Published online 27 November 2012 *Nucleic Acids Research, 2013, Vol. 41, Database issue* ***D597–D604***
doi:10.1093/nar/gks1160

The Protist Ribosomal Reference database (PR^2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy

Laure Guillou[1,2,*], Dipankar Bachar[3,4], Stéphane Audic[1,2], David Bass[5], Cédric Berney[5], Lucie Bittner[1,2], Christophe Boutte[1,2], Gaétan Burgaud[6], Colomban de Vargas[1,2], Johan Decelle[1,2], Javier del Campo[7], John R. Dolan[8], Micah Dunthorn[9], Bente Edvardsen[10], Maria Holzmann[11], Wiebe H.C.F. Kooistra[12], Enrique Lara[13], Noan Le Bescot[1,2], Ramiro Logares[7], Frédéric Mahé[1,2], Ramon Massana[7], Marina Montresor[12], Raphael Morard[1,2], Fabrice Not[1,2], Jan Pawlowski[11], Ian Probert[14,15], Anne-Laure Sauvadet[1,2], Raffaele Siano[16], Thorsten Stoeck[9], Daniel Vaulot[1,2], Pascal Zimmermann[17] and Richard Christen[3,4,*]

[1]CNRS, UMR 7144, Adaptation et Diversité en Milieu Marin, 29682 Roscoff, France, [2]UPMC Université Paris 06, UMR 7144, Station Biologique de Roscoff, 29682 Roscoff, France, [3]CNRS, UMR 7138, Systématique Adaptation Evolution, Parc Valrose, BP71. F06108 Nice cedex 02, France, [4]UMR 7138, Université de Nice-Sophia Antipolis, Systématique Adaptation Evolution, Parc Valrose, BP71. F06108 Nice cedex 02, France, [5]Department of Life Sciences, The Natural History Museum, Cromwell Road, London SW7 5BD, UK, [6]Laboratoire Universitaire de Biodiversité et Ecologie Microbienne (EA3882), ESMISAB, Technopôle Brest-Iroise, 29280 Plouzané, France, [7]Department of Marine Biology and Oceanography, Institut de Ciències del Mar (CSIC), Barcelona, Catalonia, Spain, [8]Laboratoire d'Océanographie de Villefranche, Marine Microbial Ecology, UPMC Université Paris 06 et CNRS, UMR7093, Station Zoologique, BP28, 06230 Villefranche-sur-Mer, France, [9]Department of Ecology, University of Kaiserslautern, 67663 Kaiserslautern, Germany, [10]Department of Biology, University of Oslo, Marine Biology, NO-0316 Oslo, Norway, [11]Department of Genetics and Evolution, University of Geneva, Switzerland, [12]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy, [13]Laboratory of Soil Biology, University of Neuchâtel, Rue Emile Argand 11, CH-2000, Neuchâtel, Switzerland, [14]UPMC Université Paris 06, 29682 Roscoff, France, [15]CNRS, FR2424, Station Biologique de Roscoff, 29682 Roscoff, France, [16]Ifremer, Centre de Brest, DYNECO/Pelagos BP70, 29280 Plouzané, France and [17]Point Compétence Informatique, Rue Jean-Baptiste Say, 56850 Caudan, France

Received August 14, 2012; Revised October 24, 2012; Accepted October 26, 2012

ABSTRACT

The interrogation of genetic markers in environmental meta-barcoding studies is currently seriously hindered by the lack of taxonomically curated reference data sets for the targeted genes. The Protist Ribosomal Reference database (PR^2, http://ssu-rrna.org/) provides a unique access to eukaryotic small sub-unit (SSU) ribosomal RNA and DNA sequences, with curated taxonomy. The database mainly consists of nuclear-encoded protistan sequences. However, metazoans, land plants, macrosporic fungi and eukaryotic organelles (mitochondrion, plastid and others) are also included because they are useful for the analysis of high-troughput sequencing data sets. Introns and putative chimeric sequences have been also carefully checked. Taxonomic assignation of sequences consists of eight unique taxonomic fields. In total,

*To whom correspondence should be addressed. Tel: +33 2 98 29 23 79; Fax: +33 2 98 29 23 24; Email: lguillou@sb-roscoff.fr
Correspondence may also be addressed to Richard Christen. Tel: +33 6 01 83 01 44; Email: christen@unice.fr
Present addresses:
Lucie Bittner, Department of Ecology, University of Kaiserslautern, 67663 Kaiserslautern, Germany.
Anne-Laure Sauvadet, School of Biosciences, Cardiff University, Main Building, Cardiff CF10 3AT, UK.

136 866 sequences are nuclear encoded, 45 708 (36 501 mitochondrial and 9657 chloroplastic) are from organelles, the remaining being putative chimeric sequences. The website allows the users to download sequences from the entire and partial databases (including representative sequences after clustering at a given level of similarity). Different web tools also allow searches by sequence similarity. The presence of both rRNA and rDNA sequences, taking into account introns (crucial for eukaryotic sequences), a normalized eight terms ranked-taxonomy and updates of new GenBank releases were made possible by a long-term collaboration between experts in taxonomy and computer scientists.

INTRODUCTION

The modern definition of the term 'protist' refers to unicellular eukaryotes that are either free-living or parasitic, sometimes forming colonies, but without clear differentiation into tissues. This includes all eukaryotes other than land plants (and macro-algae), animals and fungi with differentiated tissues. Protists are notoriously paraphyletic and include a wide range of microorganisms using a huge variety of reproductive, nutritional and life-history strategies. Nevertheless, the term protist has pragmatic uses and has recently gained in popularity. Large-scale analysis of protistan diversity is complicated by their heterogeneity, which reflects their extremely broad distribution and implication in multiple ecological and functional processes. This difficulty is exacerbated by the following facts: (i) species delineation is often obscure owing to lack of clear morphological criteria and paucity of knowledge concerning processes of sexual recombination; (ii) the taxonomy of protists has been radically modified in recent decades in light of new phylogenetic data; and (iii) a large proportion of protists are probably still not cultivable or yet unknown. Molecular barcoding using SSU rRNA (Small Sub-Unit Ribosomal) gene sequences consequently has become extremely popular among protistologists. Environmental barcoding has unveiled an extensive genetic diversity of protists in a wide range of ecosystems (1,2), including lineages only known by their genetic signatures (orphan environmental sequences). Recently, the use of next generation sequencing (NGS) technologies targeting selected domains of the SSU rRNA gene has permitted ecological studies of complex assemblages at ever increasing scales (3–7). However, interpretation of such data is currently seriously hindered by the lack of taxonomically curated reference data sets. Unassigned and incorrectly assigned sequences are accumulating at an increasing and alarming rate in public databases, to the extent that in early 2012, almost 20% of submitted SSU rRNA eukaryotic gene sequences had no or a very poor taxonomic assignation (see the website for more details). Undetected chimeric sequences (8), as well as the presence of introns in gene sequences (9), are also problematic.

To facilitate and increase the efficiency and accuracy of NGS data sets analyses, we here present the first comprehensive-curated database that places eukaryotic SSU rRNA gene sequences within a coherent ranked taxonomic framework covering eukaryotic diversity. Every sequence was quality checked and annotated using a multi-level taxonomic assignation. As a lot of protists are still only known by their environmental sequences, cluster names were retained when the formal taxonomy was missing [such as Syndiniales (10) and Marine STramenopiles, MAST (11)]. Although curated in less detail, sequences from metazoa, land plants and macrosporic fungi, as well as eukaryotic organelles (mitochondria, plastids, etc.), are also included in the database for their ecological interests. For example, protists may live in close association with metazoan (commensalisms, symbioses, etc.), and very small metazoan exists, inhabiting similar ecological niches. For example, copepods and polychaetes, as well as benthic animal larvae coexist with planktonic protists in aquatic systems. They may also have a great interest in ecological studies (as predators for example), even for protistologists. Even if this database is dedicated to protists, such outgroup sequences are of high relevance for extracting these groups in further analyses of NGS data sets when 'universal' eukaryotic primers are used for polymerase chain reaction (PCR) amplifications. Metazoan sequences in PR^2 allow not identifying them wrongly as new deep lineages of protists.

MATERIALS AND METHODS

The construction of this database started >10 years ago, and our procedure has been optimized over time (for more details, recent history detailed at http://ssu-rrna.org/method.html). Here, we briefly describe the present general architecture of the database.

Entries containing at least one partial SSU rRNA gene sequence of eukaryotic origin are retrieved from three public databases using keywords. Our last update retrieved 484.657, 496.462 and 123 such entries from GenBank, EMBL and WGS-EMBL, respectively. An INSDC (http://www.insdc.org/) entry as defined by its accession number in public databases may contain several rRNA gene sequences, e.g. in long genomic fragments containing several partial or complete ribosomal operons. To allow such duplicated sequences within a single entry, each sequence was given a unique identifier, acc.p1.p2, where acc is the accession number of the entry containing the sequence, and p1 and p2 are the first and last positions of the sub-sequence within the complete sequence.

A majority of extracted sequences were shorter than 100 nucleotides or around 500 nucleotides (63% of retrieved sequences), likely resulting from the recent integration of short environmental sequences derived from clone libraries. Only sequences longer than 799 nt were considered.

The first step was the identification of sequences originating from organelles. A reference database of SSU-rRNA gene sequences from chloroplasts and mitochondria was constructed using entire genomes or genomic fragments that contained a SSU-rRNA gene sequence and a

protein-coding gene specific either of mitochondria or of chloroplasts. For derived-organelle sequences such as apicoplasts, hydrogenosomes and nucleomorphs, databases were manually built, using information found in scientific publications. These databases were used to determine by sequence similarity the origin of every sequence in the database. These sequences were assigned to a reduced taxonomic framework, including their location (such as: |Organelle|chloro-SSU| or |Organelle|mito-SSU|). These sequences are not more detailed in the database.

Introns were found to be a major problem in eukaryotic rRNA sequences compared with prokaryotic sequences (1536 sequences with intron(s) described, 10 644 sequences with introns found by computation). A dedicated C++ algorithm was developed to identify the presence of introns in the remaining sequences (9). When detected, sequences with and without the intron(s) were generated (rRNA and rDNA sequences).

Sequences in the PR^2 database are assigned an identifier in the form accession.p1.p2_X, where accession is the accession number of an entry, p1 and p2 are the positions of this sequence in a larger genomic entry and X corresponding to introns treatment of the sequence [X = G: genomic sequence containing a described intron (rDNA); X = R: the previous genomic rRNA sequence, without the intron(s); X = U: no intron described, but intron(s) may be present; X = UC: introns were detected *in silico* and removed from the sequence (putative rRNA)].

Taxonomy of nuclear-encoded sequences

As all SSU-rRNA genes are orthologs, a global phylogeny can be built, and essential past speciation events can be evidenced. This property is essential to build a ranked taxonomy. For example, at rank 1, there is a world-wide agreement to recognize three clades, Bacteria, Archaea and Eukaryota. We chose to additionally use 'Organelle' as rank 1. Organelles have a eukaryote origin when they are nucleomorphs and a bacterial origin when they are mitochondrion and plastid. Because evolution of organelles and their hosts differ over time, their taxonomy is different too. In addition, scientists working on diversity are more interested in the identification of the cells that bear such organelles. Our choice was thus to allow their easy identification (and filtering out) during the first step of an analysis, targeting them as 'Organelle' at rank 1.

Nomenclature and terms of the following ranks mainly follows the classification of eukaryotes proposed by Adl *et al.* (12). Thus, the second rank describes each eukaryotic 'Super-Group' or Phylum (both terms are in use in different communities): Alveolata, Amoebozoa, Apusozoa, Archaeplastida, Excavata, Opisthonkonta, Rhizaria or stramenopiles. The taxonomic descriptions are structured by the use of eight ranks, and following ranks mainly correspond to the division, class, order, family, genus and species.

The terms used for each rank are non-ambiguous (a term cannot be found in two different clades), contain no space (that may pose problems to computers) and whenever possible retained if monophyletic. When monophyly could not be insured, the term of rank above was used, appended with suffix _X (suffix X if the above rank was already _X). As the same species name frequently occurs in different genera, the species name is composed of the genus and species, using '+' as a separator (e.g. genus = Diderma, species = Diderma + niveum). Genus and species names from public databases are stored in separate fields for comparison.

For protists and unicellular fungi, a taxonomy was proposed by the group of experts, authoring this article. For multicellular fungi, plants and metazoans, the taxonomy was built mostly using the taxonomy assigned in National Center for Biotechnology Information (NCBI)'s GenBank database entries. We first built a core reference database containing 23 116 manually analysed sequences representative of eukaryotic diversity. These analyses included reading published articles and phylogenetic analyses done by the authors of this article when necessary. This core reference database was subsequently used to automatically annotate the remaining sequences using different methods.

We are aware that for some clades such as metazoa, plants and fungi, our eight terms taxonomy is probably not as precise as it should be. Barcoding of metazoa and plants using SSU-rRNA sequences is not often used (normally only to complement Internal Transcribed Spacer (ITS) sequences). We will therefore try in a next release to propose an extended, still ranked and unified, taxonomy for fungi.

An outcrop of PR^2 is the web-based tool KeyDNAtools (http://keydnatools.com/). It uses 159 982 specific short (15 nt) oligonucleotide sequences (named keys) generated from the core reference database. Each key is a signature present in sequences of a given clade, but not in those of other clades. Besides providing a very fast taxonomic identification, it also allows for detecting putative chimeric sequences, as when different identifications are obtained from the 5′ and 3′ ends of sequences.

Specific new computer programs mostly in C, C++ and Python have been developed. First, a new parallel distributed computing Needleman–Wunsch-based C program allowing to compute pair-wise distances not taking into account terminal gaps (partially overlapping sequences) and long internal gaps (introns). This was coupled to a newly rewritten C average linkage clustering program. Second, a new parallel distributed computing Needleman–Wunsch-based C++/Python program allowing to assign a consensus taxonomy to new sequences by comparison to a reference database (Crunch_Assign).

When a conflict between taxonomies assigned using the different methods was found, it was manually solved. In the end, each nuclear encoded sequence is assigned an identifier in the form of this example:

>AY827845.1.1765_U|Eukaryota|Apusozoa|Hilomonadea|Planomonadida|Planomonadidae|Planomonadidae_Group-1|Ancyromonas|Ancyromonas + sigmoides

RESULTS

In total, we found 136 866 nuclear encoded sequences, five pseudo-genes (FJ854546, FJ854545, D14632, AF310844,

AJ404858, not included in PR^2) and 34 sequences we could only assign as putative rRNA sequences (HM538255, GU385678, AB275106, AJ628837, AY180011, CP000499, CP000499, AY256215, EU402432, AB017015, GQ330639, GU820811, JF488788, AF239231, DQ423737, DQ104596, AY835700, DQ423728, EU545797, GU072272, GU072526, GQ247249, HM174255, DQ104594, EU174762, FN598473, EU726200, EF695080, GQ483783, GQ462590, EU173354, EF567390, EF695215, HQ871039, not included in PR^2). Manual analyses of some of them allowed concluding for the presence of artefactual sequence internal or at the 5′ or 3′ end. Among nuclear-encoded sequences, we detected 1756 putative chimeric sequences, either using the KeyDNAtools and/or by manual inspection (listed on the website). For example, sequence EF023694.1.1975_U is a chimera between parent sequences of Opisthokonta, Amoebozoa and Rhizaria in position 179-471, 623-1264 and 1536-1925, respectively. Other '18S' sequences are nucleomorphs (262 sequences). In all, 9657 sequences have a chloroplastic origin, 33 051 are from mitochondria, six from hydrogenosomes (AJ237907, AJ237908, AJ871215, AJ871217 AJ871267, Y16670) and 26 from apicoplasts (U87145, AB471801, AB471802, AB471803, AB471804, AB471805, AB471806, AB471807, AB471808, AB471809, AB471810, AB471811, AB471812, AB649417, AB649418, AB649419, AB649420, AB649421, AB649422, AB649423, AB649424, HQ110105, JQ437257, JQ437258, JQ437259, U28056).

Within nuclear-encoded sequences, 54 data entries remained unassigned at the Super-Group level (Table 1), meaning that they could not be assigned to any specific taxon group within the domain Eukaryota (Eukaryota_X). The Super-Group 'Eukaryota_Mikro' was created for sequences HM563060, AF477623 and HM563061, for which no consensus has been reached for their affiliation, although Haplosporidiidae has been suggested (13). BLAST analyses conducted at NCBI against non-redundant or at DNA Data Bank of Japan (DDBJ) against all showed extremely weak sequence similarity with sequences of fungi. Using our global similarity tool (Crunch_Assign) showed no other sequence similar at ≥80% along the entire sequence. These results conducted to the creation of this new Super-Group (rank 2). For unassigned nuclear-encoded sequences (Eukaryota_X), either no other similar sequence was found or similar sequences were detected but also annotated by us as Eukaryota_X. A BLAST on NCBI non-redundant (excluding environmental sequences) and at DDBJ (all) revealed that a large number of them probably contained undescribed introns. Therefore, these sequences probably require a manual curation, but again highlight the importance of intron identification in eukaryotic sequences.

For lower taxonomic ranks, there were primarily two types of cases resulting in a failure to assign a taxonomic identity:

(1) No agreement between experts to resolve at a given rank. For example, the genus (rank 7) is assigned, the order (rank 5) is assigned, but a family (rank 6) has not yet been described, or this rank is in fact polyphyletic, with no proper descriptions of the different families.
(2) A given sequence is similar at the family level with several sequences from different families; however, they agree at the order level.

In such cases, this sequence was assigned as . . . |Order| Order_X[Genus|Genus + species. If a genus was not described (i.e. uncultured), the taxonomy becomes: . . . |Order| Order_X[Order_XX|Order_XX + sp.

More than 74 000 sequences (54% of total number of sequences in the PR2 database) belong to Opisthonkonta (Figure 1). Alveolata and Archaeplastida are second in abundances (15 and 12%, respectively). Stramenopiles and Rhizaria represent 7.2 and 5.6 %, respectively. Others SuperGroups represent less than 2.2%. Only 29.4% are complete or nearly complete. In total, 63.7% of sequences include the V4 region and only 12.1% and 11.7% include the V9 region as recognized by primers Biomarks and Wamps (see the legend of Figure 1), respectively. Apusozoa, Hacrobia, Excavata and Opisthokonta have <10% of their sequences that include the V9 region. V9 region of Amoebozoa and Archaeplastida are better represented (34% and 25%, respectively, using the Biomarks primers).

Table 1. Number of nuclear-encoded sequences in PR2 as annotated at the Super-Group taxonomic level

Super-group	n1	n2
Alveolata	20 760	20 255
Amoebozoa	1902	1880
Apusozoa	254	242
Archaeplastida	16 309	16 092
Eukaryota_Mikro	3	3
Eukaryota_X	54	54
Excavata	2871	2869
Hacrobia	2192	2132
Opisthokonta	75 056	74 484
Rhizaria	7581	7459
Stramenopiles	9884	9640
Total nuclear-encoded Eukaryota	136 866	135 110
Apicoplast	26	26
Chloroplast SSU	9657	9657
Hydrogenosome SSU	6	6
Mitochondrion SSU	36 051	36 051
Nucleomorph SSU (18S)	264	262

n1, total number; n2, excluding putative chimera; Super-Group, rank 2 taxonomy.

DOWNLOADS

We provide several different ways of downloading the database or part of it (see more explanations at http://ssu-rrna.org/downloads_eukaryotic_main_page.html).

(1) The entire database or sequences of a specific clade can be downloaded using a taxonomy browser under fasta format, with sequence identifiers as described above. Putative chimera have been removed.
(2) The entire database or sequences of major groups can be downloaded under fasta format, with only the short unique identifier. The corresponding

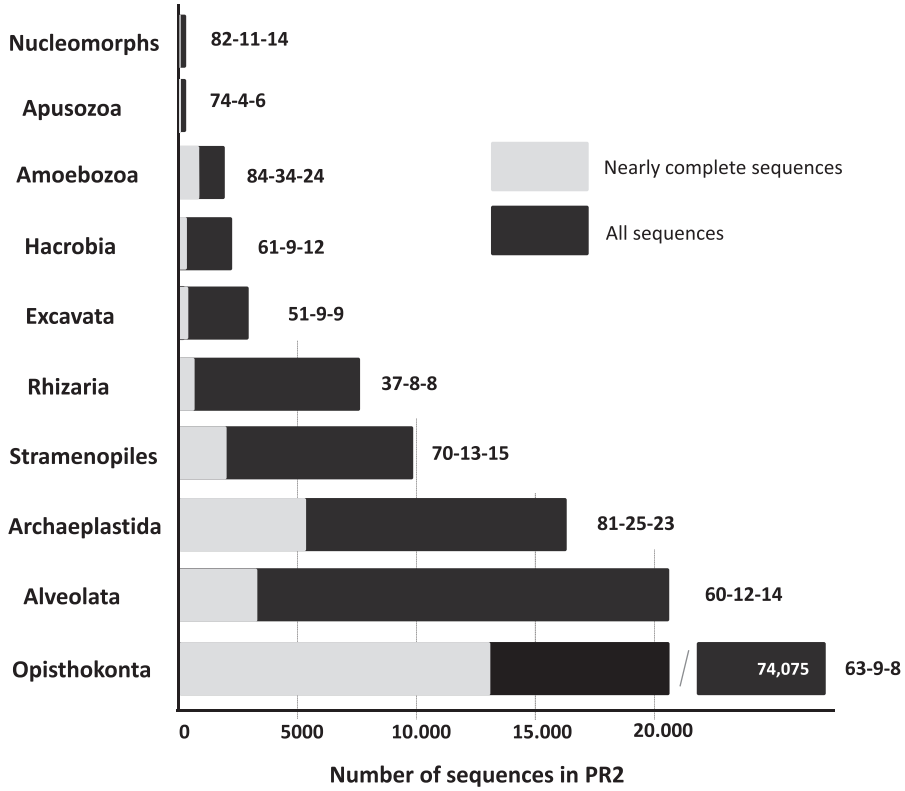

Figure 1. Total number of SSU rDNA gene sequences in the PR2 database for each main eukaryotic lineage (all sequences = grey + black, complete or nearly complete sequences in light-grey). Note that nucleomorphs were extracted from Archaeplastida. Numbers indicated after bars indicate percentages of sequences that include the following: (i) the V4 region as defined by primers forward CCAGCASCYGCGGTAATTCC and reverse ACTTTCGTTCTTGATYRA used during the European Biomarks project; (ii) the V9 region as defined by primers forward GTACACACCGCCCGTC and reverse TGATCCTTCTGCAGGTTCACCTAC used during the European Biomarks project; and (iii) the V9 region defined by primers forward TTGTACACACCGCCC and reverse CCTTCYGCAGGTTCACCTAC used by the WAMPS project. For Opithokonta, number in white = total number of sequences.

taxonomy is then downloaded as a tabulated file. This fasta format is appropriate to use in tools that do not allow for long sequence identifiers. They are also easier to use in large computations, as they spare the memory required. Finally, they are easier to use in pipelines or web sites (see below).

(3) The entire database, taxonomies and sequences under tabulated format, for easy import in relational databases.
(4) The entire database or sequences of a specific clade under fasta format, with sequence identifiers as described above, but after a clustering by sequence similarity (98, 96, 92%) and choosing only the longest sequence as representative of the cluster.
(5) Phylogenetic trees are available for the main groups. They were built using pair-wise distance computations (not taking introns as differences as explained above) and FastMe (14).
(6) Finally, we provide an 'arb' filter that allows to easily import a fasta file (with taxonomy in the identifier) into an arb database, separating sequences and taxonomy as required.
(7) *In silico* extracted domains corresponding to regions widely used in published articles and corresponding to several couples of primers.

SEARCHING THE DATABASE

We provide the following additional kinds of tools:

(1) A search by keywords, allowing to search according to taxonomy, accession number and PMID (PubMed ID: retrieval of sequences described in a given publication). Retrieven sequences can be filtered according to length, quality and when containing the variable V4 of V9 domains (often used in conjunction with deep sequencing).
(2) A search by 'sequence signature', with a link to the KeyDNAtools website (http://keydnatools.com/). This tool provides very fast results even for files containing many sequences. It also allows for detection of putative chimera as explained above.
(3) A BLAST search against the database, as usually found on most sites.
(4) A search (Crunch_Assign) using our modified global (Needleman–Wunsch based) algorithm that returns the most similar hits based on the entire alignment of the sequences, and not based on a good local alignment (high scoring pair, in BLAST). As a result, the percentage of similarity computed is more in agreement with what would be found using a Multiple Sequence Alignment [Clustal (15), Muscle

(16), MAFFT (17),...] before computing distances. It allows or does not allow accounting for introns as described above.
(5) A search of one or two primer motifs in sequences, returning every sequence that contains the primer(s) with International Union of Pure and Applied Chemistry (IUPAC) encoding allowed and also the possibility of mismatches between primer and sequence (a C program).
(6) *In silico* extracted domains corresponding to regions widely used in published articles and corresponding to several couples of primers.

Both BLAST and Crunch_Assign similarity searches are coupled to BLAST2Tree or Crunch_Assign2Tree that use our Scriptree software (18). Similarity search results can simply be copied and then pasted in the '2Tree section'; a phylogenetic tree is built and displayed on the fly, with taxonomic assignations (as chosen by the user) displayed in regard of each leaf. This section also allows downloading the sequences that have been pasted and the taxonomy as a tabulated file (19).

CONCLUSION AND PERSPECTIVES

There are presently three databases, SILVA (20), RDP (21) and Green genes (22), offering a curated taxonomy for prokaryotic SSU rRNA sequences. Only SILVA additionally provides reference sequences for SSU-rRNA sequences of eukaryotic origin, curated for sequence quality but using the NCBI taxonomy (although recently a 'SILVA' taxonomy is now proposed). Because our sequence identifier, i.e. accession.p1.p2, is similar to that used by SILVA, both databases can be easily compared.

Based on the last release 111, 1518 of the 71 787 eukaryotic SILVA reference sequences are not present in the PR^2 database. Manual checks showed that these sequences correspond to sequences extracted from entries in which no annotation allowed to identify the presence of a SSU-rRNA sequence, annotated as mRNA or annotated as prokaryotes. In all, 670 sequences identified as mitochondria were not in PR^2; none of the SILVA chloroplast sequences was absent from PR^2. Missing sequences will be soon analysed and incorporated in PR^2.On the other hand, 53 735/7774 nuclear, 31 492/29 763 mitochondrial, 462/18 chloroplastic and 133/80 other organelle sequences present in PR^2 were not in SILVA reference sequences and SILVA entire database, respectively. This can be largely explained by the use of drastic filtering steps used by SILVA both in minimal length and sequence quality. However, because we are also users of such databases to analyse NGS data sets, we detected two major reasons not to use too drastic quality filtering. First, representatives of novel environmental clades are often found within clone libraries with length of $<$1000 nt. Also, use of extreme quality filters may remove important sequences representatives of environmental groups, too short and/or having poor quality at one of the end of a sequence (one-step Sanger sequencing without enough noise treatment for example). In PR^2, sequence quality was indirectly inferred by the quality of the taxonomic assignation because bad-quality sequences became poorly assigned. Again, as sequence identifiers are similar between both databases, sequences can be easily compared between both databases.

The PR^2 database possesses several valuable complementary tools or databases lacking in other databases.

A ranked taxonomy

As for the PR^2 database, SILVA taxonomy for eukaryotes now offers a taxonomy based on the structure proposed by Adl *et al.* (12). However, contrarily to SILVA, we proposed a normalized eight terms ranked taxonomy for every sequence in the database. We proceeded to this 'normalization' from our experience in dealing with very large data sets using automated pipelines, and a depth of sequencing that revealed organisms spanning the entire spectrum of known living organisms. When considering the NCBI taxonomy for example, two sequences of Perciformes were found described using 22 ranks (AY263842 and EF470892 for Perciformes), whereas another Perciforme (AF112595) was described using only 15 ranks, and 10 360 sequences of Perciformes had between 16 and 21 ranks. Numerous examples exist for protists. A very good example is for the genus *Carpediomonas*. NCBI classify this genus within Eukaryota (rank 1), Fornicata (rank 2), Carpediomonas (rank 3). However, sequence AY117416 (*Carpediemonas membranifera*, 23) has no rank 2 taxonomy in its entry. As a result, it becomes extremely difficult using a computer and the lists of terms provided by a non-ranked taxonomy to identify for two different sequences, which members of the two lists indeed correspond to the same rank. This is the problem solved by our ranked taxonomy, thanks to a worldwide list of taxonomic experts. As an example, taxonomy of sequence AY117416 becomes Eukaryota|Excavata|Metamonada|Fornicata|Fornicata_Group-2|Carpediomonas-like|Carpediemonas|Carpediemonas + membranifera in PR^2. In SILVA, this sequence is linked to a 7 terms taxonomy, but taxonomy is seemingly not ranked and unified.

When occurring, missing ranks are automatically replaced in PR^2 (labeled as clade-i_X, where clade-i is the term for the next higher rank). This strategy allows rapidly inferring the taxonomy at the most probable higher rank and provides a rapid method for screening putative novel lineages at each taxonomic level.

Introns

Most SSU rRNA databases and biodiversity analyses of prokaryotes understandably neglect introns. Although found even in *Escherichia coli* (24,25), introns are rare in Bacteria and not very abundant in Archaea. Even when present, they have not yet been, to our knowledge, described in rRNA gene sequences. However, in Eukaryota, introns can be relatively abundant in rRNA gene sequences at least in some groups (9). This led us to incorporate in our database both the rRNA and the rDNA sequences. As most NGS (or clone library) analyses of the biodiversity are dealing with PCR amplification of extracted gDNA, introns may represent a large

part of the variability observed. Having genomic sequences, in addition to the rRNA transcript, in the database is important, not only for searching by similarity but also for the *in silico* estimation of expected amplicon lengths.

Organelles

Organelles are often poorly treated in reference databases. For hydrogenosomes (AJ237907, AJ871215, AJ871217, AJ871267, Y16670), only sequence AJ871217 can be found in SILVA labeled as 'Unclassified'. For GreenGenes, sequences were not found when searching by accession number. At RDP, the classifier resulted in every case into 'unclassified_Bacteria'. For the 26 apicoplast sequences, none was found in SILVA reference sequences or in the 'ssu-accession-parc.acs', release 111 (3 186 762 accession numbers). Even for better-known organelles, taxonomic assignation is not really better. For example, sequence AB000109 mitochondrion of *Dictyostelium discoideum* is labeled as 'Unclassified' in SILVA. Chloroplasts are generally well identified in SILVA. However, among the chloroplastic sequences detected in this study, 263 were found in SILVA reference sequences as chloroplasts. Our approach to build independent databases for these organelles allowed us to probably reach a more precise taxonomic affiliation of organelles. Having such prokaryotic organelles in our database is essential with NGS data sets of both prokaryotes and eukaryotes because the use of 'Bacteria' or 'Eukaryota' specific primers resulted in some cases in a significant proportion of amplicons that are in fact of Organelle origin (3–7). Even if Organelle sequences are simply discarded from the final analysis, this database avoids identifying these sequences as some new deep lineages.

Chimeric sequences

Chimeric sequences are PCR-generated hybrid products between multiple parent sequences that can be falsely interpreted as novel organisms, thus inflating apparent diversity (8,26). The two algorithms most widely used for 16S chimera detection are Pintail (27), included in RDP and SILVA databases, and Bellerophon (28) included in GreenGenes. In all cases, chimera are detected by comparing independent regions of a sequence alignment. The KeyDNAtools does not require the prior alignment of sequences, and it is particularly efficient to detect complex chimera having more than two parent sequences, or between two closely related parents. This tool can be used in concert with other detection methods. Our database, which has been screened for putative chimera, offers two possibilities of download: either including or excluding putative chimeric sequences.

Similarity searches

BLAST is a widely used tool that finds regions of local similarity between sequences. However, such search based on a good local high scoring pair could lead to very bad results. We thus developed two independent methods of assignation. The first one, the Crunch_Assign software is using a Needleman–Wunsch algorithm. It is also faster than BLAST and returns a score computed on the entire alignment. Because we are working on Eukaryotes, we also included the possibility of ignoring putative introns (to our knowledge, this possibility is not included in any other software). The second one, the KeyDNAtools is also very fast and offers additionally chimera detection as discussed above. In >95% of cases, both assignations provide similar results. Sequences not annotated by the KeyDNAtools likely result from the absence of the corresponding clade in the core reference database, low quality sequences or novel variants of the gene present in newly available sequences, not yet included in the core data set. Conversely, sequences not assigned by the Crunch_Assign software are often chimera or low-quality sequences. After a search by similarity, we offer the possibility to build a phylogenetic tree on the fly, using most similar sequences found by BLAST or Crunch_Assign.

Updates

We have developed a pipeline that allows to analyse a GenBank new release within a week. Most of the time spent is indeed in manual checking of conflicts after average linkage clusterings, as explained previously. As a result, updates of the PR^2 database will be done shortly after each GenBank new release. As a result, numbers provided in this article will probably differ from that available from PR2 at publication time of this manuscript.

ACKNOWLEDGEMENTS

The authors warmly thank Marion Viprey for her pioneering helps with the construction of the PR^2 database. Computations have been done on the 'Mésocentre SIGAMM' machine, hosted by Observatoire de la Côte d'Azur, Nice, France.

FUNDING

The European Union's Seventh Framework Programmes (FP7) BIOMARKS (2008-6530, ERA-net Biodiversa) and MicroB3 [287589] and the following ANR (France) projects: AQUAPARADOX, PARALEX and GIME. Funding for open access charge: ANR Paralex (French) and BIOMARKS (FP7).

Conflict of interest statement. None declared.

REFERENCES

1. López-García,P., Rodríguez-Valera,F., Pedrós-Alió,C. and Moreira,D. (2001) Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature*, **409**, 603–607.
2. Moon-van der Staay,S.Y., Watcher,R.D. and Vaulot,D. (2001) Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature*, **409**, 607–610.
3. Pawlowski,J., Christen,R., Lecroq,B., Bachar,D., Shahbakkia,H.R., Amaral-Zettler,A. and Guillou,L. (2011) Eukaryotic richness in the abyss: insights from pyrotag sequencing. *PLoS One*, **6**, e18169.
4. Hartmann,M., Howes,C.G., Vaninsberghe,D., Yu,H., Bachar,D., Christen,R., Henrik,N.R., Hallam,S.J. and Mohn,W.W. (2012)

Significant and persistent impact of timber harvesting on soil microbial communities in Northern coniferous forests. *ISME J.*, **6**, 2199–2218.
5. Lecroq,B., Lejzerowicz,F., Bachar,D., Christen,R., Esling,P., Baerlocher,L., Østerås,M., Farinelli,L. and Pawlowski,J. (2011) Ultra-deep sequencing of foraminiferal microbarcodes unveils hidden richness of early monothalamous lineages in deep-sea sediments. *Proc. Natl Acad. Sci. USA*, **108**, 13177–13182.
6. Edgcomb,V., Orsi,W., Bunge,J., Jeon,S., Christen,R., Leslin,C., Holder,M., Taylor,G.T., Suarez,P., Varela,R. *et al.* (2011) Protistan microbial observatory in the Cariaco Basin, Caribbean. I. Pyrosequencing vs Sanger insights into species richness. *ISME J.*, **5**, 1344–1356.
7. Behnke,A., Engel,M., Christen,R., Nebel,M., Klein,R.R. and Stoeck,T. (2011) Depicting more accurate pictures of protistan community complexity using pyrosequencing of hypervariable SSU rRNA gene regions. *Environ. Microbiol.*, **13**, 340–349.
8. Berney,C., Fahrni,J. and Pawlowski,J. (2004) How many novel eukaryotic 'kingdoms'? Pitfalls and limitations of environmental DNA surveys. *BMC Biol.*, **2**, 13.
9. Bachar,D., Guillou,L. and Christen,R. (2012) Detection of introns in eukaryotic small subunit ribosomal RNA gene sequences. *Dataset Pap. Biol.*, doi:10.7167/2013/854869.
10. Guillou,L., Viprey,M., Chambouvet,A., Welsh,R.M., Massana,R., Scanlan,D.J. and Worden,A.Z. (2008) Widespread occurrence and genetic diversity of marine parasitoids belonging to Syndiniales (Alveolata). *Environ. Microbiol.*, **10**, 3349–3365.
11. Logares,R., Audic,S., Santini,S., Pernice,M.C., de Vargas,C. and Massana,R. (2012) Diversity patterns and activity of uncultured marine heterotrophic flagellates unveiled with pyrosequencing. *ISME J.*, **6**, 1823–1833.
12. Adl,S.M., Simpson,A.G.B., Lane,C.E., Lukeš,J., Bass,D., Bowser,S.S., Brown,M.W., Burki,F., Dunthorn,M., Hampl,V. *et al.* (2012) The revised classification of eukaryotes. *J. Eukaryot. Microbiol.*, **59**, 429–493.
13. Carnegie,R.B., Meyer,G.R., Blackbourn,J., Cochennec-Laureau,N., Berthe,F.C. and Bower,S.M. (2003) Molecular detection of the oyster parasite *Mikrocytos mackini*, and a preliminary phylogenetic analysis. *Dis. Aquat. Organ.*, **54**, 219–227.
14. Desper,R. and Gascuel,O. (2002) Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J. Comput. Biol.*, **9**, 687–705.
15. Larkin,M.A., Blackshields,G., Brown,N.P., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.
16. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
17. Katoh,K. and Toh,H. (2010) Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics.*, **26**, 1899–900.
18. Chevenet,F., Croce,O., Hebrard,M., Christen,R. and Berry,V. (2010) ScripTree: scripting phylogenetic graphics. *Bioinformatics.*, **26**, 1125–1126.
19. Croce,O., Chevenet,F. and Christen,R. (2010) A new web server for the rapid identification of microorganisms. *J. Microbial Biochem. Technol.*, **2**, 84–88.
20. Pruesse,E., Quast,C., Knittel,K., Fuchs,B., Ludwig,W., Peplies,J. and Glöckner,F.O. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.*, **35**, 7188–7196.
21. Cole,J.R., Chai,B., Farris,R.J., Wang,Q., Kulam,S.A., McGarrell,D.M., Garrity,G.M. and Tiedje,J.M. (2005) The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res.*, **33**, 294–296.
22. DeSantis,T.Z., Hugenholtz,P., Larsen,N., Rojas,M., Brodie,E.L., Keller,K., Huber,T., Dalevi,D., Hu,P. and Andersen,G.L. (2006) Greengenes, a chimera-checked 16s rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.*, **72**, 5069–5072.
23. Simpson,A.G., Roger,A.J., Silberman,J.D., Leipe,D.D., Edgcomb,V.P., Jermiin,L.S., Patterson,D.J. and Sogin,M.L. (2002) Evolutionary history of 'early-diverging' eukaryotes: the excavate taxon *Carpediemonas* is a close relative of Giardia. *Mol. Biol. Evol.*, **19**, 1782–1791.
24. Sunde,M. (2005) Class I integron with a group II intron detected in an *Escherichia coli* strain from a free-range reindeer. *Antimicrob. Agents Chemother.*, **49**, 2512–2514.
25. Ferat,J.L., Le Gouar,M. and Michel,F. (1994) Multiple group II self-splicing introns in mobile DNA from *Escherichia coli*. *C. R. Acad. Sci. III.*, **317**, 141–148.
26. Hugenholtz,P. and Huber,T. (2003) Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. *IJSEM*, **53**, 289–293.
27. Ashelford,K.E., Chuzhanova,N.A., Fry,J.C., Jones,A.J. and Weightman,A.J. (2005) At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl. Environ. Microbiol.*, **71**, 7724–7736.
28. Huber,T., Faulkner,G. and Hugenholtz,P. (2004) Bellerophon; a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics*, **20**, 2317–2319.