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 Campo1† and Patrick J. Keeling1
1Department of Botany, University of British Columbia, Vancouver, BC, Canada.
Departments of 2Marine Sciences and 3Ecology and Evolutionary Biology, University of Connecticut, Stamford, CT, USA.
4Key Laboratory of Coastal Biology and Bioresource Utilization, Yantai Institute of Coastal Zone Research, Chinese Academy of Sciences, Yantai, China.
5School 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.

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 (Slapeta 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 PR² (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.

<table>
<thead>
<tr>
<th>Group</th>
<th>Sequences</th>
<th>Representative sequences</th>
<th>Isolates</th>
<th>Environmental</th>
<th>Curator(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Karyorelictea</td>
<td>278</td>
<td>45</td>
<td>89%</td>
<td>11%</td>
<td>VB</td>
</tr>
<tr>
<td>Heterotrichea</td>
<td>258</td>
<td>37</td>
<td>74%</td>
<td>26%</td>
<td>VB</td>
</tr>
<tr>
<td>Protrucxia</td>
<td>8</td>
<td>5</td>
<td>75%</td>
<td>25%</td>
<td>VB</td>
</tr>
<tr>
<td>Oligotrichia</td>
<td>1752</td>
<td>183</td>
<td>7%</td>
<td>93%</td>
<td>LS</td>
</tr>
<tr>
<td>Choreotrichia</td>
<td>1394</td>
<td>165</td>
<td>22%</td>
<td>78%</td>
<td>LS</td>
</tr>
<tr>
<td>Hypotrichia</td>
<td>972</td>
<td>138</td>
<td>35%</td>
<td>66%</td>
<td>QZ &amp; EG</td>
</tr>
<tr>
<td>Euplotia</td>
<td>430</td>
<td>83</td>
<td>72%</td>
<td>28%</td>
<td>VB</td>
</tr>
<tr>
<td>Other Spirotrichea</td>
<td>6</td>
<td>6</td>
<td>83%</td>
<td>17%</td>
<td>VB</td>
</tr>
<tr>
<td>Armophorea</td>
<td>133</td>
<td>39</td>
<td>63%</td>
<td>37%</td>
<td>VB</td>
</tr>
<tr>
<td>Litostomathea</td>
<td>1897</td>
<td>228</td>
<td>18%</td>
<td>82%</td>
<td>QZ &amp; EG</td>
</tr>
<tr>
<td>Cariacotrichia</td>
<td>346</td>
<td>14</td>
<td>0%</td>
<td>100%</td>
<td>VB &amp; LS</td>
</tr>
<tr>
<td>Colpodea</td>
<td>131</td>
<td>34</td>
<td>68%</td>
<td>32%</td>
<td>VB</td>
</tr>
<tr>
<td>Oligohymenophorea</td>
<td>2638</td>
<td>354</td>
<td>38%</td>
<td>62%</td>
<td>VB</td>
</tr>
<tr>
<td>Nassophorea</td>
<td>310</td>
<td>49</td>
<td>6%</td>
<td>94%</td>
<td>QZ &amp; EG</td>
</tr>
<tr>
<td>Phyllopharyngea</td>
<td>379</td>
<td>172</td>
<td>33%</td>
<td>67%</td>
<td>QZ &amp; EG</td>
</tr>
<tr>
<td>Prostomatea²</td>
<td>458</td>
<td>128</td>
<td>8%</td>
<td>92%</td>
<td>QZ &amp; EG</td>
</tr>
<tr>
<td>Plagiopylea</td>
<td>98</td>
<td>40</td>
<td>19%</td>
<td>81%</td>
<td>VB</td>
</tr>
<tr>
<td>Mesodinium</td>
<td>134</td>
<td>20</td>
<td>11%</td>
<td>89%</td>
<td>VB &amp; LS</td>
</tr>
<tr>
<td>Total</td>
<td>11622</td>
<td>1741</td>
<td>28%</td>
<td>72%</td>
<td></td>
</tr>
</tbody>
</table>

a. And closely associated CONThreeP clades.

VB, Vittorio Boscaro; LS, Luciana Santoferrara; QZ, Qianqian Zhang; EG, Eleni Gentekaki.

© 2018 Society for Applied Microbiology and John Wiley & Sons Ltd, *Environmental Microbiology*, 20, 2218–2230
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
Ciliate taxa with conflicting phylogeny remained 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 identified in-between the characterized groups. Cariacotricha (Orsi et al., 2012a) seems to be the only large clade of exclusively environmental sequences not associated with any major group of characterized ciliates. Cariacotricha 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 (Massachusetts) (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 rDNA 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...
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 distills 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 reads 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.

© 2018 Society for Applied Microbiology and John Wiley & Sons Ltd, Environmental Microbiology, 20, 2218–2230
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,
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 PR² 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 taxon 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; Shaizib 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 alphaneumeral 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 (Bottigieg 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 Glimmer 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_seqscs.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 perirhicks. 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


**EukRef-Ciliophora, a SSU rRNA database** 2229


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.