MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2015) doi: 10.1111/1755-0998.12410

PFR^2: a curated database of planktonic foraminifera 18S 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. M. WEINER,‡ AURORE ANDRÉ,‡‡§§ HEIDI A. SEEARS,¶¶*** CHRIS M. WADE,¶¶ FRÉDÉRIC QUILLÉVÉRÉ,‡‡ CHRISTOPHE J. DOUADY,†††‡‡‡ GILLES ESCARGUEL,‡‡ THIBAULT DE GARIDEL-THORON,§§§ MICHAEL SICCHA,‡ MICHAL KUCERA‡ and COLOMBAN DE VARGAS*†

*Centre National de la Recherche Scientifique, UMR 7144, EPEP, Station Biologique de Roscoff, 29680 Roscoff, France, †Sorbonne Universités, UPMC Univ Paris 06, UMR 7144, Station Biologique de Roscoff, 29680 Roscoff, France, ‡MARUM Center for Marine Environmental Sciences, University of Bremen, Leobener Strasse, 28359 Bremen, Germany, §School of GeoSciences, University of Edinburgh, Edinburgh EH9 3JW, UK, ¶School of Geography and GeoSciences, University of St Andrews, Fife KY16 9AL, UK, **Department of Ecology, Technische Universität Kaiserslautern, 67663 Kaiserslautern, Germany, ††Department of Biology, Shinshu University, Asahi3-1-1, Matsumoto, Nagano 390-8621, Japan, ‡‡CNRS UMR 5276, Laboratoire de Géologie de Lyon: Terre, Planètes, Environnement, Université Claude Bernard Lyon 1, 69622 Villeurbanne, France, §§UFR Sciences Exactes et Naturelles, Université de Reims-Champagne-Ardenne, Campus Moulin de la Housse, Batiment 18, 51100 Reims, France, ¶¶School of Life Sciences, University of Nottingham, University Park, Nottingham NG7 2RD, UK, ***Department of Biological Sciences, Lehigh University, Iacocca Hall, 111 Research Drive, Bethlehem, PA, 18105, USA, †††UMR5023 Ecologie des Hydrosystèmes Naturels et Anthropisés, Université Lyon 1, ENTPE, CNRS, Université de Lyon, 6 rue Raphaël Dubois, 69622 Villeurbanne, France, ‡‡‡Institut Universitaire de France, 103 Boulevard Saint-Michel, 75005 Paris, France, §§§Aix-Marseille Université, CNRS, CEREGE UM34, Technopôle de l'Arbois, 13545 Aix-en-Provence, France*

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 PFR^2, the *Planktonic foraminifera Ribosomal Reference* database. The first version of the database includes 3322 reference 18S rDNA sequences belonging to 32 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 PFR^2 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.

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

Received 20 January 2015; revision received 25 March 2015; accepted 27 March 2015

Correspondence: Raphaël Morard, Fax: +49 (0)-421-218 – 9865974
E-mail: rmorard@marum.de

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 behavioural 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 & Presting 2007; Hollingsworth *et al.* 2009; Nassonova *et al.* 2010; Pawlowski & Lecroq 2010; Stern *et al.* 2010; Hamsher *et al.* 2011; 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 prerequisites 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 behaviour have been thoroughly documented by analyses of plankton hauls (e.g. Bé & Hudson 1977), sediment trap series (e.g. Žarić *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 morphotaxonomy and phylogeny of the group have been largely confirmed by molecular genetic analyses (e.g. Aurahs *et al.* 2009a) based on the highly informative, ~1000-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 & 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 & Wade 2008; Wade *et al.* 1996; de Vargas *et al.* 1997, 1999, 2001, 2002; de Vargas & Pawlowski 1998; Stewart *et al.* 2001; Ujiié *et al.* 2008; Aurahs *et al.* 2009b, 2011; Morard *et al.* 2009, 2011, 2013; Ujié & Lipps 2009; Ujiié *et al.* 2012; Seears *et al.* 2012; Weiner *et al.* 2012, 2014; Quillévéré *et al.* 2013, André *et al.* 2014). 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 (Fig. 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, 1787 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

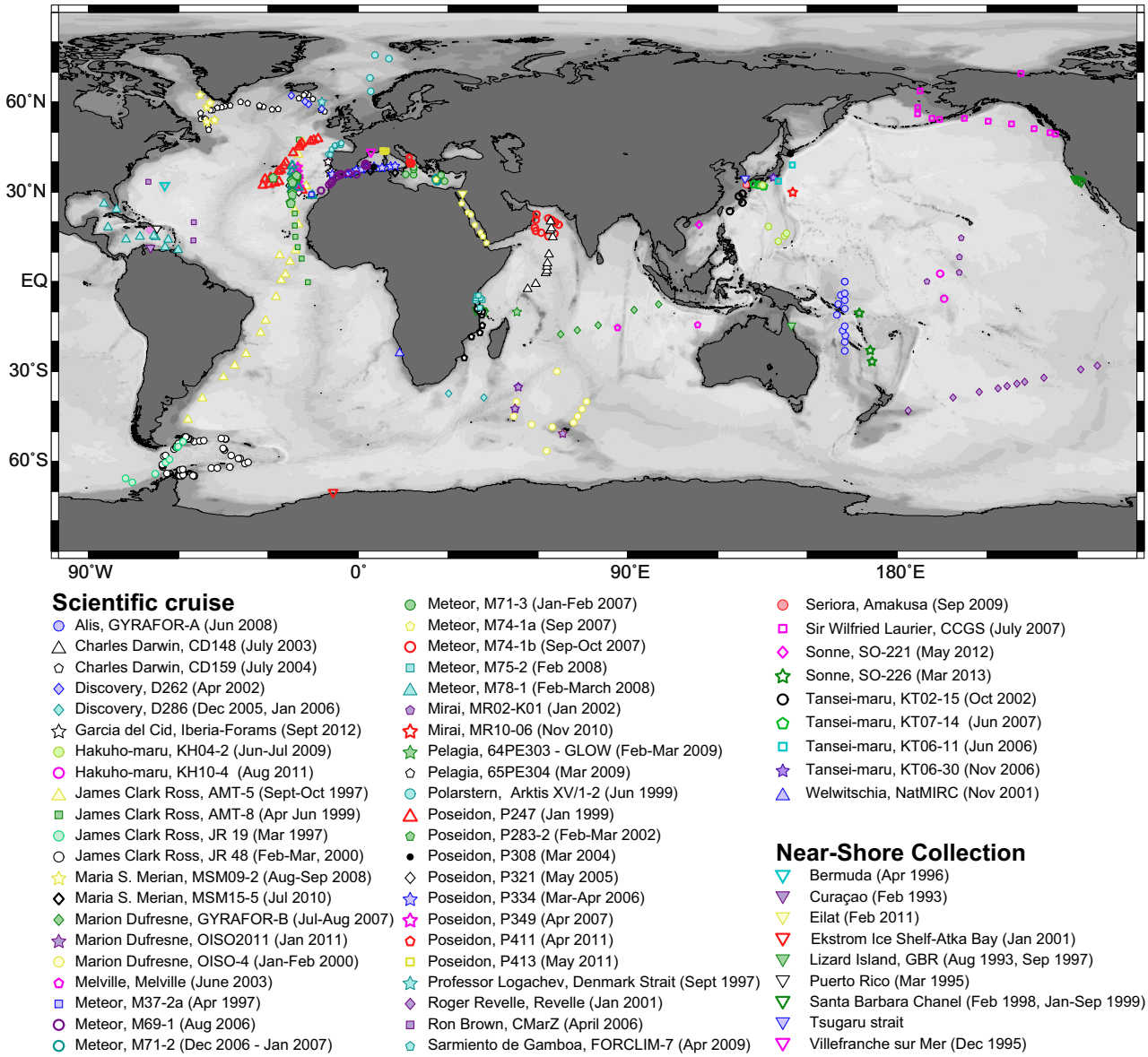

Fig. 1 Sampling Map. Location of the 460 oceanic stations sampled over 20 years for single-cell genetic studies of planktonic foraminifera. Each symbol corresponds to a scientific cruise or near shore collection site. Cruise names and dates of the collection expeditions are indicated in the legend. Grey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

collection unites the current morphological, genetic, ecological and biogeographical 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.

Materials and methods

Primary database assembly

A total of 1787 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 with each sequence consisted of: (i) their organismal origin (specimen voucher, taxonomic path, infraspecific 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 (Table S1, Supporting information).

We next compiled all unpublished 18S rDNA sequences generated by the co-authors of this study and linked them with the same suite of metadata. These sequences originate from single-cell extractions of planktonic foraminifera collected by stratified or nonstratified 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 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 analysed and annotated for this study. In addition, 925 unpublished sequences analysed in Darling *et al.* (2000, 2003, 2004, 2006, 2007), Darling & 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 nos 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 m (Fig. 1; Table S1, Supporting information).

Taxonomy

Morphological taxonomy—As the first step in the curation process, the primary taxonomic annotations of all 3532 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 mislabelled 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 catalogue 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 *G. ruber* (white) were renamed as *G. ruber*. The 113 sequences of *G. ruber* and *G. ruber* (white) attributed to the genotype II were renamed *G. elongatus* following Aurahs *et al.* (2011). The 12 sequences labelled *Globigerinella aequilateralis* were renamed *Globigerinella siphonifera* following Hemleben *et al.* (1989). The seven sequences corresponding to the right-coiled morphotype of *Neogloboquadrina pachyderma* were renamed *N. 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 Table S1 (Supporting information).

Annotation of genetic types—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 1356 sequences downloaded from NCBI were associated with a genetic type label, which was always retained. In addition, 19 sequences labelled as *G. ruber*, 15 as *Globigerinoides sacculifer*, 36 as *Globigerinita glutinata*, six as *Globigerinita uvula*, nine as *Globorotalia inflata*, 10 as *N. incompta*, six as *N. pachyderma*, five as *Orbulina universa*, five as *Pulleniatina obliquiloculata*, 30 as *Hastigerina pelagica* and 32 as *G. siphonifera* have been analysed after their first release in the public domain by Aurahs *et al.* (2009a,b), Ujié *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 PFR^2 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 *O. universa*, *Globigerina bulloides*, *N. incompta*, *Neogoboquadrina dutertrei*, *N. 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 PFR^2 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 3532 18S rDNA sequences were reassigned to 33 species names and 2276 sequences were annotated to the level of genetic types (Table S1, Supporting information). 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 and 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, nonspinose, microperforate, monolamellar and nonspiral. 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 *G. ruber*, *G. elongatus*, *G. siphonifera*, *Globigerinella calida*, *H. pelagica*, *G. bulloides*, *N. dutertrei*, *P. obliquiloculata* and *T. quinqueloba*. Genetic type attributions lacking hierarchical structure were reported in the rank 'genetic type level 1'. After this step, the Primary Reference Database (Fig. 2) of 3532 sequences contained 113 different taxonomic paths (Table S1, Supporting information).

Sequences partitioning into conserved and variable regions

Because PFR^2 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 3412 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 & 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 (Table 1). 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 (Fig. 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 analysed in this way. All regions fully covered in a sequence and containing sequence motives observed at least twice in the whole data set were labelled as 'complete'. Regions fully covered but containing a sequence motive that was observed only once in the whole data set were labelled as 'poor'. This is because we consider sequencing/PCR errors as the most likely cause for the occurrence of such unique sequence motives. We realize that using this procedure, even genuine unique sequences may be discarded from the analysis, but this would be the case only if such sequences deviated in all regions. In all other cases, the regions were labelled as 'partial' when only a part of the region was present or 'not available' if they did not contain any fragment of the sequence. As a result, we obtain the Partitioned Primary Reference Database (Fig. 2). The coverage of each individual region in the Partitioned Primary Reference Database is given in Table S1 (Supporting information), and all sequence partitions are given in Appendix S1 (Supporting information).

Semi-automated iterative curation pipeline for optimal taxonomic assignment

The consistency of taxonomic assignments within the annotated database of partitioned sequences was assessed using a semi-automated process (Figs 2 and 3). All 'complete' regions of sequences with the same taxonomic assignment at the morphospecies level were automatically aligned using global pairwise alignment (Needleman & Wunsch 1970), as implemented in the software NEEDLE from the Emboss suite of bioinformatics tools (Rice *et al.* 2000). To detect annotation inconsistencies, mean pairwise similarities were computed for each 'complete' region of each sequence against all other sequences with the same taxonomic assignment from the

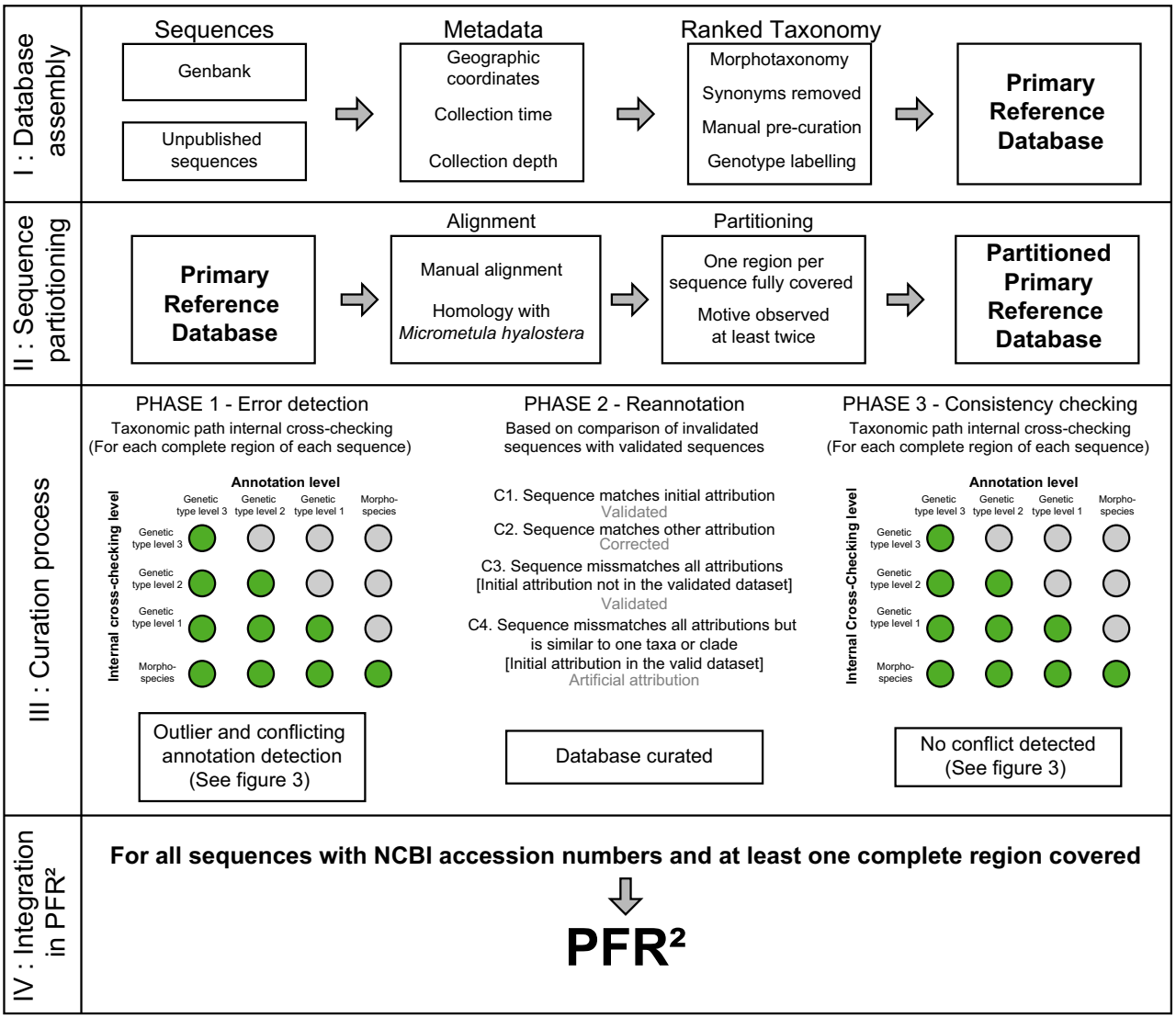

Fig. 2 Workflow to constitute PFR^2. In step I, the sequences, metadata and taxonomic information are retrieved from public databases and literature or from the internal databases of the co-authors to constitute the Primary Reference Database. In step II, the coverage of each sequence is evaluated by alignment with structural regions of the 18S RNA secondary structure derived for the species *Micrometula hyalostera* (Pawlowski & 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 Fig. 3). Sequences with wrong annotation are invalidated, compared to the validated part of the data set (Phase 2) and re-annotated depending on the best hit out of the valid data set. 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 (PFR^2). The results of all steps are given in Table S1 (Supporting information).

finest annotation level 'genetic type level 3' up to the 'species level' rank. Results are provided in Table S1 (Supporting information) and were visualized using R (R Development Core Team 2014) and the ggplot2 library (Wickham 2009). The resulting plots are given in Appendix S2 (Supporting information). If all annotations are consistent and there is no variation within taxa, each sequence within the analysed 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

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

Region	Specificity	Beginning	End	Min length	Max length	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	Eukaryotes–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

sequence will find a perfect match within the taxon but the mean pairwise similarity will also be lower than 1.

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 analysed 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 analysed 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 (Figs 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 (Fig. 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'; Fig. 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 analysed 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 analysed 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 data set. 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 data set, 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 (Fig. 2, Table S1, Supporting information).

In the third step, sequences that received new attributions were reanalysed 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 neighbour joining using only sequences that are covering the six conserved and five variable regions of the 5′ end fragment. First, each individual sequence included in the first version of

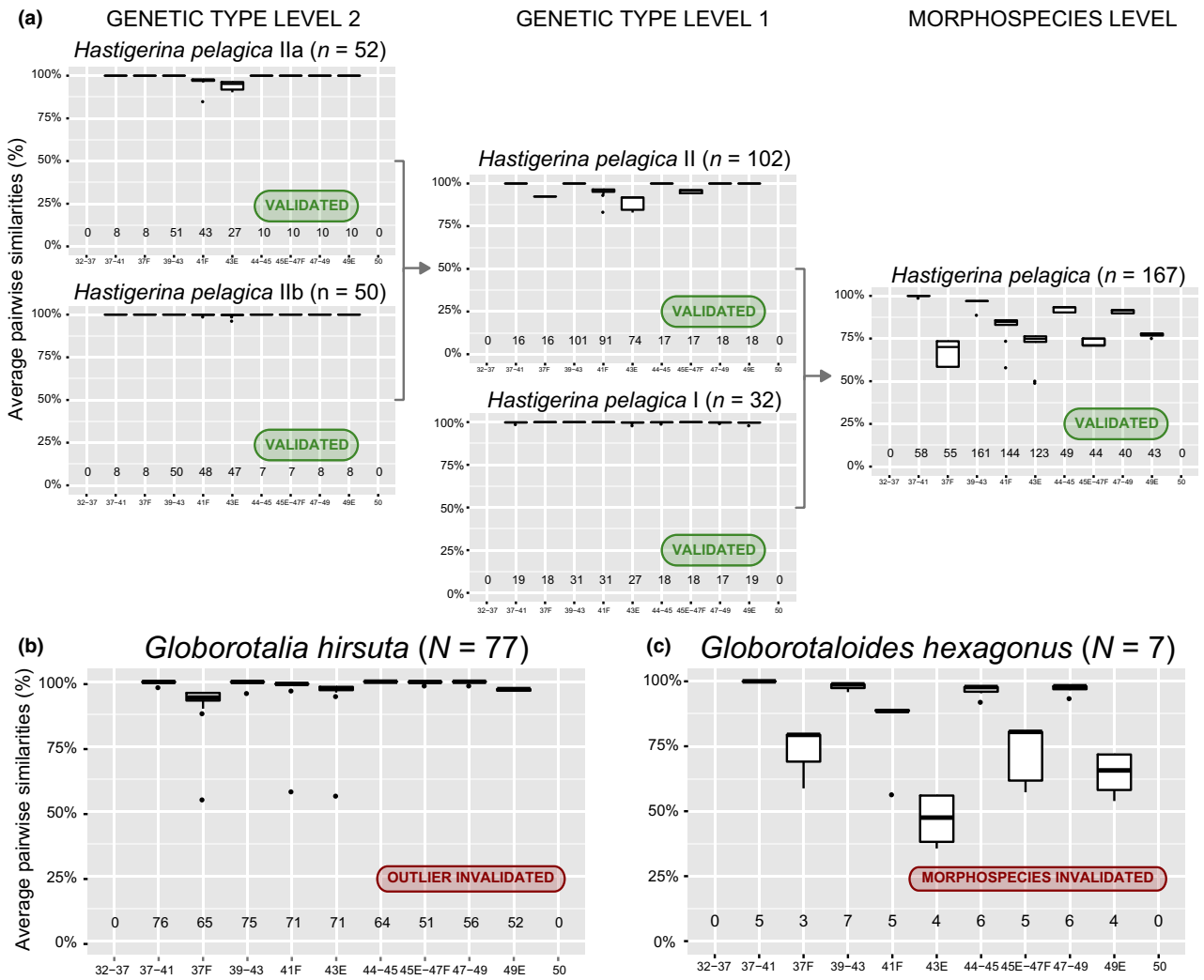

Fig. 3 Annotation inconsistency detection. The procedure followed to identify annotation inconsistencies is exemplified by three cases. Each graph represents variability in pairwise similarities observed across each region of all sequences sharing the same annotation level. The names of the taxon and 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, and 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 Appendix S2 (Supporting information), and the pairwise similarities calculated for each taxonomic level are given in Table S1 (Supporting information).

PFR^2 was blasted against the remaining part of the database including n-1 sequences using SWIPE (Rognes 2011). The sequences among the 'n-1 PFR^2 database' returning the highest score were retrieved and their taxonomic attribution compared to the one of the blasted sequence (Table S1, Supporting information). Second, we retrieved all sequences covering the five variable and six conserved regions and divided them according to their assignment to higher taxa (here simplified by the morphogroups monolamellar, nonspinose, spinose, and microperforates + benthic). Each subset was automatically aligned using MAFFT v.7 (Katoh & 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 analysed sequences for the monolamellar, nonspinose, spinose, and microperforates + nonspiral morphogroups, respectively. For each alignment, a tree was inferred using a neighbour-joining approach with Juke and Cantor distance while taking into account

gap sites as implemented in SEAVIEW 4 (Fig. S1, Supporting information) with 100 pseudoreplicates. The scripts used to perform the different curation steps are available as Appendix S3 (Supporting information).

Results

Of the 3532 planktonic foraminiferal 18S rDNA partial sequences analysed, 3347 (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 (Table S1, Supporting information). Among the 3347 curated sequences, the taxonomic assignment of 84 was initially invalidated. Of these, three 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 data set. 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 named *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 data set 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 3347 18S rDNA sequences from individual planktonic foraminifera, we generated the *Planktonic Foraminiferal Ribosomal Reference* or PFR^2 database. Of the 3347 sequences, 25 were shorter than 200 bp and could not be deposited in NCBI (see Table S1, Supporting information). The PFR^2 1.0 database thus includes 3322 reference sequences assigned to 32 morphospecies and six taxa with unresolved taxonomy (Fig. 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 PFR^2 to assess its robustness returned an identical taxonomic path for 2509 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 neighbour 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 neighbour. 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 (Fig. S1, Supporting information). This analysis confirmed the monophyly of all morphospecies, except the *G. calida*/*G. siphonifera* plexus. All clades were strongly supported except for the sister species *G. tumida* and *G. ungulata* and the monolamellar species *Hastigerina pelagica* and *H. 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 *H. 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 (Fig. 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.

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

The PFR^2 web interface

To facilitate data download and comparative sequence analyses, PFR^2 has been implemented into a dedicated web interface, available at http://pfr2.sb-roscoff.fr. The website provides:

1 a search/browse module, which allows the user to download parts of the database either by taxonomic

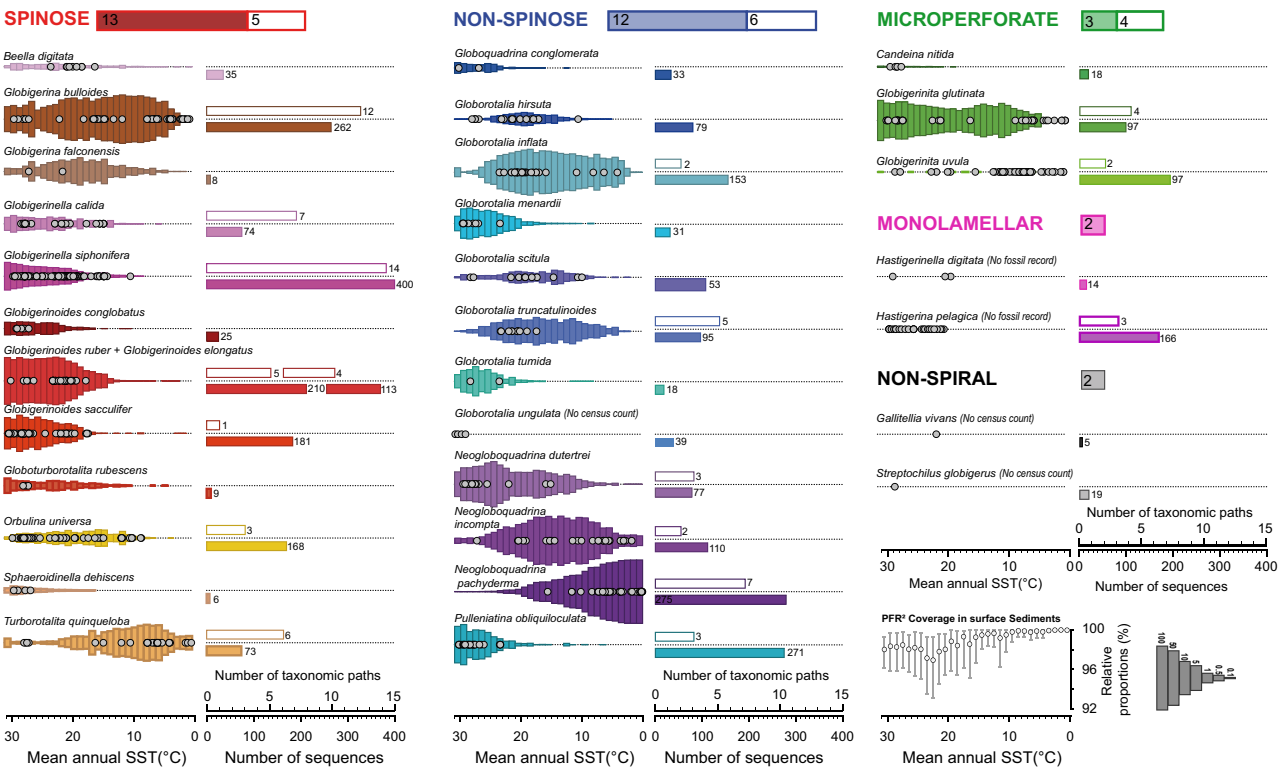

Fig. 4 Taxonomic and ecological coverage of PFR^2. For each morphogroup (spinose, nonspinose, microperforates, monolamellar and nonspiral), the number of species included in PFR^2 is given in the filled bar, while the number of species not present is indicated in the adjacent open bar. The relative abundance in the sediments of each species included in PFR^2 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 PFR^2 plotted against the mean annual SST in discrete 1 °C intervals. Vertical bars represent 95% confidence intervals for each 1 °C bin.

rank (morphogroup name, genus name, species name), geographic region (e.g. North Atlantic, Mediterranean Sea, Indian Ocean) or collection (cruise name);

2 a classical BLAST/similarity module that facilitates identification of unknown sequences;

3 a map module displaying the localities for all sequences present in PFR^2 and facilitating download of all data from each single locality; and

4 a download section with direct access to all data included in PFR^2. All sequences and sequence partitions are available in FASTA format, and the metadata are available in a tabulated file.

Discussion

Comprehensive databases of ribosomal RNA sequences with curated taxonomy are available for protists (Protist ribosomal reference database, PR^2; Guillou *et al.*, 2013) and for the major domains of life (SILVA; Yilmaz *et al.* 2013). These databases include sequences of planktonic foraminifera. However, they are used mainly as benchmarks to annotate complex environmental data sets (e.g. Logares *et al.* 2014) at the morphological species level. In contrast, PFR^2 has been designed and implemented in a way that facilitates other applications.

First, because of structural limitations, PR^2 contains 'only' 402 sequences of planktonic foraminifera (based on Released 203 of GenBank, October 2014), compared to PFR^2, which contains for now 3322 SSU rDNA sequences. Second, 2276 of the sequences present in PFR^2 have an assignation to the genetic type level, and as far as possible, the sequences are associated with metadata related to the origin of each specimen and the conditions where it was collected, thus forming a basis for ecological modelling. Third, most importantly, using planktonic foraminifera as a case study, we propose and implement an annotation scheme with unmatched accuracy and full tracking of changes. This is only possible because of the narrower focus of PFR^2 combined with high-level expert knowledge of their taxonomy. The fidelity of the annotations will facilitate

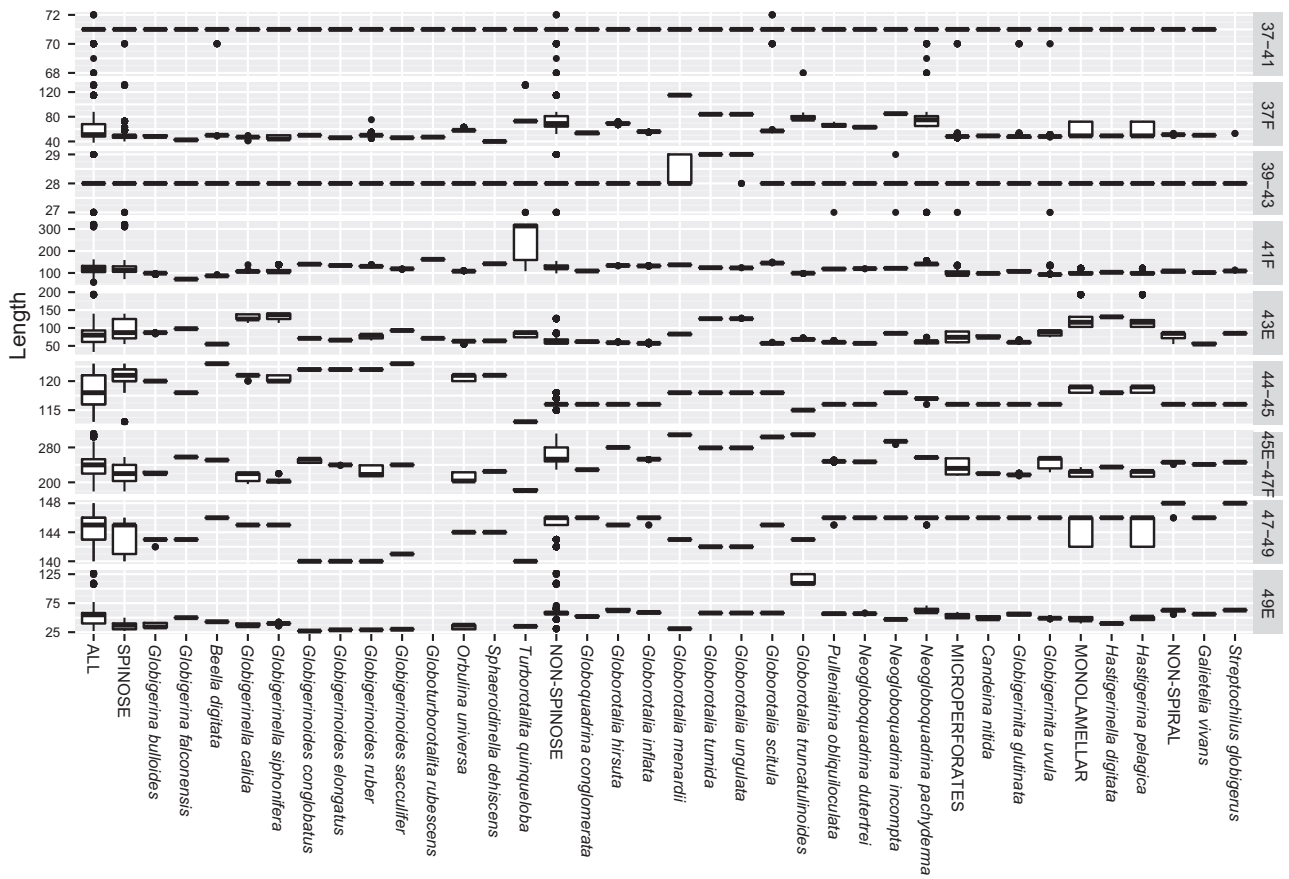

Fig. 5 Length polymorphism. Each rectangle represents the length polymorphism within each region of the analysed 18S rDNA fragment across all resolved taxonomic units in PFR^2. The regions are based on the rRNA secondary structure and are named following Pawlowski & Lecroq (2010).

a qualitatively entirely different level of analysis of eDNA libraries.

For example, the design of PFR^2 allows to incorporate advances in classical and molecular taxonomy, particularly at the level of genetic types (e.g. André *et al.* 2014), which can be re-evaluated depending on the criteria used to delineate molecular OTUs. Further, by retaining information on clone attribution to specimens (vouchers), PFR^2 allows to evaluate intragenomic polymorphism, which offers excellent opportunity to identify the taxonomically relevant level of variability (Weber and Pawlowski, 2014). Finally, the modular structure of PFR^2 (i.e. its partitioning into variable and conserved regions) is particularly suitable for the evaluation of existing barcodes or the design of new barcoding systems needed to capture total or partial 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 & Lecroq 2010), several other regions may be suitable as targets for barcoding of planktonic foraminifera (Fig. 5).

The main difference between PFR^2 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 PFR^2 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). Modelling studies showed that the integration of cryptic diversity into paleoceanographic studies will improve their accuracy (Kucera & 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, PFR^2 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 data sets. 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.

Acknowledgements

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 PFR^2 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 the effort of the SCOR/IGBP Working Group 138 'Modern planktonic foraminifera and ocean changes'.

References

André A, Weiner A, Quillévéré F *et al.* (2013) The cryptic and the apparent reversed: lack of genetic differentiation within the morphologically diverse plexus of the planktonic foraminifer *Globigerinoides sacculifer*. *Paleobiology*, **39**, 21–39.

André A, Quillévéré F, Morard R *et al.* (2014) SSU rDNA divergence in planktonic foraminifera: molecular taxonomy and biogeographic implications (V Ketmaier, Ed.). *PLoS ONE*, **9**, 1–19.

Aurahs R, Göker M, Grimm GW *et al.* (2009a) Using the multiple analysis approach to reconstruct phylogenetic relationships among planktonic foraminifera from highly divergent and length-polymorphic SSU rDNA sequences. *Bioinformatics and Biology Insights*, **3**, 155–177.

Aurahs R, Grimm GW, Hemleben V, Hemleben C, Kucera M (2009b) Geographical distribution of cryptic genetic types in the planktonic foraminifer *Globigerinoides ruber*. *Molecular Ecology*, **18**, 1692–1706.

Aurahs R, Treis Y, Darling K, Kucera M (2011) A revised taxonomic and phylogenetic concept for the planktonic foraminifer species *Globigerinoides ruber* based on molecular and morphometric evidence. *Marine Micropaleontology*, **79**, 1–14.

Aze T, Ezard THG, Purvis A *et al.* (2011) A phylogeny of Cenozoic macroperforate planktonic foraminifera from fossil data. *Biological Reviews of the Cambridge Philosophical Society*, **86**, 900–927.

Bé AWH, Hudson WH (1977) Ecology of planktonic foraminifera and biogeographic patterns of life and fossil assemblages in the Indian Ocean. *Micropaleontology*, **23**, 369–414.

Bé AWH, Tolderlund D (1971) Distribution and ecology of living planktonic foraminifera in surface waters of the Atlantic and Indian Oceans. In: *The Micropalaeontology of Oceans* (eds Funnell BM, Riedel WR), pp. 105–149, text-figs. 1–27. Cambridge University Press, London.

Brummer GA, Hemleben C, Michael S (1986) Planktonic foraminiferal ontogeny and new perspectives for micropalaeontology. *Nature*, **319**, 50–52.

Darling KF, Wade CM (2008) The genetic diversity of planktic foraminifera and the global distribution of ribosomal RNA genotypes. *Marine Micropaleontology*, **67**, 216–238.

Darling KF, Kroon D, Wade CM, Leigh Brown AJ (1996) Molecular phylogeny of the planktic foraminifera. *Journal of Foraminiferal Research*, **26**, 324–330.

Darling KF, Wade CM, Kroon D, Leigh Brown AJ (1997) Planktic foraminiferal molecular evolution and their polyphyletic origins from benthic taxa. *Marine Micropaleontology*, **30**, 251–266.

Darling KF, Wade CM, Kroon D, Leigh Brown AJ, Bijma J (1999) The diversity and distribution of modern planktic foraminiferal small subunit ribosomal RNA genotypes and their potential as tracers of present and past ocean circulations. *Paleoceanography*, **14**, 3–12.

Darling KF, Wade CM, Stewart IA *et al.* (2000) Molecular evidence for genetic mixing of Arctic and Antarctic subpolar populations of planktonic foraminifers. *Nature*, **405**, 43–47.

Darling KF, Kucera M, Wade CM, von Langen PJ, Pak DK (2003) Seasonal distribution of genetic types of planktonic foraminifer morphospecies in the Santa Barbara Channel and its paleoceanographic implications. *Paleoceanography*, **18**, 1–10.

Darling KF, Kucera M, Pudsey CJ, Wade CM (2004) Molecular evidence links cryptic diversification in polar planktonic protists to quaternary climate dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 7657–7662.

Darling KF, Kucera M, Kroon D, Wade CM (2006) A resolution for the coiling direction paradox in *Neogloboquadrina pachyderma*. *Paleoceanography*, **21**, PA2011.

Darling KF, Kucera M, Wade CM (2007) Global molecular phylogeography reveals persistent Arctic circumpolar isolation in a marine planktonic protist. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 5002–5007.

Darling KF, Thomas E, Kasemann SA *et al.* (2009) Surviving mass extinction by bridging the benthic/planktic divide. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 12629–12633.

Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution*, **27**, 221–224.

Guillou L, Bachar D, Audic S *et al.* (2012) The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, **41**, D597–D604.

Guillou L, Bachar D, Audic S, *et al.* (2013) The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic acids research*, **41**, D597–604.

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

Hemleben C, Spindler M, Anderson OR (1989) *Modern Planktonic Foraminifera*, pp. 363. Springer-Verlag New York Inc., New York, New York.

Hollingsworth PM, Forrest LL, Spouge JL *et al.* (2009) A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 12794–12797.

Holzmann M, Pawlowski J (1996) Preservation of foraminifera for DNA extraction and PCR amplification. *Journal of Foraminiferal Research*, **26**, 264–267.

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–780.

Kucera M, Darling KF (2002) Cryptic species of planktonic foraminifera: their effect on palaeoceanographic reconstructions. *Philosophical Transactions of the Royal Society of London, Series A, Mathematical, Physical, and Engineering Sciences*, **360**, 695–718.

Kucera M, Weinelt M, Kiefer T *et al.* (2005) Reconstruction of sea-surface temperatures from assemblages of planktonic foraminifera: multi-technique approach based on geographically constrained calibration data sets and its application to glacial Atlantic and Pacific oceans. *Quaternary Science Reviews*, **24**, 951–998.

Logares R, Audic S, Bass D *et al.* (2014) Patterns of rare and abundant marine microbial eukaryotes. *Current Biology*, **24**, 813–821.

Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B (2011) How many species are there on Earth and in the ocean? *PLoS Biology*, **9**, e1001127.

Morard R, Quillévéré F, Escarguel G *et al.* (2009) Morphological recognition of cryptic species in the planktonic foraminifer *Orbulina universa*. *Marine Micropaleontology*, **71**, 148–165.

Morard R, Quillévéré F, Douady CJ *et al.* (2011) Worldwide genotyping in the planktonic foraminifer *Globoconella inflata*: implications for life history and paleoceanography. *PLoS ONE*, **6**, 1–12.

Morard R, Quillévéré F, Escarguel G, de Garidel-thoron T (2013) Ecological modeling of the temperature dependence of cryptic species of planktonic foraminifera in the Southern Hemisphere. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **391**, 13–33.

Nassonova E, Smirnov A, Fahrni J, Pawlowski J (2010) Barcoding amoebae: comparison of SSU, ITS and COI genes as tools for molecular identification of naked lobose amoebae. *Protist*, **161**, 102–115.

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**, 443–453.

Nikolaev SI, Berney C, Fahrni JF *et al.* (2004) The twilight of Heliozoa and rise of Rhizaria, an emerging supergroup of amoeboid eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 8066–8071.

Pawlowski J, Lecroq B (2010) Short rDNA barcodes for species identification in foraminifera. *The Journal of Eukaryotic Microbiology*, **57**, 197–205.

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

Pawlowski J, Holzmann M (2014) A plea for DNA barcoding of foraminifera. *Journal of foraminiferal research*, **44**, 62–67.

Pawlowski J, Lejzerowicz F, Esling P (2014) Next-generation environmental diversity surveys of foraminifera: preparing the future. *Biological Bulletin*, **227**, 93–106.

Quillévéré F, Morard R, Escarguel G *et al.* (2013) Global scale same-specimen morpho-genetic analysis of *Truncorotalia truncatulinoides*: a perspective on the morphological species concept in planktonic foraminifera. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **391**, 2–12.

R Development Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available from http://www.R-project.org.

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics*, **16**, 2–3.

Rognes T (2011) Faster Smith-Waterman database searches with inter-sequence SIMD parallelisation. *BMC Bioinformatics*, **12**, 221.

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

Schoch CL, Seifert KA, Huhndorf S *et al.* (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for fungi. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 6241–6246.

Seears HA, Darling KF, Wade CM (2012) Ecological partitioning and diversity in tropical planktonic foraminifera. *BMC Evolutionary Biology*, **12**, 54.

Sherwood AR, Presting GG (2007) Universal primers amplify a 23S rDNA plastid marker in eukaryotic algae and cyanobacteria. *Journal of Phycology*, **43**, 605–608.

Sogin ML, Morrison HG, Huber JA *et al.* (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 12115–12120.

Spencer-Cervato C, Thierstein HR, Lazarus DB, Beckmann J-P (1994) How synchronous are neogene marine plankton events? *Paleoceanography*, **9**, 739.

Stern RF, Horak A, Andrew RL *et al.* (2010) Environmental barcoding reveals massive dinoflagellate diversity in marine environments. *PLoS ONE*, **5**, e13991.

Stewart IA, Darling KF, Kroon D, Wade CM, Troelstra SR (2001) Genotypic variability in subarctic Atlantic planktic foraminifera. *Marine Micropaleontology*, **43**, 143–153.

Ujiié Y, Lipps JH (2009) Cryptic diversity in planktonic foraminifera in the northwest Pacific ocean. *Journal of Foraminiferal Research*, **39**, 145–154.

Ujiié Y, Kimoto K, Pawlowski J (2008) Molecular evidence for an independent origin of modern triserial planktonic foraminifera from benthic ancestors. *Marine Micropaleontology*, **69**, 334–340.

Ujiié Y, Asami T, de Garidel-Thoron T *et al.* (2012) Longitudinal differentiation among pelagic populations in a planktic foraminifer. *Ecology and Evolution*, **2**, 1725–1737.

de Vargas C, Pawlowski J (1998) Molecular versus taxonomic rates of evolution in planktonic foraminifera. *Molecular Phylogenetics and Evolution*, **9**, 463–469.

de Vargas C, Zaninetti L, Hilbrecht H, Pawlowski J (1997) Phylogeny and rates of molecular evolution of planktonic foraminifera: SSU rDNA sequences compared to the fossil record. *Journal of Molecular Evolution*, **45**, 285–294.

de Vargas C, Norris R, Zaninetti L, Gibb SW, Pawlowski J (1999) Molecular evidence of cryptic speciation in planktonic foraminifers and their relation to oceanic provinces. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 2864–2868.

de Vargas C, Renaud S, Hilbrecht H, Pawlowski J (2001) Pleistocene adaptive radiation in *Globorotalia truncatulinoides*: genetic, morphologic, and environmental evidence. *Paleobiology*, **27**, 104–125.

de Vargas C, Bonzon M, Rees NW, Pawlowski J, Zaninetti L (2002) A molecular approach to biodiversity and biogeography in the planktonic foraminifer *Globigerinella siphonifera* (d'Orbigny). *Marine Micropaleontology*, **45**, 101–116.

Wade CM, Darling KF, Kroon D, Brown AJL (1996) Early evolutionary origin of the planktic foraminifera inferred from small subunit rDNA sequence comparisons. *Journal of Molecular Evolution*, **43**, 672–677.

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

Weiner A, Aurahs R, Kurasawa A, Kitazato H, Kucera M (2012) Vertical niche partitioning between cryptic sibling species of a cosmopolitan marine planktonic protist. *Molecular Ecology*, **21**, 4063–4073.

Weiner AKM, Weinkauf MFG, Kurasawa A *et al.* (2014) Phylogeography of the tropical planktonic foraminifera lineage *Globigerinella* reveals isolation inconsistent with passive dispersal by ocean currents. *PLoS ONE*, **9**, e92148.

Weiner AKM, Weinkauf MFG, Kurasawa A, Darling KF, Kucera M (2015) Genetic and morphometric evidence for parallel evolution of the *Globigerinella calida* morphotype. *Marine Micropaleontology*, **114**, 19–35.

Wickham H (2009) *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York, New York.

Yilmaz P, Parfrey LW, Yarza P *et al.* (2013) The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Research*, **42**, D643–D648.

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.

K.F.D., C.d.V., Y.U., R.M., T.d.G.-T., A.K.M.W., H.A.S., M.K., A.A. and M.S. participated in sample collection; C.d.V., M.K., K.F.D., C.M.W., C.J.D., F.Q., G.E. and T.d.G.-T. provided laboratory infrastructure; K.F.D., Y.U., R.M., A.K.M.W., A.A. and H.A.S. participated in laboratory work. F.M. and R.M. conceived and designed the bioinformatics pipeline, F.M. performed the computational work, and S.A. built the website. R.M. wrote the manuscript with help from M.K. and C.d.V. All authors read, edited and approved the final manuscript.

Data accessibility

Sequences, NCBI Accession nos and metadata are available in Table S1 and Appendix S1 (Supporting information) and on the PFR^2 website at http://pfr2.sb-roscoff.fr. The custom scripts used to perform the curation procedure are available in Appendix S3 (Supporting information); the results of the curation process are given in Table S1 and Appendix S1 (Supporting information).

Supporting Information

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

Table S1 Information on all consecutive steps followed to constitute the PFR^2.

Fig. S1 Neighbor-joining trees showing the monophyly of each morphospecies present in PFR^2.

Appendix S1 FASTA files of sequences used to build the PFR^2.

Appendix S2 Box plots showing pairwise similarities for each taxonomic level.

Appendix S3 Custom scripts used to perform the different curation steps.