FEMS Microbiology Ecology, 93, 2017, fix023

doi: 10.1093/femsec/fix023
Advance Access Publication Date: 1 March 2017
Research Article

RESEARCH ARTICLE

# Overview of freshwater microbial eukaryotes diversity: a first analysis of publicly available metabarcoding data

Didier Debroas[1,*], Isabelle Domaizon[2], Jean-Francois Humbert[3], Ludwig Jardillier[4], Cécile Lepère[1], Anne Oudart[1] and Najwa Taïb[1]

[1]Université Clermont Auvergne, CNRS, Laboratoire: Microorganismes: Génome et Environnement, F-63000 Clermont-Ferrand, France, [2]INRA, UMR 42 Centre Alpin de Recherche sur les Réseaux Trophiques et Ecosystèmes Limniques, F-74200 Thonon Les Bains, France, [3]INRA, UMR BIOEMCO, Site de l'ENS, 46 rue d'Ulm, 75005 Paris, France and [4]Unité d'Ecologie, Systématique et Evolution, CNRS UMR 8079, Université Paris-Sud, 91405 Orsay, France

*Corresponding author: Université Clermont Auvergne, CNRS, Laboratoire: Microorganismes: Génome et Environnement, F-63000 Clermont–Ferrand, France. Tel: +33 473407837; E-mail: didier.debroas@uca.fr
**One sentence summary:** A first overview of microorganisms (protists and Fungi) in lakes and rivers.
**Editor:** Riks Laanbroek

## ABSTRACT

Although they are widespread, diverse and involved in biogeochemical cycles, microbial eukaryotes attract less attention than their prokaryotic counterparts in environmental microbiology. In this study, we used publicly available 18S barcoding data to define biases that may limit such analyses and to gain an overview of the planktonic microbial eukaryotic diversity in freshwater ecosystems. The richness of the microbial eukaryotes was estimated to 100 798 operational taxonomic units (OTUs) delineating 1267 clusters or phylogenetic units (PUs, i.e. monophyletic groups of OTUs that are phylogenetically close). By summing the richness found in aquatic environments, we can predict the microbial eukaryotic richness to be around 200 000–250 000 species. The molecular diversity of protists in freshwater environments is generally higher than that of the morphospecies and cultivated species catalogued in public databases. Amoebozoa, Viridiplantae, Ichthyosporea, and Cryptophyta are the most phylogenetically diverse taxa, and characterisation of these groups is still needed. A network analysis showed that Fungi, Stramenopiles and Viridiplantae play central role in lake ecosystems. Finally, this work provides guidance for compiling metabarcoding data and identifies missing data that should be obtained to increase our knowledge on microbial eukaryote diversity.

**Keywords:** microbial eukaryotes; metabarcoding; freshwater; bioinformatics

## INTRODUCTION

Although microorganisms have been investigated since the 17th century, it is astounding how little we know about their diversity. Over the last few decades, estimations of the microbial richness on Earth have varied widely. For example, Mora *et al.* (2011) predicted the presence of about 611 000 fungal taxa, 36 400 protists and 9680 bacteria on Earth (including oceans), based on a predictable pattern of diversity at the highest taxonomic level, whereas another study suggested that Fungi could be

**Received:** 27 October 2016; **Accepted:** 28 February 2017

**Table 1.** Geographical location, number of ecosystems and main bibliographic sources (all data used are extensively described in the supplementary material table 1).

| Ecosystems | Geographical area | Number | Bibliographic sources |
|---|---|---|---|
| **Lakes** | Massif Central (France) | 11 | Lepère *et al.* (2013); Taib *et al.* (2013); Debroas *et al.* (2015) |
| | Alps | 6 | Lepère *et al.* (2013); Mangot *et al.* (2013); Taib *et al.* (2013); Debroas *et al.* (2015) |
| | Himalaya | 2 | Kammerlander *et al.* (2015) |
| | Arctic | 2 | Charvet *et al.* (2012, Charvet, Vincent and Lovejoy 2014) |
| | Chevreuse Valley (France) | 4 | Simon *et al.* (2014); Simon *et al.* (2015) |
| **Rivers** | Beaujolais vineyard (France) | 1 | Artigas *et al.* (2014) |
| | Massif Central (France) | 2 | Bricheux *et al.* (2013) |
| | Chevreuse Valley (France) | 1 | Simon *et al.* (2014); Simon *et al.* (2015) |

represented by as much as 5.1 million species (Blackwell 2011). However, global microbial richness is likely underestimated, since, for instance, recent studies have revealed freshwater systems are overlooked and harbour high microbial diversity (Schloss *et al.* 2016). New taxa are discovered not only because new environments are explored but also because of the rapid development of sequencing technologies. Owing to high-throughput sequencing (HTS), microbial richness is being deciphered at an unprecedented depth. As an example, whereas around 200 phylotypes were assessed by clone library techniques, HTS detected more than 1000 phylotypes in the same community (Pedrós-Alió 2012). This is partly due to the large 'rare biosphere' (Sogin *et al.* 2006) that cannot be recovered by clone libraries.

Microbial eukaryotes attract less attention than their prokaryotic counterparts in all areas of research in environmental microbiology, even though they are widespread, diverse and involved in biogeochemical cycles (Sherr and Sherr 1988; Caron *et al.* 2008; Grattepanche *et al.* 2014). This is particularly true for their diversity, which has been less explored in freshwaters (Bråte *et al.* 2010; Medinger *et al.* 2010; Nolte *et al.* 2010; Monchy *et al.* 2011; Charvet *et al.* 2012; Lepère *et al.* 2013; Mangot *et al.* 2013; Taib *et al.* 2013; Charvet, Vincent and Lovejoy 2014; Simon *et al.* 2014; Stoeck *et al.* 2014; Vick-Majors, Priscu and Amaral-Zettler 2014; Debroas, Hugoni and Domaizon 2015) than in oceans. However, even in oceans, microbial eukaryotic diversity has been underestimated by at least one order of magnitude, as recently disclosed by the Tara Ocean survey that detected up to 150 000 operational taxonomic units (OTUs) (De Vargas *et al.* 2015), when only 11 200 morphospecies had been catalogued. Several additional studies revealed important richness of groups such as Diplonemidae, Fungi, Cryptomycota, Aphelida and Perkinsozoa, both in marine and freshwater environments (Lefranc *et al.* 2005; Chen *et al.* 2008; Lefèvre *et al.* 2008; Lepère, Domaizon and Debroas 2008; Lepère *et al.* 2010; Jones *et al.* 2011; Massana 2011; Monchy *et al.* 2011; Simon *et al.* 2015). These groups, whose members can be phagotrophs, saprotrophs, parasites or symbionts, often exhibit higher relative abundances than previously observed.

Specifically, a recent investigation of microbial eukaryotic diversity in freshwater by HTS technologies (18S rRNA sequencing) revealed complex community composition, with putative roles ascribed to uncultivable taxa and biogeographical patterns that were unresolved (Lepère *et al.* 2013; Debroas, Hugoni and Domaizon 2015). Some taxa (e.g. Ichthyosporea), rarely detected by traditional molecular methods and never observed in plankton by microscopy, have been detected in an eutrophic ecosystem thanks to molecular inventories (Lepère *et al.* 2013). Similarly, typical marine lineages (e.g. Isochrydales) were recently detected in several freshwater ponds (Simon *et al.* 2014). In addition, some rare taxa are active and consist of lineages distantly related to reference taxa (e.g. Fungi and Alveolata) (Debroas, Hugoni and Domaizon 2015). Together, these data indicated that the ecology and diversity of freshwater microbial eukaryotes are certainly not well understood and are largely underestimated.

To gain an overview and global understanding of the diversity and ecology of these microorganisms, available microbial HTS data must be compiled and synthesised. Such analyses have been performed to study the diversity of marine microbial eukaryotes (Massana and Pedrós-Alió 2008) or, more specifically, the Opisthokonts (del Campo *et al.* 2015). In this study, we aimed to (i) define potential biases (e.g. due to the primer sets selected) in metabarcoding data that can limit a meta-analysis, (ii) provide an overview of the planktonic microbial eukaryotic community structure in freshwater ecosystems, using different metrics (diversity indices and network analysis) and (iii) define missing data in this research area.

## MATERIALS AND METHODS

### Data

In this work, we collected (on 1 January 2016) a set of publicly available data that were related to HTS (pyrosequencing and Illumina with MiSeq technology) of the V4 region of the gene encoding for 18S rRNA (Table 1 and Table S1, Supporting Information). These sequences were obtained from freshwater ecosystems (25 lakes and ponds, and four rivers), sampled at various depths and dates (long term or periodically), and/or obtained from various size fractions. In this analysis, we introduced external references such as V4 amplicons sequenced from a few non-freshwater ecosystems (marine ecosystems and environments characterised by salinity gradients) to compare environments and define spurious OTUs (i.e. singletons, see below)

### Cleaning procedures and clustering

All of the pyrosequencing data were examined against the following quality criteria: (i) no Ns in the nucleotide sequence, (ii) quality score $\geq$ 23 according to the PANGEA process (Giongo *et al.* 2010), (iii) a minimum sequence length of 200 bp and (iv) no sequencing errors in the forward primer. The MiSEQ data were assembled with the USEARCH tool (Edgar 2013) and examined in relation to the previous criteria as well as for the absence of errors in the reverse primer. Putative chimeras and homopolymers were detected by UCHIME (Edgar *et al.* 2011) and the script homopolymer_count.pl (http://alrlab.research.pdx.edu/aquificales/pyrosequencing.html).

The clean freshwater reads were clustered at various similarity thresholds (0.87–0.99) with USEARCH 7.0 (option: cluster_fast) (Edgar 2013) to identify representative OTUs. Clean data for the external references (e.g. sequences from microorganisms in marine environments) and selected sequences from the SILVA database named RefEUKs (see below for a detailed description) were mapped on the representative OTUs to define them. This procedure allowed us to remove the singletons. A singleton in freshwater environments was therefore defined as a read sequenced only once, regardless of the environment, and that was absent in the SILVA database.

### Taxonomic affiliations

The representative OTUs were affiliated by similarity and phylogeny with reference sequences named RefEUKs (https://github.com/panammeb/). These eukaryote references were extracted from the SSURef SILVA database (Pruesse *et al.* 2007) according to the following criteria: length > 1200 bp, alignment quality score > 75% and a pintail value > 50. In addition, the taxonomy of this reference database was modified to include typical freshwater lineages (e.g. Alveolata_1, Chrysophyceae_2, Cryptophyta_4, etc.) defined in previous studies (e.g. Debroas, Hugoni and Domaizon 2015). After a comparison of the OTUs with the RefEUKs by a similarity approach (USEARCH tool), trees of OTUs with their closest reference sequences were built in FastTree (Price, Dehal and Arkin 2010) (see the detailed pipeline in Fig. S1, Supporting Information). Taxonomic assignment was conducted according to two methods: nearest neighbour (NN) and last common ancestor (LCA) affiliations. Phylogenetic trees were used to define phylogenetic units (PUs). PUs are units of OTUs that are assigned to the same NN reference and that cluster as a monophyletic branch in the tree. The cutoff for PU delineation was dependent on the closest relative in the database and was not linked to a taxonomic rank or any threshold (Fig. S1 and Debroas, Hugoni and Domaizon 2015). This process was implemented in the pipeline PANAM (Phylogenetic Analysis of Next-generation AMplicons https://github.com/panammeb/) and is described in more detail in Taib *et al.* (2013) and in Fig. S1.

### Comparing representative OTUs with reference sequences from a public database

To compare freshwater OTUs to reference 18S rDNA sequences from the public database, we used two criteria: similarity and phylogenetic metrics. In the first approach, OTUs were compared to the SSURef SILVA database (NR 117) and were restricted to the total or cultivated eukaryotes using BLAST. In the second, different phylogenetic indices (Swenson 2009) were computed from the trees generated in the pipeline described above (Fig. S1). Mean nearest neighbour distance (MNND) is defined as the mean phylogenetic distance from each OTU to its closest relative in the PU. This index was computed by taking into account the read abundances (MNND ab) or only the presence/absence (MNND pa) of the taxa. The 'X depth/deeper' is defined as the average distance to the deepest node in the tree (Pommier *et al.* 2009) (Fig. S1). These various indices were computed using R software with the packages 'picante' (Kembel *et al.* 2010), 'Geiger' and 'ape' (Paradis, Claude and Strimmer 2004), and were implemented in PANAM.

### In silico analyses of alpha and beta diversity according to the primers used

The RefEUKs in the database were used to define OTUs by clustering in USEARCH with a threshold of 95%. The presence/absence of different primers set (without mismatches) was determined for each OTU. The final result was a 'presence/absence' table with the OTUs as rows, and the primers sets as columns examined by a correspondence analysis (COA), a multivariate method to calculate and visualise the degree of 'correspondence' between the rows and columns of a table (Legendre and Legendre 1998).

### Statistics and network

Different estimators were used to infer the taxa richness of the planktonic eukaryotes: non-parametric estimators (Chao1, ACE, jackknife) and indices based on the rank-abundance curves (log-normal and Poisson-gamma models). These estimators were computed with Vegan (Dixon 2003) and 'species' packages implemented in R. The function jackknife, implemented in the package 'species', computes the order of this estimator automatically. A network was built with CoNET (Faust *et al.* 2012), a plugin in Cytoscape software. A similarity matrix was built with different metrics (Spearman correlation, Bray-Curtis, Kullback-Leibler distances and a mutual information score) from PUs from at least 15 lakes. This initial network, with multiple edges between nodes, was redefined by randomisation. A permutation matrix, representing a null distribution, was obtained by resampling PUs as described in Faust *et al.* (2012). In the permutation step, edge-specific *P*-values were computed; however, for the final network, *P*-values of an edge were merged into one *P*-value following Brown's method (Brown 1975). In the final step, the Benjamini–Hochberg multiple testing correction was applied ($P < 0.05$).

## RESULTS

### Which data can be used to compile metabarcoding data?

Species richness measures were strongly impacted by clustering thresholds and the presence of singletons. When comparing the impact of the clustering threshold on the richness estimated from the complete sequence and V4 region of the same sequence, a difference was found only in the richness estimation when the thresholds were >95% (Fig. S2, Supporting Information).

Using this cut-off, we defined 370 488 OTUs over the 6777 514 clean reads from freshwater ecosystems. By comparing these OTUs to the sequences selected from non-freshwater environments (1189 420 reads) and from the public database (52 103 sequences), 108 896 OTUs were defined (261 592 singletons removed). Only a few singletons were retained: 370 singletons were retrieved from the comparison with the external references and 71 from the comparison with the SILVA database defining 433 OTUs (8 OTUs were common to both databases).

The taxonomic assignment showed that 8378 OTUs were classified as non-eukaryota or metazoa, and the remaining 100 518 OTUs were assigned as microbial eukaryotes. Phylogenetic affiliation delineated 1254 PUs among microbial eukaryotes. In this last step, some OTUs remained unclassified because of a low congruence between the affiliation methods (similarity, NN and LCA). After these different steps, some freshwater

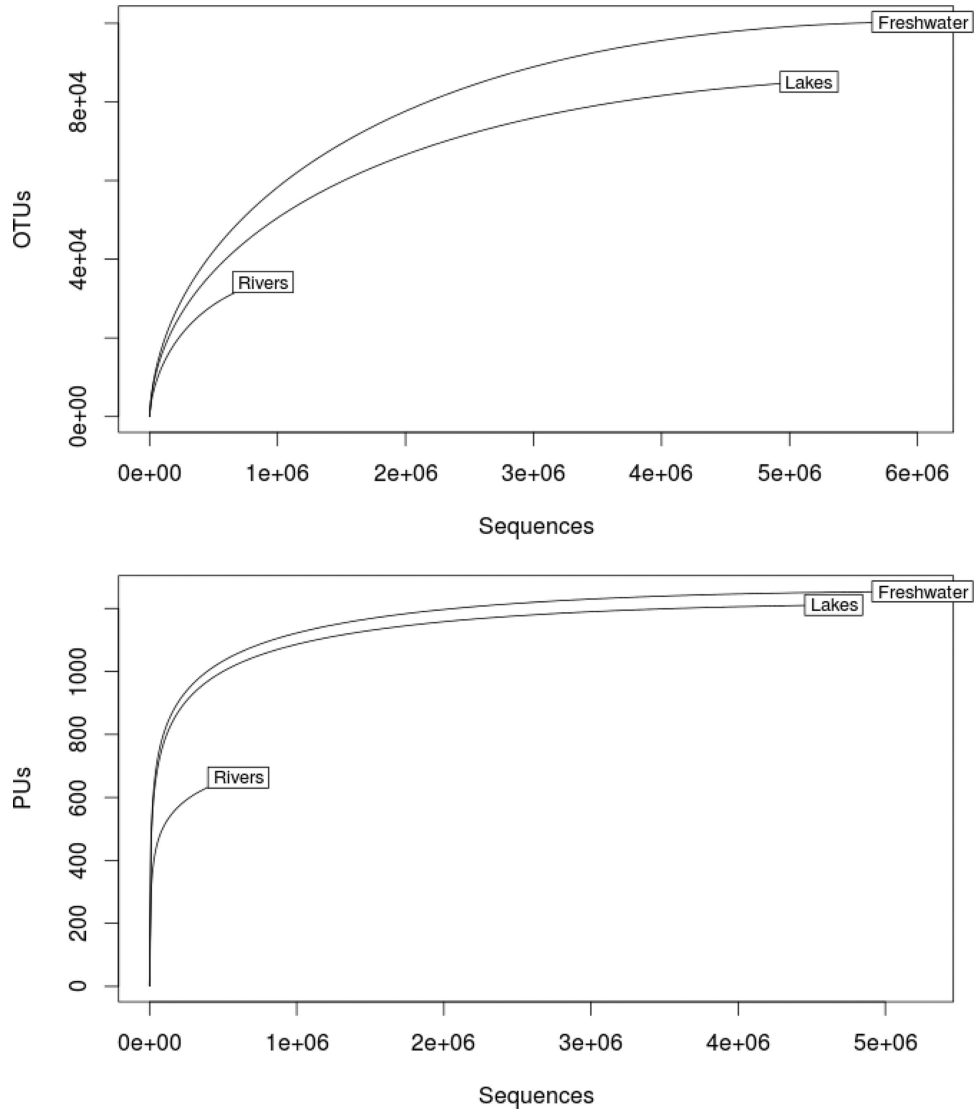


**Figure 1.** Rarefaction curves for lakes, rivers and freshwater ecosystems (rivers + lakes) computed from OTUs (top) and PUs (bottom).

ecosystems had only a few remaining reads (<300) (Table S2, Supporting Information) and were removed from further analysis such as the calculation of the beta-diversity.

The set of primers used may induce a bias in the amplification of the targeted 18S rDNA region. To test and evaluate *in silico* the putative bias in beta-diversity estimation resulting from the use of primers targeting the V4 region, a COA was performed on the OTU table (OTU × primer set) generated from the public sequence dataset. This analysis shows that the selected primer set affected the beta-diversity (Fig. S3A, Supporting Information) and richness estimations. For example, the richness was lowest with the primer set 565F-1134R (Simon *et al.* 2014) and highest with 515F-951R (e.g. Debroas, Hugoni and Domaizon 2015) ($P < 0.01$). In the same way, the OTU matrix (OTU × freshwater environments) subjected to COA discriminated between environments based on the primer set used (Fig. S3B) and therefore according to the laboratory performing the study. However, the analysis of the PU matrix (PU × freshwater environments) allowed discrimination of OTUs based on environment, independent from of the set of primers used (Fig. S3C). Finally, the PU assignments seemed independent from the primer pair used, and therefore represents the most relevant taxonomic unit, especially for ecosystem comparisons (beta-diversity); the OTU level can be used to gain an overview of the global richness (alpha-diversity).

## Microbial eukaryote richness estimation

The rarefaction curves built from OTUs showed that a plateau is reached with freshwater environments analysed together (lakes + rivers) and for lakes only (Fig. 1), but not for rivers, which are likely undersampled. The same conclusion can be drawn based on PU richness.

The estimated OTU richness in the freshwater ecosystems investigated (rivers, lakes and freshwater) varied from 100 325 (Chao1) to 111 507 (log normal), depending on the estimators (Table 2 and Table S3, Supporting Information). The PU richness in these environments was estimated to range from 1255 (Chao1) to 1315 (log normal). However, the log normal-based law did not fit the data (rank-abundance of OTUs and PUs), particularly with the rarest OTUs (results not shown). The rarefaction curves for the major taxonomic groups displayed in

**Table 2.** Observed (obs) and estimated (jackknife estimator) richness in freshwater ecosystems for the main taxonomic groups.

| | OTU obs | Estimated | PU obs | Estimated |
|---|---|---|---|---|
| **Ecosystems** | | | | |
| Freshwater | 100518 | 100798 ± 30.9 | 1254 | 1267 ± 5.1 |
| Lakes | 85109 | 95015 ± 140.8 | 1211 | 1240 ± 7.6 |
| Rivers | 34172 | 42439 ± 128.6 | 663 | 726 ± 11.2 |
| **Taxonomic groups in freshwater ecosystems** | | | | |
| Alveolata | 25965 | 26162 ± 19.8 | 162 | 162 ± 0 |
| Stramenopiles | 11012 | 11075 ± 11.2 | 319 | 319 ± 0 |
| Rhizaria | 4181 | 4201 ± 6.3 | 59 | 59 ± 0 |
| Viridiplantae | 27680 | 27745 ± 11.4 | 171 | 173 ± 2 |
| Fungi | 25713 | 25794 ± 12.7 | 416 | 424 ± 4 |
| Choanoflagellida | 990 | 993 ± 2.4 | 13 | NA |
| Amoebozoa | 229 | 230 ± 1.4 | 17 | 17 ± 0 |
| Ichthyosporea | 447 | 447 ± 0 | 5 | NA |
| Cryptophyta | 1878 | 1885 ± 3.7 | 47 | 48 ± 1.4 |
| Haptophyta | 1519 | 1555 ± 8.5 | 17 | 18 ± 1.4 |

Table 1 were saturated in OTUs and PUs (Fig. S4, Supporting Information). The major groups in freshwater environments were Viridiplantae, Alveolata and Fungi, with an estimated richness that was >25 000 OTUs. With the richness estimation based on PUs, Stramenopiles were more diverse than Viridiplantae and Alveolata.

## Community composition of the major taxonomic groups

Alveolata and Viridiplantae were dominated by Ciliophora and Chlorophyta, respectively (Fig. 2); Stramenopiles by Chrysophyceae and Diatoms; Cryptomonas by Cryptomononadales and Cryptophyta; and Fungi by Dikarya and Chytrids. Cercozoa was the only taxonomic group in the Rhizaria. Numerous reads were affiliated with freshwater clades (these clades are displayed after 'Environmental samples') that were previously delineated as Alveolata_1, LKM11 (Cryptomycota), CM1 (Cryptomycota) and Cryptophyta_4.

At this low taxonomic resolution, all of the groups detected in the rivers were also retrieved in lakes, whereas numerous groups such as the Haptophyta (e.g. *Chrysochromulina*) and Nowakowskiella environmental clades (Fungi), detected in lakes, seemed to be absent in rivers. However, since the richness is likely undersampled in rivers, we cannot exclude the possibility that these groups also occur in rivers, even though they have not been detected yet (Table S4, Supporting Information). At a finer taxonomic resolution, 630 PUs were found in both lakes and rivers, whereas 531 were restricted to lakes and 43 to the rivers.

Some taxa (4532 OTUs) were not restricted to freshwater ecosystems and were also detected in marine environments displaying different salinities. The main phyla shared by both of these environments were Alveolata (1765 OTUs), followed by Fungi (789), Stramenopiles (742) and Viridiplantae (713). More precisely, certain ecosystems, such as LacA and LacWH (more than 35% of OTUs and 62% of reads, respectively) and FAS3 (30.8% of OTUs and 57.8% of reads), presented numerous OTUs, which are also found in marine environments. The NMDS (Fig. S5, Supporting Information) analysis computed with Bray-Curtis distances could discriminate between these ecosystems and others, and grouped the four ponds and one brook together. These ecosystems shared between 21% and 24% of their OTUs with non-freshwater environments.

## Beta-diversity of the lacustrine microbial eukaryotes

Owing to the paucity of data on rivers, we focus only on lakes in this section. The number of PUs (Fig. 3) and OTUs (Fig. S6, Supporting Information) shared by different lakes decreased exponentially with an increase in the number of ecosystems considered. For example, 49 465 OTUs were found in one lake, whereas only 3 OTUs were detected in 18 ecosystems. The PU metric showed a similar pattern, with a decrease in the PUs shared between systems when the number of ecosystems increased. However, three PUs were observed in all the 25 studied lakes, whereas no OTU was shared by more than 18 lakes. There was a strong link between the most ubiquitous taxa and their abundances (i.e. number of reads), with the most widely distributed taxa being the most abundant (Fig. 3). Finally, the frequently observed taxa were also the least abundant, and therefore were the rarest in lakes. Ubiquitous PUs, detected also in rivers, were represented by two Chrysophceae and one typical freshwater Ciliate belonging to the clade Alveolata_1.

Our analyses revealed, therefore, that few PUs (or OTUs) were present in large numbers in these environments. We can thus hypothesise that the lesser known microbial eukaryotes are those present in a restricted number of ecosystems, because their probabilities of being sampled are lower than those that are ubiquitous. Indeed, the plot of phylogenetic indices (MNND and X depth/deeper) as a function of the number of locations showed a significant decrease, whereas the BLAST identity of OTUs increased (Fig. 4). Thus, the mean BLAST identity varied between 94.2% (1 location) and 98.2% (25 locations). The BLAST identity computed from a database restricted to cultivated organisms will always be lower than ones calculated from the total database. The MNND computed from abundance (ab) or incidence matrix (pa) decreased with the number of locations, varying from 0.24 (present in one lake) to 0.05 and 0.07 (pa: incidence matrix) for the PUs present in all lakes. A similar pattern was also observed with the X depth/deeper metric. The comparisons between indices for the rarest PUs (present in most of the five lakes) indicated that Amoebozoa, Viridiplantae, Ichthyosporea and Cryptophyta were present with the highest phylogenetic values (Fig. S7, Supporting Information) and represent the most intriguing microbial eukaryotes in freshwater ecosystems.

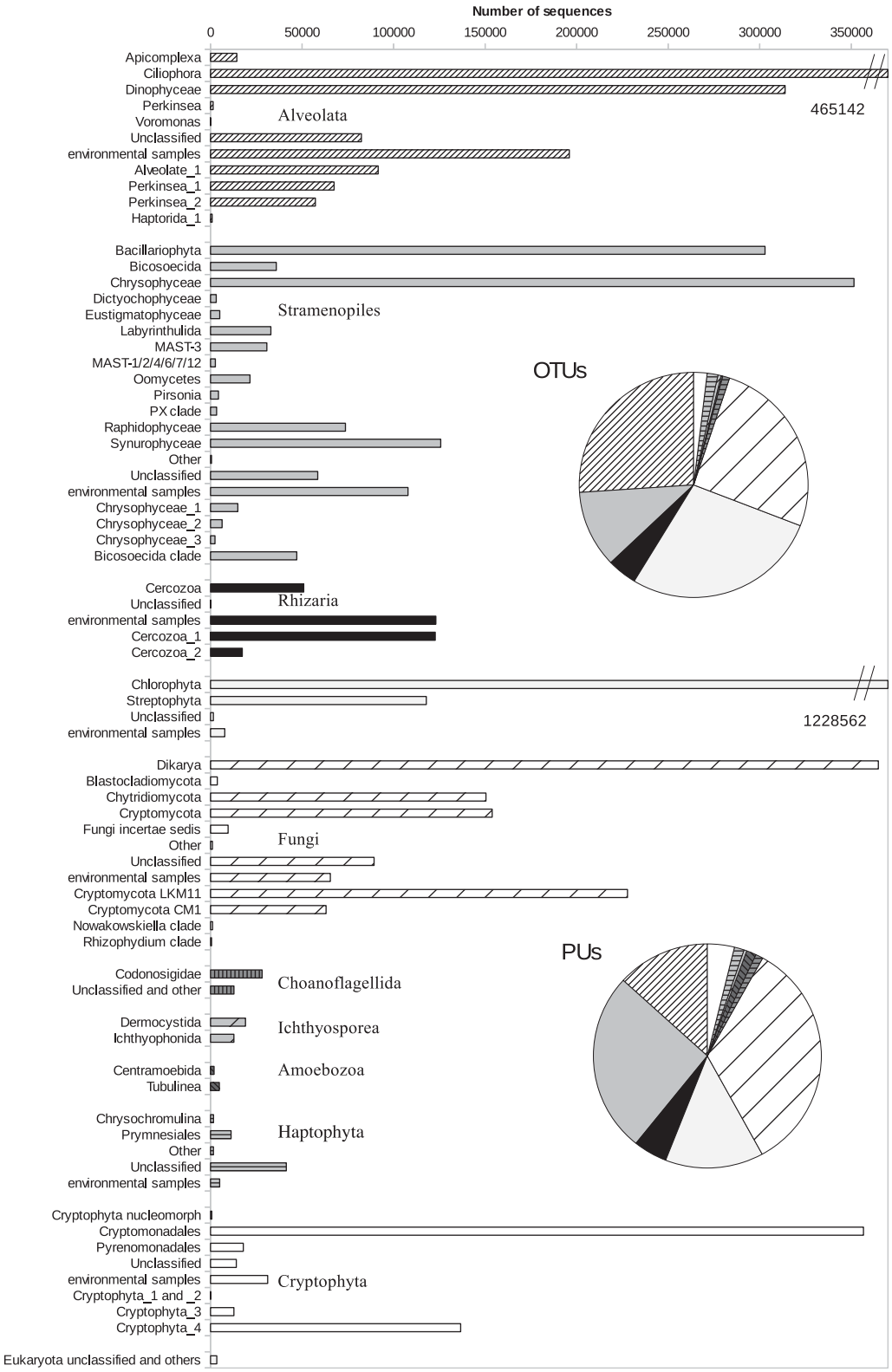


**Figure 2.** A general overview of microbial eukaryotic community composition in freshwater ecosystems. Taxa displayed after 'environmental samples' (e.g. Chrysophyceae_1) correspond to freshwater lineages.

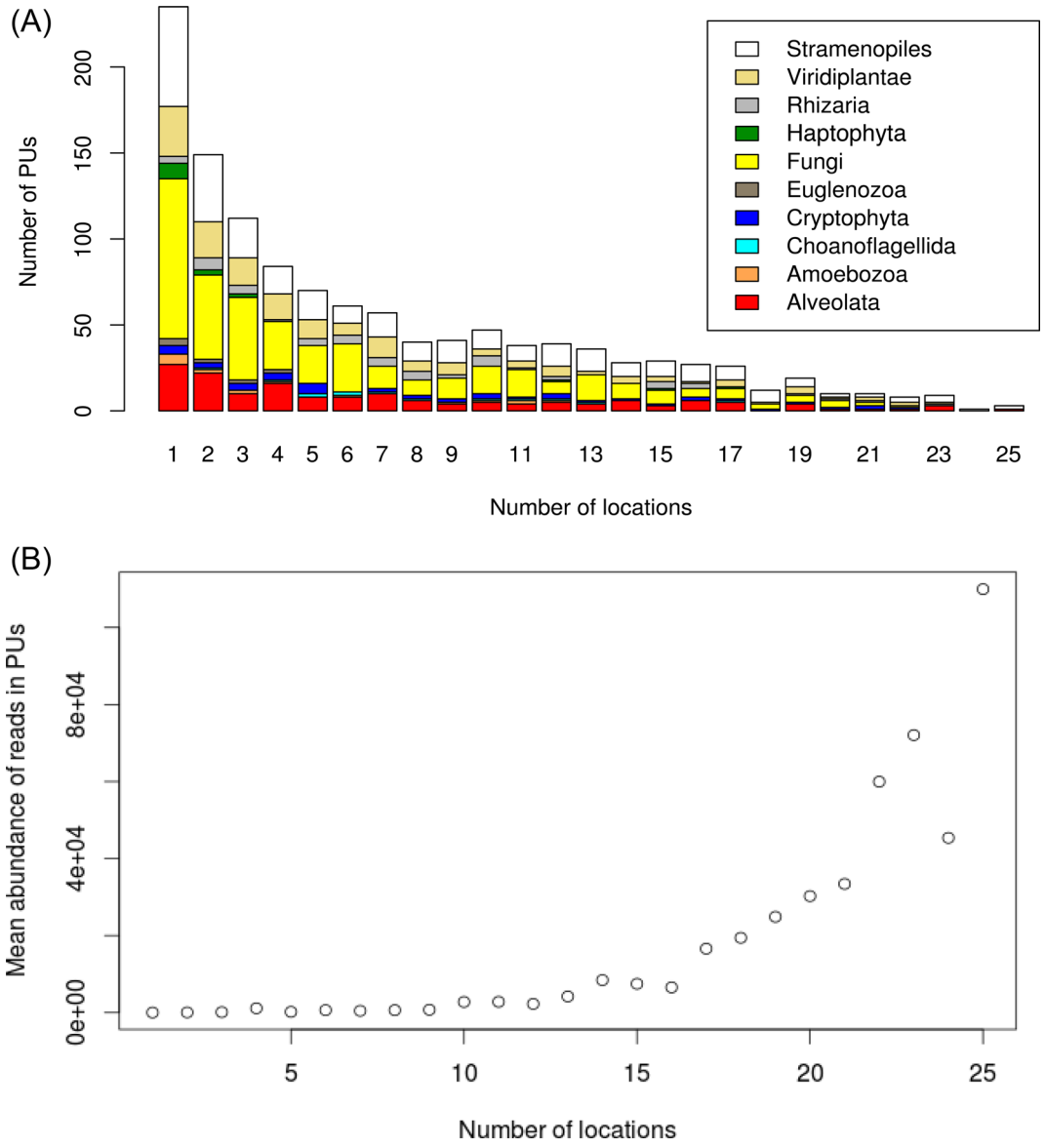


**Figure 3.** Community composition (**A**) and mean abundances of reads (**B**) inferred from PUs as a function of the occupation range (i.e. number of lakes occupied).

### Lacustrine network of microbial eukaryotes

A network was built with the 158 PUs (i.e. nodes) that were detected in at least 15 lakes, based on the relationship between the mean abundance of reads per PU and number of locations (Fig. 3B). The number of nodes with a given degree follows a power law (gamma = 1.16), showing the non-random organisation of this network. It consists of one major cluster with 124 nodes and an average number of neighbours equal to 4.2 (two nodes pair connected with one edge are not shown) (Fig. 5). The clustering coefficient of this network is equal to 0.25 and the diameter to 9.

The highest mean edges, computed from the correlations between PUs, were obtained for Fungi (4.7), Stramenopiles (4.6) and Viridiplantae (4.2) (Table S5, Supporting Information). The closeness and centrality of a node were the highest for Fungi and Viridiplantae. This number reflects the amount of control that these taxa exert over interactions with of other nodes in the network. Haptophyceae were characterised by the lowest number of neighbours (2.0), showing only a few connections with other microbial eukaryotes. The majority of interaction types in the network were co-occurrences (green colour). However, some nodes were characterised by mutual exclusion from other taxa (red lines), such as PU-73 (Dikarya), PU-75 (Alveolata unclassified), PU-1061 (Chrysophyceae) and PU-54 (Alveolata_1) (Fig. 4). Overall, the mutual exclusion principle was mainly associated with Alveolata and an average negative degree equal to 1.24 (Table S5).

## DISCUSSION

### Which methodological aspects matter most in comparisons of metabarcoding data?

With the avalanche of metabarcoding data, a meta-analysis or secondary analysis can be a powerful tool (ArchMiller *et al.* 2015) to decipher the structure and ecology of microbial communities. HTS technologies allow the retrieval of only a small portion of the gene coding for the SSU rRNA, and the targeted zone varies according to the study (V3, V4 or V9). The portion of the

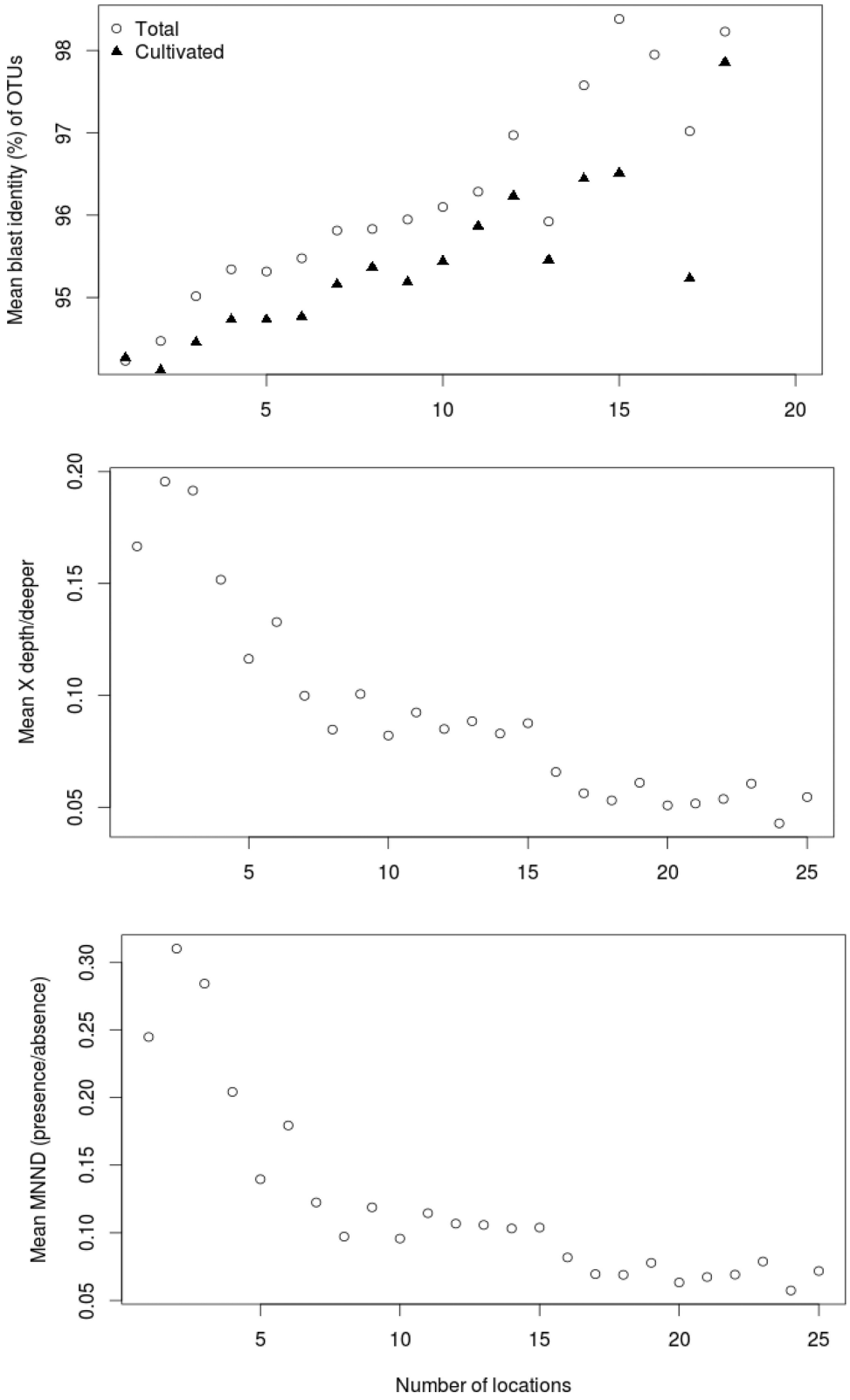


**Figure 4.** Mean OTU BLAST identity (**A**), X depth/deeper (**B**) and MNND pa (C) as a function of the occupation range (i.e. number of lakes occupied).

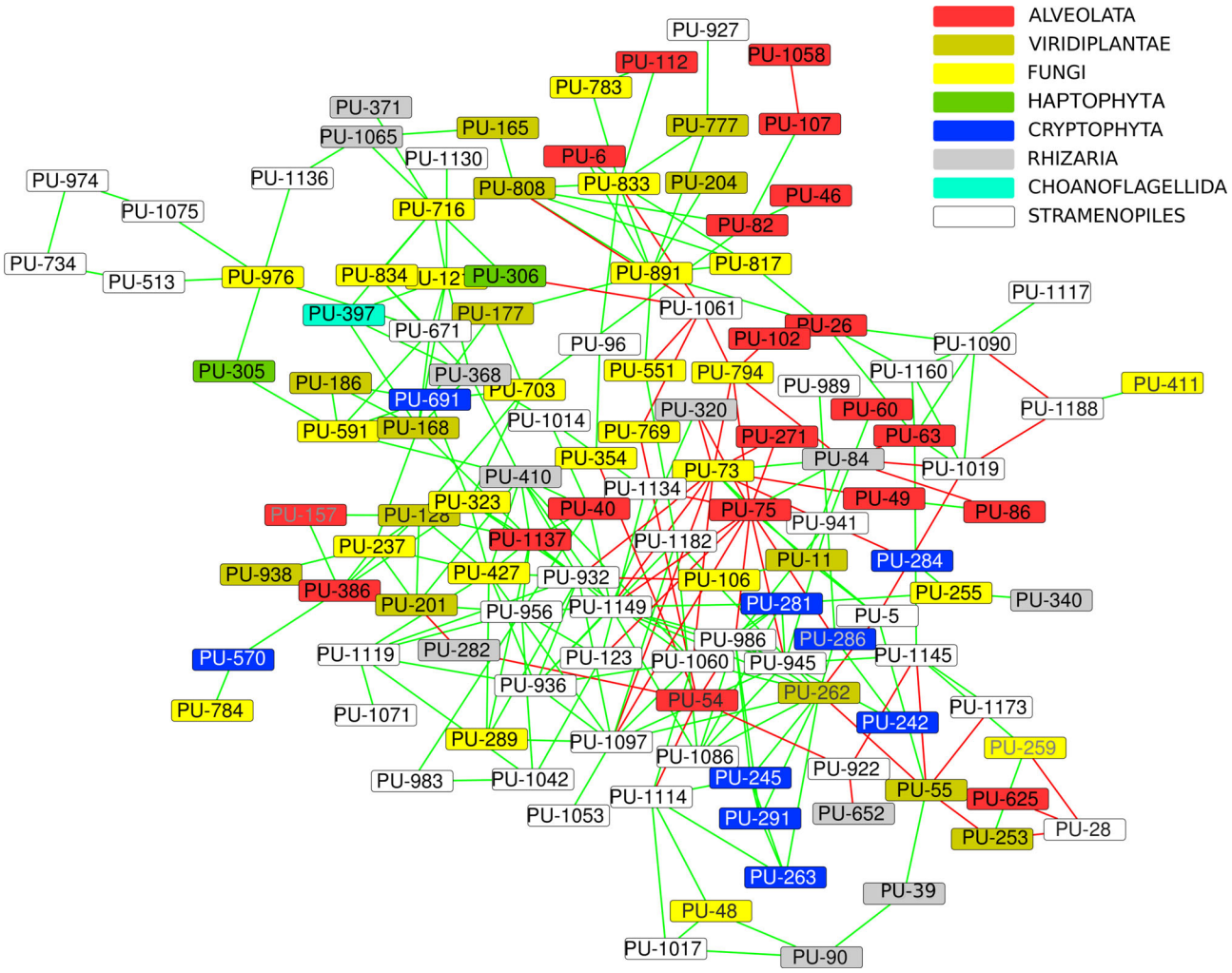


**Figure 5.** Network of PUs in lacustrine ecosystems. The PUs chosen in this analysis were found in at least 15 lakes, allowing us to examine core interactions in these ecosystems.

variable zone targeted influences the assessment of richness, diversity and microbial community composition, including both prokaryotes and eukaryotes (Schloss 2010). In freshwaters, the main phylogenetic markers used are the V4 (Table 1) and V9 regions (Korajkic *et al.* 2015) of the 18S rRNA and, more rarely, the V3 zone (Nolte *et al.* 2010). For easier and more accurate taxonomic identification, we chose to focus on the V4 sequences, because this is a larger dataset and also because this variable zone is present in almost all Sanger sequences deposited in GenBank. In addition, pairwise distances from the V4 region more closely matched the near full-length of the 18S rDNA than the V9 region (Dunthorn *et al.* 2012).

The denoising process and/or clustering threshold also have an impact on the measured richness. In this study, the pyrosequencing data were checked for homopolymer errors that were higher for the fragment of the V4 region compared to that of the V9 region (Behnke *et al.* 2010). We chose a clustering threshold of 95% for two main reasons. First, our study shows that the richness inferred from the V4 region was congruent with that of the full-length 18S rDNA gene at a cut-off inferior or equal to 95% (Fig. S2). Second, this conservative threshold also allowed us to take into account potential sequencing errors that escaped the quality filters, as evidenced by an internal standard (Mangot *et al.* 2013) or when analysing a mock eukaryote communities (Behnke *et al.* 2010). However, this threshold may aggregate some organisms that can have different ecologies. The singletons that are typically discarded in the sequence dataset also caused a problem because they likely represent spurious OTUs, but they can also represent rare species. To avoid removing putative true OTUs, the singletons were confirmed by checking whether they were present in the HTS dataset from non-freshwater ecosystems and from the public sequence database. Once the 'true' singletons were removed, the richness was divided by a factor of 2 at a threshold of 95% (Fig. S2). Notably, another important bias was the effect of 'universal' primers that target the same V4 region of the 18S rDNA. Our *in silico* analysis showed that the alpha - and beta-diversity were strongly dependent on which of the five different primer sets was used (Fig. S3A). It is therefore difficult to disentangle the bias due to the primers used from the effects of biogeography or environmental parameters on discriminating between environments (Fig. S2A). By analysing the sequence dataset at the PU level, we eliminated the biases related to classical OTU clustering that can affect the characterisation of beta-diversity (Fig. S2C). Indeed, richness estimated from OTUs can be biased because they are defined based on sequence identities that can vary according to taxa (Behnke *et al.* 2010). In

addition, PUs allow the elimination of a potential bias introduced by polymorphisms generated by multiple copies of the SSU rRNA genes in microbial cells. Therefore, a PU can include different species but also the same species that presents a high degree of intragenomic variation. Thus, we used both metrics, namely OTUs and PUs, to gain an overview of eukaryotic diversity, and only used PUs to compare ecosystems (i.e. beta-diversity). We suggest that the use of PUs (i.e. monophyletic groups of OTUs that are phylogenetically close) can complement OTUs-based analyses (sequence divergence) in comparative HTS studies.

## First estimation of microbial eukaryotic richness

We estimated the richness at 100 798 OTUs and 1267 PUs, and we can consider that 13 PUs and 476 OTUs were veiled in this study (the difference between observed and estimated in Table 2). To reach 95% of the OTU and PU extrapolated richness, ∼4 million and ∼2 million reads are needed, respectively. These two metrics confirm that the richness of the protists inhabiting freshwater systems was likely captured by this analysis even though some taxa are not detected by V4 primers such as freshwater foraminiferans and Reticulamoeba. The use of diverse primer sets can cause several biases, as discussed, but this represents also a valuable approach to obtain a precise understanding of community diversity, especially when this approach is coupled with a multi-site investigation. Finally, biases resulting from the sets of PCR primers (Jeon *et al.* 2008) can also be minimised using a multiple-primer approach as outlined by Stoeck *et al.* (2006).

To our knowledge, the total richness of the planktonic microbial eukaryotic community has been determined only on rare occasions. De Vargas *et al.* (2015) showed an extrapolated richness of about 150 000 OTUs in marine systems, mainly based on the pico- and nano-size fractions (i.e. protists). According to our study, the truncated Preston log-normal measure displayed a weak fit with the OTU distribution because the rarest OTUs contributed greatly to the *alpha*- and *beta*-diversities (Lynch and Neufeld 2015). Our estimations of the microbial eukaryote richness in aquatic ecosystems using the metabarcoding approach suggest that the work of Mora *et al.* (2011) probably underestimated the true species number. Richness certainly depends on the number of ecosystems sampled, because few OTUs or PUs were detected in all lakes. This observation can be explained either by restricted dispersal capacity (Richards *et al.* 2005; Lepère *et al.* 2013; Forster *et al.* 2015) or the role played by environmental factors in shaping microbial community composition (Simon *et al.* 2015). Examining the role of geographical barriers versus abiotic parameters would require an experimental design that was perfectly controlled (distance between ecosystems × environmental factors) (Martiny *et al.* 2006), as well as an exhaustive sampling of the different habitats and a temporal survey to take into account seasonal variations in community composition. Finally, we found, as expected, that very few species were shared between freshwater and marine environments, because one of the main boundaries that microbes face in their dispersion is salinity (Logares *et al.* 2009). However, MAST, discovered in ocean (Massana *et al.* 2002) and thought to be restricted to marine environments, was also detected in these freshwater ecosystems (Fig. 2, Table S4). It is possible that these microorganisms are dead cells as a result of a requirement for very different conditions to those found in marine environments. However, although their activity in lacustrine plankton has still to be proved, their strong temporal dynamics suggest their adaptation to freshwater environments (Massana *et al.* 2014; Simon *et al.* 2015).

The microbial eukaryote richness in aquatic environment could be around 200 000–250 000 species by taking account our data and those obtained by De Vargas *et al.* (2015). However, this estimation is based on different phylogenetic markers (V4 vs V9) and different bioinformatic pipelines (i.e. clustering methods). Nevertheless, this estimation is close to the lowest number of protist species estimated by Adl *et al.* (2007): $1.4 \times 10^5$ to $1.6 \times 10^6$. However, in this study, Adl *et al.* (2007) arbitrarily estimated the number of predicted species for several groups to be twice the number of described species. For instance, the richness of Haptophytes (< 400) was certainly underestimated when compared to molecular species estimation, as we found 1555 OTUs in freshwater ecosystems and De Vargas et *al.* (2015) detected 713 OTUs of these algae in oceans.

## Are there new species to be found in freshwater ecosystems?

The phylogenetic indices determined in our study highlight the fact that some PUs detected in a small set of locations (i.e. rarest taxa) are less known (high MNND and X depth/deeper, low BLAST identity), shedding light on putative novel species. The link between rarity and sequence similarity in public databases is seldom investigated (Debroas, Hugoni and Domaizon 2015), although this unreferenced diversity likely represents unknown lineages that may have unique metabolic pathways and physiological properties. A group of archaeal OTUs with low identities (Hugoni *et al.* 2013) was shown to correspond to rare archaea in coastal surface waters. In the present work, the unknown lineages were found within the Viridiplantae (Chlorophyta) and Cryptophyta, and to a lesser extent within Amoebozoa and Ichthyosporea. To explore these lineages, an alternative approach could be to use specific primers designed to target the V4 region associated with universal primers in order to obtain the full SSU sequence (Lynch, Bartram and Neufeld 2012).

Finally, novel taxa present in freshwater ecosystems are summarised in Fig. 6, with the number of freshwater OTUs generally higher than (i) the morphospecies catalogued among the taxa that can be detected in freshwater systems, with 53 818 (Pawlowski *et al.* 2012), and (ii) OTUs (V4 clustered at 95% similarity) from the cultivated species. However, the richness of diatom morphospecies (included in Stramenopiles in Fig. 6) fits their estimates of molecular richness in lakes and rivers. Diatoms have been extensively studied because of their large size and characteristic ornamentation, which make them easy to identify morphologically, but also because of their important ecological roles and value as bioindicators of water quality. With this exception, the large disagreement between morphological and molecular-based affiliations can be explained by the lack of morphological features to distinguish small cells, the occurrence of different morphotypes in the life cycle of one species and the presence of cryptic species adapted to fine-scale environmental patchiness (Grattepanche *et al.* 2014). Santoferrara *et al.* (2014) found a 10-fold to 100-fold difference in the diversity between ciliate morphospecies and corresponding phylogenetic species based on pyrosequencing data. For example, *Strombidium oculatum*, an easily recognisable grass-green ciliate is composed of at least 11 different phylogenetic clades. These results also highlight the absence of reference barcodes for morpho- and/or cryptic species. The development of single-cell genomic technologies (genome amplification of sorted cells) along with the amplification a more

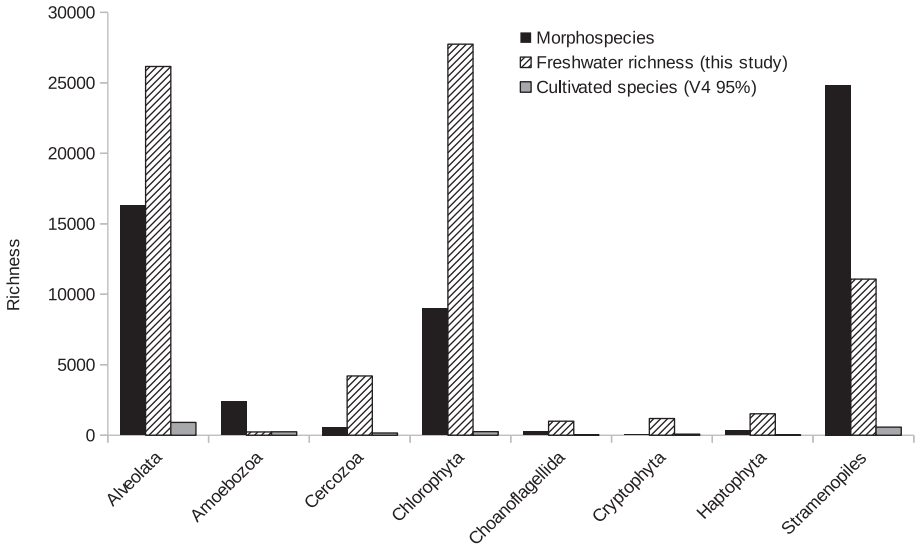


**Figure 6.** Richness computed from morphospecies (Pawlowski *et al.* 2012), freshwater OTUs (Sobs in table 2), and cultivated species (region V4 from the Silva database clustered at 95%).

variable phylogenetic marker (i.e. ITS) or group-specific could be used for this purpose

## The lacustrine core taxa viewed through network inference

Microorganisms interact in many ways, for example, by competing for resources or through predation and parasitism, but they can also cooperate by transferring metabolites or through quorum sensing (e.g. Stocker and Seymour 2012). Over the last few years, co-occurrence networks have been applied to microbial communities to understand their putative interactions and topological features of the networks. These studies have been very useful to deciphering trophic networks, mainly in marine systems. In addition, these networks appear to be composed of highly interconnected nodes (i.e. species) (Berry and Widder 2014). If these networks were built to reveal microbial interactions in a wide range of ecosystems such as soil (Barberán *et al.* 2012), the human gut (Coyte, Schluter and Foster 2015) and oceans (Lima-Mendez *et al.* 2015), this was never the case for freshwater microbial eukaryotes. The PUs that include rare and abundant taxa allowed us to examine putative interactions in these ecosystems. The average clustering coefficient of this network was quite similar to that found in marine ecosystems (0.229) (Lima-Mendez *et al.* 2015) and some soils (Deng *et al.* 2012). This low coefficient (with a range between 0 and 1) means that, on average, the nodes are strongly connected.

Based on their average number of neighbours, Fungi, Stramenopiles and Viridiplantae seem to play central roles in the planktonic eukaryote network and may contribute to ecosystem stability. Fungi was the third most diverse group when examining OTUs (after Alveolata and Viridiplantae) in freshwater and the most diverse and abundant when considering PUs. Molecular analyses of environmental DNA samples have indeed revealed unexpected diversity in Fungi from aquatic environments, including a wide range of habitats such as hydrothermal vent ecosystems (Calvez *et al.* 2009), coastal regions (Gutierrez *et al.* 2011), anoxic regions (Jebaraj *et al.* 2010), lakes (Monchy *et al.* 2011) and rivers (Duarte *et al.* 2015). The majority of the sequences found in our dataset were affiliated to undescribed Fungi, i.e. the dark matter fungi (Grossart *et al.* 2015). These fungi are particularly common in the early diverging fungal branches of the tree of life, presumably occurring at zoosporic stages, and the vast majority remaining uncultured. They are represented by Chytridyomycota (Kagami, Miki and Takimoto 2014) and Cryptomycota (Jones *et al.* 2011) in freshwater. For a long time, fungi were thought to have negligible ecological functions in aquatic systems, and it is only recently that the need to consider aquatic fungi in modelling of plankton food webs has arisen (Niquil *et al.* 2011). These fungi can be saprotrophic and/or parasitic. However, there is still very little knowledge about their ecological functions, such as their role in food web dynamics and biogeochemical cycling of organic matter, nutrients and energy. The central role of Stramenopiles in the network can be explained by the great diversity of ecophysiological properties characterising this taxonomic group; some are strictly autotrophs (e.g. Diatoms), whereas other are mixotrophs (e.g. some Chrysophyceae), heterotrophs (e.g. Bicosoecida) or parasites (some oomycetes). Only the lineages MAST 1, 2 and 12 were part of the network, and the number of neighbours varied between 1 (MAST-12) and 6 (MAST-2). The latter is quite surprising, as they may have colonised the freshwater ecosystems only recently (Massana *et al.* 2014). Their presence in at least 15 lakes and their many neighbours suggest that these lineages may play a significant role in ecosystem functioning, although this will need to be confirmed. One of the three main taxonomic groups in the network is Viridiplantae, represented mainly by the Chlorophyceae. Chlorophyceae are a diverse assemblage of mostly freshwater green algae. They are ecologically significant as primary producers and participate to the global carbon cycle. In this network, Alveolates were characterised by their strong negative interactions with each other. However, the ecological interpretation of such interactions in a network is not straightforward. For instance, a negative correlation between two OTUs can mean parasitism, predation or competitive exclusion is occurring. Among the putative parasites Apicomplexa and Perkinsozoa (Mangot, Debroas and Domaizon 2011), only one edge is negative (between Perkinsozoa and Dinophyceae). The majority of the negative edges involved a Ciliophora classified as Alveolate_1 (100% negative edges), a typically freshwater clade and an unclassified Alveolate (77% negative edges). These microorganisms are connected to various taxa (e.g. Fungi, Stramenopiles

and Cryptophyta), and their interactions can be interpreted as predation (passive filtration) rather than parasitism (specific interactions with few hosts). Indeed, ciliates exert a major link between pico- and nanoplankton and higher trophic levels (Sherr and Sherr 1988). Their role as phagotrophs has been illuminated by numerous studies, with feeding pressure on bacteria, auto- and heterotrophic pico- and nanoplankton (e.g. Pfister, Auer and Arndt 2002; Gaedke and Wickham 2004).

## CONCLUSION

As for Bacteria and Archaea (Schloss *et al.* 2016), the majority of studies have focused their sequencing effort on the same environments. For freshwater microbial eukaryotes, most of the lakes studied have been located in Northern Europe. Exploring ecosystems in a small geographical area restricts the environmental diversity studies. In addition, a limited sequencing effort directed at particular microbial eukaryotes compared to their prokaryotic counterparts likely means the rarest species have been undersampled. Since this rare biosphere may be composed of novel taxa, there is a great need to explore its composition to resolve microbial eukaryote diversity. In addition, rivers are an undersampled ecosystem, although they are inhabited by particular species adapted, for example, to life in biofilms. Thus, investigating the molecular diversity of freshwater microbial eukaryotes, using a metabarcoding and/or a single-cell genome sequencing approach (del Campo *et al.* 2014), should be a priority in microbial ecology, so that a 'Genomic Encyclopaedia of Microbes', which so far has been restricted to the two other domains of life, can be assembled (Wu *et al.* 2009).

## SUPPLEMENTARY DATA

Supplementary data are available at *FEMSEC* online.

## FUNDING

This work was partly supported by the SENDEFO project funded by the French ANR (National Research Agency) Contaminants-Ecosystems-Health (CES-2009).

***Conflict of interest.*** None declared.

## REFERENCES

Adl SM, Leander BS, Simpson AGB *et al.* Diversity, nomenclature, and taxonomy of protists. *Syst Biol* 2007;**56**:684–9.

ArchMiller AA, Bauer EF, Koch RE *et al.* Formalizing the definition of meta-analysis in molecular ecology. *Mol Ecol* 2015;**24**:4042–51.

Artigas J, Pascault N, Bouchez A *et al.* Pes Comparative sensitivity to the fungicide tebuconazole of biofilm and plankton microbial communities in freshwater ecosystems. *Sci Total Environ* 2014;**468–9**:326–36.

Barberán A, Bates ST, Casamayor EO *et al.* Using network analysis to explore co-occurrence patterns in soil microbial communities. *ISME J* 2012;**6**:343–51.

Behnke A, Engel M, Christen R *et al.* Depicting more accurate pictures of protistan community complexity using pyrosequencing of hypervariable SSU rRNA gene regions. *Environ Microbiol* 2010;**13**:340–9.

Berry D, Widder S. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Microb Symbioses* 2014;**5**:219.

Blackwell M. The Fungi: 1, 2, 3 … 5.1 million species? *Am J Bot* 2011;**98**:426–38.

Bråte J, Logares R, Berney C *et al.* Freshwater Perkinsea and marine-freshwater colonizations revealed by pyrosequencing and phylogeny of environmental rDNA. *ISME J* 2010;**4**:1144–53.

Bricheux G, Morin L, Le Moal G *et al.* Pyrosequencing assessment of prokaryotic and eukaryotic diversity in biofilm communities from a French river. *MicrobiologyOpen* 2013;**2**:402–14.

Brown MB. A method for combining non-independent, one-sided tests of significance. *Biometrics* 1975;**31**:987–92.

Calvez TL, Burgaud G, Mahé S *et al.* Fungal diversity in deep-sea hydrothermal ecosystems. *Appl Environ Microb* 2009;**75**:6415–21.

Caron DA, Worden AZ, Countway PD *et al.* Protists are microbes too: a perspective. *ISME J* 2008;**3**:4–12.

Charvet S, Vincent WF, Comeau A *et al.* Pyrosequencing analysis of the protist communities in a High Arctic meromictic lake: DNA preservation and change. *Front Extreme Microbiol* 2012;**3**:422.

Charvet S, Vincent WF, Lovejoy C. Effects of light and prey availability on Arctic freshwater protist communities examined by high-throughput DNA and RNA sequencing. *FEMS Microbiol Ecol* 2014;**88**:550–64.

Chen M, Chen F, Yu Y *et al.* Genetic diversity of eukaryotic microorganisms in Lake Taihu, a large shallow subtropical lake in China. *Microb Ecol* 2008;**56**:572–83.

Coyte KZ, Schluter J, Foster KR. The ecology of the microbiome: networks, competition, and stability. *Science* 2015;**350**:663–6.

del Campo J, Mallo D, Massana R *et al.* Diversity and distribution of unicellular opisthokonts along the European coast analysed using high-throughput sequencing. *Environ Microbiol* 2015;**17**:3195–207.

del Campo J, Sieracki ME, Molestina R *et al.* The others: our biased perspective of eukaryotic genomes. *Trends Ecol Evol* 2014;**29**:252–9.

DeVargas C, Audic S, Henry N *et al.* Eukaryotic plankton diversity in the sunlit ocean. *Science* 2015;**348**:1261605.

Debroas D, Hugoni M, Domaizon I. Evidence for an active rare biosphere within freshwater protists community. *Mol Ecol* 2015;**24**:1236–47.

Deng Y, Jiang Y-H, Yang Y *et al.* Molecular ecological network analyses. *BMC Bioinformatics* 2012;**13**:113.

Dixon P. VEGAN, a package of R functions for community ecology. *J Veg Sci* 2003;**14**:927–30.

Duarte S, Barlocher F, Trabulo J *et al.* Stream-dwelling fungal decomposer communities along a gradient of eutrophication unraveled by 454 pyrosequencing. *Fungal Divers* 2015;**70**:127–48.

Dunthorn M, Klier J, Bunge J *et al.* Comparing the hyper-variable V4 and V9 regions of the small subunit rDNA for assessment of ciliate environmental diversity. *J Eukaryot Microbiol* 2012;**59**:185–7.

Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods* 2013;**10**:996–8.

Edgar RC, Haas BJ, Clemente JC *et al.* UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 2011;**27**:2194–200.

Faust K, Sathirapongsasuti JF, Izard J *et al.* Microbial co-occurrence relationships in the human microbiome. *PLoS Comput Biol* 2012;**8**:e1002606.

Forster D, Bittner L, Karkar S *et al.* Testing ecological theories with sequence similarity networks: marine ciliates exhibit

similar geographic dispersal patterns as multicellular organisms. *BMC Biol* 2015;**13**, doi:10.1186/s12915-015-0125-5.

Gaedke U, Wickham SA. Ciliate dynamics in response to changing biotic and abiotic conditions in a large, deep lake (Lake Constance). *Aquat Microb Ecol* 2004;**34**:247–61.

Giongo A, Crabb DB, Davis-Richardson AG *et al.* PANGEA: pipeline for analysis of next generation amplicons. *ISME J* 2010;**4**:852–61.

Grattepanche J-D, Santoferrara LF, McManus GB *et al.* Diversity of diversity: conceptual and methodological differences in biodiversity estimates of eukaryotic microbes as compared to bacteria. *Trends Microbiol* 2014;**22**:432–7.

Gutiérrez MH, Pantoja S, Tejos E *et al.* Role of fungi in processing marine organic matter in the upwelling ecosystem off Chile. *Mar Biol* 2011;**158**:205–19.

Grossart HP, Wurzbacher C, James TY *et al.* Discovery of dark matter fungi in aquatic ecosystems demands a reappraisal of the phylogeny and ecology of zoosporic fungi. *Fungal Ecol* 2015;**19**:28–38.

Hugoni M, Taib N, Debroas D *et al.* Structure of the rare archaeal biosphere and seasonal dynamics of active ecotypes in surface coastal waters. *Proc Natl Acad Sci* U.S.A. 2013;**110**:6004–9.

Jebaraj CS, Raghukumar C, Behnke A *et al.* Fungal diversity in oxygen-depleted regions of the Arabian Sea revealed by targeted environmental sequencing combined with cultivation. *FEMS Microbiol Ecol* 2010;**71**:399–412.

Jeon S, Bunge J, Leslin C *et al.* Environmental rRNA inventories miss over half of protistan diversity. *BMC Microbiol* 2008;**8**:222.

Jones MDM, Forn I, Gadelha C *et al.* Discovery of novel intermediate forms redefines the fungal tree of life. *Nature* 2011;**474**:200–3.

Kagami M, Miki T, Takimoto G. Mycoloop: chytrids in aquatic food webs. *Front Microbiol* 2014;**5**. doi:10.3389/fmicb.2014.00166.

Kammerlander B, Breiner H-W, Filker S *et al.* High diversity of protistan plankton communities in remote high mountain lakes in the European Alps and the Himalayan mountains. *FEMS Microbiol Ecol* 2015;**91**:fiv010.

Kembel SW, Cowan PD, Helmus MR *et al.* Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* 2010;**26**:1463–4.

Korajkic A, Parfrey LW, McMinn BR *et al.* Changes in bacterial and eukaryotic communities during sewage decomposition in Mississippi river water. *Water Res* 2015;**69**:30–9.

Lefèvre E, Roussel B, Amblard C *et al.* The molecular diversity of freshwater picoeukaryotes reveals high occurrence of putative parasitoids in the plankton. *PLoS One* 2008;**3**: e2324.

Lefranc M, Thénot A, Lepère C *et al.* Genetic diversity of small eukaryotes in lakes differing by their trophic status. *Appl Environ Microb* 2005;**71**:5935–42.

Legendre P, Legendre L. *Numerical Ecology*, 2nd ed. Amsterdam: Elsevier, 1998.

Lepère C, Domaizon I, Debroas D. Unexpected importance of potential parasites in the composition of the freshwater small-eukaryote community. *Appl Environ Microbiol* 2008;**74**: 2940–9.

Lepère C, Domaizon I, Taïb N *et al.* Geographic distance and ecosystem size determine the distribution of smallest protists in lacustrine ecosystems. *FEMS Microbiol Ecol* 2013;**85**:85–94.

Lepère C, Masquelier S, Mangot J-F *et al.* Vertical structure of small eukaryotes in three lakes that differ by their trophic status: a quantitative approach. *ISME J* 2010;**4**:1509–19.

Lima-Mendez G, Faust K, Henry N *et al.* Determinants of community structure in the global plankton interactome. *Science* 2015;**348**:1262073.

Logares R, Bråte J, Bertilsson S *et al.* Infrequent marine–freshwater transitions in the microbial world. *Trends Microbiol* 2009;**17**:414–22.

Lynch MDJ, Bartram AK, Neufeld JD. Targeted recovery of novel phylogenetic diversity from next-generation sequence data. *ISME J* 2012;**6**:2067–77.

Lynch MDJ, Neufeld JD. Ecology and exploration of the rare biosphere. *Nat Rev Microbiol* 2015;**13**:217–29.

Mangot J-F, Debroas D, Domaizon I. Perkinsozoa, a well-known marine protozoan flagellate parasite group, newly identified in lacustrine systems: a review. *Hydrobiologia* 2011;**659**:37–48.

Mangot J-F, Domaizon I, Taib N *et al.* Short-term dynamics of diversity patterns: evidence of continual reassembly within lacustrine small eukaryotes. *Environ Microbiol* 2013; **15**:1745–58.

Martiny JB, Bohannan BJ, Brown JH *et al.* Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol* 2006;**4**:102–12.

Massana R. Eukaryotic picoplankton in surface oceans. *Annu Rev Microbiol* 2011;**65**:91–110.

Massana R, del Campo J, Sieracki ME *et al.* Exploring the uncultured microeukaryote majority in the oceans: reevaluation of ribogroups within stramenopiles. *ISME J* 2014;**8**:854–66.

Massana R, Guillou L, Díez B *et al.* Unveiling the organisms behind novel eukaryotic ribosomal DNA sequences from the ocean. *Appl Environ Microb* 2002;**68**:4554–8.

Massana R, Pedrós-Alió C. Unveiling new microbial eukaryotes in the surface ocean. *Curr Opin Microbiol* 2008;**11**:213–8.

Medinger R, Nolte V, Pandey RV *et al.* Diversity in a hidden world: potential and limitation of next-generation sequencing for surveys of molecular diversity of eukaryotic microorganisms. *Mol Ecol* 2010;**19**:32–40.

Monchy S, Sanciu G, Jobard M *et al.* Exploring and quantifying fungal diversity in freshwater lake ecosystems using rDNA cloning/sequencing and SSU tag pyrosequencing. *Environ Microbiol* 2011;**13**:1433–53.

Mora C, Tittensor DP, Adl S *et al.* How many species are there on Earth and in the ocean? *PLoS Biol* 2011;**9**:e1001127.

Niquil N, Kagami M, Urabe J *et al.* Potential role of fungi in plankton food web functioning and stability: a simulation analysis based on Lake Biwa inverse model. *Hydrobiologia* 2011;**659**:65–79.

Nolte V, Pandey RV, Jost S *et al.* Contrasting seasonal niche separation between rare and abundant taxa conceals the extent of protist diversity. *Mol Ecol* 2010;**19**:2908–15.

Nowack ECM, Melkonian M. Endosymbiotic associations within protists. *Philos T Roy Soc B* 2010;**365**:699–712.

Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 2004;**20**:289–90.

Pawlowski J, Audic S, Adl S *et al.* CBOL Protist Working Group: barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLoS Biol* 2012;**10**:e1001419.

Pedrós-Alió C. The rare bacterial biosphere. *Annu Rev Mar Sci* 2012;**4**:449–66.

Pfister G, Auer B, Arndt H. Pelagic ciliates (Protozoa, Ciliophora) of different brackish and freshwater lakes—a community analysis at the species level. *Limnol - Ecol Manag Inland Waters* 2002;**32**:147–68.

Pommier T, Canbäck B, Lundberg P *et al.* RAMI: a tool for identification and characterization of phylogenetic clusters in microbial communities. *Bioinformatics* 2009;**25**:736–42.

Price MN, Dehal PS, Arkin AP. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* 2010;**5**:e9490.

Pruesse E, Quast C, Knittel K *et al*. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 2007;**35**:7188–96.

Richards TA, Vepritskiy AA, Gouliamova DE *et al*. The molecular diversity of freshwater picoeukaryotes from an oligotrophic lake reveals diverse, distinctive and globally dispersed lineages. *Environ Microbiol* 2005;**7**:1413–25.

Santoferrara LF, Grattepanche J-D, Katz LA *et al*. Pyrosequencing for assessing diversity of eukaryotic microbes: analysis of data on marine planktonic ciliates and comparison with traditional methods. *Environ Microbiol* 2014;**16**:2752–63.

Sherr EB, Sherr BF. Role of microbes in pelagic food webs: a revised concept. *Limnol Oceanogr* 1988;**33**:1225–7.

Schloss PD. The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. *PLoS Comput Biol* 2010;**6**:e1000844.

Schloss PD, Girard R, Martin T *et al*. The status of the microbial census: an update. *MBio* 2016;**7**:e00201-16.

Simon M, Jardillier L, Deschamps P *et al*. Complex communities of small protists and unexpected occurrence of typical marine lineages in shallow freshwater systems. *Environ Microbiol* 2014;**17**:3610–27.

Simon M, López-García P, Deschamps P *et al*. Marked seasonality and high spatial variability of protist communities in shallow freshwater systems. *ISME J* 2015. doi:10.1038/ismej.2015.6.

Sogin ML, Morrison HG, Huber JA *et al*. Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *P Natl Acad Sci USA* 2006;**103**:12115–20.

Stocker R, Seymour JR. Ecology and physics of bacterial chemotaxis in the ocean. *Microbiol Mol Biol R* 2012;**76**:792–812.

Stoeck T, Breiner H-W, Filker S *et al*. A morphogenetic survey on ciliate plankton from a mountain lake pinpoints the necessity of lineage-specific barcode markers in microbial ecology. *Environ Microbiol* 2014;**16**:430–44.

Stoeck T, Hayward B, Taylor GT *et al*. A multiple PCR-primer approach to access the microeukaryotic diversity in environmental samples. *Protist* 2006;**157**:31–43.

Swenson NG. Phylogenetic resolution and quantifying the phylogenetic diversity and dispersion of communities. *PLOS One* 2009;**4**:e4390.

Taib N, Mangot J-F, Domaizon I *et al*. Phylogenetic affiliation of SSU rRNA genes generated by massively parallel sequencing: new insights into the freshwater protist diversity. *PLoS One* 2013;**8**:e58950.

Vick-Majors TJ, Priscu JC, Amaral-Zettler LA. Modular community structure suggests metabolic plasticity during the transition to polar night in ice-covered Antarctic lakes. *ISME J* 2014;**8**:778–89.

Wu D, Hugenholtz P, Mavromatis K *et al*. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 2009;**462**:1056–60.