Station Biologique de Roscoff
UMR7144 - CNRS and Sorbonne Université

PR2

release 4.11

Daniel Vaulot
vaulot@sb-roscoff.fr

November 3, 2018

Content

What is PR^2 ?

What is PR^2 ?

- PR^2 = **P**rotist **R**ibosomal **R**eference database.
  Open access database of eukaryotic 18S rRNA sequences.

What is PR^2 ?

- PR^2 = **P**rotist **R**ibosomal **R**eference database.
  Open access database of eukaryotic 18S rRNA sequences.
- All sequences originate from GenBank.

What is PR^2 ?

- PR^2 = **P**rotist **R**ibosomal **R**eference database.
  Open access database of eukaryotic 18S rRNA sequences.
- All sequences originate from GenBank.
- Sequences receive a detailed taxonomic assignment (8 levels).
  Taxonomic annotation for both strain and environmental sequences.

What is PR^2 ?

- PR^2 = **P**rotist **R**ibosomal **R**eference database.
  Open access database of eukaryotic 18S rRNA sequences.
- All sequences originate from GenBank.
- Sequences receive a detailed taxonomic assignment (8 levels).
  Taxonomic annotation for both strain and environmental sequences.
- 176,818 sequences

What is PR^2 ?

- PR^2 = **P**rotist **R**ibosomal **R**eference database.
  Open access database of eukaryotic 18S rRNA sequences.
- All sequences originate from GenBank.
- Sequences receive a detailed taxonomic assignment (8 levels).
  Taxonomic annotation for both strain and environmental sequences.
- 176,818 sequences
- 47,000 species

What is PR^2 used for ?

What is PR^2 used for ?

- Annotation of metabarcoding data

What is PR^2 used for ?

- Annotation of metabarcoding data
- Biogeography

What is PR^2 used for ?

- Annotation of metabarcoding data
- Biogeography
- Sequence analysis

What is PR^2 used for ?

- Annotation of metabarcoding data
- Biogeography
- Sequence analysis
- Links between phylogeny and functional traits

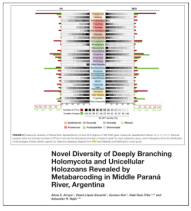
What is PR^2 used for ?

- Annotation of metabarcoding data
- Biogeography
- Sequence analysis
- Links between phylogeny and functional traits
- 220 papers citing PR^2.

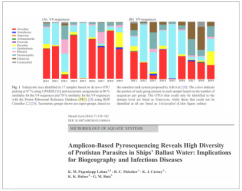
What is PR^2 used for ?

Metabarcoding

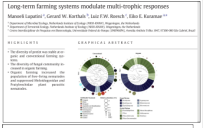
- Marine ecosystems



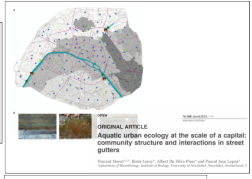
Novel Diversity of Deeply Branching Holomycota and Unicellular Holozoans Revealed by Metabarcoding in Middle Paraná River, Argentina



Amplicon-Based Pyrosequencing Reveals High Diversity of Protistan Parasites in Ships' Ballast Water: Implications for Biogeography and Infectious Diseases



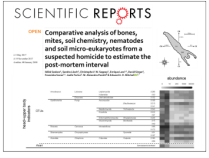
Long-term farming systems modulate multi-trophic responses



ORIGINAL ARTICLE

Aquatic urban ecology at the scale of a capital: community structure and interactions in street gutters

The ecology and diversity of microbial eukaryotes in geothermal springs



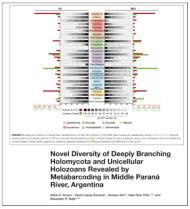
SCIENTIFIC REPORTS

Comparative analysis of bones, mites, soil chemistry, nematodes and soil micro-eukaryotes from a suspected homicide to estimate the post-mortem interval

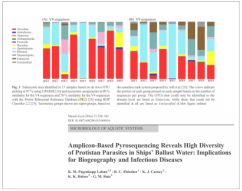
What is PR^2 used for ?

Metabarcoding

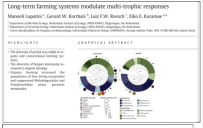
- Marine ecosystems
- Ballast waters



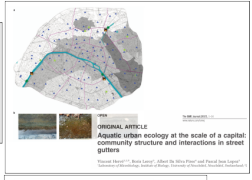
Novel Diversity of Deeply Branching Holomycota and Unicellular Holozoans Revealed by Metabarcoding in Middle Paraná River, Argentina



Amplicon-Based Pyrosequencing Reveals High Diversity of Protistan Parasites in Ships' Ballast Water: Implications for Biogeography and Infectious Diseases

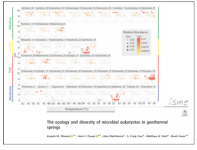


Long-term farming systems modulate multi-trophic responses

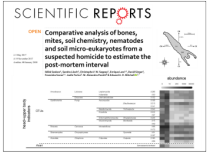


ORIGINAL ARTICLE

Aquatic urban ecology at the scale of a capital: community structure and interactions in street gutters



The ecology and diversity of microbial eukaryotes in geothermal springs

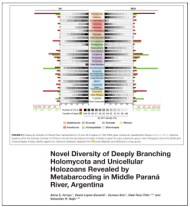


Comparative analysis of bones, mites, soil chemistry, nematodes and soil micro-eukaryotes from a suspected homicide to estimate the post-mortem interval

What is PR^2 used for ?

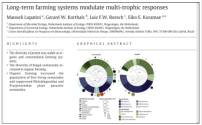
Metabarcoding

- Marine ecosystems
- Ballast waters
- River systems

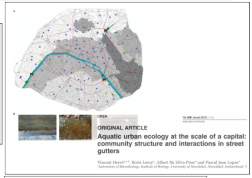


Novel Diversity of Deeply Branching Holomycota and Unicellular Holozoans Revealed by Metabarcoding in Middle Paraná River, Argentina

Amplicon-Based Pyrosequencing Reveals High Diversity of Protistan Parasites in Ships' Ballast Water: Implications for Biogeography and Infectious Diseases

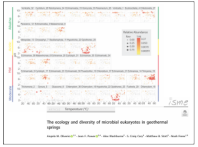


Long-term farming systems modulate multi-trophic responses

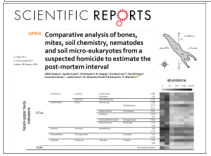


ORIGINAL ARTICLE

Aquatic urban ecology at the scale of a capital: community structure and interactions in street gutters



The ecology and diversity of microbial eukaryotes in geothermal springs



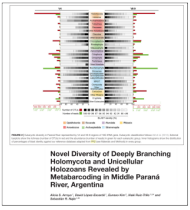
SCIENTIFIC REPORTS

Comparative analysis of bones, mites, soil chemistry, nematodes and soil micro-eukaryotes from a suspected homicide to estimate the post-mortem interval

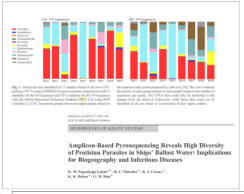
What is PR^2 used for ?

Metabarcoding

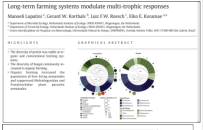
- Marine ecosystems
- Ballast waters
- River systems
- Hot Springs



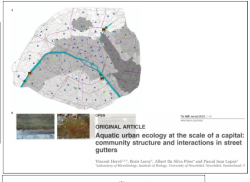
Novel Diversity of Deeply Branching Holomycota and Unicellular Holozoans Revealed by Metabarcoding in Middle Paraná River, Argentina



Amplicon-Based Pyrosequencing Reveals High Diversity of Protistan Parasites in Ships' Ballast Water: Implications for Biogeography and Infectious Diseases



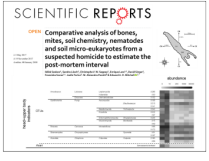
Long-term farming systems modulate multi-trophic responses



ORIGINAL ARTICLE

Aquatic urban ecology at the scale of a capital: community structure and interactions in street gutters

The ecology and diversity of microbial eukaryotes in geothermal springs



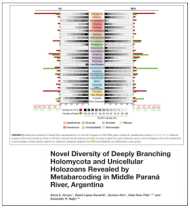
SCIENTIFIC REPORTS

Comparative analysis of bones, mites, soil chemistry, nematodes and soil micro-eukaryotes from a suspected homicide to estimate the post-mortem interval

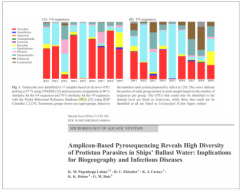
What is PR^2 used for ?

Metabarcoding

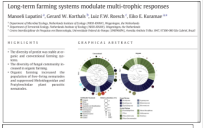
- Marine ecosystems
- Ballast waters
- River systems
- Hot Springs
- Soil



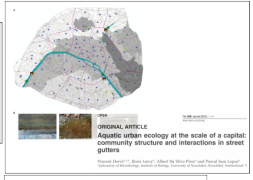
Novel Diversity of Deeply Branching Holomycota and Unicellular Holozoans Revealed by Metabarcoding in Middle Paraná River, Argentina



Amplicon-Based Pyrosequencing Reveals High Diversity of Protistan Parasites in Ships' Ballast Water: Implications for Biogeography and Infectious Diseases

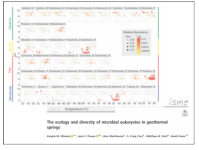


Long-term farming systems modulate multi-trophic responses

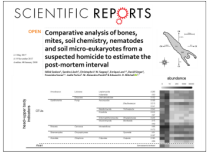


ORIGINAL ARTICLE

Aquatic urban ecology at the scale of a capital: community structure and interactions in street gutters



The ecology and diversity of microbial eukaryotes in geothermal springs



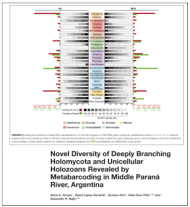
SCIENTIFIC REPORTS

Comparative analysis of bones, mites, soil chemistry, nematodes and soil micro-eukaryotes from a suspected homicide to estimate the post-mortem interval

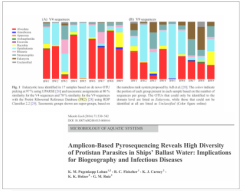
What is PR^2 used for ?

Metabarcoding

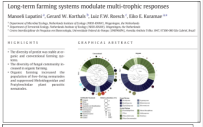
- Marine ecosystems
- Ballast waters
- River systems
- Hot Springs
- Soil
- Farming systems



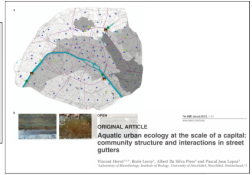
Novel Diversity of Deeply Branching Holomycota and Unicellular Holozoans Revealed by Metabarcoding in Middle Paraná River, Argentina



Amplicon-Based Pyrosequencing Reveals High Diversity of Protistan Parasites in Ships' Ballast Water: Implications for Biogeography and Infectious Diseases



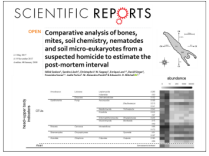
Long-term farming systems modulate multi-trophic responses



ORIGINAL ARTICLE

Aquatic urban ecology at the scale of a capital: community structure and interactions in street gutters

The ecology and diversity of microbial eukaryotes in geothermal springs



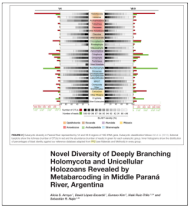
SCIENTIFIC REPORTS

Comparative analysis of bones, mites, soil chemistry, nematodes and soil micro-eukaryotes from a suspected homicide to estimate the post-mortem interval

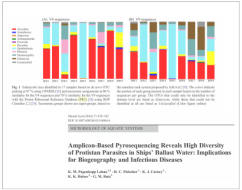
What is PR^2 used for ?

Metabarcoding

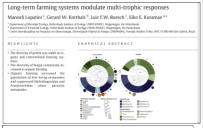
- Marine ecosystems
- Ballast waters
- River systems
- Hot Springs
- Soil
- Farming systems
- Urban ecology



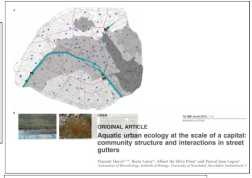
Novel Diversity of Deeply Branching Holomycota and Unicellular Holozoans Revealed by Metabarcoding in Middle Paraná River, Argentina



Amplicon-Based Pyrosequencing Reveals High Diversity of Protistan Parasites in Ships' Ballast Water: Implications for Biogeography and Infectious Diseases

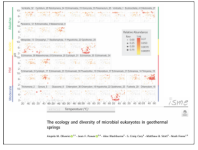


Long-term farming systems modulate multi-trophic responses

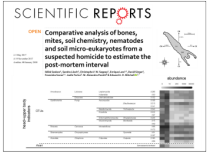


ORIGINAL ARTICLE

Aquatic urban ecology at the scale of a capital: community structure and interactions in street gutters



The ecology and diversity of microbial eukaryotes in geothermal springs



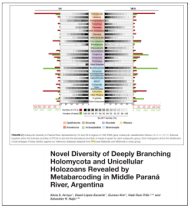
SCIENTIFIC REPORTS

Comparative analysis of bones, mites, soil chemistry, nematodes and soil micro-eukaryotes from a suspected homicide to estimate the post-mortem interval

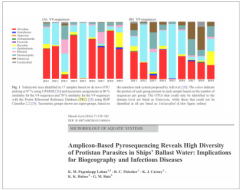
What is PR^2 used for ?

Metabarcoding

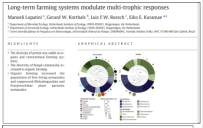
- Marine ecosystems
- Ballast waters
- River systems
- Hot Springs
- Soil
- Farming systems
- Urban ecology
- Criminology



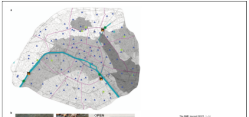
Novel Diversity of Deeply Branching Holomycota and Unicellular Holozoans Revealed by Metabarcoding in Middle Paraná River, Argentina



Amplicon-Based Pyrosequencing Reveals High Diversity of Protistan Parasites in Ships' Ballast Water: Implications for Biogeography and Infectious Diseases

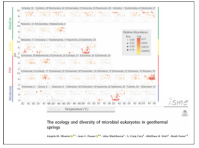


Long-term farming systems modulate multi-trophic responses

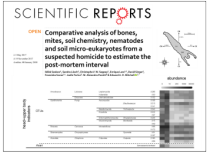


ORIGINAL ARTICLE

Aquatic urban ecology at the scale of a capital: community structure and interactions in street gutters



The ecology and diversity of microbial eukaryotes in geothermal springs

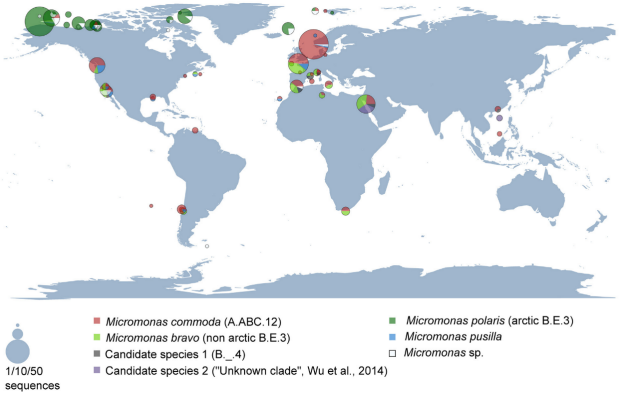


SCIENTIFIC REPORTS

Comparative analysis of bones, mites, soil chemistry, nematodes and soil micro-eukaryotes from a suspected homicide to estimate the post-mortem interval

What is PR^2 used for ?

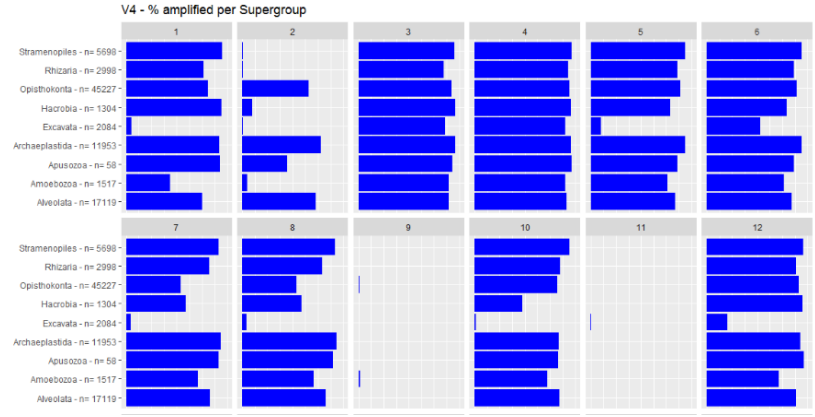
Biogeography



Simon, N. et al. 2017. Revision of the Genus Micromonas Manton et Parke (Chlorophyta, Mamiellophyceae), of the Type Species *M. pusilla* (Butcher) Manton & Parke and of the Species *M. commoda* van Baren, Bachy and Worden and Description of Two New Species. Protist. 168:612–35.

What is PR^2 used for ?

Sequence analysis



Primer analysis - *On-going work with S. Geisen and D. Bass.*

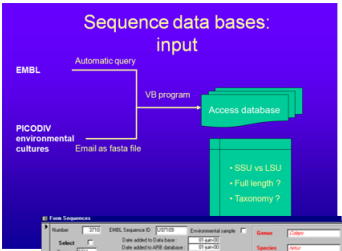
History of PR^2

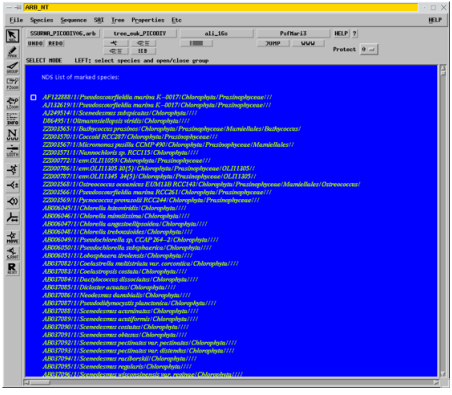
PR^2 history

1997	**Excel file** created by D. Vaulot during L. Guillou thesis
2000-2003	Access/ARB database maintained by D. Vaulot during **PICODIV**
2006-2010	**KeyDNAtools** developed by L. Guillou
2010-2013	Project **BioMarks**: creation of PR^2 by L. Guillou Database maintained by R. Christen : ssu-rrna.org
mid-2016	Web site died
2016	D. Vaulot takes over maintenance Raw data deposited to **Figshare**
2017	Database moved to MySQL Development of R scripts to manage the database Repository on **GitHub**

PR^2 history
PICODIV

PR² history
PICODIV



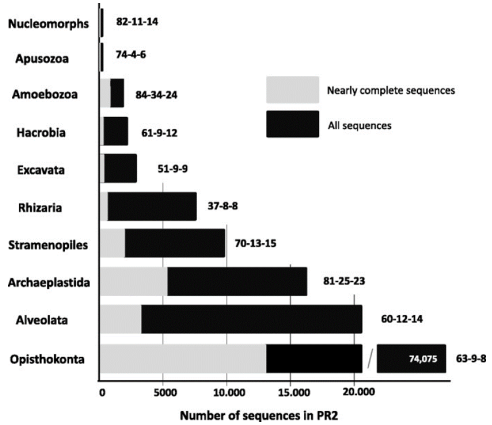


PR^2 history
Guillou et al. 2013 paper

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

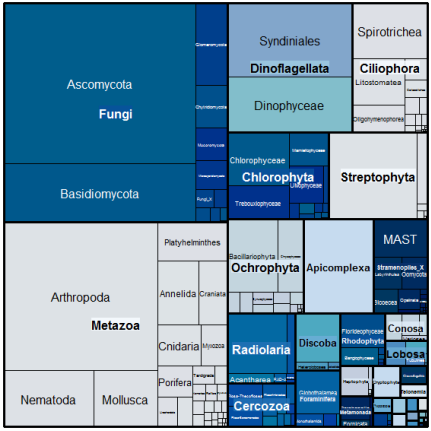
The Protist Ribosomal Reference database (PR^2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy

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,*]



PR^2 Now

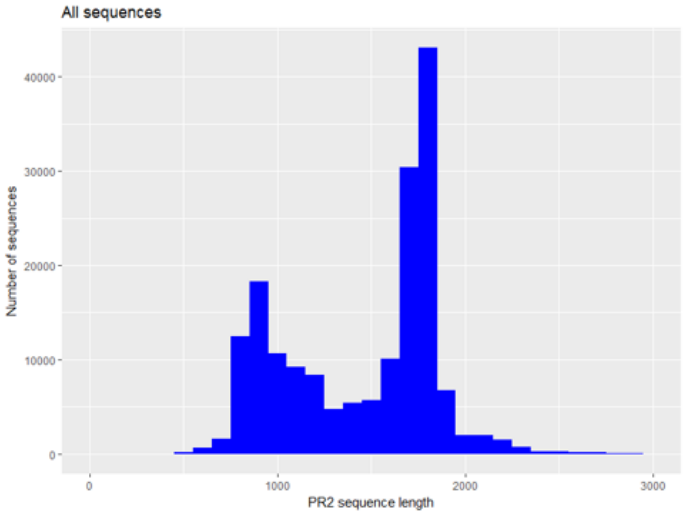
Statistics
Taxonomic distribution



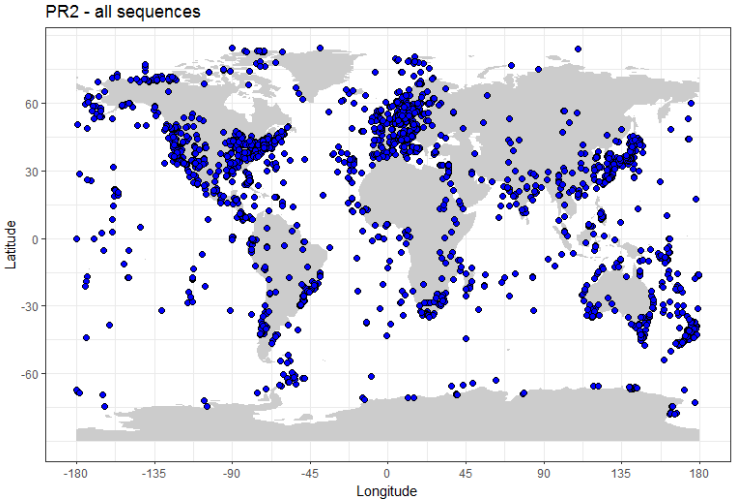
Number of taxa - all sequences

kingdom	supergroup	division	class	order	family	genus	species
1	11	38	227	506	1297	22708	45799

Statistics
Sequence size distribution



Statistics
Geographical distribution



Recent updates

Version	Date	Who	Major group updated
4.11	30/10/2018	D. Vaulot, A. Lopes	Chloropicophyceae, Mamiellophyceae
		EukRef	Ciliates
4.9	20/02/2018	S. Mordret, R. Piredda, D. Sarno	Dinophyceae
4.7	27/09/2017	C. Bachy, W.-T. Chen	Cilates (Sprirotrichea)
4.4	10/11/2016	D. Vaulot	Bolidophyceae
4.0	21/10/2015	B. Edvardsen	Haptophyta
3.0	31/8/2015	M. Tragin	Chlorophyta
2.0	07/02/2015	T. Biard	Rhizaria

Recent updates
Dinoflagellates

Received: 3 November 2017 | Revised: 15 February 2018 | Accepted: 24 February 2018
DOI: 10.1111/1755-0998.12781

RESOURCE ARTICLE

WILEY MOLECULAR ECOLOGY RESOURCES

DINOREF: A curated dinoflagellate (Dinophyceae) reference database for the 18S rRNA gene

Solenn Mordret[1] | Roberta Piredda[1] | Daniel Vaulot[2] | Marina Montresor[1] | Wiebe H. C. F. Kooistra[1] | Diana Sarno[1]

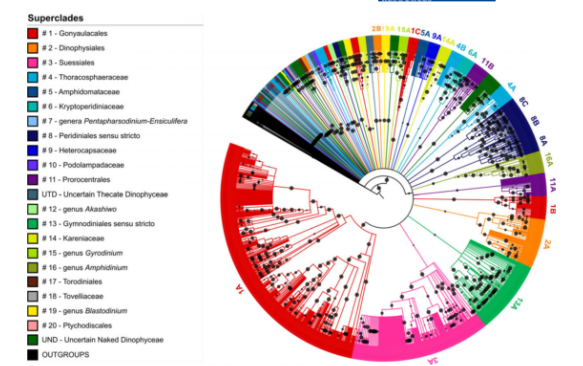


FIGURE 2 Consensus phylogenetic tree (RAXML, GTR model) based on 1,540 unique 18S rRNA sequences in the DINOREF. Alignment of 2,153 bp with three sequences of Ciliates (U97109; X56165 and X03772) and three sequences of Apicomplexa (M97703; AF236097 and AF291427) used as outgroup. Clades are ordered according to their size and are supported by bootstrap values ≥50%; black dots are proportional to bootstrap values. The colours of the Superclades and clades correspond to those in Table 1. Clades within each Superclade have been marked (A, B, C, etc.), along the outer rim of the tree, corresponding to their assignment in this figure. The Superclades "Uncertain Naked Dinophyceae" and "Uncertain Thecate Dinophyceae" have not been marked and neither have the small clades on the upper left of the tree. The tree can be visualized on ITOL version 3—Interactive Tree of Life (Letunic and Bork, 2016, at https://itol.embl.de/tree/1932052318357911479398328) in which all clades are marked

Recent updates

Ciliates - Eukref

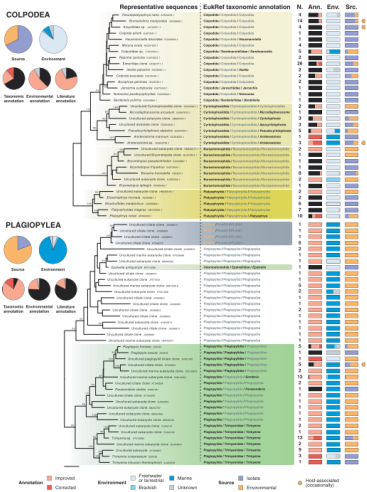
environmental microbiology

Environmental Microbiology (2018) **20**(6), 2218–2230 doi:10.1

EukRef-Ciliophora: a manually curated, phylogeny-based database of small subunit rRNA gene sequences of ciliates

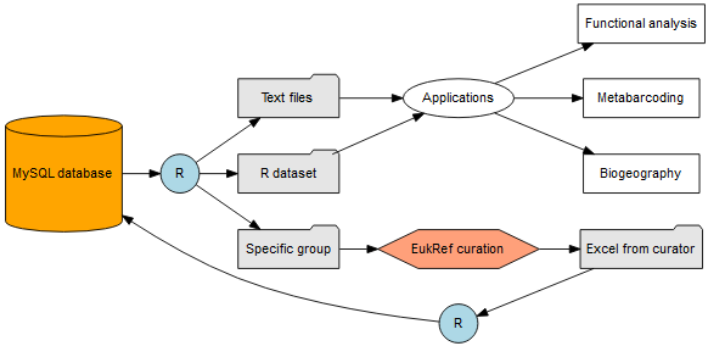
Vittorio Boscaro,[1*] Luciana F. Santoferrara,[2,3] Qianqian Zhang,[4] Eleni Gentekaki,[5] Mitchell J. Syberg-Olsen,[1] Javier del Campo[1†] and Patrick J. Keeling[1]

30%. The performance of EukRef-C
rior to the current SILVA database
reads from a global marine surve
outputs are publicly available to ma



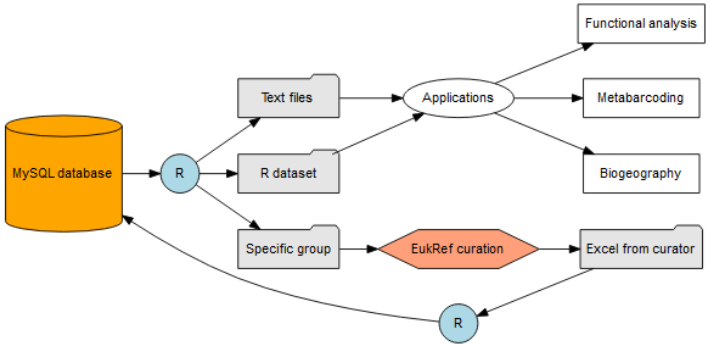
How is PR^2 implemented and maintained ?

Implementation



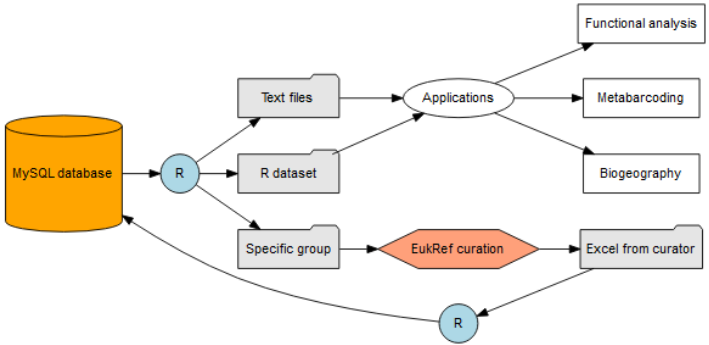
- MySQL database

Implementation



- MySQL database
- Processing done with R (tidyr libraries)

Implementation

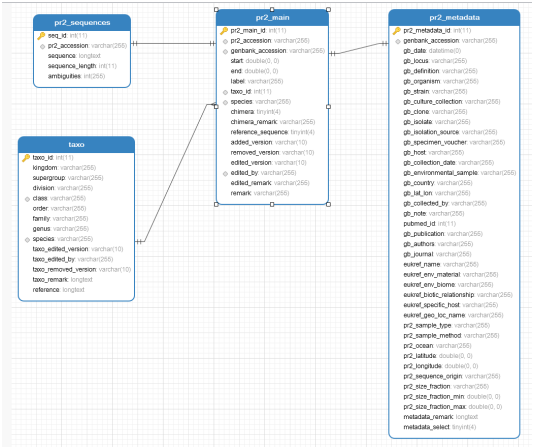


- MySQL database
- Processing done with R (tidyr libraries)
- Data available on GitHub (and Figshare - DOI number)

MySQL database

Tables

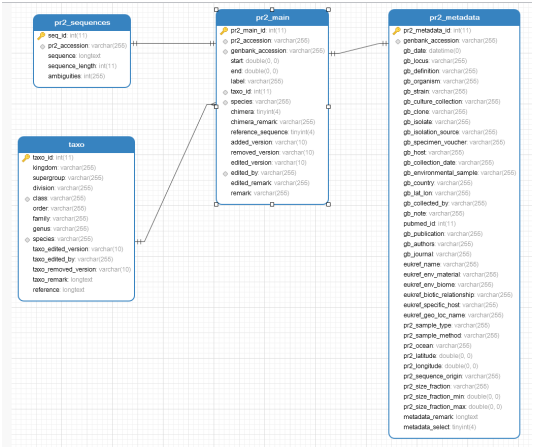
- pr2_main : sequences assigned to species



MySQL database

Tables

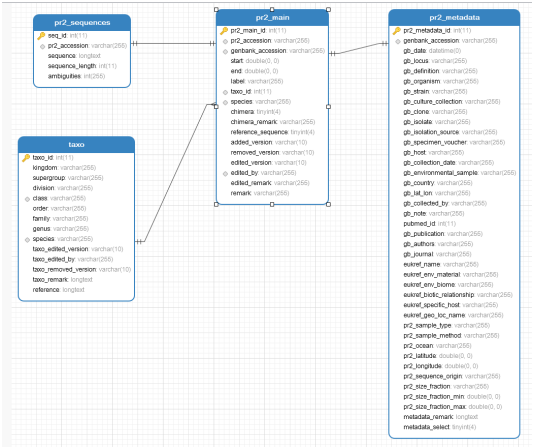
- pr2_main : sequences assigned to species
- pr2_sequence : sequence of each entry



MySQL database

Tables

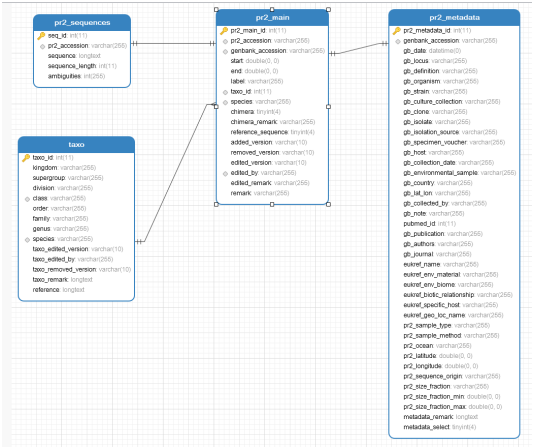
- pr2_main : sequences assigned to species
- pr2_sequence : sequence of each entry
- pr2_metadata : metadata for each entry



MySQL database

Tables

- pr2_main : sequences assigned to species
- pr2_sequence : sequence of each entry
- pr2_metadata : metadata for each entry
- pr2_taxonomy : one line per species



MySQL database

Table: pr2_main

- Each entry has a PR2 accession number
  (2 entries may correspond to the same Genbank accession number, e.g. for genomes)
- Sequences are linked to taxonomy by species name
- Annotation of Chimera (removed when PR2 is exported)

pr2_main_id	pr2_accession	genbank_accession	start	end	label	taxo_id	species	chimera	chimera_remark	added_date	added_version	removed_date	removed_version	edited_date
86	FJ831626.1.1132_U	FJ831626	1	1132	U	10910	Glomus_sp.	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)
87	EF527080.1.1343_U	EF527080	1	1343	U	1765	Dinophyceae_XXX_sp.	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)
88	FJ971827.1.959_U	FJ971827	1	959	U	10371	Stephanoecidae_Group_D_X_sp.	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)
89	EF024076.1.1760_U	EF024076	1	1760	U	1597	Oxytrichidae_X_sp.	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)
90	KC146315.1.919_U	KC146315	1	919	U	23185	Caprella_sp.	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)
91	HQ870051.1.844_U	HQ870051	1	844	U	1085	Scuticociliatia_XX_sp.	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)
92	AJ810741.1.1881_U	AJ810741	1	1881	U	25651	Acrotrichis_sp.	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)
93	AJ236016.1.1768_U	AJ236016	1	1768	U	7639	Nicotiana_tabacum	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)
94	AB534503.1.1611_UC	AB534503	1	1611	UC	13475	Pezizomycetes_X_sp.	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)
95	AB016507.1.1750_U	AB016507	1	1750	U	15756	Tetrapisispora_arboricola	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)
96	AF164355.1.1064_UC	AF164355	1	1064	UC	10837	Didymella_exigua	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)
97	DQ396545.1.1663_U	DQ396545	1	1663	U	375	Hepatocystis_sp.	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)
98	GU825500.1.1074_U	GU825500	1	1074	U	2321	Dino-Group-II-Clade-21_X_sp.	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)
99	AJ430853.1.1680_U	AJ430853	1	1680	U	18222	Glomus_arbuscular	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)
100	EF689841.1.3147_G	EF689841	1	3147	G	12180	Placidiopsis_cartilaginea	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)
101	D78331.1.1734_U	D78331	1	1734	U	17255	Dermomyces_huiaensis	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)
102	DQ459973.1.1689_U	DQ459973	1	1689	U	20459	Stylaria_lacustris	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)
103	AB208110.1.1742_U	AB208110	1	1742	U	14995	Tolypocladium_cylindrosporum	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)

MySQL database

Table: pr2_metadata

- Genbank annotations (gb_ fields)
- Some gb fields have been manually edited (e.g. gb_strain and gb_clone)
- Manually curated annotations (eg. sample_ fields)
- Fields computed from gb fields such as longitude and latitude
- Phenotypic information (auto vs. hetero, mixotroph etc. . .)

pr2_metadata_id	genbank_accession	gb_def	gb_locus	gb_definition	pubmed_id	gb_strain	gb_culture_collect	gb_clone	gb_isolate	gb_isolation_source	gb_specimen_voucher	gb_host
175028	KT860996	NULL	PLN	Pelagophyceae sp. RCC2505 18S ribos...	NULL	RCC2505	NULL	NULL	NULL	NULL	NULL	NULL
175029	KT861000	NULL	PLN	Pelagophyceae sp. RCC2511 18S ribos...	NULL	RCC2511	NULL	NULL	NULL	NULL	NULL	NULL
175030	KT861026	NULL	PLN	Pelagococcus sp. RCC3069 18S riboso...	NULL	RCC3069	NULL	NULL	NULL	NULL	NULL	NULL
175031	KT861062	NULL	PLN	Pelagomonas calceolata strain RCC105...	NULL	RCC1050	NULL	NULL	NULL	NULL	NULL	NULL
175032	KT861063	NULL	PLN	Pelagomonas calceolata strain RCC105...	NULL	RCC1051	NULL	NULL	NULL	NULL	NULL	NULL
175033	KT861109	NULL	PLN	Pelagomonas calceolata strain RCC969...	NULL	RCC969	NULL	NULL	NULL	NULL	NULL	NULL
175034	KT861111	NULL	PLN	Pelagophyceae sp. RCC1024 18S ribos...	NULL	RCC1024	NULL	NULL	NULL	NULL	NULL	NULL
175035	KU743777	NULL	ENV	Uncultured eukaryote clone 08D03P04 ...	NULL	NULL	NULL	08D03P04	NULL	subtropical coastal ecosystem	NULL	NULL
175036	KU743797	NULL	ENV	Uncultured eukaryote clone 08D03P54 ...	NULL	NULL	NULL	08D03P54	NULL	subtropical coastal ecosystem	NULL	NULL
175037	KU743799	NULL	ENV	Uncultured eukaryote clone 08D03P58 ...	NULL	NULL	NULL	08D03P58	NULL	subtropical coastal ecosystem	NULL	NULL
175038	KU743806	NULL	ENV	Uncultured eukaryote clone 08D30P22 ...	NULL	NULL	NULL	08D30P22	NULL	subtropical coastal ecosystem	NULL	NULL
175039	KU743842	NULL	ENV	Uncultured eukaryote clone 10D03P49 ...	NULL	NULL	NULL	10D03P49	NULL	subtropical coastal ecosystem	NULL	NULL
175040	KX014627	NULL	PLN	Pelagomonas calceolata strain RCC454...	NULL	RCC454B	NULL	NULL	NULL	NULL	NULL	NULL
175041	KX014630	NULL	PLN	Pelagophyceae sp. RCC4552 18S ribos...	NULL	RCC4552	NULL	NULL	NULL	NULL	NULL	NULL
175042	KX523139	NULL	PLN	Pelagomonas calceolata culture-collect...	NULL	RCC4419	NULL	NULL	NULL	NULL	NULL	NULL
175043	KX523144	NULL	PLN	Pelagomonas calceolata culture-collect...	NULL	RCC4425	NULL	NULL	NULL	NULL	NULL	NULL

MySQL database

Table: pr2_taxonomy

- 8 taxonomic levels (kingdom -> species)
- Follows PR2 convention (_X, _XX etc..)
- Contains 47 000 species
- Each name is unique (i.e. does not appear in different columns or different lines).
- Any daughter taxon has a unique mother taxon.

taxo_id	kingdom	supergroup	division	class	order	family	genus	species
947	Eukaryota	Alveolata	Ciliophora	Oligohymenophorea	Penicula	Paramecidae	Paramecium	Paramecium_polycaryum
948	Eukaryota	Alveolata	Ciliophora	Oligohymenophorea	Penicula	Paramecidae	Paramecium	Paramecium_primaurelia
949	Eukaryota	Alveolata	Ciliophora	Oligohymenophorea	Penicula	Paramecidae	Paramecium	Paramecium_putrinum
950	Eukaryota	Alveolata	Ciliophora	Oligohymenophorea	Penicula	Paramecidae	Paramecium	Paramecium_schewiakoffi
951	Eukaryota	Alveolata	Ciliophora	Oligohymenophorea	Penicula	Paramecidae	Paramecium	Paramecium_sp.
952	Eukaryota	Alveolata	Ciliophora	Oligohymenophorea	Penicula	Paramecidae	Paramecium	Paramecium_tetraurelia
953	Eukaryota	Alveolata	Ciliophora	Oligohymenophorea	Penicula	Paramecidae	Paramecium	Paramecium_woodruffi
955	Eukaryota	Alveolata	Ciliophora	Oligohymenophorea	Penicula	Penicula_X	Paranassula	Paranassula_sp.
956	Eukaryota	Alveolata	Ciliophora	Oligohymenophorea	Penicula	Penicula_X	Penicula_XX	Penicula_XX_sp.
957	Eukaryota	Alveolata	Ciliophora	Oligohymenophorea	Penicula	Stokesiidae	Stokesia	Stokesia_sp.
958	Eukaryota	Alveolata	Ciliophora	Oligohymenophorea	Penicula	Stokesiidae	Stokesia	Stokesia_vernalis
959	Eukaryota	Alveolata	Ciliophora	Oligohymenophorea	Penicula	Urocentridae	Urocentrum	Urocentrum_sp.
960	Eukaryota	Alveolata	Ciliophora	Oligohymenophorea	Penicula	Urocentridae	Urocentrum	Urocentrum_turbo
961	Eukaryota	Alveolata	Ciliophora	Oligohymenophorea	Peritrichia	Astylozoidae	Astylozoon	Astylozoon_enriquesi
962	Eukaryota	Alveolata	Ciliophora	Oligohymenophorea	Peritrichia	Epistylididae	Campanella	Campanella_umbellaria
963	Eukaryota	Alveolata	Ciliophora	Oligohymenophorea	Peritrichia	Epistylididae	Epistylis	Epistylis_chrysemydis
964	Eukaryota	Alveolata	Ciliophora	Oligohymenophorea	Peritrichia	Epistylididae	Epistylis	Epistylis_galea
965	Eukaryota	Alveolata	Ciliophora	Oligohymenophorea	Peritrichia	Epistylididae	Epistylis	Epistylis_hentscheli

R scripts

R scripts - uses tidyr universe

- Add new sequences from GenBank
- Correct taxonomy of existing sequences (EukRef output)
- Extract metadata from Genbank entries
- Check sequences problems (short sequences, sequences with ambiguities)
- Analyze taxonomy
- Export data to a variety of format (fasta, R data)

R scripts
Sequence processing

3.1 Read the data and reformat

3.1.1 Read the data

- Read the Tab delimited file because problems with strange characters and CR in some fields
- The first line of the file is skipped, the second line contains the name of the fields to be used in PR2
- The last column is empty and discarded (named "empty")
- Number of sequences = 11622

```
dir_pr2_update <- "../updates/2018 Ciliates Eukref"
pr2.env$editor <- "Eukref - Boscaro V."

full_path <- function(file_name) {
    str_c(dir_pr2_update, "/", file_name)
}

file_pr2_update_excel <- full_path("EukRef Ciliates 2018 2.0.xlsx")
# Import the tsv skipping the first line

pr2_update_raw <- read_tsv(full_path("EUKREF-CILIOPHORA_corrected_v2.0.tsv"),
    col_names = TRUE, guess_max = 2e+05, na = c("", "NA"), skip = 1)

problems(pr2_update_raw)
```

```
# tibble [0 x 4]
# ... with 4 variables: row <int>, col <int>, expected <chr>, actual <chr>
```

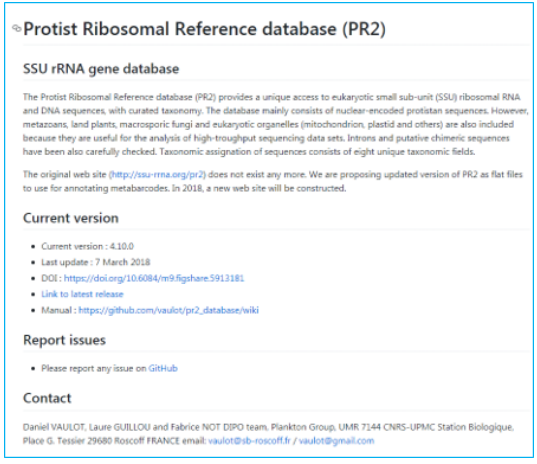
```
spec(pr2_update_raw)
```

Access to PR^2

GitHub

GitHub

- https://github.com/vaulot/pr2database
- Releases - current version 4.11.0
- Wiki
- Issues



Protist Ribosomal Reference database (PR2)

SSU rRNA gene database

The Protist Ribosomal Reference database (PR2) 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.

The original web site (http://ssu-rrna.org/pr2) does not exist any more. We are proposing updated version of PR2 as flat files to use for annotating metabarcodes. In 2018, a new web site will be constructed.

Current version

- Current version : 4.10.0
- Last update : 7 March 2018
- DOI : https://doi.org/10.6084/m9.figshare.5913181
- Link to latest release
- Manual : https://github.com/vaulot/pr2_database/wiki

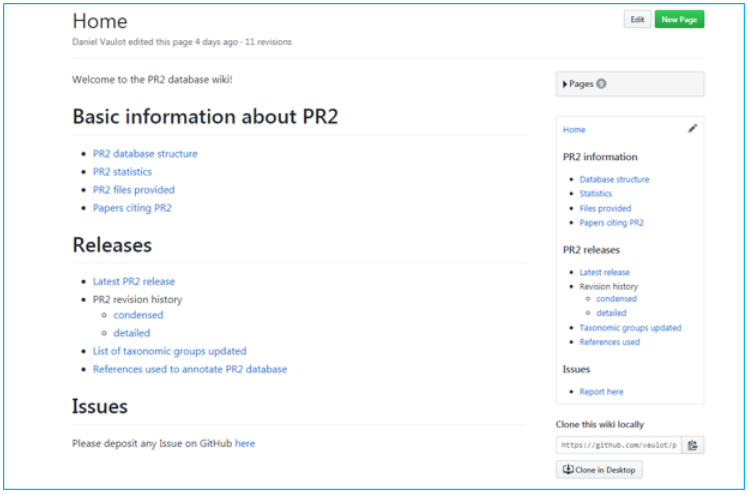
Report issues

- Please report any issue on GitHub

Contact

Daniel VAULOT, Laure GUILLOU and Fabrice NOT DIPO team, Plankton Group, UMR 7144 CNRS-UPMC Station Biologique, Place G. Tessier 29680 Roscoff FRANCE email: vaulot@sb-roscoff.fr / vaulot@gmail.com

GitHub
Wiki


Home
Daniel Vaulot edited this page 4 days ago · 11 revisions
Edit
New Page
Welcome to the PR2 database wiki!
Basic information about PR2
PR2 database structure
PR2 statistics
PR2 files provided
Papers citing PR2
Releases
Latest PR2 release
PR2 revision history
condensed
detailed
List of taxonomic groups updated
References used to annotate PR2 database
Issues
Please deposit any Issue on GitHub here
Pages
Home
PR2 information
Database structure
Statistics
Files provided
Papers citing PR2
PR2 releases
Latest release
Revision history
condensed
detailed
Taxonomic groups updated
References used
Issues
Report here
Clone this wiki locally
https://github.com/vaulot/p
Clone in Desktop

GitHub
Download formats

Export formats

- metabarcode annotation
  - mothur
  - Qiime
  - dada2
  - USEARCH, VSEARCH
- BLAST - fasta files
- metadata
- **R dataset** - new

PR2 files

Daniel Vaulot edited this page just now · 12 revisions

List of files are provided for each release

Important note : All files are in UNIX format (end of lines are indicated Line Feed-LF only). Please see at the bottom of the file how to converts the files to Windows format (end of lines are CR+LF). See at the bottom of this page how to convert.

- Two files for use with Qiime or Mothur.
  - **pr2_mothur.fasta.gz** contains all sequences in fasta format with the accession in the description line
  - **pr2_mothur.tax.gz** contains the taxonomy of each sequence separated from the accession number by a tabulation
    - Note :Qiime only use 7 taxonomical levels by default.
- **pr2_UTAX.fasta.gz** contains one fasta file with the accession number of the sequence and its full taxonomy on the description line in the UTAX format. It is suitable to use with USEARCH and VSEARCH.
- **pr2_dada2.fasta.gz** contains a dada2 format compatible training file
- **pr2_taxo_long.fasta.gz** contains one fasta file with the accession number of the sequence, the name of the sequence and its full taxonomy on the description line. It is suitable to build a local database for BLAST search
- **pr2_metadata.csv.gz** contains a tabulation separated file with all the metadata from genbank as well as annotation made to the PR2 database.
- **pr2_merged.csv.gz** contains a tabulation separated file the full PR2 database including sequences, taxonomy and metadata.
- **R dataset.** See detailed instructions on how to install and use

The R pr2database package

Daniel Vaulot edited this page 4 days ago · 7 revisions

The PR2 database is now provided as a R package

Installation

1

Install from the GitHub web site using the devtools package

```
install.packages(devtools)
devtools::install_github("vaulot/pr2database")
```

Selecting sequences from a specific taxon

2

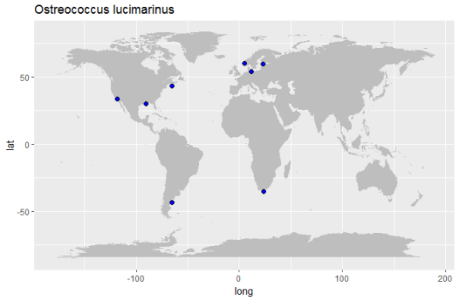
Let us select all the available sequences for the Mamiellophyceae *Ostreococcus*

```
# Filter only the sequences for which the column genus contains Ostreococcus
pr2_ostreo <- pr2 %>% dplyr::filter(genus == "Ostreococcus")

# Select only the columns of interest
pr2_ostreo <- pr2_ostreo %>% dplyr::select( genbank_accession, species,
                                            pr2_sample_type, gb_strain, gb_clone,
                                            pr2_latitude, pr2_longitude,
                                            sequence_length, sequence  )
```

Drawing a map of sequence locations

3

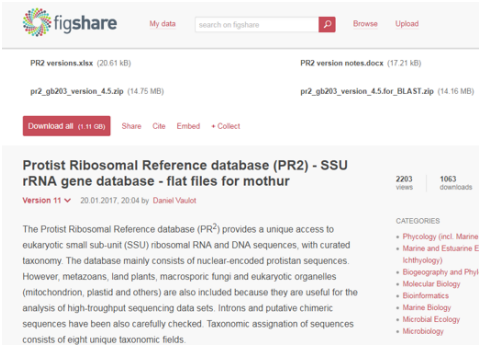
```
library(maps)
world <- map_data("world")

ggplot() +
  geom_polygon(data = world, aes(x=long, y = lat, group = group), fill="grey") +
  coord_fixed(1.3) +
  geom_point(data=pr2_ostreo, aes(x=pr2_longitude, y=pr2_latitude), fill="blue", size=2, sh
  ggtitle("Ostreococcus lucimarinus")
```



Figshare

Figshare

- https://doi.org/10.6084/m9.figshare.5913181
- PhytoRef (16S plastid) is also on Figshare .



What is next ?

Database

- Coordinate with **EukRef**

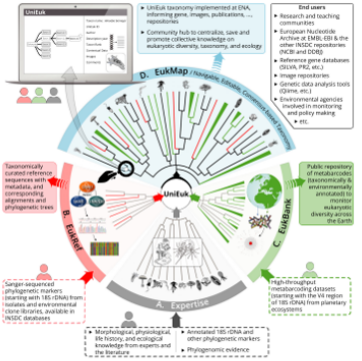


Figure 1 The UniEuk workflow. Bottom-up, community-based information on eukaryotic biodiversity from (A) classical knowledge, (B) phylogenetic diversity, and (C) environmental 'omics' surveys, converge and synergize through the UniEuk modules to inform the navigable and editable, consensus-based taxonomic framework (D). Dotted and colored frames indicate input and output information, respectively. Line drawings of eukaryotes adapted with permission from https://genev.unige.ch/system/pavlovskii/tree.png.

Database

- Coordinate with **EukRef**
- Reference sequences

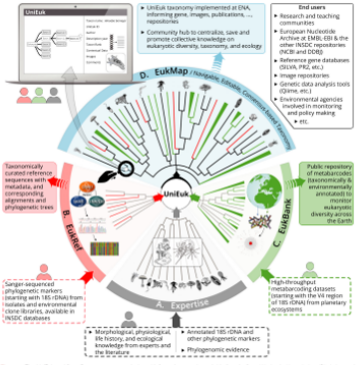


Figure 1 The UniEuk workflow. Bottom-up, community-based information on eukaryotic biodiversity from (A) classical knowledge, (B) phylogenetic diversity, and (C) environmental 'omics' surveys, converge and synergize through the UniEuk modules to inform the navigable and editable, consensus-based taxonomic framework (D). Dotted and colored frames indicate input and output information, respectively. Line drawings of eukaryotes adapted with permission from https://genev.unige.ch/system/pavlovskii/tree.png.

Database

- Coordinate with **EukRef**
- Reference sequences
- Chimeras

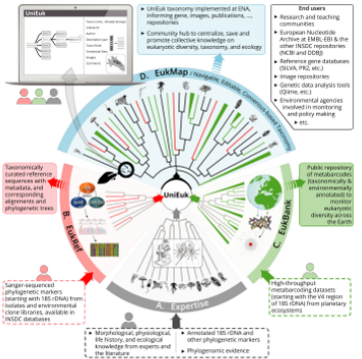


Figure 1 The UniEuk workflow. Bottom-up, community-based information on eukaryotic biodiversity from (A) classical knowledge, (B) phylogenetic diversity, and (C) environmental 'omics' surveys, converge and synergize through the UniEuk modules to inform the navigable and editable, consensus-based taxonomic framework (D). Dotted and colored frames indicate input and output information, respectively. Line drawings of eukaryotes adapted with permission from https://genev.unige.ch/system/pavlovskii/tree.png.

Database

- Coordinate with **EukRef**
- Reference sequences
- Chimeras
- Reannotate environmental sequences (Wang/DECIPHER)

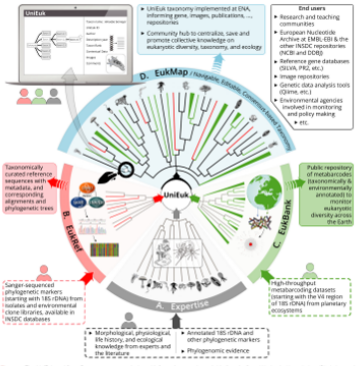


Figure 1 The UniEuk workflow. Bottom-up, community-based information on eukaryotic biodiversity from (A) classical knowledge, (B) phylogenetic diversity, and (C) environmental 'omics' surveys, converge and synergize through the UniEuk modules to inform the navigable and editable, consensus-based taxonomic framework (D). Dotted and colored frames indicate input and output information, respectively. Line drawings of eukaryotes adapted with permission from https://genev.unige.ch/system/pavlovski/tol/tree.png.

Database

- Coordinate with **EukRef**
- Reference sequences
- Chimeras
- Reannotate environmental sequences (Wang/DECIPHER)
- Import more recent GenBank sequences

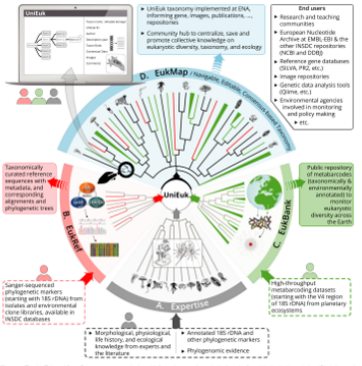


Figure 1 The UniEuk workflow. Bottom-up, community-based information on eukaryotic biodiversity from (A) classical knowledge, (B) phylogenetic diversity, and (C) environmental 'omics' surveys, converge and synergize through the UniEuk modules to inform the navigable and editable, consensus-based taxonomic framework (D). Dotted and colored frames indicate input and output information, respectively. Line drawings of eukaryotes adapted with permission from https://genev.unige.ch/system/pavlovski/tol/tree.png.

Database

- Coordinate with **EukRef**
- Reference sequences
- Chimeras
- Reannotate environmental sequences (Wang/DECIPHER)
- Import more recent GenBank sequences
- Incorporate new metadata types (e.g. mixotrophs)

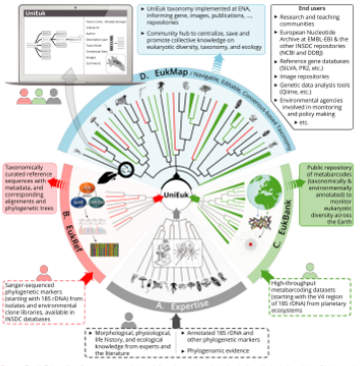


Figure 1 The UniEuk workflow. Bottom-up, community-based information on eukaryotic biodiversity from (A) classical knowledge, (B) phylogenetic diversity, and (C) environmental 'omics' surveys, converge and synergize through the UniEuk modules to inform the navigable and editable, consensus-based taxonomic framework (D). Dotted and colored frames indicate input and output information, respectively. Line drawings of eukaryotes adapted with permission from https://genev.unige.ch/system/pavlovskaia/tree.png.

Database

- Coordinate with **EukRef**
- Reference sequences
- Chimeras
- Reannotate environmental sequences (Wang/DECIPHER)
- Import more recent GenBank sequences
- Incorporate new metadata types (e.g. mixotrophs)
- Incorporate 16S plastid, ITS, SSU

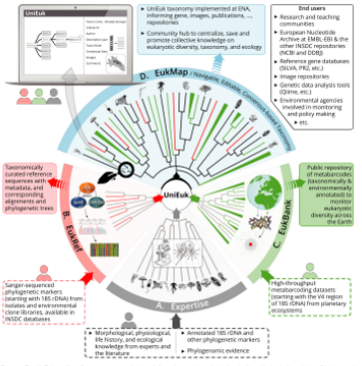


Figure 1 The UniEuk workflow. Bottom-up, community-based information on eukaryotic biodiversity from (A) classical knowledge, (B) phylogenetic diversity, and (C) environmental 'omics' surveys, converge and synergize through the UniEuk modules to inform the navigable and editable, consensus-based taxonomic framework (D). Dotted and colored frames indicate input and output information, respectively. Line drawings of eukaryotes adapted with permission from https://genev.unige.ch/system/pavlovskii/tree.png.

Database

- Coordinate with **EukRef**
- Reference sequences
- Chimeras
- Reannotate environmental sequences (Wang/DECIPHER)
- Import more recent GenBank sequences
- Incorporate new metadata types (e.g. mixotrophs)
- Incorporate 16S plastid, ITS, SSU
- Provide alignments for specific groups

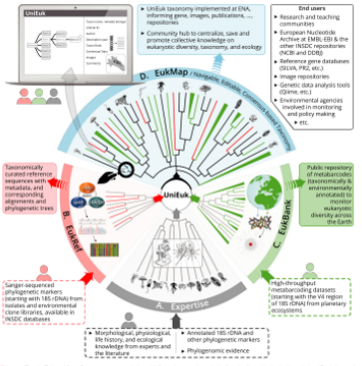


Figure 1 The UniEuk workflow. Bottom-up, community-based information on eukaryotic biodiversity from (A) classical knowledge, (B) phylogenetic diversity, and (C) environmental 'omics' surveys, converge and synergize through the UniEuk modules to inform the navigable and editable, consensus-based taxonomic framework (D). Dotted and colored frames indicate input and output information, respectively. Line drawings of eukaryotes adapted with permission from https://genev.unige.ch/system/pawlowski/files/tree.png.

Web site

In the coming years, we will try to provide users with new functionalities.

However this already can be done easily using R and the **pr2database library**.

- Specific datasets

Web site

In the coming years, we will try to provide users with new functionalities.

However this already can be done easily using R and the **pr2database library**.

- Specific datasets
  - Reference sequences (e.g.for alignements)

Web site

In the coming years, we will try to provide users with new functionalities.

However this already can be done easily using R and the **pr2database library**.

- Specific datasets
  - Reference sequences (e.g.for alignements)
  - Chimeras

Web site

In the coming years, we will try to provide users with new functionalities.

However this already can be done easily using R and the **pr2database library**.

- Specific datasets
  - Reference sequences (e.g.for alignements)
  - Chimeras
  - Taxonomic groups (e.g. diatoms . . .)

Web site

In the coming years, we will try to provide users with new functionalities.

However this already can be done easily using R and the **pr2database library**.

- Specific datasets
  - Reference sequences (e.g.for alignements)
  - Chimeras
  - Taxonomic groups (e.g. diatoms . . .)
- BLAST search

Web site

In the coming years, we will try to provide users with new functionalities.

However this already can be done easily using R and the **pr2database library**.

- Specific datasets
  - Reference sequences (e.g.for alignements)
  - Chimeras
  - Taxonomic groups (e.g. diatoms . . .)
- BLAST search
- Automatic metabarcode annotation using Wang classifier/DECIPHER

Web site

In the coming years, we will try to provide users with new functionalities.

However this already can be done easily using R and the **pr2database library**.

- Specific datasets
  - Reference sequences (e.g.for alignements)
  - Chimeras
  - Taxonomic groups (e.g. diatoms . . .)
- BLAST search
- Automatic metabarcode annotation using Wang classifier/DECIPHER
- Primer and Probe specificity (cf. work with S.Geisen)

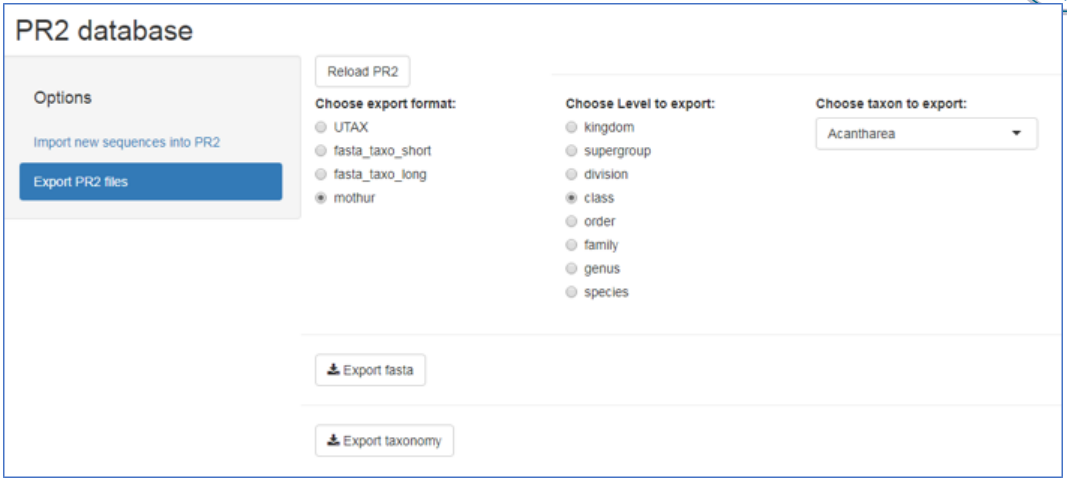
Web site

In the coming years, we will try to provide users with new functionalities.

However this already can be done easily using R and the **pr2database library**.

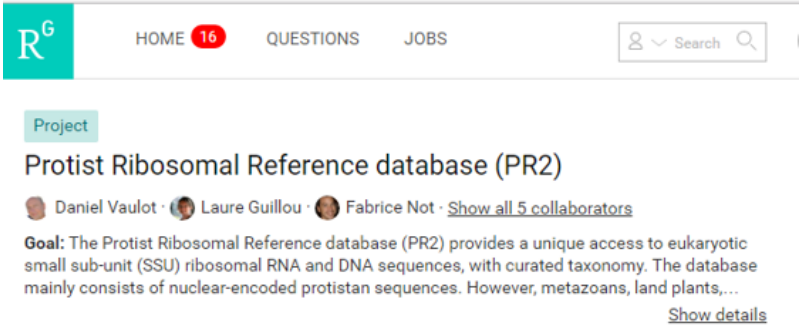
- Specific datasets
  - Reference sequences (e.g.for alignements)
  - Chimeras
  - Taxonomic groups (e.g. diatoms . . .)
- BLAST search
- Automatic metabarcode annotation using Wang classifier/DECIPHER
- Primer and Probe specificity (cf. work with S.Geisen)
- Visualisation of metadata (position . . .)

Web site



Example of interactive download of sequences.

Research Gate



Follow PR^2 on Research Gate

Thank you for your attention