environmental microbiology

Environmental Microbiology (2018) **20**(6), 2218–2230 doi:10.1111/1462-2920.14264

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]

[1]*Department of Botany, University of British Columbia, Vancouver, BC, Canada.*

Departments of [2]Marine Sciences and [3]Ecology and Evolutionary Biology, University of Connecticut, Stamford, CT, USA.

[4]*Key Laboratory of Coastal Biology and Bioresource Utilization, Yantai Institute of Coastal Zone Research, Chinese Academy of Sciences, Yantai, China.*

[5]*School of Science, Mae Fah Luang University, Chiang Rai, Thailand.*

Summary

High-throughput sequencing (HTS) surveys, among the most common approaches currently used in environmental microbiology, require reliable reference databases to be correctly interpreted. The EukRef Initiative (eukref.org) is a community effort to manually screen available small subunit (SSU) rRNA gene sequences and produce a public, high-quality and informative framework of phylogeny-based taxonomic annotations. In the context of EukRef, we present a database for the monophyletic phylum Ciliophora, one of the most complex, diverse and ubiquitous protist groups. We retrieved more than 11 500 sequences of ciliates present in GenBank (28% from identified isolates and 72% from environmental surveys). Our approach included the inference of phylogenetic trees for every ciliate lineage and produced the largest SSU rRNA tree of the phylum Ciliophora to date. We flagged approximately 750 chimeric or low-quality sequences, improved the classification of 70% of GenBank entries and enriched environmental and literature metadata by 30%. The performance of EukRef-Ciliophora is superior to the current SILVA database in classifying HTS reads from a global marine survey. Comprehensive outputs are publicly available to make the new tool a useful guide for non-specialists and a quick reference for experts.

Received 5 February, 2018; revised 13 April, 2018; accepted 27 April, 2018. *For correspondence. E-mail vittorio.boscaro@botany.ubc.ca; Tel. +1 604 822 2845; Fax +1 (604) 822-6089. †Present address: Department of Marine Biology and Oceanography, Institut de Ciències del Mar – CSIC, Spain.

Introduction

Microbial eukaryotic communities are essential components of virtually all ecosystems, from deep sea (López-García *et al.*, 2001; Sauvadet *et al.*, 2010; Scheckenbach *et al.*, 2010; Pernice *et al.*, 2016) to marine coasts (Massana *et al.*, 2015), from freshwater (Šlapeta *et al.*, 2005; Mangot *et al.*, 2013; Simon *et al.*, 2016) and soil (Mahé *et al.*, 2017), to the gut and skin microbiomes of animals (Williams and Coleman, 1992; Wegener Parfrey *et al.*, 2014). Understanding the diversity and structure of these communities is an essential requirement to fully comprehend any major ecological process, including global carbon cycling and food webs, symbiotic relationships and the distribution and spread of parasites and invasive species (Worden *et al.*, 2015). High-throughput sequencing (HTS) surveys based on the small subunit (SSU) rRNA gene are currently the most widely used approach to address these questions, and in the last decade have confirmed that we still know only a fraction of the microbial diversity and biotic interactions on Earth (Sogin *et al.*, 2006; Worden *et al.*, 2015).

Many successes notwithstanding, HTS community characterizations are also prone to several biases. Nucleic acids extraction yields, PCR amplification efficiency and gene copy numbers differ among organisms, and their impacts on environmental studies have been reviewed many times (e.g., Schloss *et al.*, 2011; Fonseca *et al.*, 2012; Esling *et al.*, 2015; Schmidt *et al.*, 2015). Sequencing errors, for a long time mitigated by clustering and consensus methods, are now tackled by more complex error-recognition models (Callahan *et al.*, 2016; Amir *et al.*, 2017). Nonetheless, one crucial component in the HTS pipeline is especially impervious to automated improvements: the reliance on reference databases. The identification of

any environmental sequence is dependent on its closest relatives in extant repositories being themselves correctly annotated. Surprisingly often, this is not the case. New taxa are continuously described, and taxonomic classifications are periodically overhauled. Moreover, public repositories like GenBank contain a large number of misidentified, poorly contextualized or even artefactual sequences. The combined effect of these issues makes it increasingly difficult to maintain reliable reference databases.

The EukRef Initiative (eukref.org; del Campo *et al*., 2018) is a community effort to provide a publicly available reference dataset of eukaryotic SSU rRNA gene sequences. Phylogenetic trees and taxonomic annotations, curated by experts of each taxonomic group, will be freely shared online and integrated with other commonly used databases, such as SILVA (Quast *et al*., 2013) and PR^2 (Guillou *et al*., 2013). In this article, we describe the public release of EukRef-Ciliophora, a database of annotated ciliate sequences and associated outputs (reference phylogenetic trees, alignments and classification framework, available at https://github.com/eukref/curation).

Ciliates (phylum Ciliophora) are among the most well-known, charismatic and ubiquitous protists (Hausmann and Bradbury, 1996; Lynn, 2008). Morphologically complex and comparatively large single-celled eukaryotes, ciliates are major contributors in the trophic networks of aquatic environments (Weisse *et al*., 2016), but are also found in terrestrial environments (Foissner, 1998), and as commensals or parasites of animals, including livestock and humans (Zaman, 1978; Newbold *et al*., 2015). Many recent papers have used this focal group to test hypotheses and detect patterns in HTS surveys (e.g., Bachy *et al*., 2013; Stoeck *et al*., 2014; Forster *et al*., 2015; Gimmler *et al*., 2016; Santoferrara *et al*., 2016; Boscaro *et al*., 2017), and even more studies have evaluated them as part of microbial communities (e.g., Edgcomb *et al*., 2011; Charvet *et al*., 2012; Lie *et al*., 2014; de Vargas *et al*., 2015; Grossmann *et al*., 2016; Hu *et al*., 2016). While different sequencing techniques and analysis pipelines have been tested, taxa identification often relies on outdated and potentially misleading reference databases. Finding and characterizing ciliates is an essential task in microbial ecology, and traditional approaches cannot keep the pace of HTS techniques. However, sequences alone do not carry information in a vacuum, and need solid foundations to be used. With the release of EukRef-Ciliophora, we expect to provide an important tool for many researchers in (and especially out of) the field, making downstream analyses easier, quicker and more reliable.

Results

General characteristics of the database

Groups within the phylum Ciliophora (broadly corresponding to the traditional rank of class or, for spirotrichs, subclass) were assigned to one or two curators (Table 1). Group-level outputs, obtained following the procedures outlined in Fig. 1, were then combined into the final database. The EukRef-Ciliophora database includes more than

Table 1. List of individually curated ciliate groups, showing the number of sequences in the final database, the number of representative sequences used in phylogenetic analyses, the percentage of sequences from the environment versus from isolated organisms and the curators.

Group	Sequences	Representative sequences	Isolates	Environmental	Curator(s)
Karyorelictea	278	45	89%	11%	VB
Heterotrichea	258	37	74%	26%	VB
Protocruzia	8	5	75%	25%	VB
Oligotrichia	1752	183	7%	93%	LS
Choreotrichia	1394	165	22%	78%	LS
Hypotrichia	972	138	35%	66%	QZ & EG
Euplotia	430	83	72%	28%	VB
Other Spirotrichea	6	6	83%	17%	VB
Armophorea	133	39	63%	37%	VB
Litostomatea	1897	228	18%	82%	QZ & EG
Cariacotrichea	346	14	0%	100%	VB & LS
Colpodea	131	34	68%	32%	VB
Oligohymenophorea	2638	354	38%	62%	VB
Nassophorea	310	49	6%	94%	QZ & EG
Phyllopharyngea	379	172	33%	67%	QZ & EG
Prostomatea[a]	458	128	8%	92%	QZ & EG
Plagiopylea	98	40	19%	81%	VB
Mesodinium	134	20	11%	89%	VB & LS
Total	**11622**	**1741**	**28%**	**72%**	

a. And closely associated CONThreeP clades.
VB, Vittorio Boscaro; LS, Luciana Santoferrara; QZ, Qianqian Zhang; EG, Eleni Gentekaki.

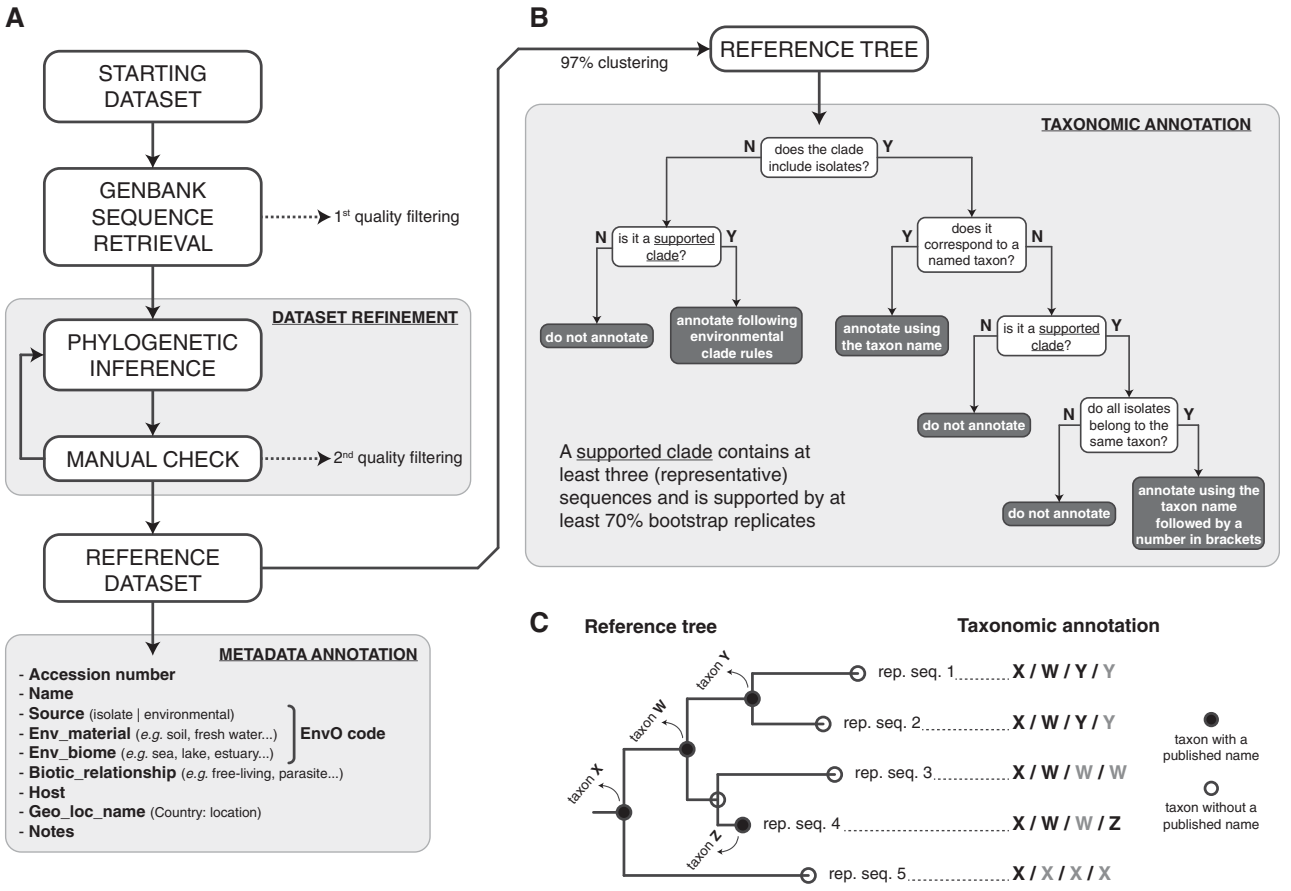

Fig. 1. EukRef-Ciliophora database generation pipeline.
A. Workflow summary: the EukRef pipeline (described in detail on http://eukref.org) was performed on each individually annotated ciliate group. Quality filtering steps, both automatic and manual, were used to produce a complete reference dataset of available ciliate sequences longer than 500 bp, purged of chimeric and other types of low-quality data.
B. Summary of the rules followed to annotate nodes in reference trees.
C. Schematic representation of the correspondence between node-level annotation and final taxonomic string attached to representative sequences. All taxonomic strings in EukRef have the same number of elements; the names of broader taxa are propagated to fill in gaps when needed (shown in grey). Only taxa existing in literature (named) were used.

11 500 annotated SSU rRNA sequences, most of which (72%) are of environmental origin (Table 1; Fig. 2A). Almost half of the sequences (45%) come from marine environments, and about another 30% from freshwater or terrestrial systems (less than 4% were collected in transitional, brackish areas such as estuaries or lagoons). About 13% of the sequences originated from metazoan gut, faecal or skin microbiomes.

Almost 70% of the EukRef-Ciliophora taxonomic annotations have a higher resolution compared with the corresponding GenBank entries (e.g., sequences assigned at the phylum level in GenBank are annotated up to the genus level in our database; Fig. 2B). The annotation of some sequences (0.8%) was corrected, as they were mislabelled in GenBank, either as the wrong ciliate taxa or as belonging to non-ciliate groups (e.g., Dinophyceae, Metazoa, Bacteria). The annotation of environmental and literature metadata also increased the amount of information over that in GenBank (25%–30% of entries are more informative in our database as a result of manually inspecting the original publications). However, some metadata (about 10%) remain unavailable, largely due to entries from unpublished work. In addition to incomplete or incorrect taxonomic labels, poor environmental metadata and outdated or missing literature metadata in GenBank, 748 of the sequences originally retrieved were removed because chimeric or of poor quality (54% discovered by the UCHIME algorithm and 46% manually), which suggests that more than 5% of ciliate sequences deposited in GenBank are methodological artefacts.

Group-level annotation

Sequence alignments, phylogenetic trees and detailed comments on phylogeny and taxonomic annotation of

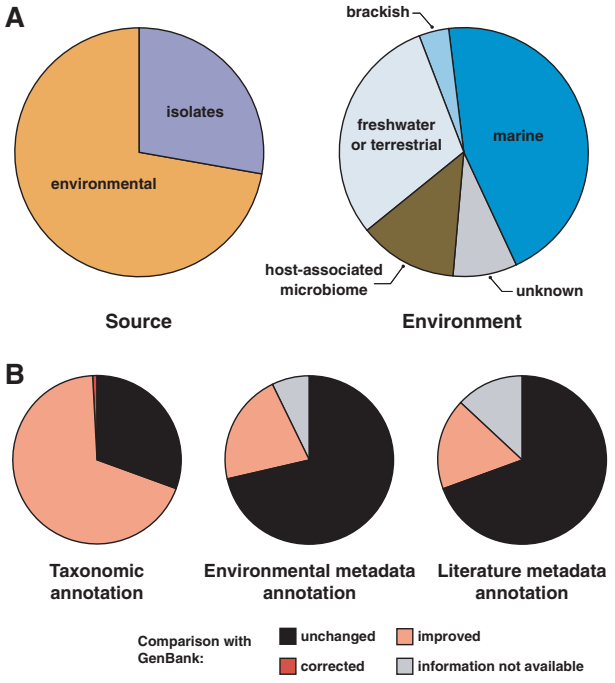

Fig. 2. General features of the EukRef-Ciliophora database (11 622 sequences).
A. Sequence origin in terms of source (environmental vs. from isolated organisms; left) and collection environment (right).
B. Information content in the EukRef database compared with corresponding GenBank entries. From left to right: depth of taxonomic annotation, availability of environmental metadata (calculated on the three main EukRef entries *environmental material*, *environmental biome*, and *geographic location name*), reference papers information ('improved' from GenBank if at least one entry among *publication*, *authors*, and *journal* was added or updated).

each group are publicly available on the EukRef website. Examples of the information content included in EukRef-Ciliophora are shown for the classes Colpodea and Plagiopylea (Fig. 3). In general, the ratio of environmental sequences versus sequences from isolated organisms varied widely among groups. Sequences from isolated and identified organisms were obviously labelled with more detailed classifications in GenBank. As a result, the improvement in the taxonomic annotation provided by EukRef was more pronounced in groups rich in environmental sequences (such as Plagiopylea or Oligotrichia), than it was in groups with many morphologically-identified sequences (such as Colpodea or Karyorelictea). Conversely, environmental sequences tend to have more diligently reported metadata, making the update of information of EukRef-Ciliophora especially noticeable for isolates.

Ciliate taxa with conflicting phylogeny remained intentionally unannotated (see details in the Taxonomic Note files). For example, many of the genera with molecularly characterized representatives were not recovered as monophyletic in our trees. Most cases are due to known artificial lineages or to misidentified sequences. In particular, representatives of many 'common', flagship genera (e.g., *Strombidium* in Oligotrichia, *Vorticella* and *Cyclidium* in Oligohymenophorea, *Amphisiella* and *Oxytricha* in Hypotrichia) are scattered among other genera in the trees. In other cases, problems may be related to the relatively conserved nature of the SSU rRNA gene and the clustering strategy used during the annotation (i.e., sequences belonging to different genera were often found to be mixed in 97%-similarity clusters; see 'Experimental procedures'). Taxa too similar at the molecular level were not annotated separately, as they would be even harder to discriminate using HTS short-sequence data.

Phylum-level phylogenetic tree

The coherence of the overall database was confirmed by phylogenetic inference on the representative SSU rRNA gene sequences, including the newly-annotated environmental sequences. This taxon-rich tree recovered seven of the twelve Ciliophora classes as monophyletic (Fig. 4): Karyorelictea, Heterotrichea, Litostomatea (with the exclusion of *Mesodinium*, *Cyclotrichium*, *Askenasia* and *Paraspathidium*, which have long been known to branch separately), Cariacotrichea, Colpodea, Oligohymenophorea and Plagiopylea. The monophyly of Phyllopharyngea was recovered, but with very low support for the current class definition (i.e., including the synhymeniids, which were formerly assigned to the Nassophorea; Gong *et al*., 2009); the clade of phyllopharyngeans with the exclusion of synhymeniids was fully supported. Other groups are instead split in multiple clades, but without significant statistical support. Most Armophorea sequences clustered apart from caenomorphids in our tree, as reported elsewhere (Vďačný *et al*., 2010; da Silva Paiva *et al*., 2013; Li *et al*., 2017), but this is strongly at odds with morphology (Lynn, 2008). Nassophorea was split in three well-characterized lineages (Nassulida, Microthoracida and Discotrichida), but recent phylogenomic analyses have shown that at least Nassulida and Microthoracida do form a monophyletic group (Lynn *et al*., 2018). Notably, the discotrichid clade includes mostly previously unassigned environmental sequences. The non-monophyly or poorly supported monophyly of Spirotrichea and its subclasses is also dubious, although the divergence of at least *Phacodinium* is compatible with morphology (Lynn, 2008). Phylogenomic analyses have recently confirmed that another atypical former spirotrich, *Protocruzia*, indeed branches separately (Gentekaki *et al*., 2014). Prostomatea represents the most complex situation, as members of this class not only appear as basal (instead of sister) to Plagiopylea but are also entangled with various other *incertae sedis* lineages (including *Cyclotrichium* and *Paraspathidium*).

The genus *Mesodinium*, formally belonging to Litostomatea (Lynn, 2008), is extremely divergent and branches separately from all other ciliates (Johnson *et al*., 2004; Gao *et al*., 2016). For this reason, it was annotated separately (Table 1) and excluded from the main phylogenetic inference (Fig. 4). Running a second tree including *Mesodinium* confirmed previous observations and did not influence the rest of the topology (data not shown).

No major environmental clades were detected in-between the characterized groups. Cariacotrichea (Orsi *et al*., 2012a) seems to be the only large clade of exclusively environmental sequences not associated with any major group of characterized ciliates. Cariacotrichea were originally described from the anoxic Cariaco Basin (Venezuela), but many of the 346 sequences associated to this class in EukRef-Ciliophora come from other anoxic marine sites, including the Saanich Inlet in British Columbia (Orsi *et al*., 2012b), the Hydrate Ridge offshore of Oregon (Pasulka *et al*., 2016) and the Framvaren Fjord in Norway (Behnke *et al*., 2010), as well as the estuarine Great Sippewisset Salt Marsh in Cape Cod (Massachussets) (Stoeck and Epstein, 2003). Large environmental lineages are also present within many traditional classes, especially the Oligohymenophorea.

HTS read analyses

To examine the performance of the new reference database, 73 474 HTS ciliate sequences collected by the *Tara* Oceans expedition (de Vargas *et al*., 2015) were analysed with Naive Bayes classifiers trained on either EukRef-Ciliophora or SILVA (Fig. 5A and B). On a broader level (down to traditional subclasses), the two databases provided similar classifications, largely in accordance with previous analyses (Gimmler *et al*., 2016). The most diverse groups in the sampled marine biomes were oligotrichs, choreotrichs and oligohymenophoreans. Colpodeans, usually reported in freshwater and terrestrial environments, were also confirmed to be quite numerous, although the vast majority of colpodean sequences were assigned by EukRef-Ciliophora to a single genus, *Aristerostoma*, which includes known marine species (Dunthorn *et al*., 2009). In total, 26% of the HTS reads were classified to a higher degree of resolution (i.e., to less inclusive taxa) by EukRef-Ciliophora as compared with SILVA (Fig. 5B). About 18% of the reads were inaccurately assigned by SILVA to non-monophyletic genera intentionally not annotated in our database (see above). Most of the remaining sequences (approximately 43%) were similarly identified by EukRef-Ciliophora and SILVA, and only 0.2% were classified in entirely different groups.

HTS reads were also mapped on the ciliate tree (including *Mesodinium*) using the Evolutionary Placement Algorithm (Supporting Information S4; Berger *et al*., 2011) to manually assess any discrepancies. All sequences with conflicting classifications were confirmed to be correctly placed by the EukRef-trained Naive Bayes classifier. Approximately 2.5% sequences assigned to genera by SILVA but not by EukRef-Ciliophora did cluster outside the boundaries of those genera in the tree (Fig. 5C). The low proportion of HTS reads where the SILVA classification provided greater resolution than EukRef-Ciliophora (10%; Fig. 5B) appear to arise for a variety of reasons. In some cases (e.g., the *Mesodinium* clade), the EPA tree confirmed the more accurate SILVA annotation. But in most cases the discrepancy was due to differing interpretations of taxon boundaries. For example, 5.8% of the HTS reads clustered within a large environmental clade in Oligohymenophorea, named here OLIGO5. This clade is not nested within any of the known oligohymenophorean subclasses, but it is loosely associated with divergent representatives of Scuticociliatia (a lineage which is itself not recovered as monophyletic, see the Taxonomic Note for Oligohymenophorea). The sequences are assigned by the SILVA classifier to Scuticociliatia, probably because in the absence of annotated environmental sequences, the closest references are classified within this group. Until more data are available, the most conservative approach is to use the Oligohymenophorea/OLIGO5 identification provided by the EukRef classifier.

Discussion

EukRef-Ciliophora as a reference tool

More than 11 500 publicly available SSU rRNA gene sequences have been manually screened by ciliate specialists and compiled into a single EukRef-Ciliophora database. The process has confirmed and quantified several problems among the sequences deposited in GenBank, such as: (a) the non-negligible portion of low-quality sequences, especially chimeras; (b) the absence of third-party control on taxonomic classifications, which in a small but relevant fraction of cases are demonstrably wrong; (c) the common lack of basic metadata, which are often present only in the associated publication; (d) the relatively common practice of releasing sequences in the absence of any peer-reviewed associated work, or alternatively to omit updating the literature information once a publication is available (or in the most confusing cases to provide contrasting information in GenBank and published articles). The curation process also confirmed the abundance of ‘flagship’ genera whose sequences are so scattered in the tree that they convey no meaningful information. Taxonomic experts in any particular group are usually aware of these and other related issues, but it is unlikely that researchers interested in broader questions, such as most environmental ecologists, would be able to easily navigate through literature and data in order to decipher such chaotic information. This issue is compounded

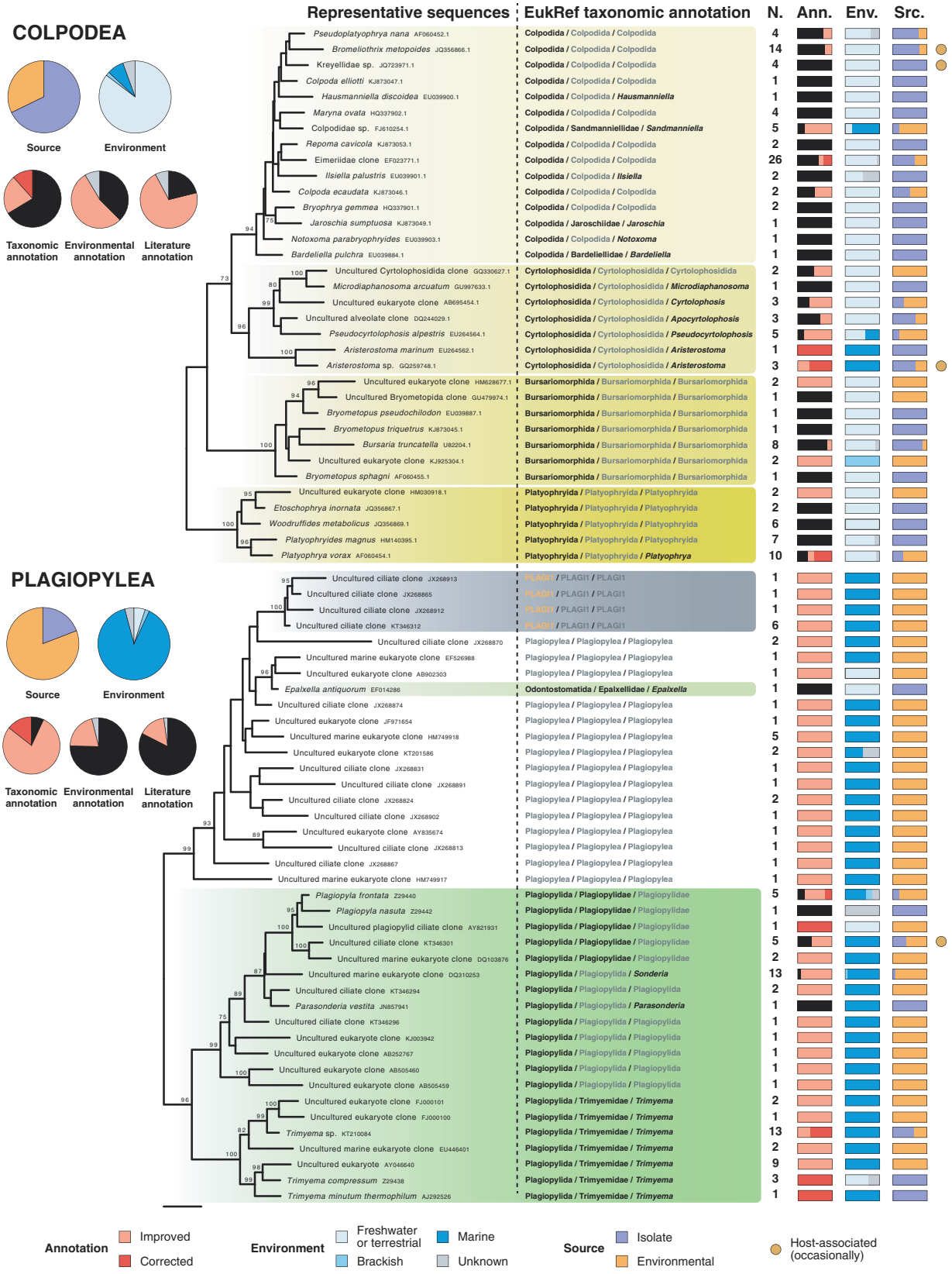
Representative sequences
EukRef taxonomic annotation
N.
Ann.
Env.
Src.
COLPODEA
Source
Environment
Taxonomic annotation
Environmental annotation
Literature annotation
Pseudoplatyophrya nana AF060452.1
Bromeliothrix metopoides JQ356866.1
Kreyellidae sp. JQ723971.1
Colpoda elliotti KJ873047.1
Hausmanniella discoidea EU039900.1
Maryna ovata HQ337902.1
Colpodidae sp. FJ610254.1
Repoma cavicola KJ873053.1
Eimeriidae clone EF023771.1
Ilsiella palustris EU039901.1
Colpoda ecaudata KJ873046.1
Bryophrya gemmea HQ337901.1
Jaroschia sumptuosa KJ873049.1
Notoxoma parabryophryides EU039903.1
Bardeliella pulchra EU039884.1
Uncultured Cyrtolophosidida clone GQ330627.1
Microdiaphanosoma arcuatum GU997633.1
Uncultured eukaryote clone AB695454.1
Uncultured alveolate clone DQ244029.1
Pseudocyrtolophosis alpestris EU264564.1
Aristerostoma marinum EU264562.1
Aristerostoma sp. GQ259748.1
Uncultured eukaryote clone HM628677.1
Uncultured Bryometopida clone GU479974.1
Bryometopus pseudochilodon EU039887.1
Bryometopus triquetrus KJ873045.1
Bursaria truncatella U82204.1
Uncultured eukaryote clone KJ925304.1
Bryometopus sphagni AF060455.1
Uncultured eukaryote clone HM030918.1
Etoschophrya inornata JQ356867.1
Woodruffides metabolicus JQ356869.1
Platyophryides magnus HM140395.1
Platyophrya vorax AF060454.1
Colpodida / Colpodida / Colpodida
Colpodida / Sandmanniellidae / Sandmanniella
Colpodida / Colpodida / Ilsiella
Colpodida / Jaroschiidae / Jaroschia
Colpodida / Colpodida / Notoxoma
Colpodida / Bardeliellidae / Bardeliella
Cyrtolophosidida / Cyrtolophosidida / Cyrtolophosidida
Cyrtolophosidida / Cyrtolophosidida / Microdiaphanosoma
Cyrtolophosidida / Cyrtolophosidida / Cyrtolophosis
Cyrtolophosidida / Cyrtolophosidida / Apocyrtolophosis
Cyrtolophosidida / Cyrtolophosidida / Pseudocyrtolophosis
Cyrtolophosidida / Cyrtolophosidida / Aristerostoma
Bursariomorphida / Bursariomorphida / Bursariomorphida
Platyophryida / Platyophryida / Platyophryida
Platyophryida / Platyophryida / Platyophrya
PLAGIOPYLEA
Source
Environment
Taxonomic annotation
Environmental annotation
Literature annotation
Uncultured ciliate clone JX268913
Uncultured ciliate clone JX268865
Uncultured ciliate clone JX268912
Uncultured ciliate clone KT346312
Uncultured ciliate clone JX268870
Uncultured marine eukaryote clone EF526988
Uncultured eukaryote clone AB902303
Epalxella antiquorum EF014286
Uncultured ciliate clone JX268874
Uncultured eukaryote clone JF971654
Uncultured marine eukaryote clone HM749918
Uncultured eukaryote clone KT201586
Uncultured ciliate clone JX268831
Uncultured ciliate clone JX268891
Uncultured ciliate clone JX268824
Uncultured ciliate clone JX268902
Uncultured eukaryote clone AY835674
Uncultured ciliate clone JX268813
Uncultured ciliate clone JX268867
Uncultured marine eukaryote clone HM749917
Plagiopyla frontata Z29440
Plagiopyla nasuta Z29442
Uncultured plagiopylid ciliate clone AY821931
Uncultured ciliate clone KT346301
Uncultured marine eukaryote clone DQ103876
Uncultured marine eukaryote clone DQ310253
Uncultured ciliate clone KT346294
Parasonderia vestita JN857941
Uncultured ciliate clone KT346296
Uncultured eukaryote clone KJ003942
Uncultured eukaryote clone AB252767
Uncultured eukaryote clone AB505460
Uncultured eukaryote clone AB505459
Uncultured eukaryote clone FJ000101
Uncultured eukaryote clone FJ000100
Trimyema sp. KT210084
Uncultured marine eukaryote clone EU446401
Uncultured eukaryote AY046640
Trimyema compressum Z29438
Trimyema minutum thermophilum AJ292526
PLAGI / PLAGI1 / PLAGI1
Plagiopylea / Plagiopylea / Plagiopylea
Odontostomatida / Epalxellidae / Epalxella
Plagiopylida / Plagiopylidae / Plagiopylidae
Plagiopylida / Plagiopylida / Sonderia
Plagiopylida / Plagiopylida / Parasonderia
Plagiopylida / Plagiopylida / Plagiopylida
Plagiopylida / Trimyemidae / Trimyema
Annotation
Improved
Corrected
Environment
Freshwater or terrestrial
Marine
Brackish
Unknown
Source
Isolate
Environmental
Host-associated (occasionally)

Fig. 3.

by broader and broader analyses, where the group in question is only one of many being investigated.

Over three quarters of the relatively long (>500 bp) SSU rRNA gene sequences of ciliates obtained to date are environmental, and most are deposited with minimal (or incorrect) taxonomic labels. For these sequences, EukRef-Ciliophora provides a huge boost in classification accuracy and depth. For sequences of isolated organisms, the new database reports a variety of metadata missing from GenBank and corrects many mistakes, including species and genera deposited with an incorrect label. In both instances, the new database has culled artefactual entries and provides a phylogeny-based, taxonomically-informative classification. In summary, EukRef-Ciliophora distils the extensive, time-consuming process of 'data-cleaning' (that is usually performed repeatedly and independently by different researchers) into a useful tool to speed up the work of specialists and guide the analyses of non-specialists.

EukRef-Ciliophora as a classification tool for environmental HTS reads

The main goal of EukRef-Ciliophora and the entire EukRef Initiative is to provide a reliable reference database for high-throughput environmental sequencing projects. Using a large set of marine HTS reads as a test, we found that a classifier trained on EukRef-Ciliophora fared better in most respects than an alternative SILVA-trained classifier. Almost half (47%) of the sequences were better annotated (e.g., providing a greater resolution of classification, correcting mistaken attributions or avoiding misleadingly detailed assignments to unreliable taxa) by EukRef-Ciliophora, while 43% of the assignments were identical with this database or SILVA. In addition to a comprehensive sequence database, EukRef-Ciliophora provides curated phylogenetic trees that can be used to map HTS reads less automatically but more accurately than Naive Bayes classifiers.

The key features that make the EukRef-Ciliophora annotations reliable and informative are their foundations in phylogenetic trees, avoiding the naming of uninformative taxa and the inclusion of otherwise poorly-identified environmental sequences. Clearly polyphyletic genera as well as taxa which are too similar at the molecular level to be differentiated by common clustering strategies are a nuisance when classifying HTS reads. While more targeted molecular surveys might be able to discriminate these organisms, and even resolve their relationships, single regions of the SSU rRNA gene do not permit such a fine resolution. Hence, assigning broader taxonomic classifications to such clusters is another way to avoid uncertain attributions. The identification and cataloguing of exclusively environmental clades is also an important factor. We did not find any completely new lineages outside of the traditional classes (either due to a potential limitation in our algorithms or because ciliates have been relatively well-surveyed), but we did detect many environmental clades within characterized ciliate classes and subclasses. An example of why this matters can be seen in a recent report (Pasulka *et al.*, 2016) suggesting the discovery of a novel lineage of environmental sequences, which when re-analysed in the framework provided here, is in fact shown to be nested within the discotrichid clade (Nassophorea).

SSU rRNA-based phylogeny and systematics

Taxonomic changes and classification revisions were completely avoided during the preparation of this database. Nevertheless, the phylogenetic analyses performed are taxon-rich and provide a few new insights into every ciliate group, as detailed in the Taxonomic Notes accompanying the database (https://github.com/eukref/curation). The SSU rRNA tree of the phylum is the most comprehensive to date and depicts the state of knowledge as well as the limits reached by this marker. In particular, it should be noted that the usefulness of the SSU rRNA gene for phylum-level systematics and resolution of deep nodes has probably reached its limit. The current classification of ciliates was heavily influenced by SSU rRNA phylogenies to begin with (Lynn, 2008), and many tenets, such as the monophyly of traditional classes, have been repeatedly tested using other gene loci (e.g., Gao *et al.*, 2016). But many if not all of the remaining problems are unlikely to be solved by a more extensive use of this marker, especially where contrasts between the inferred phylogeny and morphological

Fig. 3. Details of the annotation of two ciliate groups, Colpodea (top) and Plagiopylea (bottom).
Characterization pie charts (i.e., source, environment, and improvement of the EukRef annotation compared with GenBank) are shown. Maximum Likelihood reference trees are built on SSU rRNA representative sequences (defined as the longest sequence of 97%-similarity clusters), but the annotation was performed taking into consideration all sequences. Broader taxa were propagated to fill gaps in the taxonomy (shaded in grey). Bar plots of stats are also associated to each cluster. The class Colpodea exemplifies a group with relatively few environmental sequences and whose classification has been recently revised based on SSU rRNA phylogeny (Foissner *et al.*, 2011), making the annotation of main clades straightforward. The scarcity of annotated low-rank taxa is due to the high sequence similarity between many colpodean genera, often merged in the same cluster. The class Plagiopylea is instead mostly represented by environmental sequences. While the structure of the SSU rRNA tree is generally well-known (Modeo *et al.*, 2013), the annotation of order Odontostomatida is hindered by the existence of a single molecularly investigated species (Stoeck *et al.*, 2007). One annotated environmental clade is also visible on the top of the tree. N., number of sequences per cluster; Ann., improvements in taxonomic annotation; Env., environment; Src., source. Outgroups are not shown. Bootstrap support values 70% or higher are associated to nodes. The black bar, shared by both trees, stands for an inferred evolutionary distance of 0.05.

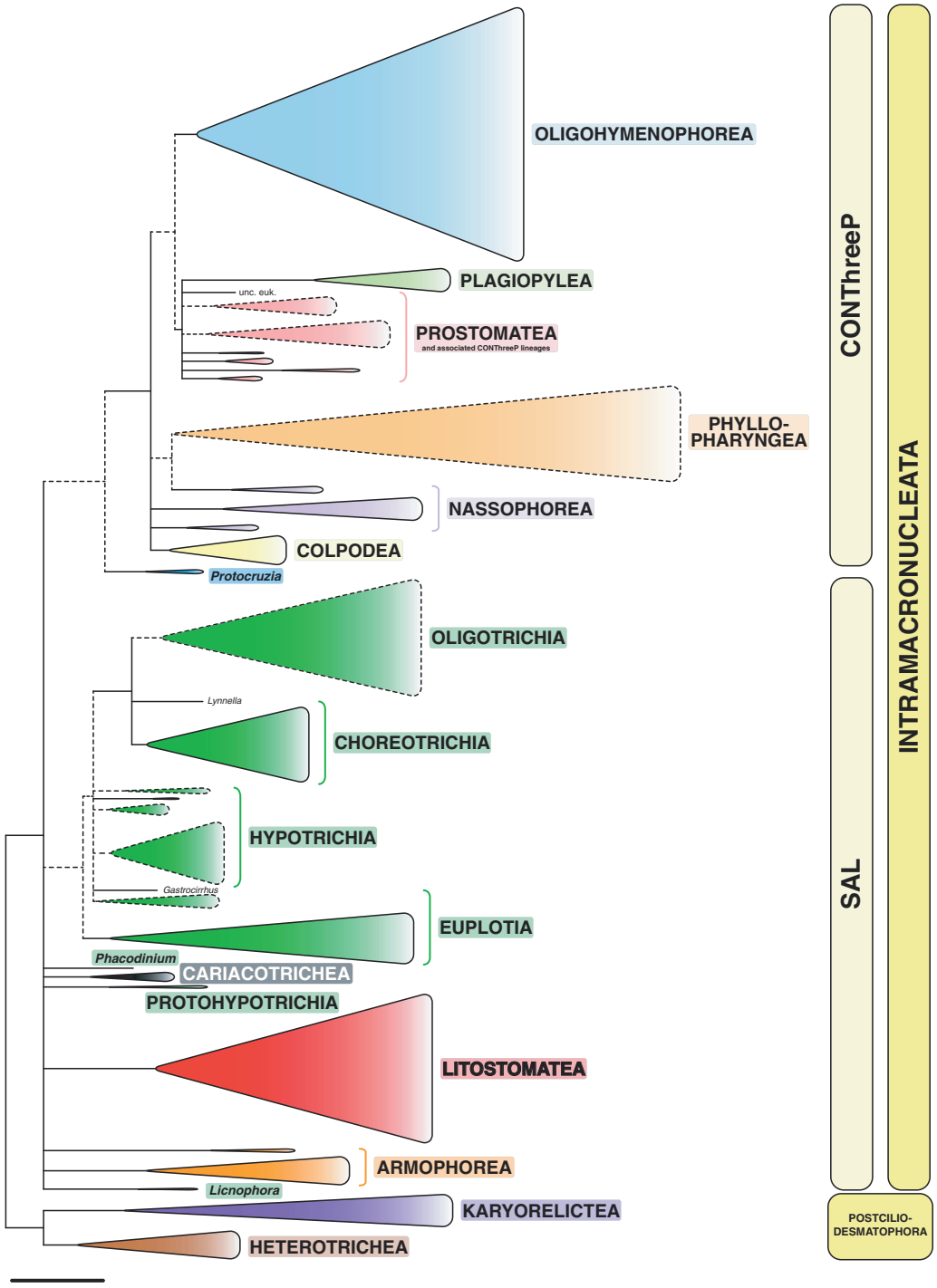

Fig. 4. Maximum Likelihood tree of the phylum Ciliophora (exluding the genus *Mesodinium*) based on the 1246 representative sequences from the EukRef-Ciliophora database longer than 1000 bp.
The root was placed between subphyla Postciliodesmatophora and Intramacronucleata. Individually annotated groups are labelled and highlighted in different colours (with the exception of spirotrich taxa, which are all in green). The recently proposed taxa SAL (Gentekaki *et al.*, 2014; not recovered here as monophyletic) and CONthreeP (Adl *et al.*, 2012) are labelled. Nodes with less than 50% bootstrap support were collapsed into polytomies, and remaining nodes with less than 70% boostrap support are shown in dashed lines. The bar stands for an inferred evolutionary distance of 0.20.

classifications exist (Gentekaki *et al.*, 2017; Lynn and Kolisko, 2017; Lynn *et al.*, 2018). Our phylum-level inference including more than 1200 good-quality representative sequences left many nodes uncertain (Fig. 4), suggesting that any future development will probably come from a phylogenomic approach (Gentekaki *et al.*, 2014). Conversely,

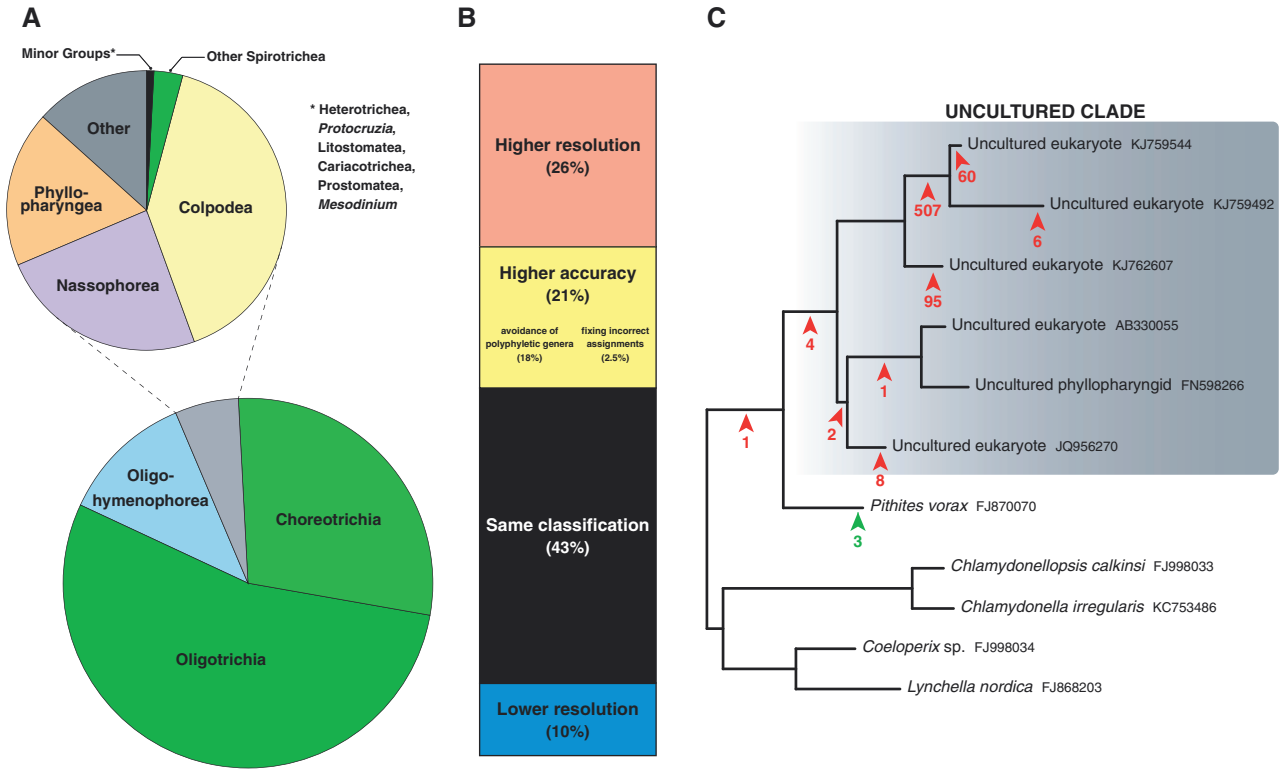

Fig. 5. Classification of 73 474 HTS environmental reads (clustered at 99%) obtained by the *Tara* Oceans survey (de Vargas *et al.*, 2015) and previously classified as ciliates (Gimmler *et al.*, 2016).
A. Pie charts of sequences assigned to different ciliate groups by a EukRef-trained Naive Bayes classifier.
B. Comparison between the taxonomic assignment of the EukRef-trained classifier and a SILVA-trained classifier. EukRef-Ciliophora provided better classifications for almost half of the reads (either due to higher resolution or increased accuracy, for example avoiding the annotation of artificial genera), and virtually identical results to SILVA for most of the remaining reads. Only 10% of the environmental sequences were classified with a lower resolution using EukRef-Ciliophora.
C. Phylogenetic positions, obtained using the Evolutionary Placement Algorithm on the tree of Ciliophora, of 686 HTS reads assigned by the SILVA-trained classifier to the genus *Pithites* (arrowheads shows positions, numbers correspond to the number of reads). Only 3 reads (green) actually cluster with this genus; the remaining 683 (red) are closer to unassigned uncultured sequences, as distant to *Pithites* as *Pithites* is from other described genera. The EukRef-trained Naive Bayes classifier does not commit this error.

SSU rRNA phylogenies are still very useful at lower (less inclusive) taxonomic levels (i.e., within classes). For this reason, the annotations provided here were based on reference trees from each group, and not the full Ciliophora phylogeny. It is essential that traditional systematic efforts continue to provide a solid framework in which the growing body of otherwise meaningless environmental sequences can be linked to taxonomic, ecological and functional categories.

Experimental procedures

Sequence retrieval and curation

Following EukRef guidelines (eukref.org; del Campo *et al.*, 2018; Fig. 1A), a raw SSU rRNA gene sequence dataset for each group was obtained using the eukref_gbretrieve.py script. Briefly, the script requires as input a small but comprehensive set of reliable sequences belonging to the target group. It then performs a cyclical BLAST search against the GenBank database. During each cycle, the best 100 hits are retrieved, subjected to preliminary chimera checking (using the UCHIME software; Edgar *et al.*, 2011) and length filtering (keeping only sequences at least 500 bp long), and then added to the growing dataset, which is then used as input for the subsequent cycle. The process ends when no new sequence shares (a) 80% or higher average identity with the input sequences, and (b) 70% or higher identity with at least one recognized representative of the group (according to the PR^2 database).

The raw dataset was then refined manually. Sequences were aligned with MAFFT (Katoh and Standley, 2013); ambiguously aligned regions were trimmed with trimAl (parameters: -gt 0.3 -st 0.001; Capella-Gutierrez *et al.*, 2009). A phylogenetic tree was inferred with RAxML (Stamatakis, 2014) (100 starting trees; GTRCAT or GTRGAMMA models used for groups with more or less than 100 sequences respectively) and rooted using an appropriate set of outgroups. Tree topologies and the primary sequence structure in the alignments were checked by eye, inspecting suspicious long branches and misaligned sequences. Chimeric sequences identified by

manual BLAST, poor-quality sequences with many ambiguous bases and outlier sequences collected by the script (i.e., those that clustered outside of the target group) were removed. Tree inference and manual inspection were repeated multiple times until no further sequence was discarded.

Taxonomic annotation

Sequences from the refined dataset were clustered with USEARCH (Edgar, 2010) at 97% similarity, a commonly used threshold in ciliates and other protists (e.g., Behnke *et al*., 2011; Stoeck *et al*., 2014; Grossmann *et al*., 2016; Fig. 1B and C). The longest sequence in each cluster was retained as representative. A phylogenetic tree was built on this set plus outgroups (RAxML; 100 starting trees; 100 bootstrap pseudoreplicates; GTRCAT or GTRGAMMA models used for groups with more or less than 100 sequences respectively) and used as guide for taxonomic annotation. Only nodes in this reference tree could be associated with taxa names. The taxonomic annotation of each sequence is the combination of names of all nodes the sequence belongs to (Fig. 1C). If necessary, the names of hierarchically higher (broader) taxa were propagated so that each taxonomic string had the same number of elements (for practical reasons, traditional taxonomic ranks were not used in building the database; del Campo *et al*., 2018).

A taxon could only be associated to a single node, and vice versa. Hence, taxa that were not monophyletic in the reference tree could not be annotated directly. No novel taxon was established or proposed: existing names were used whenever possible, following Lynn (2008), Adl and colleagues (2012) and recent reviews for each group (e.g., Foissner *et al*., 2011; Xu *et al*., 2013; Fan *et al*., 2014; Huang *et al*., 2014; Shazib *et al*., 2014; Zhang *et al*., 2014; Santoferrara *et al*., 2017). Rules have been implemented for EukRef-Ciliophora to produce a taxonomic annotation that is both conservative and information-rich (Fig. 1B): (i) If a node corresponds to an existing taxon, especially one corroborated in recent literature, it is annotated with the name of the taxon; (ii) If a node contains only, but not all, the representatives of an existing taxon (i.e., the taxon is paraphyletic in the tree), it may be annotated with the name of the taxon followed by a number in square brackets if and only if it includes at least three representative sequences and is supported by $>$70% bootstraps; (iii) Environmental-only clades may be annotated using an alphanumerical code if they meet the same criteria (i.e., three or more representative sequences, $>$70% bootstrap support); (iv) No annotation is applied at or below the species level. These guidelines could be overlooked only if the result was a broader annotation (e.g., some clades were not annotated despite meeting the criteria, if considered non-informative or unreliable by the curator). Even if the annotation was based on the representative sequences included in the reference tree, all sequences in every cluster were taken into consideration. Taxonomic annotations were expanded to all the sequences included in a final, tabular database.

Metadata annotation

Each entry in the final database was associated with metadata. GenBank's *accession numbers* were used as unique identifiers, and the deposited *name* of each sequence was also recorded. Environmental metadata included *source* (isolated organism vs. environmental sequences), *environmental material* (the material from which the sample came from, e.g., soil, freshwater...), *environmental biome* (the biome the sample was taken from, e.g., lake, hydrothermal vent, rhizosphere...), *biotic relationship* (free living, parasite, commensal...), *host* (if applicable) and *geographic location name* [in the format (Country or Ocean: location)]. Entries in the *environmental material* and *environmental biome* columns are labelled, whenever possible, using terms and numerical identifiers according to the Environment Ontology (EnvO) code (Buttigieg *et al*., 2013). Information used to fill metadata columns was compiled by consulting both GenBank entries and approximately 750 associated papers. Literature metadata (*publication*, *authors* and *journal*) were also updated manually. A *note* column was used for any other relevant information (e.g., to highlight discrepancies between the information deposited in GenBank and in the corresponding publication).

Phylum-level analyses

To confirm that no major environmental clade outside of the traditional ciliate classes was missed by the group-level curation, sequences from all groups were combined and used as input for the eukref_gbretrieve.py script, in this case targeting the whole phylum. A phylogenetic tree of Ciliophora was also inferred using all representative sequences longer than 1000 bp (RAxML; 100 starting trees; 100 bootstrap pseudoreplicates; GTRCAT model). The phylogenetic analysis was performed both with and without the extremely divergent genus *Mesodinium*.

Database testing

The final EukRef-Ciliophora database was tested and compared with the SILVA reference database using HTS reads (V9 region of the SSU rRNA gene) collected by the *Tara* Oceans survey (de Vargas *et al*., 2015). Reads identified as ciliates by Gimmler and colleagues (2016) were clustered at 99% similarity, then classified using the platform QIIME 2 v2017.10 (https://qiime2.org; Caporaso *et al*., 2010). A Naive Bayes classifier was trained on the 97% clustered SILVA reference sequences (release 128), trimmed to the V9 region following the protocol suggested by Werner and colleagues (2012) and the SILVA 'majority' classification framework (all levels). Sequences confirmed to belong to Ciliophora by this analysis were then classified with the same pipeline but using as references the representative sequences and corresponding annotations of the EukRef-Ciliophora database, formatted as in SILVA (Supporting Information S1–2; a brief guide on its usage is presented in Supporting Information S3). To judge any discrepancy between the two outputs, *Tara* Oceans reads were also mapped on our phylogenetic tree of Ciliophora (see above): the reads were aligned using the global ciliate alignment as reference and trimmed in QIIME (using the align_seqs.py script and the filter_alignment.py script respectively), then added to the tree using the Evolutionary Placement Algorithm (EPA) of RAxML (parameters: -f v -G 0.2 –m GTRCAT).

Acknowledgements

Dr. Denis Lynn (University of British Columbia) is gratefully acknowledged for his support, helpful suggestions to the text and for assistance in the annotation of peritrichs. We thank Dr. Martin Kolisko (Czech Academy of Sciences) for technical assistance with the eukref_gbretrieve.py script. Members of the International Research Coordination Network for Biodiversity of Ciliates are acknowledged for useful discussions during the network meeting held in Washington DC on September 2016. QZ thanks to Dr. Jun Gong (Sun Yat-sen University) for assistance in chimera checking and annotations. This work was supported by a grant (RGPIN-2014-03994) from the Natural Sciences and Engineering Research Council of Canada to PJK. VB and JdC were supported by grants to the Centre for Microbial Diversity and Evolution from the Tula Foundation. LFS was supported by grant OCE1435515 from the U.S. National Science Foundation to Dr. George McManus (University of Connecticut). QZ was supported by the grant 31672251 from the National Natural Science Foundation of China. JdC was supported by a Marie Curie International Outgoing Fellowship grant (FP7-PEOPLE-2012-IOF-331450 CAARL).

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

References

Adl, S.A., Simpson, A.G.B., Lane, C.E., Lukeš, J., Bass, D., Bowser, S.S. *et al.* (2012) The revised classification of eukaryotes. *J Eukaryot Microbiol* **59:** 429–493.

Amir, A., McDonald, D., Navas-Molina, J.A., Kopylova, E., Morton, J.T., Xu, Z.Z. *et al.* (2017) Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems* **2:** e00191-16.

Bachy, C., Dolan, J.R., López-García, P., Deschamps, P., and Moreira, D. (2013) Accuracy of protist diversity assessments: morphology compared with cloning and direct pyrosequencing of 18S rRNA genes and ITS regions using the conspicuous tintinnid ciliates as a case study. *ISME J* **7:** 244–255.

Behnke, A., Barger, K.J., Bunge, J., and Stoeck, T. (2010) Spatio-temporal variations in protistan communities along an O_2/H_2S gradient in the anoxic Framvaren Fjord (Norway). *FEMS Microbiol Ecol* **72:** 89–102.

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.

Berger, S.A., Krompass, D., and Stamatakis, A. (2011) Performance, accuracy, and web server for evolutionary placement of short sequence reads under Maximum Likelihood. *Syst Biol* **60:** 291–302.

Boscaro, V., Rossi, A., Vannini, C., Verni, F., Fokin, S.I., and Petroni, G. (2017) Strengths and biases of high-throughput sequencing data in the characterization of freshwater ciliate microbiomes. *Microb Ecol* **73:** 865–875.

Buttigieg, P.L., Morrison, N., Smith, B., Mungall, C.J., and Lewis, S.E.; the ENVO Consortium. (2013) The environment ontology: contextualising biological and biomedical entities. *J Biomed Semantics* **4:** 43.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., and Holmes, S.P. (2016) DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods* **13:** 581–583.

Capella-Gutierrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25:** 1972–1973.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., *et al.* (2010) QIIME allows analysis of high-throughput community sequence data. *Nat Methods* **7:** 335–336.

Charvet, S., Vincent, W.F., Comeau, A., and Lovejoy, C. (2012) Pyrosequencing analysis of the protist communities in a High Arctic meromictic lake: DNA preservation and change. *Front Microbiol* **3:** 422.

da Silva Paiva, T., do Nascimento Borges, B., and da Silva-Neto, I.D. (2013) Phylogenetic study of Class Armophorea (Alveolata, Ciliophora) based on 18S-rDNA data. *Genet Mol Biol* **36:** 571–585.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., *et al.* (2015) Eukaryotic plankton diversity in the sunlit ocean. *Science* **348:** 1261605.

del Campo, J., Kolisko, M., Boscaro, V., Santoferrara, L.F., Massana, R., Guillou, L., *et al.* (2018) EukRef: phylogenetic curation of ribosomal RNA to enhance understanding of eukaryotic diversity and distribution. *bioRxiv.* doi:10.1101/278085

Dunthorn, M., Eppinger, M., Schwarz, M.V.J., Schweikert, M., Boenigk, J., Katz, L.A., and Stoeck, T. (2009) Phylogenetic placement of the Cyrtolophosididae Stokes, 1888 (Ciliophora; Colpodea) and neotypification of *Aristerostoma marinum* Kahl, 1931. *Int J Syst Evol Microbiol* **59:** 167–180.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26:** 2460–2461.

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., and Knight, R. (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27:** 2194–2200.

Edgcomb, V., Orsi, W., Bunge, J., Jeon, S., Christen, R., Leslin, C., *et al.* (2011) Protistan microbial observatory in the Cariaco Basin, Caribbean. I. Pyrosequencing vs Sanger insights into species richness. *ISME J* **5:** 1344–1356.

Esling, P., Lejzerowicz, F., and Pawlowski, J. (2015) Accurate multiplexing and filtering for high-throughput amplicon-sequencing. *Nucleic Acids Res* **43:** 2513–2524.

Fan, X., Pan, H., Li, L., Jiang, J., Al-Rasheid, K.A.S., and Gu, F. (2014) Phylogeny of the poorly known ciliates, Microthoracida, a systematically confused taxon (Ciliophora), with morphological reports of three species. *J Eukaryot Microbiol* **61:** 227–237.

Foissner, W. (1998) An updated compilation of world soil ciliates (Protozoa, Ciliophora), with ecological notes, new records, and descriptions of new species. *Eur J Protistol* **34:** 195–235.

Foissner, W., Stoeck, T., Agatha, S., and Dunthorn, M. (2011) Intraclass evolution and classification of the Colpodea (Ciliophora). *J Eukaryot Microbiol* **58:** 397–415.

Fonseca, V.G., Nichols, B., Lallias, D., Quince, C., Carvalho, G.R., Power, D.M., and Creer, S. (2012) Sample richness

and genetic diversity as drivers of chimera formation in nSSU metagenetic analyses. *Nucleic Acids Res* **40:** e66.

Forster, D., Bittner, L., Karkar, S., Dunthorn, M., Romac, S., Audic, S., *et al*. (2015) Testing ecological theories with sequence similarity networks: marine ciliates exhibit similar geographic dispersal patterns as multicellular organisms. *BMC Biol* **13:** 16.

Gao, F., Warren, A., Zhang, Q., Gong, J., Miao, M., Sun, P., *et al*. (2016) The all-data-based evolutionary hypothesis of ciliated protists with a revised classification of the Phylum Ciliophora (Eukaryota, Alveolata). *Sci Rep* **6:** 24874.

Gentekaki, E., Kolisko, M., Boscaro, V., Bright, K.J., Dini, F., Di Giuseppe, G., *et al*. (2014) Large-scale phylogenomic analysis reveals the phylogenetic position of the problematic taxon *Protocruzia* and unravels the deep phylogenetic affinities of the ciliate lineage. *Mol Phylogenet Evol* **78:** 36–42.

Gentekaki, E., Kolisko, M., Gong, Y., and Lynn, D.H. (2017) Phylogenomics solves a long-standing evolutionary puzzle in the ciliate world: the subclass Peritrichia is monophyletic. *Mol Phylogenet Evol* **106:** 1–5.

Gimmler, A., Korn, R., de Vargas, C., Audic, S., and Stoeck, T. (2016) The *Tara* Oceans voyage reveals global diversity and distribution patterns of marine planktonic ciliates. *Sci Rep* **6:** 33555.

Gong, J., Stoeck, T., Yi, Z., Miao, M., Zhang, Q., Roberts, D.M.L., *et al*. (2009) Small subunit rRNA phylogenies show that the Class Nassophorea is not monophyletic (Phylum Ciliophora). *J Eukaryot Microbiol* **56:** 339–489.

Grossmann, L., Jensen, M., Heider, D., Jost, S., Glücksman, E., Hartikainen, H., *et al*. (2016) Protistan community analysis: key findings of a large-scale molecular sampling. *ISME J* **10:** 2269–2279.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., *et al*. (2013) The Protist Ribosomal Reference database (PR^2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Res* **41:** D597–D604.

Hausmann, K., and Bradbury, P.C. (1996) *Ciliates: Cells as Organisms*. Stuttgart, Germany: Gustav Fischer Verlag.

Hu, S.K., Campbell, V., Connell, P., Gellene, A.G., Liu, Z., Terrado, R., and Caron, D.A. (2016) Protistan diversity and activity inferred from RNA and DNA at a coastal ocean site in the eastern North Pacific. *FEMS Microbiol Ecol* **92:** fiw050.

Huang, J., Chen, Z., Song, W., and Berger, H. (2014) Three-gene based phylogeny of the Urostyloidea (Protista, Ciliophora, Hypotrichia), with notes on classification of some core taxa. *Mol Phylogenet Evol* **70:** 337–347.

Johnson, M.D., Tengs, T., Oldach, D.W., Delwiche, C.F., and Stoecker, D.K. (2004) Highly divergent SSU rRNA genes found in the marine ciliates *Myrionecta rubra* and *Mesodinium pulex*. *Protist* **155:** 347–359.

Katoh, K., and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30:** 772–780.

Li, S., Bourland, W.A., Al-Farraj, S.A., Li, L., and Hu, X. (2017) Description of two species of caenomorphid ciliates (Ciliophora, Armophorea): morphology and molecular phylogeny. *Eur J Protistol* **61:** 29–40.

Lie, A.A.Y., Liu, Z., Hu, S.K., Jones, A.C., Kim, D.Y., Countway, P.D., *et al*. (2014) Investigating microbial eukaryotic diversity from a global census: insights from a comparison of pyrotag and full-length sequences of 18S rRNA genes. *Appl Environ Microbiol* **80:** 4363–4373.

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.

Lynn, D.H. (2008) *The Ciliated Protozoa. Characterization, Classification, and Guide to the Literature* (3rd ed.). Dordrecht, Netherlands: Springer.

Lynn, D.H., and Kolisko, M. (2017) Molecules illuminate morphology: phylogenomics confirms convergent evolution among 'oligotrichous' ciliates. *Int J Syst Evol Microbiol* **67:** 3676–3682.

Lynn, D.H., Kolisko, M., and Bourland, W. (2018) Phylogenomic analysis of *Nassula variabilis* n. sp., *Furgasonia blochmanni*, and *Pseudomicrothorax dubius* confirms a nassophorean clade. *Protist* **169**: 180–189.

Mahé, F., de Vargas, C., Bass, D., Czech, L., Stamatakis, A., Lara, E., *et al*. (2017) Parasites dominate hyperdiverse soil protist communities in Neotropical rainforests. *Nat Ecol Evol* **1:** 0091.

Mangot, J.-F., Domaizon, I., Taib, N., Marouni, N., Duffaud, E., Bronner, G., and Debroas, D. (2013) Short-term dynamics of diversity patterns: evidence of continual reassembly within lacustrine small eukaryotes. *Environ Microbiol* **15:** 1745–1758.

Massana, R., Gobet, A., Audic, S., Bass, D., Bittner, L., Boutte, C., *et al*. (2015) Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. *Environ Microbiol* **17:** 4035–4049.

Modeo, L., Fokin, S.I., Boscaro, V., Andreoli, I., Ferrantini, F., Rosati, G., *et al*. (2013) Morphology, ultrastructure, and molecular phylogeny of the ciliate *Sonderia vorax* with insights into the systematics of order Plagiopylida. *BMC Microbiol* **13:** 40.

Newbold, C.J., de la Fuente, G., Belanche, A., Ramos-Morales, E., and McEwan, N.R. (2015) The role of ciliate protozoa in the rumen. *Front Microbiol* **6:** 1313.

Orsi, W., Edgcomb, V., Faria, J., Foissner, W., Fowle, W.H., Hohmann, T., *et al*. (2012) Class Cariacotrichea, a novel ciliate taxon from the anoxic Cariaco Basin, Venezuela. *Int J Syst Evol Microbiol* **62:** 1425–1433.

Orsi, W., Song, Y.C., Hallam, S., and Edgcomb, V. (2012) Effect of oxygen minimum zone formation on communities of marine protists. *ISME J* **6:** 1586–1601.

Pasulka, A.L., Levin, L.A., Steele, J.A., Case, D.H., Landry, M.R., and Orphan, V.J. (2016) Microbial eukaryotic distributions and diversity patterns in a deep-sea methane seep ecosystem. *Environ Microbiol* **18:** 3022–3043.

Pernice, M.C., Giner, C.R., Logares, R., Perera-Bel, J., Acinas, S.G., Duarte, C.M., *et al*. (2016) Large variability of bathypelagic microbial eukaryotic communities across the world's oceans. *ISME J* **10:** 945–958.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F.O. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41:** D590–D596.

Santoferrara, L.F., Grattepanche, J.-D., Katz, L.A., and McManus, G.B. (2016) Patterns and processes in microbial biogeography: do molecules and morphologies give the same answers? *ISME J* **10:** 1779–1790.

Santoferrara, L.F., Alder, V.V., and McManus, G.B. (2017) Phylogeny, classification and diversity of Choreotrichia and Oligotrichia (Ciliophora, Spirotrichea). *Mol Phylogenet Evol* **112:** 12–22.

Sauvadet, A.L., Gobet, A., and Guillou, L. (2010) Comparative analysis between protist communities from the deep-sea pelagic ecosystem and specific deep hydrothermal habitats. *Environ Microbiol* **12:** 2946–2964.

Scheckenbach, F., Hausmann, K., Wylezich, C., Weitere, M., and Arndt, H. (2010) Large-scale patterns in biodiversity of microbial eukaryotes from the abyssal sea floor. *Proc Natl Acad Sci USA* **107:** 115–120.

Schloss, P.D., Gevers, D., and Westcott, S.L. (2011) Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One* **6:** e27310.

Schmidt, T.S.B., Rodrigues, J.F.M., and von Mering, C. (2015) Limits to robustness and reproducibility in the demarcation of operational taxonomic units. *Environ Microbiol* **17:** 1689–1706.

Shazib, S.U.A., Vd'ačny, P., Kim, J.H., Jang, S.W., and Shin, M.K. (2014) Phylogenetic relationships of the ciliate class Heterotrichea (Protista, Ciliophora, Postciliodesmatophora) inferred from multiple molecular markers and multifaceted analysis strategy. *Mol Phylogenet Evol* **78:** 118–135.

Simon, M., López-García, P., Deschamps, P., Restoux, G., Bertolino, P., Moreira, D., and Jardillier, L. (2016) Resilience of freshwater communities of small microbial eukaryotes undergoing severe drought events. *Front Microbiol* **7:** 812.

Šlapeta, J., Moreira, D., and López, G. (2005) The extent of protist diversity: insights from molecular ecology of freshwater eukaryotes. *Proc Biol Sci* **272:** 2073–2081.

Sogin, M.L., Morrison, H.G., Huber, J.A., Welch, M.D., Huse, S.M., Neal, P.R., Arrieta, J.M., and Herndl, G.J. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere." *Proc Natl Acad Sci USA* **103:** 12115–12120.

Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30:** 1312–1313.

Stoeck, T., and Epstein, S. (2003) Novel eukaryotic lineages inferred from small-subunit rRNA analyses of oxygen-depleted marine environments. *Appl Environ Microbiol* **69:** 2657–2663.

Stoeck, T., Foissner, W., and Lynn, D.H. (2007) Small-subunit rRNA phylogenies suggest that *Epalxella antiquorum* (Penard, 1922) Corliss, 1960 (Ciliophora, Odontostomatida) is a member of the Plagiopylea. *J Eukaryot Microbiol* **54:** 436–442.

Stoeck, T., Breiner, H.-W., Filker, S., Ostermaier, V., Kammerlander, B., and Sonntag, B. (2014) A morphogenetic survey on ciliate plankton from a mountain lake pinpoints the necessity of lineage-specific barcode markers in microbial ecology. *Environ Microbiol* **16:** 430–444.

Vďačný, P., Orsi, W., and Foissner, W. (2010) Molecular and morphological evidence for a sister group relationship of the classes Armophorea and Litostomatea (Ciliophora, Intramacronucleata, Lamellicorticata infraphyl. nov.), with an account on basal litostomateans. *Eur J Protistol* **46:** 298–309.

Wegener Parfrey, L., Walters, W.A., Lauber, C.L., Clemente, J.C., Berg-Lyons, D., Teiling, C., *et al.* (2014) Communities of microbial eukaryotes in the mammalian gut within the context of environmental eukaryotic diversity. *Front Microbiol* **5:** 298.

Weisse, T., Anderson, R., Arndt, H., Calbet, A., Hansen, P.J., and Montagnes, D.J.S. (2016) Functional ecology of aquatic phagotrophic protists – Concepts, limitations, and perspectives. *Eur J Protistol* **55:** 50–74.

Werner, J.J., Koren, O., Hugenholtz, P., DeSantis, T.Z., Walters, W.A., Caporaso, J.G., *et al.* (2012) Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys. *ISME J* **6:** 94–103.

Williams, A.G., and Coleman, G.S. (1992) *The Rumen Protozoa*. New York, NY: Springer-Verlag.

Worden, A.Z., Follows, M.J., Giovannoni, S.J., Wilken, S., Zimmerman, A.E., and Keeling, P.J. (2015) Rethinking the marine carbon cycle: factoring in the multifarious lifestyles of microbes. *Science* **347:** 1257594.

Xu, Y., Li, J., Song, W., and Warren, A. (2013) Phylogeny and establishment of a new ciliate family, Wilbertomorphidae fam. nov. (Ciliophora, Karyorelictea), a highly specialized taxon represented by *Wilbertomorpha colpoda* gen. nov., spec. nov. *J Eukaryot Microbiol* **60:** 480–489.

Zaman, V., (1978) *Balantidium coli*. In *Parasitic Protozoa*. Kreyer, P. (ed). New York, NY: Academic Press, Vol. **2**, pp. 633–653.

Zhang, Q., Yi, Z., Fan, X., Warren, A., Gong, J., and Song, W. (2014) Further insights into the phylogeny of two ciliate classes Nassophorea and Prostomatea (Protista, Ciliophora). *Mol Phylogenet Evol* **70:** 162–170.

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

S1. EukRef-Ciliophora representative sequences.

S2. EukRef-CIliophora taxonomic annotation of representative sequences.

S3. EukRef-Ciliophora usage guidelines.

S4. EPA tree.