



PFR²: 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 Sears, Chris M Wade, Frédéric Quillévéré, et al.

► To cite this version:

Raphaël Morard, Kate F Darling, Frédéric Mahé, Stéphane Audic, Yurika Ujiié, et al.. PFR²: a curated database of planktonic Foraminifera18S ribosomal DNA as a resource for studies of plankton ecology, biogeography, and evolution. *Molecular Ecology Resources*, 2015, in press. 10.1111/1755-0998.12410 . hal-01149023v1

HAL Id: hal-01149023

<https://hal.sorbonne-universite.fr/hal-01149023v1>

Submitted on 6 May 2015 (v1), last revised 11 May 2015 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1PFR²: a curated database of planktonic Foraminifera 18S ribosomal DNA as a resource for studies
2 of plankton ecology, biogeography, and evolution.

3 Raphaël Morard^{1,2,3}, Kate F. Darling^{4,5}, Frédéric Mahé⁶, Stéphane Audic^{1,2}, Yurika Ujiie⁷, Agnes
4 K. F. Weiner³, Aurore André^{8,9}, Heidi Seears^{10,11}, Chris M. Wade¹⁰, Frédéric Quillévéré⁸,
5 Christophe J. Douady^{12,13}, Gilles Escarguel⁸, Thibault de Garidel-Thoron¹⁴, Michael Siccha³,
6 Michal Kucera³ and Colomban de Vargas^{1,2}

7¹*Centre National de la Recherche Scientifique, UMR 7144, EPEP, Station Biologique de Roscoff,*
8 *France*

9²*Sorbonne Universités, UPMC Univ Paris 06, UMR 7144, Station Biologique de Roscoff, France*

10 ³*MARUM Center for Marine Environmental Sciences, University of Bremen, Leobener Strasse,*
11 *28359 Bremen, Germany*

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

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

14⁶*Department of Ecology, Technische Universität Kaiserslautern, 67663 Kaiserslautern, Germany*

15 ⁷*Department of Biology, Shinshu University, Matsumoto, Japan*

16 ⁸*CNRS UMR 5276, Laboratoire de Géologie de Lyon: Terre, Planètes, Environnement,*
17 *Université Lyon 1, Villeurbanne, France*

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

20 ¹⁰*School of Life Sciences, University of Nottingham, University Park, Nottingham, NG7 2RD,*
21 *UK*

22 ¹¹*Department of Biological Sciences, Lehigh University, Bethlehem, USA*

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

25 ¹³*Institut Universitaire de France, Paris, France*

26 ¹⁴*Centre Européen de Recherche et d'Enseignement de Géosciences de l'Environnement, Centre*
27 *National de la Recherche Scientifique, et Aix-Marseille Université, Aix-en-Provence, France*

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

30 Corresponding Author: Raphaël Morard, MARUM Center for Marine Environmental Sciences,
31 University of Bremen, Leobener Strasse, 28359 Bremen, Germany, Fax: +49 (0) 421 218 –
32 9865974, rmorard@marum.de.

33 Abstract

34 Planktonic Foraminifera (Rhizaria) are ubiquitous marine pelagic protists producing
35 calcareous shells with conspicuous morphology. They play an important role in the marine
36 carbon cycle and their exceptional fossil record serves as the basis for past climate
37 reconstructions. A major worldwide sampling effort over the last two decades has resulted in the
38 establishment of multiple large collections of cryopreserved individual planktonic foraminifera
39 samples. Thousands of 18S rDNA partial sequences have been generated, representing all major
40 known morphological taxa across their worldwide oceanic range. This comprehensive data

41coverage provides an opportunity to assess patterns of molecular ecology and evolution in a
42holistic way for an entire group of planktonic protists. We combined all available published and
43unpublished genetic data to build PFR², the *Planktonic Foraminifera Ribosomal Reference*
44database. The first version of the database includes 3,322 reference 18S rDNA sequences
45belonging to 32 of the 47 known morphospecies of planktonic Foraminifera, collected from 460
46oceanic stations. All sequences have been rigorously taxonomically curated using a six-rank
47annotation system fully resolved to the level of morphological species and linked to a series of
48metadata. The PFR² website, available at <http://pfr2.sb-roscoff.fr>, allows downloading the entire
49database or specific sections, as well as the identification of new planktonic Foraminiferal
50sequences. Its novel, fully documented curation process integrates advances in morphological
51and molecular taxonomy. It allows for an increase in its taxonomic resolution and assures that
52integrity is maintained by including a complete contingency tracking of annotations and assuring
53that the annotations remain internally consistent.

54Introduction

55 Despite their ubiquity and the critical role they play in global biogeochemical cycles,
56unicellular eukaryotes (protists) remain the most poorly known domain of life (e.g. Pawlowski et
57al., 2012). Because of their extreme morphological and behavioral diversity, the study of even
58relatively narrow lineages requires a high degree of taxonomic expertise (e.g. Guillou et al.,
592012, Pawlowski and Holzmann, 2014). As a result, the knowledge of protistan ecology and
60evolution is limited by the small number of taxonomists resulting in scarcity of taxonomically
61well-resolved ecological data. As an alternative approach, numerous studies have demonstrated
62the potential of identification of protists by means of short DNA sequences or barcodes (e.g.,
63Saunders, 2005; Sherwood et al., 2007; Hollingsworth et al., 2009; Nossonova et al., 2010;

64Pawlowski and Lecroq, 2010; Hamsher et al., 2011; Stern et al., 2010; Schoch et al., 2012), both
65at the single-cell and metacommunity levels (e. g., Sogin et al., 2006; Logares et al., 2014, de
66Vargas et al., 2015). Such barcoding/metabarcoding approaches critically rely on the fidelity of
67the marker gene with respect to specificity (avoiding ambiguity in identification),
68comprehensiveness (assuring all taxa in the studied group are represented in the reference
69barcode database) and accuracy (assuring that barcode assignments are consistent with a
70coherent, phenotypic taxonomic framework; e. g. Zimmermann et al., 2014). These three pre-
71requisites are rarely found in protists, where classical morphological taxonomy is often
72challenging, DNA extraction and sequencing from a single cell is prone to contamination, and a
73large portion of the diversity in many groups remains unknown (e.g. Mora et al., 2011). In this
74respect, planktonic Foraminifera represent a rare exception.

75 Planktonic Foraminifera are ubiquitous pelagic marine protists with reticulated
76pseudopods, clustering within the Rhizaria (Nikolaev et al., 2004). The group is marked by a
77rather low number of morphospecies (47; Hemleben et al., 1989), which can be distinguished
78using structural characteristics of their calcite shells. Their global geographic distribution,
79seasonal dynamics, vertical habitats and trophic behavior have been thoroughly documented by
80analyses of plankton hauls (e.g., Bé and Hudson, 1977), sediment trap series (e.g., Zaric et al.,
812005) and thousands of surface sediment samples across the world oceans (e.g., Kucera et al.,
822005). Their outstanding preservation in marine sediments resulted in arguably the most
83complete fossil record, allowing comprehensive reconstruction of the evolutionary history of the
84group (Aze et al., 2011). The morpho-taxonomy and phylogeny of the group have been largely
85confirmed by molecular genetic analyses (e.g., Aurahs et al., 2009a) based on the highly
86informative, ~1,000 bp fragment at the 5'end of the 18S rDNA gene. These analyses confirmed

87that the morphological characters used to differentiate planktonic Foraminifera taxa are
88phylogenetically valid both at the level of morphological species and at the level of higher taxa.
89The studied gene fragment contains six hypervariable expansion segments, some unique to
90Foraminifera, providing excellent taxonomic resolution (Pawlowski and Lecroq, 2010). Analyses
91of this fragment revealed the existence of genetically distinct lineages within most of the
92morphospecies, which likely represent reproductively isolated units (Darling et al., 1996, 1997,
931999, 2000, 2003, 2004, 2006, 2007, 2008, 2009; Wade et al., 1996; de Vargas et al., 1997, 1999,
942001, 2002, de Vargas and Pawlowski, 1998; Stewart et al., 2001; Aurahs et al., 2009b, 2011;
95Ujiié et al., 2008, 2009, 2012; Morard et al., 2009, 2011, 2013; Sears et al., 2012; Weiner, 2012,
962014; André et al., 2014). In order to assess the ecology and biogeography of such cryptic
97species, large numbers of rDNA sequences from single-cell extractions collected across the
98world oceans have been generated for most morphospecies (Figure 1). Due to this extensive
99single-cell rDNA sequencing throughout the last decades, the genetic and morphological
100diversity of planktonic foraminifera have been linked together to a degree that now allows for
101transfer of taxonomic expertise. The knowledge of the genetic and morphological taxonomy of
102the group allows the establishment of an exceptionally comprehensive reference genetic database
103that can be further used to interpret complex data from plankton metagenomic studies with a
104high level of taxonomic resolution. Because planktonic Foraminifera are subject to the same
105ecological forcing as other microplankton, including the dominance of passive transport in a
106relatively unstructured environment, huge population sizes, and basin-scale distribution of
107species, they can potentially serve as a model for the study of global ecological patterns in other
108groups of pelagic protists, whose diversity remains largely undiscovered (Mora et al., 2011).

109 By early 2014, 1,787 partial 18S rDNA sequences from single-cell extractions of
110 planktonic Foraminifera were available in public databases. However, their NCBI taxonomy is
111 often inconsistent, lacking standardization. It includes (and retains) obvious identification errors,
112 as discussed by Aurahs et al. (2009) and André et al. (2014), and their annotation lacks critical
113 metadata. In addition, an equivalent number of rDNA sequences not deposited in public
114 databases have been generated by the co-authors of the present study. Collectively, the existing
115 rDNA sequences from single cells collected throughout the world oceans cover the entire
116 geographic and taxonomic range of planktonic Foraminifera. This collection unites the current
117 morphological, genetic, ecological, and biogeographic knowledge of the group and may serve as
118 a *Rosetta Stone/Philae Obelisk* for interpreting metabarcoding data (Pawlowski et al., 2014). To
119 pave the way for future exploitation of this resource, we combined all published and unpublished
120 planktonic Foraminifera rDNA sequence data and curated the resulting database with a semi-
121 automated bioinformatics pipeline. The resulting “Planktonic Foraminifera Ribosomal Reference
122 database” (PFR²) is a highly resolved, fully annotated and internally entirely consistent collection
123 of 18S rDNA sequences of planktonic Foraminifera, aligned and evaluated in a way that
124 facilitates direct assessment of barcoding markers.

125 **Material and Methods**

126 *Primary database assembly*

127 A total of 1,787 18S rDNA sequences of planktonic Foraminifera were downloaded from the
128 GenBank query portal (<http://www.ncbi.nlm.nih.gov/>; release 201) on the 14th of May 2014. The
129 taxonomic path and metadata for these sequences were extracted from NCBI and supplemented
130 by information in original papers when available. The metadata associated to each sequence

131consisted of: (i) their organismal origin (specimen voucher, taxonomic path, infra specific
132genetic type assignment), (ii) their methodological origin (direct sequencing or cloning), and (iii)
133their spatio-temporal origin (geographic coordinates, depth, and time of collection). Metadata
134were described using standard vocabularies and data formats. For 47 sequences, the coordinates
135of the collection site could not be recovered, in which case the locality was described in words
136(Supplementary Material 1).

137We next compiled all unpublished 18S rDNA sequences generated by the authors of this paper
138and linked them with the same suite of metadata. These sequences originate from single-cell
139extractions of planktonic Foraminifera collected by stratified or non-stratified plankton net hauls,
140in-situ water pumping, as well as SCUBA diving. After collection, the specimens were
141individually picked under a stereomicroscope, cleaned, taxonomically identified and transferred
142into DNA extraction buffer or air-dried on cardboard slides and stored at -20°C or -80°C. DNA
143extractions were performed following the DOC (Holzmann & Pawlowski, 1996), the GITC*
144(Morard et al., 2009), or the Urea (Weiner et al., 2014) protocols. Sequences located at the 5' end
145of the 18S rDNA were obtained following the methodology described in de Vargas et al. (1997),
146Darling et al. (1996, 1997), Aurahs et al. (2009b), Morard et al. (2011) and Weiner et al., (2014).
147In total, 820 new planktonic Foraminiferal sequences were analyzed and annotated for this study.
148In addition, 925 unpublished sequences analyzed in Darling et al. (2000, 2003, 2004, 2006,
1492007), Darling and Wade (2008), Sears et al. (2012) and Weiner et al. (2014) were also
150included. All unpublished sequences, except 177 sequences shorter than 200bp, were deposited
151in GenBank under the accession numbers KM19301 to KM194582. Overall, PFR² contains data
152from 460 sites sampled during 54 oceanographic cruises and 15 near shore collection campaigns

153between 1993 and 2013. It covers all oceanic basins, all seasons, and water depths ranging
154between the surface and 700 meters (Figure 1; Supplementary Material 1).

155*Taxonomy*

156Morphological taxonomy

157As the first step in the curation process, the primary taxonomic annotations of all 3,532 18S
158rDNA sequences gathered from NCBI and our internal databases were harmonized. The
159identification of planktonic Foraminifera is challenging especially for juvenile individuals, which
160often lack diagnostic characters (Brummer et al., 1986). Thus, many of the published and
161unpublished 18S rDNA sequences were mislabelled or left in open nomenclature. In some cases
162the same taxon has been recorded under different names, reflecting inconsistent usage of generic
163names, synonyms and misspelling. To harmonize the taxonomy, we first carried out a manual
164curation of the original annotations to remove the most obvious taxonomic conflicts in the
165primary database. To this end, the sequence annotations were aligned with a catalog of 47 species
166names based on the taxonomy used in Hemleben et al. (1989), but adding *Globigerinoides*
167*elongatus* following Aurahs et al. (2011) and treating *Neogloboquadrina incompta* following
168Darling et al. (2006). Thus, the 109 sequences labelled as *Globigerinoides ruber* (pink) and the
16963 labelled as *Globigerinoides ruber* (white) were renamed as *Globigerinoides ruber*. The 113
170sequences of *Globigerinoides ruber* and *Globigerinoides ruber* (white) attributed to the
171genotypes II were renamed *Globigerinoides elongatus*. The 12 sequences labelled *Globigerinella*
172*aequilateralis* were renamed *Globigerinella siphonifera* following Hemleben et al. (1989). The 7
173sequences corresponding to the right-coiled morphotype of *Neogloboquadrina pachyderma* were
174renamed *Neogloboquadrina incompta*. All taxonomic reassignments were checked by sequence

175similarity analyses to the members of the new group. Next, we attempted to resolve the
176attribution of sequences with unresolved taxonomy and searched manually for obviously
177misattributed sequences. This refers to sequences which are highly divergent from other
178members of their group but identical to sequences of other well resolved taxa. Overall, these first
179steps of manual curation led to taxonomic reassignment of 124 sequences. All corrections and
180their justification are documented in the Supplementary Material 1.

181Molecular taxonomy

182In order to preserve the information on the attribution of 18S rDNA sequences to genetic types
183(potential cryptic species), we harmonized the existing attributions at this level for species where
184extensive surveys have been carried out and published. A total of 1,356 sequences downloaded
185from NCBI were associated with a genetic type label, which was always retained. In addition, 19
186sequences labelled as *Globigerinoides ruber*, 15 as *Globigerinoides sacculifer*, 36 as
187*Globigerinita glutinata*, 6 as *Globigerinita uvula*, 9 as *Globorotalia inflata*, 10 as
188*Neogloboquadrina incompta*, 6 as *Neogloboquadrina pachyderma*, 5 as *Orbulina universa*, 5 as
189*Pulleniatina obliquiloculata*, 30 as *Hastigerina pelagica* and 32 as *Globigerinella siphonifera*
190have been analyzed after their first release in the public domain by Aurahs et al. (2009), Ujiie et
191al. (2012), Weiner et al. (2012, 2014) and André et al. (2013, 2014), and were attributed to a
192genetic type by these authors. These attributions differ from those in the NCBI label, but were
193retained in the PFR² database. In case of multiple attributions of the same sequence to different
194genetic types by several authors, we retained the molecular taxonomy that was based on the
195study presenting the most resolved and comprehensive attribution. In addition, 877 unpublished
196sequences belonging to *Orbulina universa*, *Globigerina bulloides*, *Neogloboquadrina incompta*,
197*Neogloboquadrina dutertrei*, *Neogloboquadrina pachyderma*, and *Turborotalita quinqueloba*

198received a genotypic attribution following de Vargas et al. (1999) and Darling et al. (2004, 2006,
1992007, 2008). Most of these sequences have been produced and identified within earlier studies,
200but were not originally deposited on NCBI. Their PFR² genotypic assignment is therefore
201entirely consistent with the attribution of the representative sequences of the same genetic type
202that were deposited on NCBI.

203PFR² final taxonomic framework

204As a result of the first manual curation and annotation to the level of genetic type, the original
2053,532 18S rDNA sequences were re-assigned to 33 species names and 2,276 sequences were
206annotated to the level of genetic types (Supplementary Material 1). For all sequences, we
207established a ranked taxonomy with six levels: 1- Morphogroup, 2-Genus, 3-Species, 4-Genetic
208type level 1, 5-Genetic type level 2, 6-Genetic type 3. For the “Morphogroup” rank we used the
209taxonomical framework of Hemleben et al. (1989), dividing the extant planktonic Foraminifera
210species into five clades based on the ultrastructure of the calcareous shell: Spinose, Nonspinose,
211Microperforate, Monolamellar and Non-spiral. The “Genus” and “Species” ranks follow the
212primary annotation as described above. For the “Genetic type level 1”, “Genetic type level 2”
213and “Genetic type level 3” ranks, we used the hierarchical levels presented in the labels of the
214genetic types of *Globigerinoides ruber*, *Globigerinoides elongatus*, *Globigerinella siphonifera*,
215*Globigerinella calida*, *Hastigerina pelagica*, *Globigerina bulloides*, *Neogloboquadrina dutertrei*,
216*Pulleniatina obliquiloculata* and *Turborotalita quinqueloba*. Genetic type attributions lacking
217hierarchical structure were reported in the rank “Genetic type level 1”. After this step, the
218Primary Reference Database (Figure 2) of 3,532 sequences contained 113 different taxonomic
219paths (Supplementary Material 1).

220 Sequences partitioning into conserved and variable regions

221 Because PFR² is a resource not only for taxonomic assignment but also for ecological and
222 biogeographical studies, all planktonic Foraminiferal 18S rDNA sequences were included
223 irrespective of length, as long as they contained taxonomically relevant information. As a result,
224 the length of the sequences included in the annotated primary database ranges between 33 and
225 3,412 bp. To evaluate their coverage and information content, all sequences were manually
226 aligned using Seaview 4 (Gouy et al., 2010) to the borders of each variable region of the 18S
227 rDNA fragment. The positions of the borders were determined according to the SSU rDNA
228 secondary structure of the monothalamous Foraminifera *Micrometula hyalostera* presented by
229 Pawlowski and Lecroq (2010), except for the region 37/f where a strict homology was difficult to
230 establish for all sequences. Instead, we defined the end of this region by the occurrence of a
231 pattern homologous to the series of nucleotides “CUUUCACAUGA” located at the 3’ end of
232 Helix 37. We also noticed that the short conserved fragment located between the variable regions
233 45/e and 47/f was difficult to identify across all sequences. We thus merged the regions 45/e, 46
234 and 47/f into a single region that we named 45E-47F (Table1). As a result, the position and
235 length of six conserved (32-37, 37-41, 39-43, 44-45, 47-49, 50) and five variable (37F, 41F, 43E,
236 45E-47F, 49E) regions were identified for all sequences (Figure 2). The remaining part of the
237 18S rDNA sequence, only present in sequences EU199447, EU199448 and EU199449 and
238 located before the motive “AAGGGCACCACAAGA” has not been analyzed in this way. All
239 regions fully covered in a sequence and containing sequence motives observed at least twice in
240 the whole dataset were labelled as “complete”. Regions fully covered but containing a sequence
241 motive that was observed only once in the whole dataset were labelled as “poor”. This is because
242 we consider sequencing/PCR errors as the most likely cause for the occurrence of such unique

243sequence motives. We realize that using this procedure, even genuine unique sequences may be
244discarded from the analysis, but this would be the case only if such sequences deviated in all
245regions. In all other cases, the regions were labelled as “partial” when only a part of the region
246was present or “not available” if they did not contain any fragment of the sequence. As a result
247we obtain the Partitioned Primary Reference Database (Figure 2). The coverage of each
248individual region in the Partitioned Primary Reference Database is given in Supplementary
249Material 1, and all sequence partitions are given in Supplementary Material 2.

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

251The consistency of taxonomic assignments within the annotated database of partitioned
252sequences was assessed using a semi-automated process (Figure 2 and 3). All “complete” regions
253of sequences with the same taxonomic assignment at the morphospecies level were automatically
254aligned using global pairwise alignment (Needleman & Wunsch 1970), as implemented in the
255software *needle* from the Emboss suite of bioinformatics tools (Rice et al., 2000). To detect
256annotation inconsistencies, mean pairwise similarities were computed for each “complete”
257region of each sequence against all other sequences with the same taxonomic assignment from
258the finest annotation level “Genetic type level 3” to the rank “Species level”. Results are
259provided in Supplementary Material 1 and were visualized using R (R Development Core Team,
2602014) and the ggplot2 library (Wickham, 2009). The resulting plots are given in Supplementary
261Material 3. If all annotations are consistent and there is no variation within taxa, each sequence
262within the analyzed taxon should only find an exact match and the mean pairwise similarity for
263that taxon should be 1. However, there are several reasons why the mean pairwise similarity
264within a taxon may be lower. First, if a sequence has been assigned the wrong name, its
265similarity to all other sequences labelled with that name will be low and the resulting mean

pairwise similarity decreases. 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 may 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 at the “Species level”, always starting with the full database (Figure 2 and 3A). Thus, at each level, we expected a misidentified sequence to have a lower pairwise similarity from the mean than any pairwise similarity 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. Once “validated”, sequences cannot be “invalidated” during analyses of higher rank taxa, because they represent known 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 level of genetic type 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) and 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, to evaluate the robustness and potential limitations of the curated taxonomy, we performed a leave-one-out BLAST analysis and a monophyly validation by NJ on long sequences. 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

312their taxonomic attribution compared to the one of the blasted sequence (Supplementary Material
3131). Second, we retrieved all sequences covering the 5 variable and 6 conserved regions and
314divided them according to their assignment to higher taxa (here simplified by the morphogroups
315Monolamellar, Non-Spinose, Spinose and Microperforates + Benthic). Each subset was
316automatically aligned using MAFFT v.7 (Kato et al., 2013) and the subsequent alignments were
317trimmed off on the edge to conserve only homologous fragments. For each alignment, a
318phylogenetic tree was inferred using a Neighbor-Joining approach with Juke and Cantor distance
319while taking into account gap sites as implemented in SEAVIEW 4 (Supplementary Material 4)
320with 100 pseudo-replicates. The scripts used to perform the different curation steps are available
321as Supplementary Material 5.

322Results

323Of the 3,532 planktonic Foraminiferal 18S rDNA partial sequences analyzed, 3,347 contained at
324least one gene region that was considered “complete” and could be subjected to the curation
325process. The remaining 185 sequences included 33 singletons (rare motives or poor quality
326sequences) and 152 sequences that were too short to cover at least one region (Supplementary
327Material 1). Amongst the 3,347 curated sequences, the taxonomic assignment of 84 was initially
328invalidated. Of these, 3 represent cases where the morphospecies attribution was correct, but the
329attribution to a genetic type was erroneous. In 46 cases, the invalidated sequences found a perfect
330match with a different taxon and thus their taxonomic assignment was changed. In all of these
331cases, the novel taxonomic assignment corresponded to a morphologically similar
332morphospecies, explaining the original misidentification of the sequenced specimen. In 14 cases,
333the original assignment was retained because the sequences did not find any match and their
334original attribution did not appear in the validated part of the dataset. All of these sequences were

335labelled as *Hastigerinella digitata*. This species name had been entirely invalidated in the first
336step because of inconsistent use of the homonymous species name *Beella digitata*. Finally, 17
337sequences received an unresolved artificial assignment. These represent six different sequence
338motives diverging substantially from all sequences in the validated part of the database and also
339between each other. Because the original attribution upon collection was obviously wrong, we
340could not reassign these sequences to the species level. In two cases, we could identify the most
341likely generic attribution, but four sequences are left with an entirely unresolved path. Finally,
342our procedure captured one sequence with a spelling error in its path and three sequences that
343appear to have been attributed correctly but represent small variants within species. After
344resolution of the 84 conflicts described above, the re-annotated dataset was subjected to a second
345round of the curation process for verification. All sequences were validated.

346Having established an internally consistent taxonomic annotation for all 3,347 18S rDNA
347sequences from individual planktonic Foraminifera, we generated the *Planktonic Foraminiferal*
348*Ribosomal Reference* or PFR² database. Of the 3,347 sequences, 25 were shorter than 200 bp, and
349could not be deposited in NCBI (see Supplementary Material 1). The PFR²1.0 database thus
350includes 3,322 reference sequences assigned to 32 species and 6 taxa with unresolved taxonomy
351(Figure 2), and contains 119 unique taxonomic paths when including all three levels of genetic
352types.

353The leave-one-out BLAST evaluation applied on the first version of PFR² to assess its robustness
354returned an identical taxonomic path for 2,509 sequences. For 614 sequences, the BLAST-
355determined taxonomic paths were identical between the “morphogroup” and “species” rank but
356displayed a different resolution between the ranks “genetic type level 1” and “genetic type level
3573”. 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 assigned to the closest genetic type within the same species by the leave-one-out procedure. Thus, 94.5 % of the sequences in the PFR² database find a nearest neighbor with a correct taxonomic assignment at the target level of species. For the remaining 180 sequences, the returned taxonomic path was inconsistent at the level of species. In two cases, the sequences were assigned to a sister species, which is morphologically and phylogenetically close (*Globorotalia unguolata* 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 mutually interchangeably (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 PFR² and confirms that the curated taxonomy is internally entirely consistent. To further confirm the validity of morphospecies level taxonomy, we constructed NJ phylogenies for the four major 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 unguolata* and the monolamellar species *Hastigerina pelagica* and *Hastigerinella digitata*. In the first case, the poor support reflects the lack of differentiation between the two species in the conserved region of the gene which decreases the bootstrap score and in the second case the extreme divergence of

380the two genetic lineage of *Hastigerina pelagica* renders the phylogenetic reconstruction difficult
381(Weiner et al., 2012).

382An analysis of the taxonomic annotations retained in PFR² reveals that the database covers at
383least 70-80% of the traditionally recognized planktonic Foraminiferal species in each clade. The
384species represented in PFR² constitute the dominant part of planktonic Foraminifera assemblages
385in the world oceans. Compared with a global database of census counts from surface sediments
386(MARGO database, Kucera et al., 2005), the species covered by PFR² account globally for >90%
387of shells larger than 150 µm found in surface sediments (Figure 4). In cold and temperate
388provinces, PFR² species account for almost the entire assemblages, while in warmer subtropical
389and tropical waters, only up to 4% of the sedimentary assemblages are not represented in PFR².
390Evidently, PFR² reference sequences cover most of the ecologically relevant portion of the
391morphological diversity and the taxa that are not yet represented in PFR² are small, rare or
392taxonomically obscure. It is possible that some of these taxa may correspond to the six sequences
393with unresolved taxonomy. If so, PFR² may be considered to cover up to 38 of the 47 recognized
394species.

395Finally, for each species present in PFR², we evaluated the ecological coverage of the global
396sampling effort (Figure 4). Morphospecies of planktonic Foraminifera are known to be
397distributed zonally across the world oceans, reflecting the latitudinal distribution of sea surface
398temperature (e.g., Bé and Tolderlund, 1971). A comparison between the temperature range of
399each species as indicated by their relative abundance in surface sediment samples (Kucera et al.,
4002005) and the temperatures measured at sampling localities shows that a large portion of the
401ecological range of the species is covered by the reference sequences in PFR² (Figure 4).

402The PFR² web interface

403To facilitate data download and comparative sequence analyses, PFR² has been implemented into

404a dedicated web interface, available at <http://pfr2.sb-roscoff.fr>. The website provides:

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

406 taxonomic rank (morphogroup name, genus name, species name), geographic region (e.g.,

407 North Atlantic, Mediterranean Sea, Indian Ocean) or collection (cruise name) ;

408 (2) a classical BLAST/Similarity module that facilitates identification of unknown sequences;

409 (3) a map module displaying the localities for all sequences present in PFR² and facilitating

410 download of all data from each single locality;

411 (4) a download section with direct access to all data included in PFR². All sequences and

412 sequence partitions are available in FASTA format and the metadata are available in a

413 tabulated file.

414

415Discussion

416Comprehensive databases of ribosomal RNA sequences with curated taxonomy are available for

417Protists (Protist ribosomal reference database, *PR*²; Guillou et al., 2013) and for the major

418domains of life (SILVA, Yilmaz et al., 2013), and these databases also include sequences of

419planktonic Foraminifera. However, those databases are used mainly as benchmarks to annotate

420complex environmental datasets (e.g. de Vargas et al., 2015) at the level of morphological

421species. In contrast, PFR² has been designed and implemented in a way that facilitates other

422applications.

423First, we note that because of the structural limitations, PFR² contains “only” 402 sequences of

424planktonic Foraminifera (Based on Released 203 of GenBank, October 2014), compared to

425PFR², which contains for now 3322 SSU rDNA sequences. Second, 2276 of the sequences

426present in PFR² have an assignation to the level of the genetic type and as far as possible, the
427sequences are associated with metadata related to the origin of each specimen and the conditions
428where it was collected, thus forming a basis for ecological modelling. Third, very importantly,
429using planktonic Foraminifera as a case study, we propose and implement an annotation scheme
430with unmatched accuracy and full tracking of changes. This is only possible because of the
431relatively “small” size of PFR² combined with high-level expert knowledge of their taxonomy.
432The fidelity of the annotations will facilitate a qualitatively entirely different level of analysis of
433eDNA libraries.

434For example, the design of PFR² allows to incorporate advances in classical and molecular
435taxonomy, particularly at the level of genetic types (e. g. André et al., 2014), which can be re-
436evaluated depending of the criteria used to delineate molecular OTUs. Further, by retaining
437information on clone attribution to specimens (vouchers), PFR² allows to evaluate intra-genomic
438polymorphism, which offers excellent opportunity to identify the phylogenetically relevant level
439of variability (Weber and Pawlowski, 2014). Finally, the modular structure of PFR² (i.e., its
440partitioning into variable and conserved regions) is particularly suitable for the evaluation of
441existing barcodes or the design of new barcoding systems needed to capture total or partial
442planktonic foraminiferal diversity within complex plankton assemblages. An examination of the
443length polymorphism in the 11 regions of the 18S rDNA fragment that have been aligned for all
444PFR² sequences reveals that next to the variable 37F region identified as a barcode for benthic
445Foraminifera (Pawlowski and Lecroq, 2010), several other regions would be suitable as targets
446for barcoding of planktonic Foraminifera (Figure 5).

447The main difference between PFR² and classical databases is in the association of sequence data
448with environmental and collection data. Such level of annotation is not feasible in large

449databases, which have to rely on the completeness and level of detail of metadata provided in
450GenBank. The association of metadata to PFR² sequences facilitates an assessment of
451biogeography and ecology of genetic types (potential cryptic species). This is important for
452studies of evolutionary processes in the open ocean such as speciation and gene flow at basin
453scale, but also for paleoceanography, which exploits ecological preferences of planktonic
454Foraminifera species to reconstruct climate history of earth (e. g. Kucera et al., 2005). Modeling
455studies showed that the integration of cryptic diversity into paleoceanographic studies may
456improve their accuracy (Kucera and Darling, 2002; Morard et al., 2013). Together with the
457MARGO database (Kucera et al., 2005) which records the occurrence of morphospecies of
458planktonic Foraminifera in surface sediments and the CHRONOS/NEPTUNE database (Spencer-
459Cervato et al., 1994; <http://www.chronos.org/>) which records their occurrence through geological
460time, PFR² represents the cornerstone to connect genetic diversity to the fossil record in an entire
461group of pelagic protists.

462

463**Conclusion and perspectives**

464The PFR² database represents the first geographically and taxonomically comprehensive
465reference barcoding system for an entire group of pelagic protists. Therefore it constitutes a
466pivotal tool to investigate the diversity, ecology, biogeography, and evolution in planktonic
467Foraminifera as a model system for pelagic protists. In addition, the database constitutes an
468important resource allowing reinterpretation and refinement of the use of Foraminifera as
469markers for stratigraphy and paleoceanography. In particular, PFR² can be used to: (i) annotate
470and classify newly generated 18S rDNA sequences from single individuals; (ii) study the

471biogeography of cryptic genetic types; (iii) design rank-specific primers and probes to target any
472group of planktonic Foraminifera in natural communities; (iv) assign accurate taxonomy to
473environmental sequences from metabarcoding or metagenomic datasets. This last point is
474particularly important. Future global metabarcoding of planktonic Foraminifera covering
475comprehensive spatio-temporal scales will likely reveal the full extent and complexity of species
476diversity and ecology in the group, serving as a model system for studies of the dynamics of the
477plankton and its interaction with the Earth system.

478**Acknowledgments:**

479We would like to thank all crew members and scientist for their help in the collection of
480planktonic Foraminifera that were used to generate the database. We would like to thank Erica de
481Leau for her help in gathering the data and Dominique Boeuf and ABIMS for their help in
482designing and hosting the PFR² website. This work was supported by grants from ANR-11-
483BTBR-0008 OCEANOMICS, ANR-09-BLAN-0348 POSEIDON, ANR-JCJC06-0142-PALEO-
484CTD, from Natural Environment Research Council of the United Kingdom (NER/J/S2000/00860
485and NE/D009707/1), the Leverhulme Trust and the Carnegie Trust for the Universities of
486Scotland, from DFG-Research Center/Cluster of Excellence “The Ocean in the Earth System”
487and from the Deutsche Forschungsgemeinschaft KU2259/19 and DU1319/1-1. This study is a
488contribution to the effort of the SCOR/IGBP Working Group 138 “Modern planktonic
489Foraminifera and ocean changes”.

490**References**

- 491
492André A, Weiner A, Quillévéré F *et al.* (2013) The cryptic and the apparent reversed : lack of
493 genetic differentiation within the morphologically diverse plexus of the planktonic
494 foraminifer *Globigerinoides sacculifer*. *Paleobiology*, **39**, 21–39.

495 André A, Quillévéré F, Morard R *et al.* (2014) SSU rDNA Divergence in Planktonic
 496 Foraminifera: Molecular Taxonomy and Biogeographic Implications (V Ketmaier, Ed.).
 497 *PLoS ONE*, **9**, 1–19.
 498 Aurahs R, Göker M, Grimm GW *et al.* (2009a) Using the Multiple Analysis Approach to
 499 Reconstruct Phylogenetic Relationships among Planktonic Foraminifera from Highly
 500 Divergent and Length-polymorphic SSU rDNA Sequences. *Bioinformatics and biology*
 501 *insights*, **3**, 155–177.
 502 Aurahs R, Grimm GW, Hemleben V, Hemleben C, Kucera M (2009b) Geographical distribution
 503 of cryptic genetic types in the planktonic foraminifer *Globigerinoides ruber*. *Molecular*
 504 *ecology*, **18**, 1692–1706.
 505 Aurahs R, Treis Y, Darling K, Kucera M (2011) A revised taxonomic and phylogenetic concept
 506 for the planktonic foraminifer species *Globigerinoides ruber* based on molecular and
 507 morphometric evidence. *Marine Micropaleontology*, **79**, 1–14.
 508 Aze T, Ezard THG, Purvis A *et al.* (2011) A phylogeny of Cenozoic macroperforate planktonic
 509 foraminifera from fossil data. *Biological reviews of the Cambridge Philosophical Society*,
 510 **86**, 900–27.
 511 Bé A.W.H., Tolderlund, D., (1971) Distribution and ecology of living planktonic foraminifera in
 512 surface waters of the Atlantic and Indian Oceans. In: Funnell, B. M., and Riedel, W. R..
 513 Eds., The micropalaeontology of oceans. London: Cambridge Univ. Press, pp. 105-149,
 514 text-figs. 1-27.
 515 Bé, A.W.H, Hudson WH (1977) Ecology of planktonic foraminifera and biogeographic patterns
 516 of life and fossil assemblages in the Indian Ocean. *Micropaleontology* **23**, 369–414.
 517 Brummer GA, Hemleben C, Michael S (1986) Planktonic foraminiferal ontogeny and new
 518 perspectives for micropalaeontology. *Nature*, **319**, 50–52.
 519 Darling KF, Kroon D, Wade CM, Leigh J (1996) Molecular Phylogeny of the planktic
 520 foraminifera. *Journal of foraminiferal research*, **26**, 324–330.
 521 Darling KF, Wade CM, Kroon D, Brown AJL (1997) Planktic foraminiferal molecular evolution
 522 and their polyphyletic origins from benthic taxa. *Marine Micropaleontology*, **30**, 251–266.
 523 Darling KF, Wade CM, Kroon D, Brown AJL, Bijma J (1999) The Diversity and Distribution of
 524 Modern Planktic Foraminiferal Small Subunit Ribosomal RNA Genotypes and their
 525 Potential as Tracers of Present and Past Ocean Circulations. *Paleoceanography*, **14**, 3–12.
 526 Darling KF, Wade CM, Stewart I a *et al.* (2000) Molecular evidence for genetic mixing of Arctic
 527 and Antarctic subpolar populations of planktonic foraminifers. *Nature*, **405**, 43–7.
 528 Darling KF, Kucera M, Wade CM, von Langen PJ, Pak DK (2003) Seasonal distribution of
 529 genetic types of planktonic foraminifer morphospecies in the Santa Barbara Channel and its
 530 paleoceanographic implications. *Paleoceanography*, **18**, 1–10.
 531 Darling KF, Kucera M, Pudsey CJ, Wade CM (2004) Molecular evidence links cryptic
 532 diversification in polar planktonic protists to Quaternary climate dynamics. *Proceedings of*
 533 *the National Academy of Sciences of the United States of America*, **101**, 7657–62.
 534 Darling KF, Kucera M, Kroon D, Wade CM (2006) A resolution for the coiling direction paradox
 535 in *Neoglobobulimina*. *Paleoceanography*, **21**, PA2011.
 536 Darling KF, Kucera M, Wade CM (2007) Global molecular phylogeography reveals persistent
 537 Arctic circumpolar isolation in a marine planktonic protist. *Proceedings of the National*
 538 *Academy of Sciences of the United States of America*, **104**, 5002–5007.
 539 Darling KF, Wade CM (2008) The genetic diversity of planktic foraminifera and the global
 540 distribution of ribosomal RNA genotypes. *Marine Micropaleontology*, **67**, 216–238.

541Darling KF, Thomas E, Kasemann S a *et al.* (2009) Surviving mass extinction by bridging the
542 benthic/planktic divide. *Proceedings of the National Academy of Sciences of the United*
543 *States of America*, **106**, 12629–33.

544de Vargas C, Zaninetti L, Hilbrecht H, Pawlowski J (1997) Phylogeny and rates of molecular
545 evolution of planktonic foraminifera: SSU rDNA sequences compared to the fossil record.
546 *Journal of molecular evolution*, **45**, 285–294.

547de Vargas C, Pawlowski J (1998) Molecular versus taxonomic rates of evolution in planktonic
548 foraminifera. *Molecular phylogenetics and evolution*, **9**, 463–469.

549de Vargas C, Norris R, Zaninetti L, Gibb SW, Pawlowski J (1999) Molecular evidence of cryptic
550 speciation in planktonic foraminifers and their relation to oceanic provinces. *Proceedings of*
551 *the National Academy of Sciences of the United States of America*, **96**, 2864–2868.

552de Vargas C, Renaud S, Hilbrecht H, Pawlowski J (2001) Pleistocene adaptive radiation in
553 Globorotaliatruncatulinoidea: genetic, morphologic, and environmental evidence.
554 *Paleobiology*, **27**, 104–125.

555de Vargas C, Bonzon M, Rees NW, Pawlowski J, Zaninetti L (2002) A molecular approach to
556 biodiversity and biogeography in the planktonic foraminifer Globigerinellasinifera
557 (d'Orbigny). *Marine Micropaleontology*, **45**, 101–116.

558De Vargas et al., 2015

559Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: A multiplatform graphical user
560 interface for sequence alignment and phylogenetic tree building. *Molecular biology and*
561 *evolution*, **27**, 221–4.

562Guillou L, Bachar D, Audic S et al. (2012) The Protist Ribosomal Reference database (PR2): a
563 catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy.
564 *Nucleic acids research*, **41**, D597–604.

565Hamsher SE, Evans KM, Mann DG, Pouličková A, Saunders GW (2011) Barcoding diatoms:
566 exploring alternatives to COI-5P. *Protist*, **162**, 405–22.

567Hemleben C, Spindler M, & Anderson OR (1989) Modern Planktonic Foraminifera. Springer-
568 Verlag New York Inc. pp. 363.

569Hollingsworth, PM, Forrest, LL, Spouge JL, et al. (2009) A DNA barcode for land plants.
570 *Proceedings of the National Academy of Sciences of the USA*, **106**, 12,794–12,797.

571Holzmann M, Pawlowski J (1996) Preservation of foraminifera for DNA extraction and PCR
572 amplification. *Journal of foraminiferal research*, **26**, 264–267.

573Kennett, JP, & Srinivasan, MS (1983) Neogene Planktonic Foraminifera. A Phylogenetic Atlas.
574 Hutchinson Ross Publishing Company, Stroudsburg, Pennsylvania. pp. 265.

575Kato K, Standley DM (2013) MAFFT multiple sequence alignment software version 7:
576 improvements in performance and usability. *Molecular biology and evolution*, **30**, 772–80.

577Kucera M, Weinelt M, Kiefer T *et al.* (2005) Reconstruction of sea-surface temperatures from
578 assemblages of planktonic foraminifera: multi-technique approach based on geographically
579 constrained calibration data sets and its application to glacial Atlantic and Pacific Oceans.
580 *Quaternary Science Reviews*, **24**, 951–998.

581Kucera M, Darling KF (2002) Cryptic species of planktonic foraminifera: their effect on
582 palaeoceanographic reconstructions. *Philosophical transactions. Series A, Mathematical,*
583 *physical, and engineering sciences*, **360**, 695–718.

584Logares R, Audic S, Bass D et al. (2014) Patterns of rare and abundant marine microbial
585 eukaryotes. *Current biology : CB*, **24**, 813–21.

586Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B (2011) How many species are there on
587 Earth and in the ocean? PLoS biology, 9, e1001127.

588Morard R, Quillévéré F, Escarguel G et al. (2009) Morphological recognition of cryptic species
589 in the planktonic foraminifer *Orbulina universa*. Marine Micropaleontology, 71, 148–165.

590Morard R, Quillévéré F, Douady CJ et al. (2011) Worldwide genotyping in the planktonic
591 foraminifer *Globoconella inflata*: implications for life history and paleoceanography. PLoS
592 ONE, 6, 1–12.

593Morard R, Quillévéré F, Escarguel G, Garidel-thoron T de (2013) Ecological modeling of the
594 temperature dependence of cryptic species of planktonic foraminifera in the Southern
595 Hemisphere. Palaeogeography, Palaeoclimatology, Palaeoecology, 391, 13–33.

596R Development Core Team (2014) R: a language and environment for statistical computing. R
597 Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org>.

598Nassonova E, Smirnov A, Fahrni J, Pawlowski J (2010) Barcoding amoebae: comparison of
599 SSU, ITS and COI genes as tools for molecular identification of naked lobose amoebae.
600 Protist, 161, 102–15.

601Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in
602 the amino acid sequence of two proteins. Journal of molecular biology, 48, 443–53.

603Nikolaev SI, Berney C, Fahrni JF et al. (2004) The twilight of Heliozoa and rise of Rhizaria, an
604 emerging supergroup of amoeboid eukaryotes. Proceedings of the National Academy of
605 Sciences of the United States of America, 101, 8066–71.

606Pawlowski J, Lecroq B (2010) Short rDNA barcodes for species identification in foraminifera.
607 The Journal of eukaryotic microbiology, 57, 197–205.

608Pawlowski J, Audic S, Adl S et al. (2012) CBOL protist working group: barcoding eukaryotic
609 richness beyond the animal, plant, and fungal kingdoms. PLoS biology, 10, e1001419.

610Pawlowski J, Lejzerowicz F, Esling P (2014) Next-Generation Environmental Diversity Surveys
611 of Foraminifera : Preparing the Future. Biol. Bull., 227, 93–106.

612Quillévéré F, Morard R, Escarguel G et al.(2013) Global scale same-specimen morpho-genetic
613 analysis of *Truncorotalia truncatulinoides*: A perspective on the morphological species
614 concept in planktonic foraminifera. Palaeogeography, Palaeoclimatology, Palaeoecology,
615 391, 2–12.

616Rice P (2000) The European Molecular Biology Open Software Suite EMBOSS : The European
617 Molecular Biology Open Software Suite. Trends in Genetics, 16, 2–3.

618Rognes T (2011) Faster Smith-Waterman database searches with inter-sequence SIMD
619 parallelisation. BMC bioinformatics, 12, 221.

620Saunders GW (2005) Applying DNA barcoding to red macroalgae: a preliminary appraisal holds
621 promise for future applications. Philosophical transactions of the Royal Society of London.
622 Series B, Biological sciences, 360, 1879–88.

623Schoch CL, Seifert K a, Huhndorf S et al. (2012) Nuclear ribosomal internal transcribed spacer
624 (ITS) region as a universal DNA barcode marker for Fungi. Proceedings of the National
625 Academy of Sciences of the United States of America, 109, 6241–6.

626Sears H a, Darling KF, Wade CM (2012) Ecological partitioning and diversity in tropical
627 planktonic foraminifera. BMC Evolutionary Biology, 12, 54.

628Sherwood AR, Presting GG (2007) Universal primers amplify a 23S rDNA plastid marker in
629 eukaryotic algae and cyanobacteria. Journal of Phycology, 43, 605–608.

630Spencer-Cervato C, Thierstein HR, Lazarus DB, Beckmann J-P (1994) How synchronous are
631 neogene marine plankton events? Paleoceanography, 9, 739.

632 Stern RF, Horak A, Andrew RL et al. (2010) Environmental barcoding reveals massive
633 dinoflagellate diversity in marine environments. *PloS one*, 5, e13991.
634 Stewart IA, Darling KF, Kroon D, Wade CM, Troelstra SR (2001) Genotypic variability in
635 subarctic Atlantic planktic foraminifera. , **43**, 143–153.
636 Sogin ML, Morrison HG, Huber J a *et al.* (2006) Microbial diversity in the deep sea and the
637 underexplored “rare biosphere”. *Proceedings of the National Academy of Sciences of the*
638 *United States of America*, **103**, 12115–20.
639 Ujiie Y, Kimoto K, Pawlowski J (2008) Molecular evidence for an independent origin of modern
640 triserial planktonic foraminifera from benthic ancestors. *Marine Micropaleontology*, **69**,
641 334–340.
642 Ujiie Y, Lipps JH (2009) Cryptic diversity in planktonic foraminifera in the northwest Pacific
643 Ocean. *Journal of foraminiferal research*, **39**, 145–154.
644 Ujiie Y, Asami T, de Garidel-Thoron T *et al.* (2012) Longitudinal differentiation among pelagic
645 populations in a planktic foraminifer. *Ecology and evolution*, **2**, 1725–37.
646 Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer New York.
647 Wade CM, Darling KF, Kroon D, Brown AJL (1996) Early Evolutionary Origin of the Planktic
648 Foraminifera Inferred from Small Subunit rDNA Sequence Comparisons. *Journal of*
649 *molecular evolution*, 43, 672–677.
650 Weber AA-T, Pawlowski J (2014) Wide occurrence of SSU rDNA intragenomic polymorphism
651 in foraminifera and its implications for molecular species identification. *Protist*, 165, 645–
652 61.
653 Weiner A, Aurahs R, Kurasawa A, Kitazato H, Kucera M (2012) Vertical niche partitioning
654 between cryptic sibling species of a cosmopolitan marine planktonic protist. *Molecular*
655 *ecology*, 21, 4063–73.
656 Weiner AKM, Weinkauf MFG, Kurasawa A *et al.* (2014) Phylogeography of the tropical
657 planktonic foraminifera lineage *Globigerinella* reveals isolation inconsistent with passive
658 dispersal by ocean currents. *PloS one*, **9**, e92148.
659 Weiner AKM, Weinkauf MFG, Kurasawa A, Darling KF, Kucera M (2015) Genetic and
660 morphometric evidence for parallel evolution of the *Globigerinella calida* morphotype.
661 *Marine Micropaleontology*, **114**, 19–35.
662 Yilmaz P, Parfrey LW, Yarza P et al. (2014) The SILVA and “All-species Living Tree Project
663 (LTP)” taxonomic frameworks. *Nucleic acids research*, 42, D643–8.
664 Žarić S, Donner B, Fischer G, Mulitza S, Wefer G (2005) Sensitivity of planktic foraminifera to
665 sea surface temperature and export production as derived from sediment trap data. *Marine*
666 *Micropaleontology*, 55, 75–105.
667 Zimmermann J, Abarca N, Enk N et al. (2014) Taxonomic reference libraries for environmental
668 barcoding: a best practice example from diatom research. *PloS one*, 9, e108793.

669
670

671 Author contribution

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

678Data Accessibility

679Sequences, NCBI accession numbers and metadata are available in Supplementary Material 1
680and 2 and on PFR² website at <http://pfr2.sb-roscoff.fr>. The custom scripts used to perform the
681curation procedure are available in Supplementary Material 5, the results of the curation process
682are given in Supplementary Material 1 and 2.

683Figures

684Figure 1

685**Sampling Map.** Location of the 460 oceanic stations sampled over 20 years for single-cell
686genetic studies of planktonic Foraminifera. Each symbol corresponds to a scientific cruise or
687near shore collection site. Cruise names and dates of the collection expeditions are indicated in
688the legend. Grey shading shows ocean bathymetry.

689Figure 2

690**Workflow to constitute PFR².** In step “I” the sequences, metadata and taxonomic information
691are retrieved from public databases and literature or from the internal databases of the authors to
692constitute the Primary Reference Database. In step “II”, the coverage of each sequence is
693evaluated by alignment with structural regions of the 18S RNA secondary structure derived for
694the species *Micrometula hyalostera* (Pawlowski and Lecroq, 2010). In step “III”, the consistency
695of the annotation is checked from the most exclusive level of annotation “genetic type 3” until
696the species level (Phase 1) to detect annotation inconsistency (See Figure 3). Sequences with
697wrong annotation are invalidated, compared to the validated part of the dataset (Phase 2) and re-
698annotated depending on the best hit out of the valid dataset. The consistency of all annotations is
699then checked again following the same procedure as in Phase 1 (Phase 3), to ensure that no
700taxonomic inconsistency remains. In step IV, all sequences which have been subjected to the
701curation process are integrated in the Planktonic Foraminifera Ribosomal Reference database
702(PFR²). The results of all steps are given in Supplementary Material 1.

703Figure 3

704**Annotation inconsistency detection.** The procedure followed to identify annotation
705inconsistency is exemplified by three cases. Each graph represents variability in pairwise
706similarities observed across each region of all sequences sharing the same annotation level. The
707names of the taxon and annotation level are given above the plot with the number of sequences
708in parenthesis. Each vertical line represents one region with the variability represented as dot
709plot, the number of “complete” regions is given at the bottom of the line. The case “A” describes
710the annotation validation process starting from the most exclusive rank of “genetic type level 3”
711to the “species” rank. After the validation at one rank level, the sequences with valid annotation
712are merged in a taxonomic unit of a higher rank. This now includes multiple sequence motifs
713decreasing the level of identity in each region, leading to a high variability in higher ranks. Case
714“B” represents the occurrence of obvious outliers at the species level, which are invalidated.
715Case “C” represents the co-occurrence of divergent sequences under the same taxonomic
716attribution, which are consequently all invalidated. The dot plots for all ranks can be found in
717Supplementary Material 4 and the pairwise similarities calculated for each taxonomic level are
718given in Supplementary Material 1.

719Figure 4

720**Taxonomic and ecological coverage of PFR².** For each morphogroup (Spinose, Non-Spinose,
721Microperforates, Monolamellar and Non-Spiral) the number of species included in PFR² is given
722in the filled bar while the number of species not present is indicated in the adjacent open bar. The
723relative abundance in the sediments of each species included in PFR² is given in log value
724against mean Sea Surface Temperature (SST) at the sampling station. Relative abundances in
725sediments are derived from the MARGO database (Kucera et al., 2005) and the mean annual
726SST from the World Ocean Atlas (Locarnini, 2005). The grey dots highlight the mean annual
727SST at the location where the living planktonic foraminifera yielding sequences were sampled.
728The number of sequences available for each species as well as the number of taxonomic paths
729above the species level is shown next to the graphs. Also shown is the cumulative mean relative
730abundance in the sediments of all species included in PFR² plotted against the mean annual SST
731in discrete 1°C intervals. Vertical bars represent 95% confidence intervals for each 1°C bin.

732Figure 5

733**Length polymorphism.** Each rectangle represents the length polymorphism within each region
734of the analyzed 18S rDNA fragment across all resolved taxonomic units in PFR². The regions are
735based on the rRNA secondary structure and are named following Pawlowski and Lecroq (2010).

736Supplementary Material.

737Supplementary Material 1.

738Information on all consecutive steps followed to constitute the PFR². All fields are explained in
739the file.

740Supplementary Material 2

741FASTA files of sequences used to build the PFR². FASTA files are provided for the full
742sequences and individual partitions.

743Supplementary Material 3

744Dot plots showing pairwise similarities for each taxonomic level. See Figure 3 for explanations
745of the content of the plots.

746Supplementary Material 4

747Neighbor-joining trees showing the monophyly of each morphospecies.

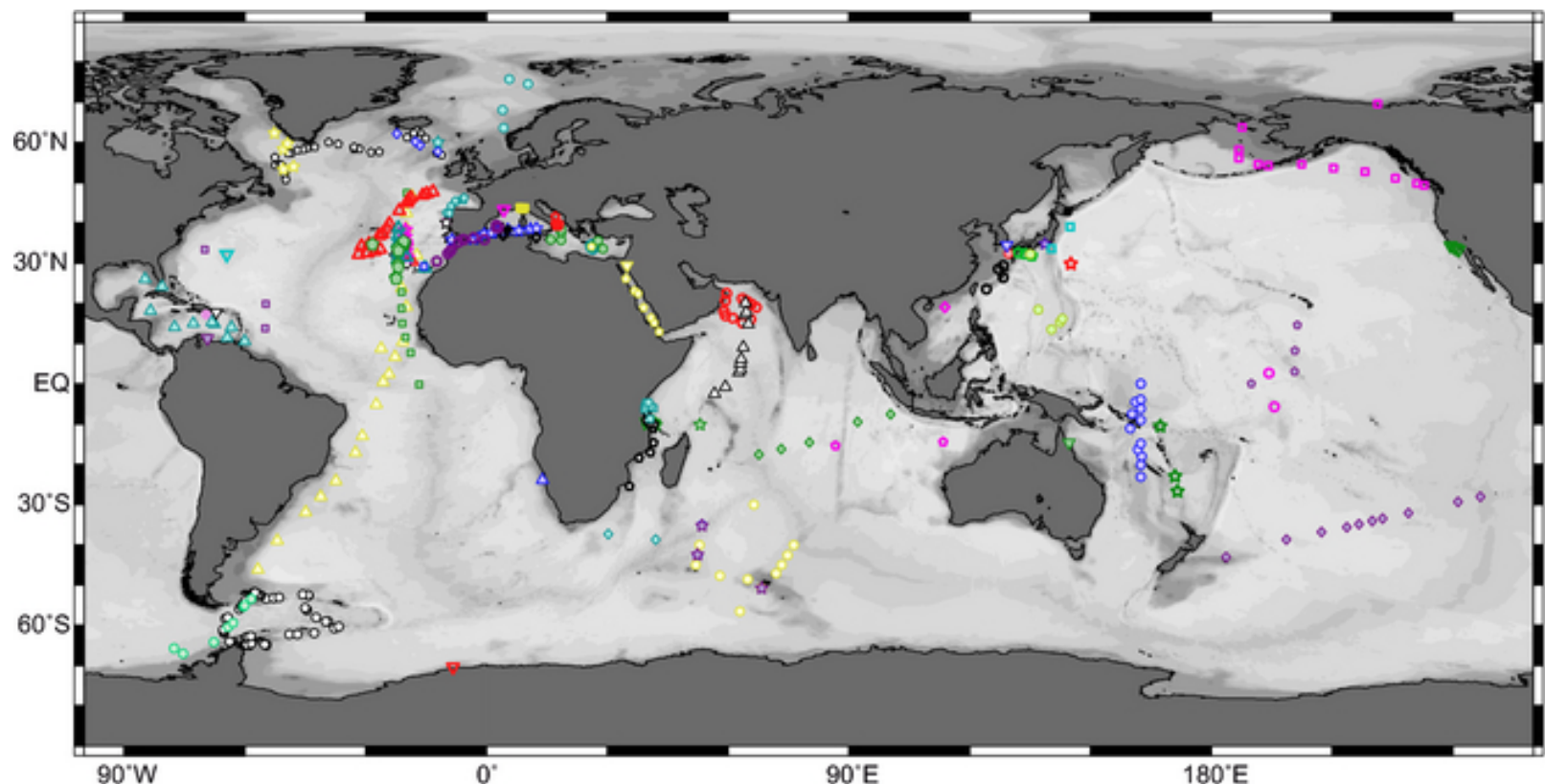
748Supplementary Material 5

749Custom scripts used to perform the different curation steps.

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

Fig 1



Scientific cruise

- Alis, GYRATOR-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, GYRATOR-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 Wilfrid 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)

Fig 2

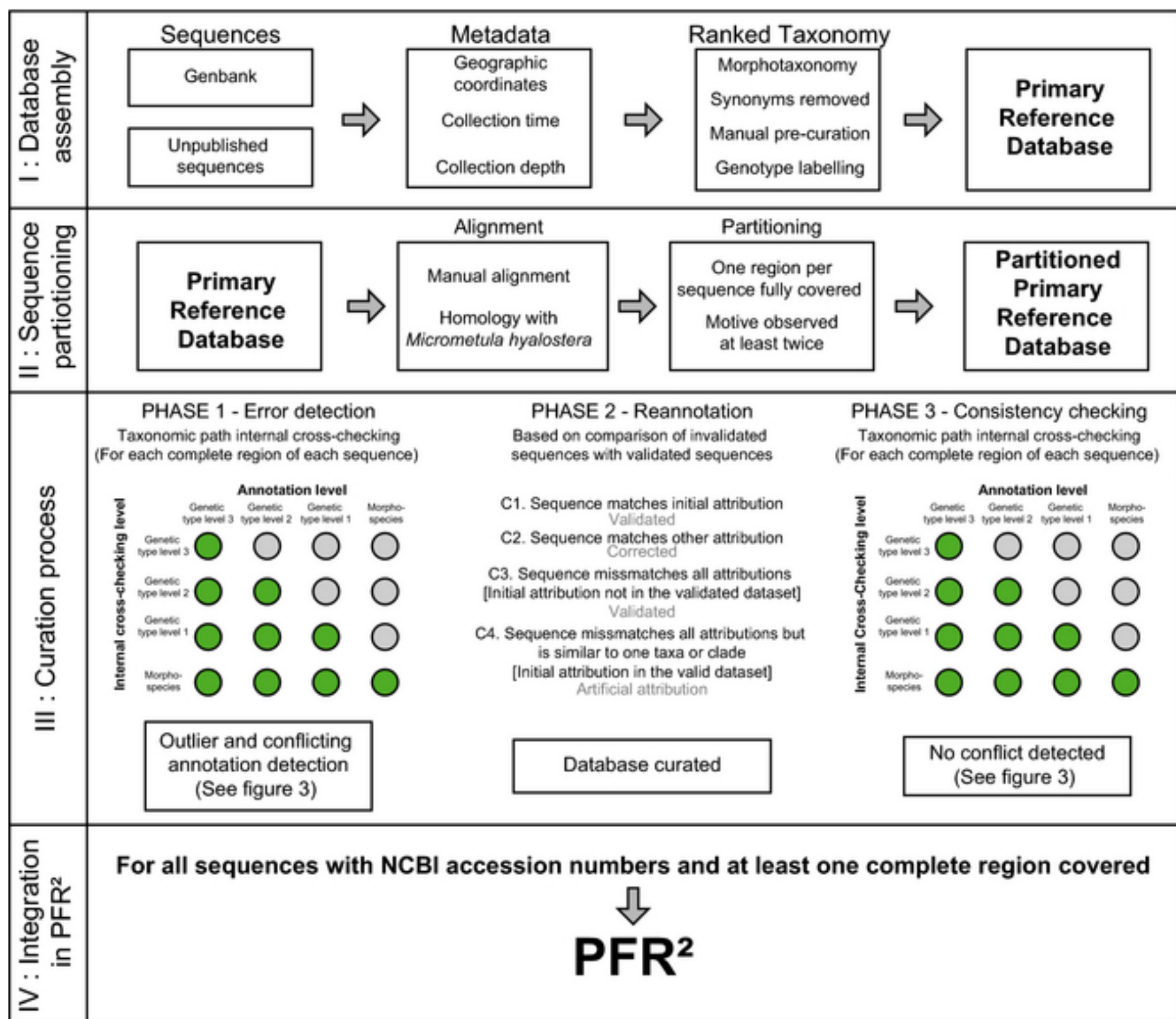


Fig 3

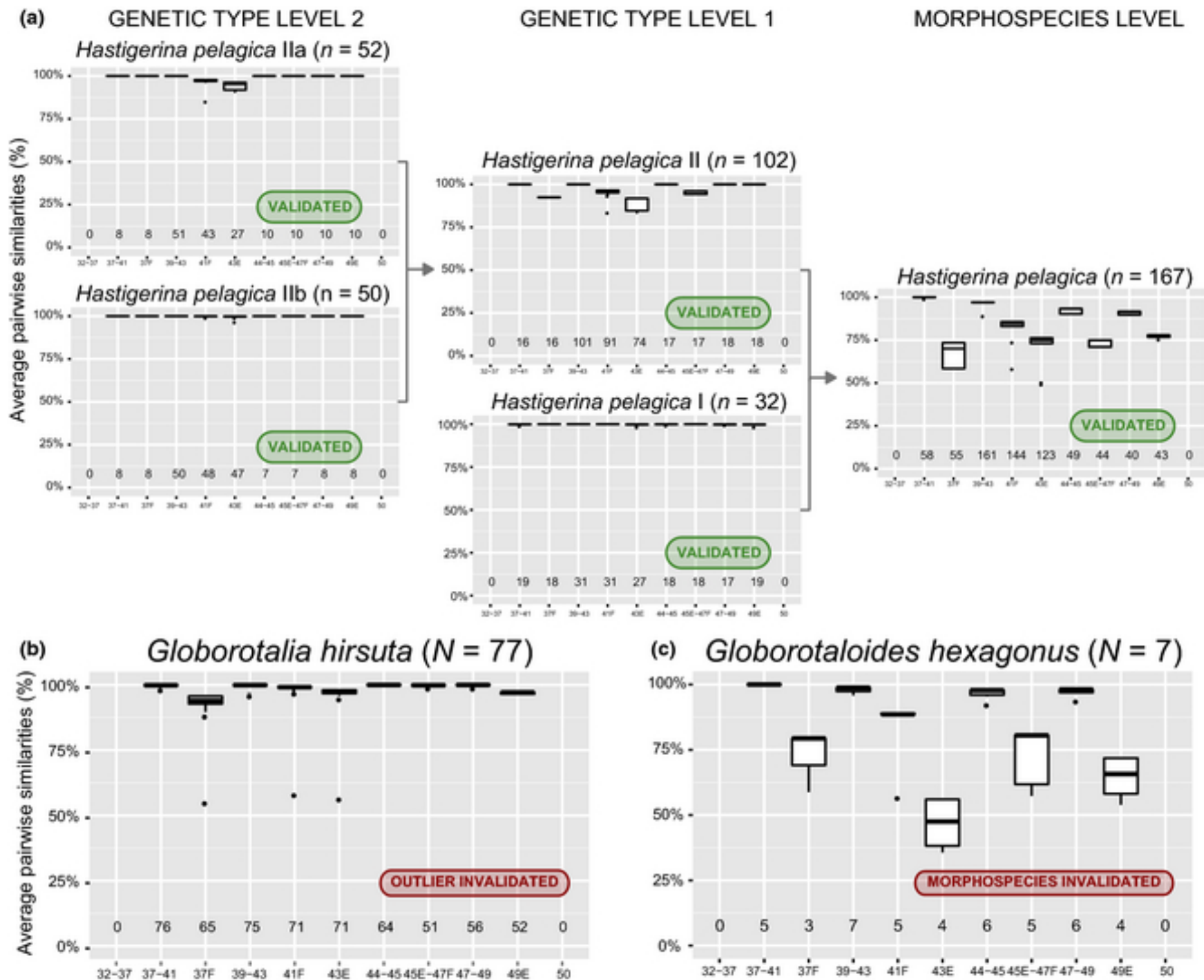


Fig 4

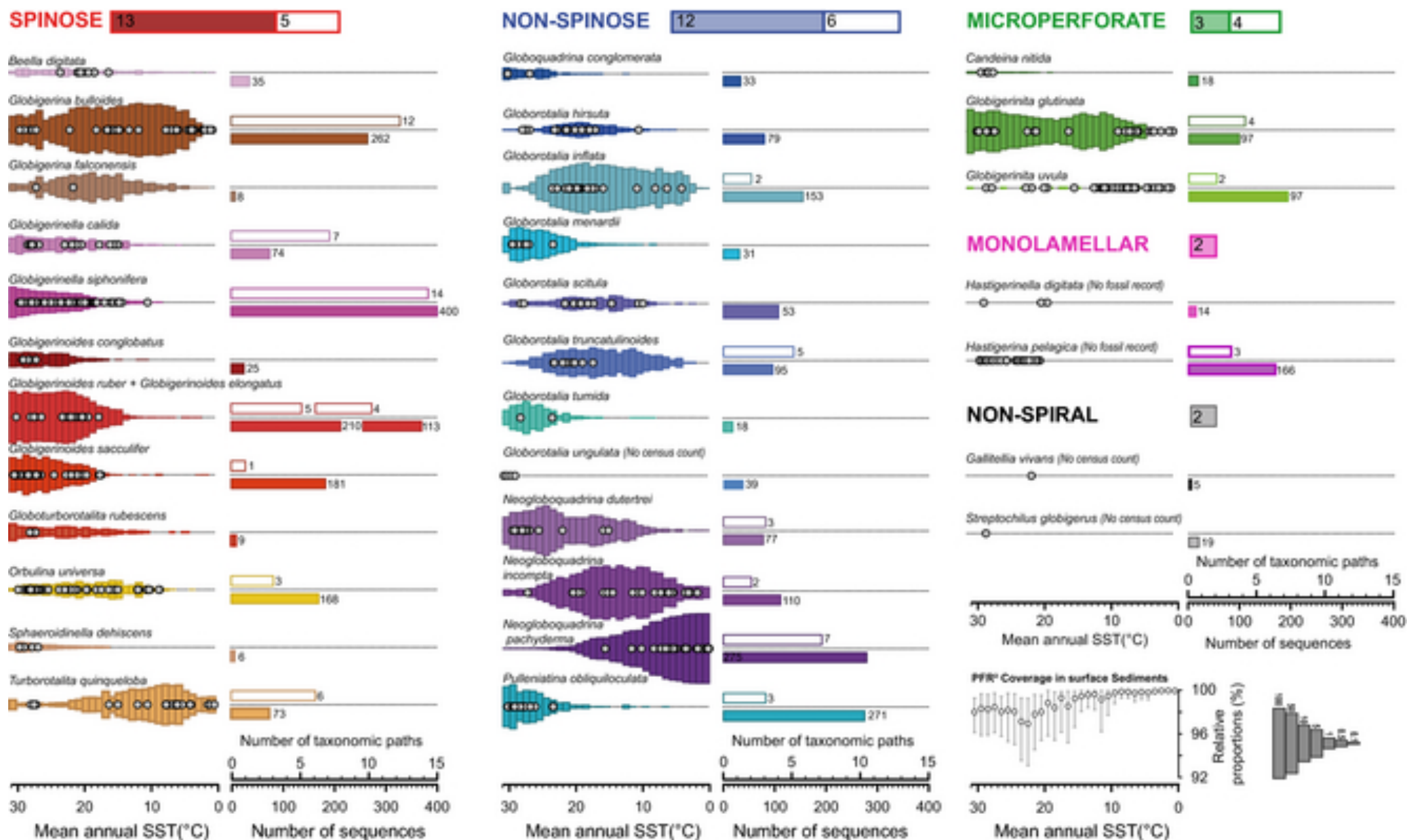


Fig 5

