (See Figure 3). Sequences with wrong annotation are invalidated, compared to the validated part of the dataset (Phase 2) and re-annotated depending on the best hit out of the valid dataset. The consistency of all annotations is then checked again following the same procedure as in Phase 1 (Phase 3), to ensure that no taxonomic inconsistency remains. In step IV, all sequences which have been subjected to the curation process are integrated in the *Planktonic Foraminifera Ribosomal Reference* database (PFR²). The results of all steps are given in Supplementary Material 1.

Figure 3

Annotation inconsistency detection. The procedure followed to identify annotation inconsistencies is exemplified by three cases. Each graph represents variability in pairwise similarities observed across each region of all sequences sharing the same annotation level. The names of the taxon and annotation level are given above the plot with the number of sequences in parenthesis. Each vertical line represents one region with the variability represented as box plot, the number of "complete" regions is given at the bottom of the line. The case "A" describes the annotation validation process starting from the most exclusive rank of "genetic type level 3" to the "species" rank. After the validation at one rank level, the sequences with valid annotation are merged into a taxonomic unit of a higher rank, this now including multiple sequence motifs which decreases the average similarity level of each region, thus leading to higher variability in higher ranks. Case "B" represents the occurrence of obvious outliers at the species level, which are invalidated. Case "C" represents the co-occurrence of divergent sequences under the same taxonomic attribution, which are consequently all invalidated. Box plots for all ranks can be found in Supplementary Material 3 and the pairwise similarities calculated for each taxonomic level are given in Supplementary Material 1.

Figure 4

Taxonomic and ecological coverage of PFR². For each morphogroup (Spinose, Non-Spinose, Microperforates, Monolamellar and Non-Spiral) the number of species included in PFR² is given in the filled bar while the number of species not present is indicated in the adjacent open bar. The relative abundance in the sediments of each species included in PFR² is given in a log-scale value against mean Sea Surface Temperature (SST) at the sampling station. Relative abundances in sediments are derived from the MARGO database (Kucera et al., 2005) and the mean annual SST (MODIS Aqua, NASA, Greenbelt, MD, USA). The grey dots highlight the mean annual SST at the location where the living planktonic foraminifera yielding sequences were sampled. The number of sequences available for each species as well as the number of taxonomic paths above the species level is shown next to the graphs. Also shown is the cumulative mean relative abundance in the sediments of all species included in PFR² plotted against the mean annual SST in discrete 1°C intervals. Vertical bars represent 95% confidence intervals for each 1°C bin.

Figure 5

Length polymorphism. Each rectangle represents the length polymorphism within each region of the analyzed 18S rDNA fragment across all resolved taxonomic units in PFR². The regions are based on the rRNA secondary structure and are named following Pawlowski and Lecroq (2010).

Supplementary Material.

Supplementary Material 1

Information on all consecutive steps followed to constitute the PFR². All fields are explained in the file.

Supplementary Material 2

FASTA files of sequences used to build the PFR². FASTA files are provided for the full sequences and individual partitions.

Supplementary Material 3

Box plots showing pairwise similarities for each taxonomic level. See Figure 3 for explanations of the content of the plots.

Supplementary Material 4

Neighbor-joining trees showing the monophyly of each morphospecies present in PFR².

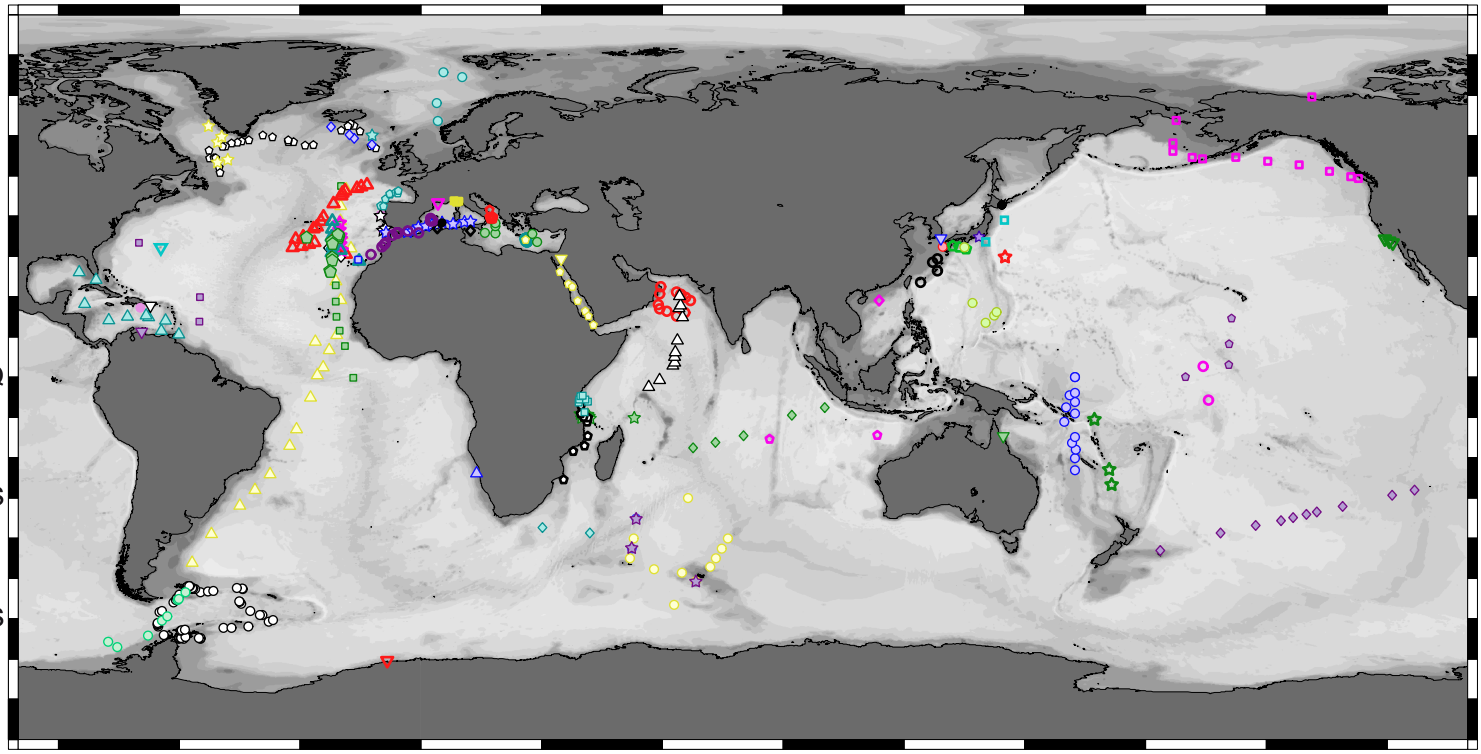
Supplementary Material 5

Custom scripts used to perform the different curation steps.

Table 1. Flanking conserved sequences of the 5 variable regions in planktonic foraminifera. The minimum and maximum lenght of each region are given as well as their coverage in the database (See details in the text).

Region	Specificity	Begining	End	Min lenght	Max lenght	Not available	Partial	Poor	Complete
32-37	Eukaryotes	-	-	-	-	949	2583	0	0
37F	Foraminifera	5'-GGAUUGACA	CUUUCACAUGA-3'	38	132	800	272	249	2211
37-41	Eukaryotes	-	-	68	72	547	403	138	2444
41F	Foraminifera	5'-AAUUGCG	GCAACGAA-3'	58	322	349	346	282	2555
39-43	Eukaryotes	-	-	27	29	460	34	57	2981
43E	Eukaryotes	5'-CUUGUU	AACUAGAGGG-3'	33	195	401	263	265	2603
44-45	Eukaryotes	-	-	113	123	487	1288	136	1621
45E-47F	Euk - Forams	5'-CAGUGAG	GGUGGGG-3'	179	312	1660	187	386	1299
47-49	Eukaryotes	-	-	140	148	1827	425	152	1128
49E	Eukaryotes	5'-GUGAG	CGAACAG-3'	27	127	2251	130	125	1026
50	Eukaryotes	-	-	-	-	2389	1143	0	0

Figure 1. Sampling Map. Location of the 460 oceanic stations sampled over 20 years for single-cell genetic studies of planktonic Foraminifera. Each symbol corresponds to a scientific cruise or near shore collection site. Cruise names and dates of the collection expeditions are indicated in the legend. Grey shading shows ocean bathymetry.



Scientific cruise

- Alis, GYRAFOR-A (Jun 2008)
- Charles Darwin, CD148 (July 2003)
- Charles Darwin, CD159 (July 2004)
- Discovery, D262 (Apr 2002)
- Discovery, D286 (Dec 2005, Jan 2006)
- Garcia del Cid, Iberia-Forams (Sept 2012)
- Hakuho-maru, KH04-2 (Jun-Jul 2009)
- Hakuho-maru, KH10-4 (Aug 2011)
- James Clark Ross, AMT-5 (Sept-Oct 1997)
- James Clark Ross, AMT-8 (Apr Jun 1999)
- James Clark Ross, JR 19 (Mar 1997)
- James Clark Ross, JR 48 (Feb-Mar, 2000)
- Maria S. Merian, MSM09-2 (Aug-Sep 2008)
- Maria S. Merian, MSM15-5 (Jul 2010)
- Marion Dufresne, GYRAFOR-B (Jul-Aug 2007)
- Marion Dufresne, OISO2011 (Jan 2011)
- Marion Dufresne, OISO-4 (Jan-Feb 2000)
- Melville, Melville (June 2003)
- Meteor, M37-2a (Apr 1997)
- Meteor, M69-1 (Aug 2006)
- Meteor, M71-2 (Dec 2006 - Jan 2007)
- Meteor, M71-3 (Jan-Feb 2007)
- Meteor, M74-1a (Sep 2007)
- Meteor, M74-1b (Sep-Oct 2007)
- Meteor, M75-2 (Feb 2008)
- Meteor, M78-1 (Feb-March 2008)
- Mirai, MR02-K01 (Jan 2002)
- Mirai, MR10-06 (Nov 2010)
- Pelagia, 64PE303 - GLOW (Feb-Mar 2009)
- Pelagia, 65PE304 (Mar 2009)
- Polarstern, Arktis XV/1-2 (Jun 1999)
- Poseidon, P247 (Jan 1999)
- Poseidon, P283-2 (Feb-Mar 2002)
- Poseidon, P308 (Mar 2004)
- Poseidon, P321 (May 2005)
- Poseidon, P334 (Mar-Apr 2006)
- Poseidon, P349 (Apr 2007)
- Poseidon, P411 (Apr 2011)
- Poseidon, P413 (May 2011)
- Professor Logachev, Denmark Strait (Sept 1997)
- Roger Revelle, Revelle (Jan 2001)
- Ron Brown, CMarZ (April 2006)
- Sarmiento de Gamboa, FORCLIM-7 (Apr 2009)
- Seriora, Amakusa (Sep 2009)
- Sir Wilfried Laurier, CCGS (July 2007)
- Sonne, SO-221 (May 2012)
- Sonne, SO-226 (Mar 2013)
- Tansei-maru, KT02-15 (Oct 2002)
- Tansei-maru, KT07-14 (Jun 2007)
- Tansei-maru, KT06-11 (Jun 2006)
- Tansei-maru, KT06-30 (Nov 2006)
- Welwitschia, NatMIRC (Nov 2001)

Near-Shore Collection

- Bermuda (Apr 1996)
- Curaçao (Feb 1993)
- Eilat (Feb 2011)
- Ekstrom Ice Shelf-Atka Bay (Jan 2001)
- Lizard Island, GBR (Aug 1993, Sep 1997)
- Puerto Rico (Mar 1995)
- Santa Barbara Chanel (Feb 1998, Jan-Sep 1999)
- Tsugaru strait
- Villefranche sur Mer (Dec 1995)

Figure 2. Workflow to constitute PFR². In step I the sequences, metadata and taxonomic information are retrieved from public databases and literature or from the internal databases of the co-authors to constitute the Primary Reference Database. In step II, the coverage of each sequence is evaluated by alignment with structural regions of the 18S RNA secondary structure derived for the species *Micrometula hyalostera* (Pawlowski and Lecroq, 2010). In step III, the consistency of the annotation is checked from the most exclusive level of annotation "genetic type 3" up to the species level (Phase 1) to detect annotation inconsistencies (See Figure 3). Sequences with wrong annotation are invalidated, compared to the validated part of the dataset (Phase 2) and re-annotated depending on the best hit out of the valid dataset. The consistency of all annotations is then checked again following the same procedure as in Phase 1 (Phase 3), to ensure that no taxonomic inconsistency remains. In step IV, all sequences which have been subjected to the curation process are integrated in the *Planktonic Foraminifera Ribosomal Reference* database (PFR²). The results of all steps are given in Supplementary Material 1.

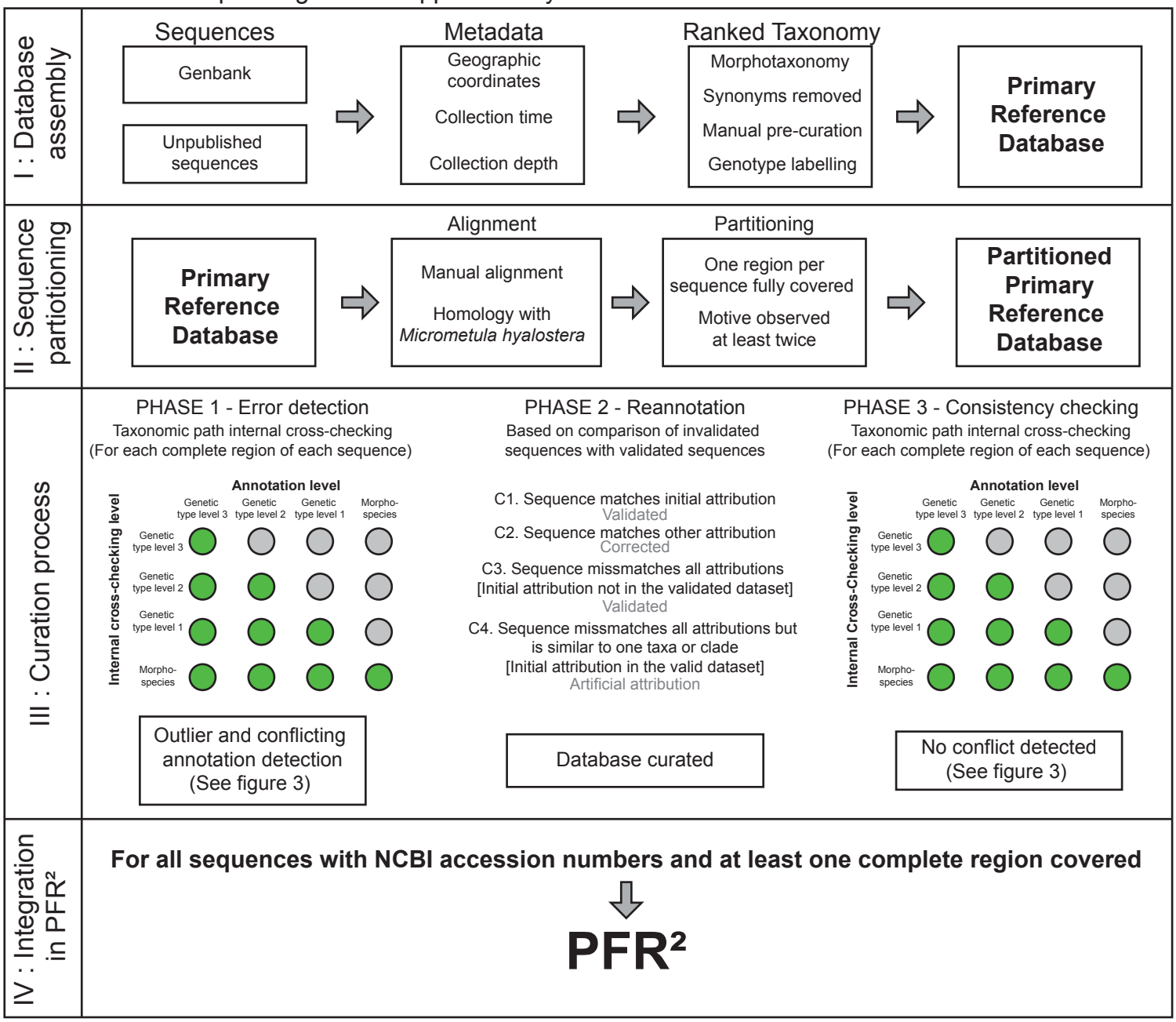


Figure 3. Annotation inconsistency detection. The procedure followed to identify annotation inconsistencies is exemplified by three cases. Each graph represents variability in pairwise similarities observed across each region of all sequences sharing the same annotation level. The names of the taxon and annotation level are given above the plot with the number of sequences in parenthesis. Each vertical line represents one region with the variability represented as box plot, the number of "complete" regions is given at the bottom of the line. The case "A" describes the annotation validation process starting from the most exclusive rank of "genetic type level 3" to the "species" rank. After the validation at one rank level, the sequences with valid annotation are merged into a taxonomic unit of a higher rank, this now including multiple sequence motifs which decreases the average similarity level of each region, thus leading to higher variability in higher ranks. Case "B" represents the occurrence of obvious outliers at the species level, which are invalidated. Case "C" represents the co-occurrence of divergent sequences under the same taxonomic attribution, which are consequently all invalidated. Box plots for all ranks can be found in Supplementary Material 3 and the pairwise similarities calculated for each taxonomic level are given in Supplementary Material 1.

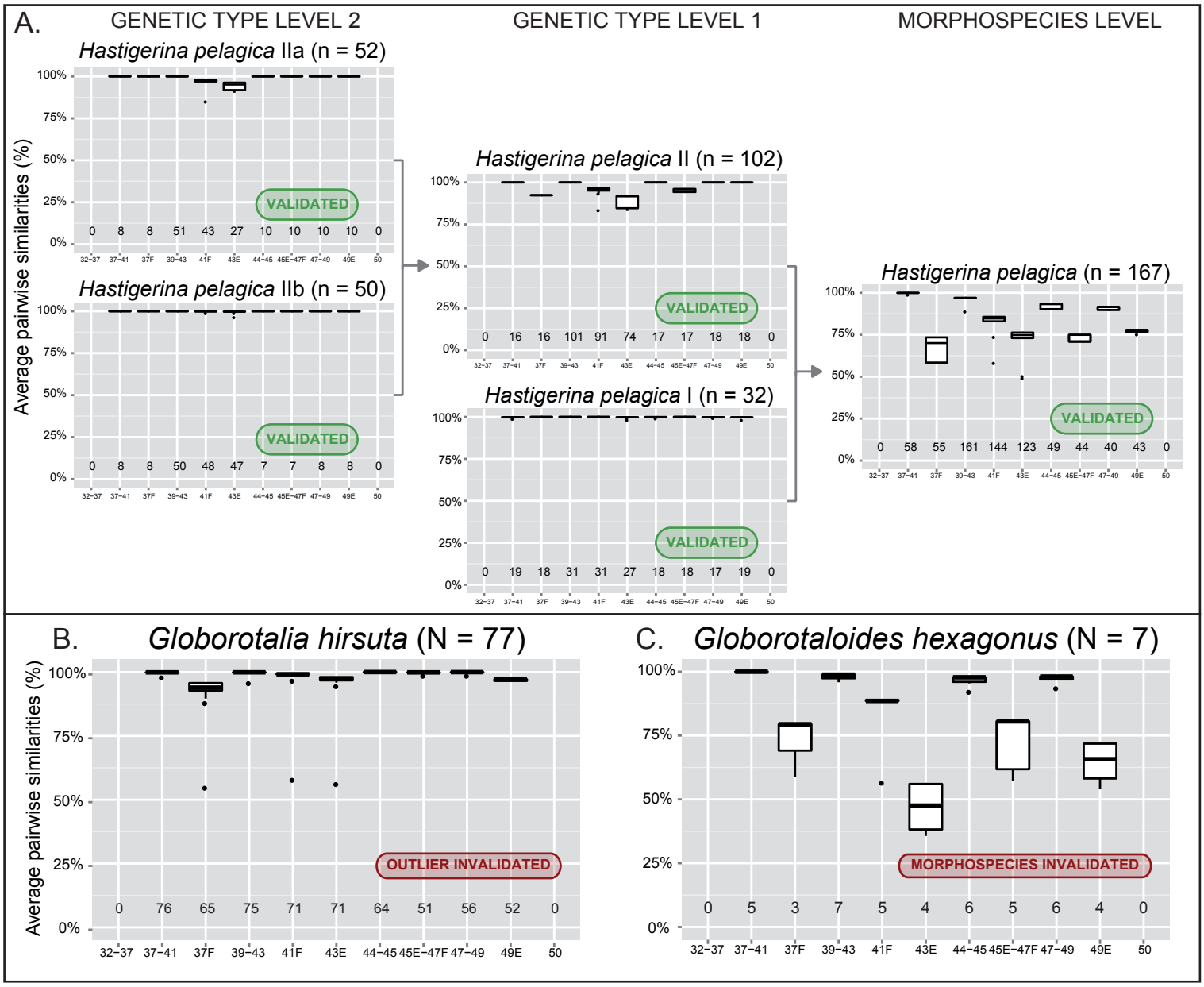


Figure 4. Taxonomic and ecological coverage of PFR². For each morphogroup (Spinose, Non-Spinose, Microperforates, Monolamellar and Non-Spiral) the number of species included in PFR² is given in the filled bar while the number of species not present is indicated in the adjacent open bar. The relative abundance in the sediments of each species included in PFR² is given in a log-scale value against mean Sea Surface Temperature (SST) at the sampling station. Relative abundances in sediments are derived from the MARGO database (Kucera et al., 2005) and the mean annual SST from the World Ocean Atlas (Locarnini, 2005). The grey dots highlight the mean annual SST at the location where the living planktonic Foraminifera yielding sequences were sampled. The number of sequences available for each species as well as the number of taxonomic paths above the species level is shown next to the graphs. Also shown is the cumulative mean relative abundance in the sediments of all species included in PFR² plotted against the mean annual SST in discrete 1°C intervals. Vertical bars represent 95% confidence intervals for each 1°C bin.

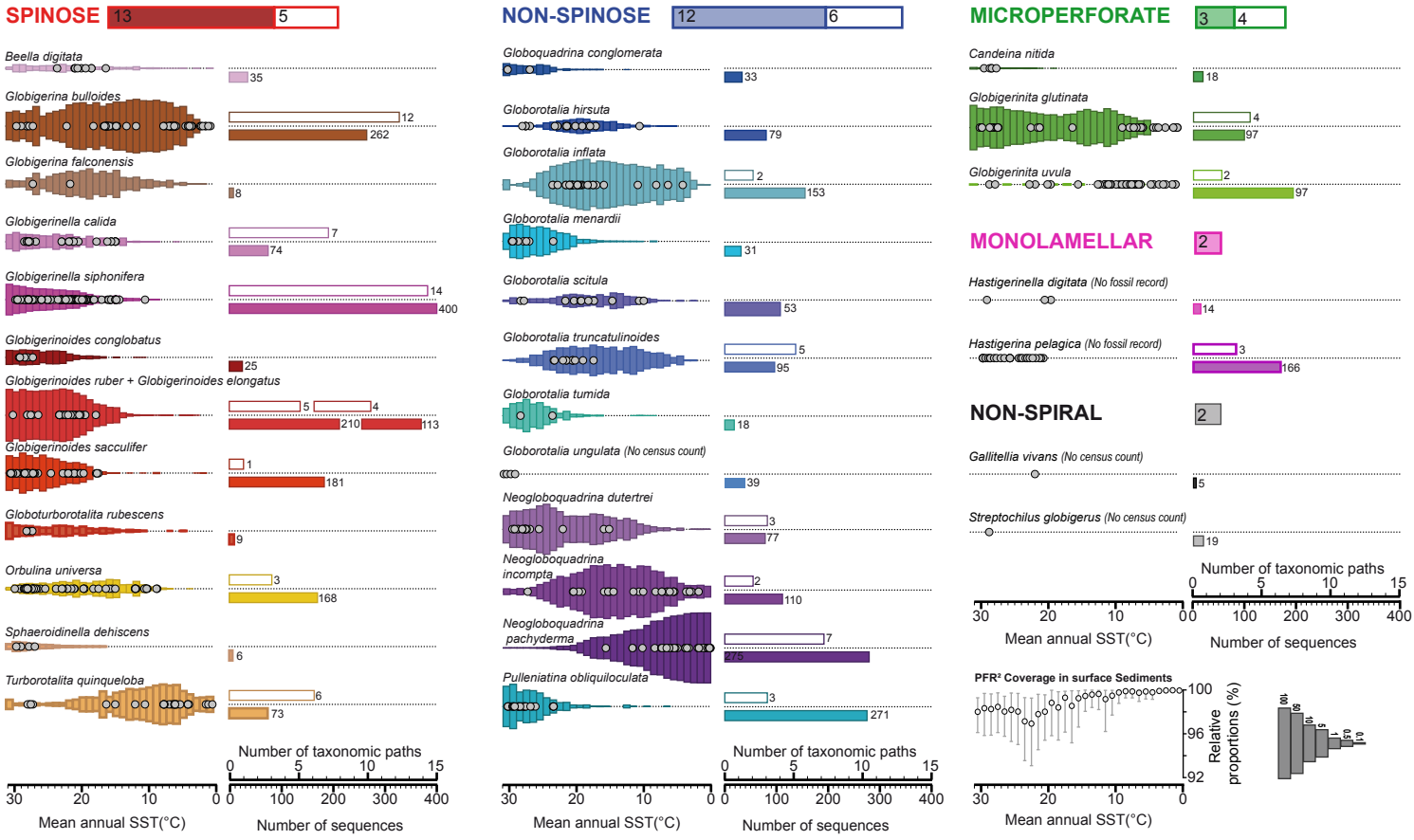


Figure 5. **Length polymorphism**. Each rectangle represents the length polymorphism within each region of the analyzed 18S rDNA fragment across all resolved taxonomic units in PFR². The regions are based on the rRNA secondary structure and are named following Pawlowski and Lecroq (2010).