BMC Biology

Research article

How many novel eukaryotic 'kingdoms'? Pitfalls and limitations of environmental DNA surveys

Cédric Berney*, José Fahrni and Jan Pawlowski

Address: Department of Zoology and Animal Biology, University of Geneva, CH – 1211 Geneva 4, Switzerland

Email: Cédric Berney* - cedric.berney@zoo.unige.ch; José Fahrni - jose.fahrni@zoo.unige.ch; Jan Pawlowski - jan.pawlowski@zoo.unige.ch

* Corresponding author

Published: 04 June 2004

BMC Biology 2004, **2**:13

Received: 05 February 2004

Accepted: 04 June 2004

This article is available from: http://www.biomedcentral.com/1741-7007/2/13

Abstract

Background: Over the past few years, the use of molecular techniques to detect cultivation-independent, eukaryotic diversity has proven to be a powerful approach. Based on small-subunit ribosomal RNA (SSU rRNA) gene analyses, these studies have revealed the existence of an unexpected variety of new phylotypes. Some of them represent novel diversity in known eukaryotic groups, mainly stramenopiles and alveolates. Others do not seem to be related to any molecularly described lineage, and have been proposed to represent novel eukaryotic kingdoms. In order to review the evolutionary importance of this novel high-level eukaryotic diversity critically, and to test the potential technical and analytical pitfalls and limitations of eukaryotic environmental DNA surveys (EES), we analysed 484 environmental SSU rRNA gene sequences, including 81 new sequences from sediments of the small river, the Seymaz (Geneva, Switzerland).

Results: Based on a detailed screening of an exhaustive alignment of eukaryotic SSU rRNA gene sequences and the phylogenetic re-analysis of previously published environmental sequences using Bayesian methods, our results suggest that the number of novel higher-level taxa revealed by previously published EES was overestimated. Three main sources of errors are responsible for this situation: (1) the presence of undetected chimeric sequences; (2) the misplacement of several fast-evolving sequences; and (3) the incomplete sampling of described, but yet unsequenced eukaryotes. Additionally, EES give a biased view of the diversity present in a given biotope because of the difficult amplification of SSU rRNA genes in some taxonomic groups.

Conclusions: Environmental DNA surveys undoubtedly contribute to reveal many novel eukaryotic lineages, but there is no clear evidence for a spectacular increase of the diversity at the kingdom level. After re-analysis of previously published data, we found only five candidate lineages of possible novel high-level eukaryotic taxa, two of which comprise several phylotypes that were found independently in different studies. To ascertain their taxonomic status, however, the organisms themselves have now to be identified.

Background

Over the past few years, cultivation-independent identification of microbial organisms by PCR amplification and sequencing of small-subunit ribosomal RNA (SSU rRNA) genes revealed a huge diversity of eubacterial and archaeal phylotypes in environmental samples, many of which are not represented by cultured organisms [1,2]. Recently, the same techniques have been applied to surveys of

eukaryotic diversity in different marine and freshwater biotopes, including planktonic [3,4] and some extreme, anoxic [5], acidic and iron-rich [6] or deep-sea hydrothermal vent [7,8] environments. All these studies revealed an unexpectedly high diversity of new eukaryotic phylotypes at three distinct taxonomic levels. Some of them can be attributed to novel species in already known genera, families or orders. Others represent novel lineages within already known eukaryotic groups, such as fungi, stramenopiles, alveolates, and kinetoplastids [8-11]. Finally, some of these new phylotypes do not seem to be related to any described lineage, and have been proposed to represent novel high-level taxonomic diversity in eukaryotes [5,7,8].

Here we report 81 new partial SSU rRNA gene sequences of eukaryotes from sediments of the small river, the Seymaz (Geneva, Switzerland). We analyze these sequences together with 403 complete or nearly complete environmental eukaryotic sequences available in GenBank. We point out some of the pitfalls that can impede a correct interpretation of the results of eukaryotic environmental DNA surveys (EES), and evaluate the candidature of some phylotypes to represent novel higher-level eukaryotic lineages. We discuss the impact of an accurate assessment of the environmental diversity on our view of eukaryote megaevolution, in light of recent hypotheses about the shape of the eukaryotic tree and the position of its root.

Results

Sequencing of 81 clones from an EES of the small river, the Seymaz (Geneva, Switzerland) yielded 58 distinct SSU rRNA phylotypes. The size of the sequences varies from 760 to 900 base pairs, which corresponds to the average size expected for the amplified fragment (helices 27 to 50 of the SSU rRNA secondary structure). Size variations occur mainly in the variable region V7, but expansions were observed in the variable region V8 for some sequences. The newly obtained SSU rRNA phylotypes were added to a general alignment of eukaryotes, including most complete or nearly complete sequences from EES available in GenBank. Sequences from cultured organisms were selected so that all major taxonomic groups of eukaryotes were represented; only extremely divergent lineages such as microsporidia and metamonads were omitted. Manual alignment of our sequences allowed the identification of 10 chimeras, which were initially detected because different regions of the same sequence contained rare substitutions and/or indels that are specific for different groups of eukaryotes. Distance analyses based on different subsets of unambiguously aligned regions (partial treeing analysis [12]) were then used to confirm the chimeric nature of these sequences (see Additional file 1 for detailed examples of how we detected chimeric sequences).

The phylogenetic position of the 48 non-chimeric phylotypes from our samples was assessed by minimum evolution analyses. Results are illustrated in Figure 1 (see Additional file 2 for a summary of the identification of all 81 sequences). The tree shown in Figure 1A is the result of an analysis of 86 partial eukaryotic SSU rRNA gene sequences, including five selected environmental phylotypes from previous studies. A total of 670 unambiguously aligned positions were included, and the GTR + G model of evolution was used (alpha = 0.37). Because of the short size of the amplified fragment, some phylogenetic signal was lost and the monophyly of cercozoans and fungi was not retrieved. Almost all phylotypes belong to already known eukaryotic groups. Their relative proportions are illustrated in Figure 1B. Only two phylotypes (Sey010 and Sey017, represented by ten and two sequences, respectively) belong to a yet undetermined, fast-evolving eukaryotic lineage (Figure 1A). They clearly correspond to already published environmental sequences from deep-sea Antarctic plankton (DH148-5-EKD18 [3]), from the Guaymas Basin hydrothermal vent (CS_R003 [7]), and from anoxic, marine sediments collected in Bolinas Tidal Flat (BOL1 cluster [5]). These phylotypes were screened by eye in search for rare sequence signatures that would support their inclusion in already known eukaryotic groups, but none could be detected, suggesting that this lineage might represent a novel high-level taxon.

In the second part of this work, we re-analysed 403 complete or nearly complete published environmental sequences, representing 289 distinct phylotypes. We focused on 28 phylotypes that could not be attributed to known groups of eukaryotes. First, our general alignment was screened by eye for the presence of specific sequence signatures, as described above. It is noteworthy that several previously undetected chimeras were identified in that way, among which three phylotypes were considered as novel high-level taxa, and this result was confirmed by partial treeing analysis. The phylogenetic position of all non-chimeric phylotypes was analysed using Bayesian methods (Figures 2, 3, and 4; see Additional file 3 for a summary of the identification of all 403 sequences). In order to avoid the loss of important informative sites, none of our partial sequences were included in these analyses. The tree shown in Figure 2 is the result of a Bayesian analysis of 125 eukaryotic SSU rRNA gene sequences, including a selection of 56 phylotypes from environmental surveys. A total of 1,175 unambiguously aligned positions were included, and the GTR + G model of evolution was used (alpha = 0.44). Since resolution within alveolates and opisthokonts was poor (using only 1,175 sites), two additional datasets were designed to refine evolutionary relationships within these supergroups. Figure 3 presents the result of a Bayesian analysis of 77 alveolate

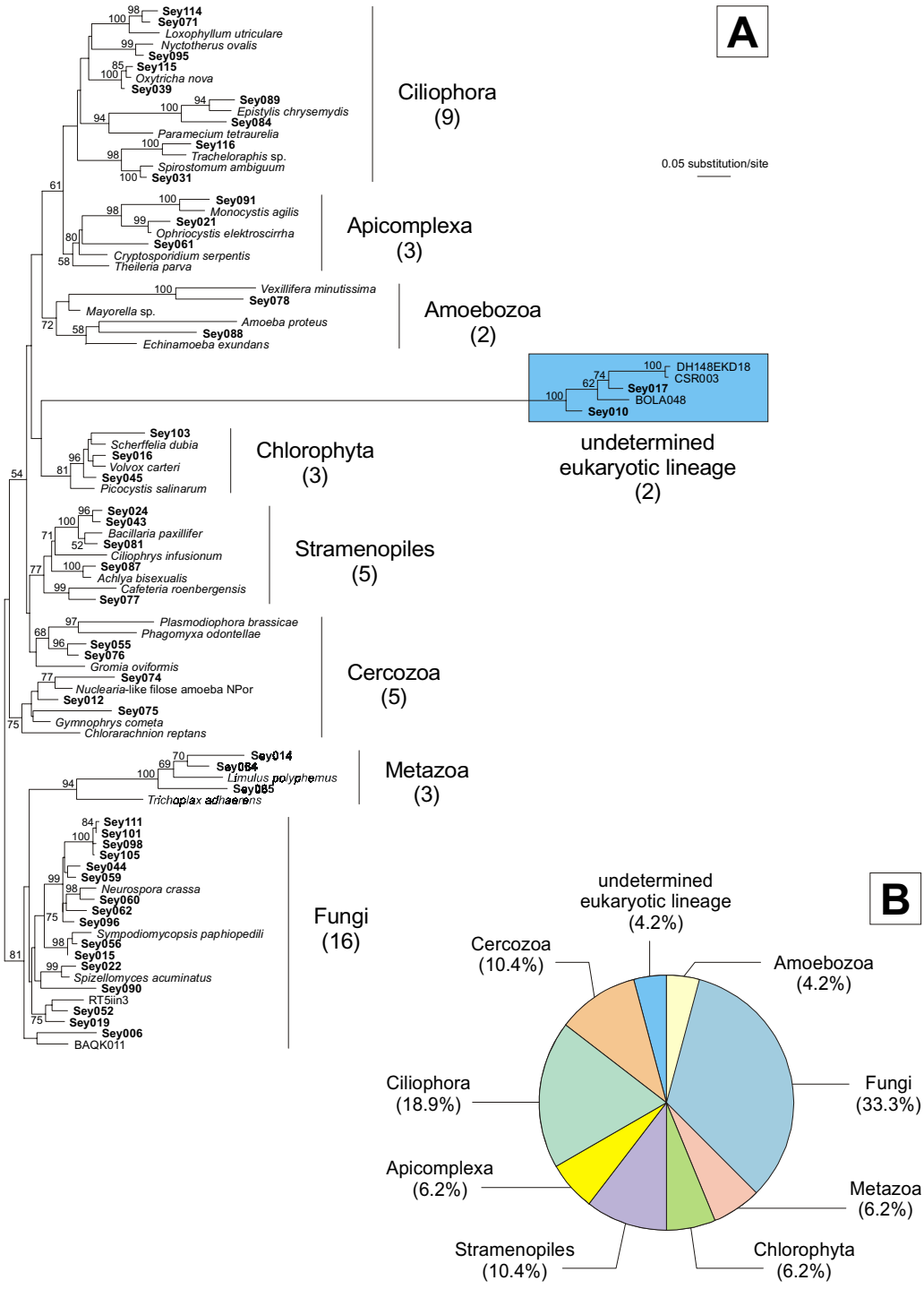

Figure 1
Identification of the 48 distinct, non-chimeric eukaryotic phylotypes we obtained from our samples of the small river, the Seymaz (Geneva, Switzerland). **(A)** Phylogenetic positions of the 48 eukaryotic phylotypes we obtained. The tree shown is the result of a minimum evolution analysis of 68 partial SSU rRNA gene sequences, using the GTR + G model of evolution (see text). The number of phylotypes belonging to each higher-level eukaryotic group is indicated in brackets under the clade name. A fast-evolving lineage of undetermined taxonomic position is highlighted in blue. The tree was arbitrarily rooted on opisthokonts. Numbers at nodes are bootstrap support values following 10,000 replicates. All branches are drawn to scale. **(B)** Relative proportion of phylotypes belonging to each higher-level eukaryotic group.

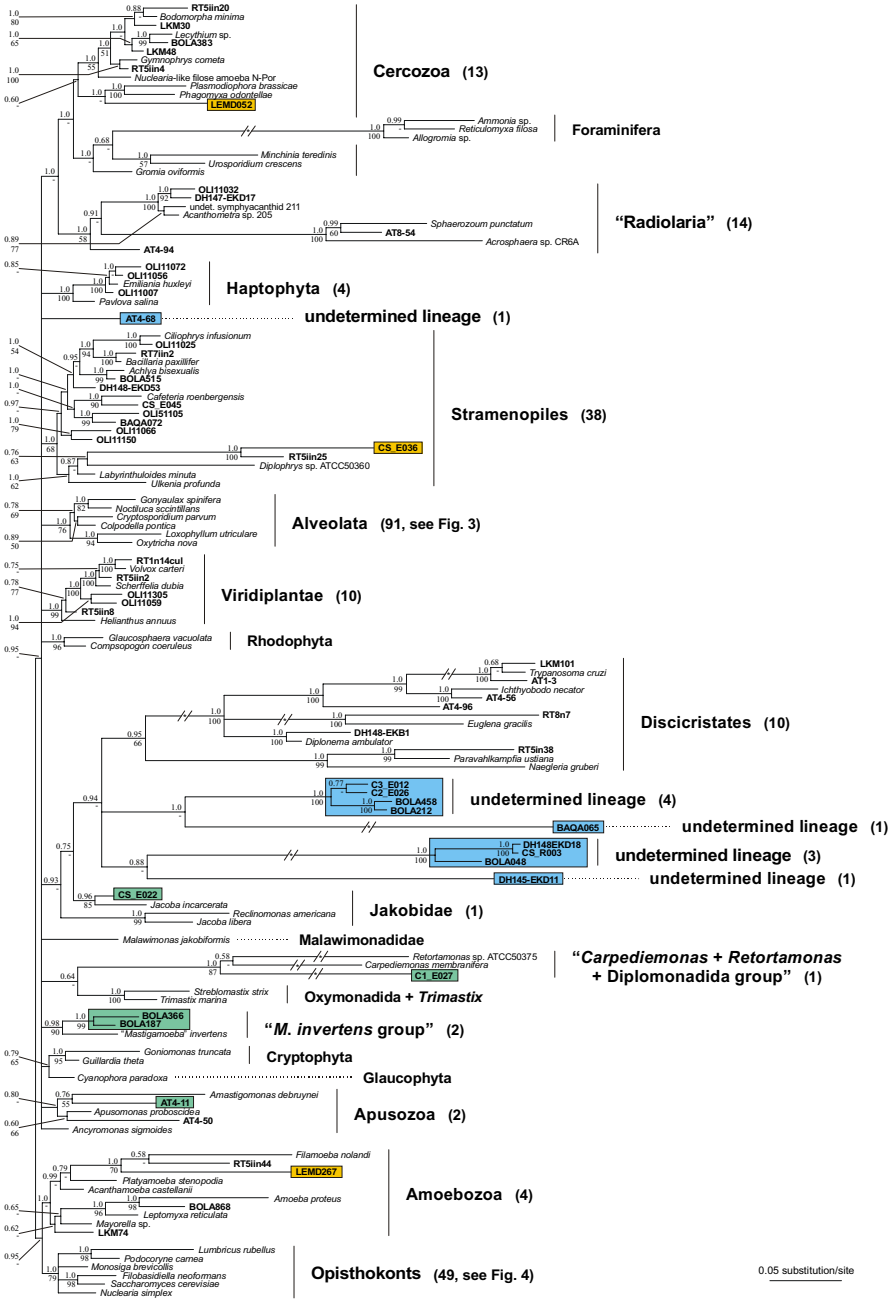

Figure 2

Bayesian phylogeny of eukaryotes based on the analysis of 125 complete or nearly complete SSU rRNA gene sequences (1,175 positions), including 56 selected environmental phylotypes (indicated in bold). The number of phylotypes belonging to each higher-level eukaryotic group is indicated in brackets next to the clade name. Phylotypes previously considered as novel eukaryotic lineages, which are in fact fast-evolving members of known groups, are highlighted in orange. Phylotypes that could be identified thanks to an increasing taxon sampling are highlighted in green. The remaining phylotypes of undetermined taxonomic position are highlighted in blue. The tree is presented with a basal bifurcation between unikonts (Amoebozoa + opisthokonts) and bikonts. The GTR + G model of evolution was used, and the topology shown is a Bayesian consensus of 20,000 sampled trees (see text). The posterior probability of each resolved node is indicated above branches, while numbers under branches represent bootstrap support following 10,000 replicates of a minimum evolution analysis of the same dataset, using maximum likelihood-corrected estimates of the distances (dashes indicate bootstrap values under 50%). Branches are drawn to scale, except those marked with an asterisk (*), which were reduced by half for clarity.

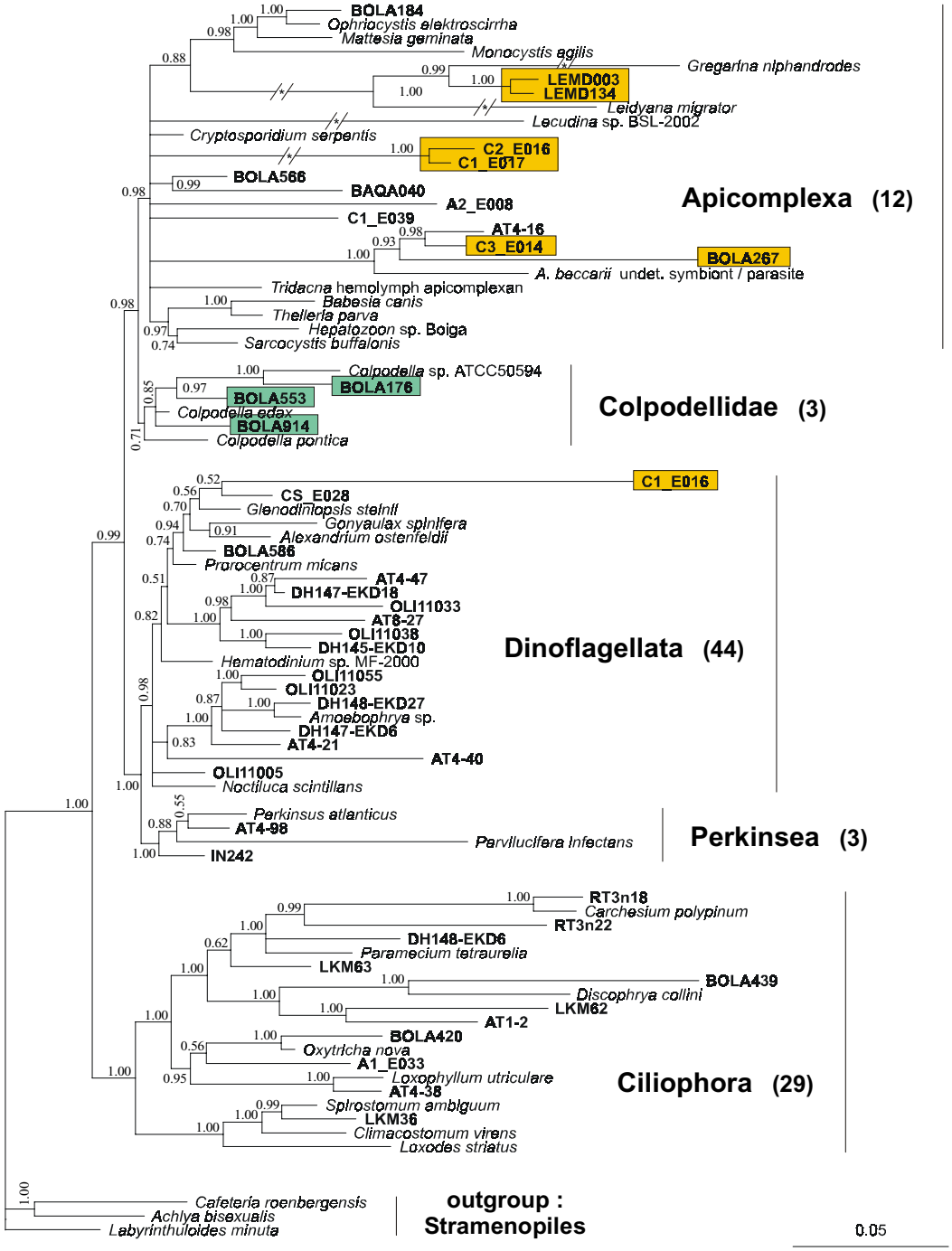

Figure 3
Bayesian phylogeny of alveolates based on the analysis of 80 complete or nearly complete SSU rRNA gene sequences (1,325 positions), including 44 selected environmental phylotypes (indicated in bold). The number of phylotypes belonging to each of the five main alveolate lineages is indicated in brackets next to the clade name. Phylotypes previously considered as novel eukaryotic lineages, which are in fact fast-evolving members of known groups are highlighted in orange. Phylotypes that could be identified thanks to an increasing taxon sampling are highlighted in green. The tree is rooted with three stramenopile sequences. The GTR + G model of evolution was used, and the topology shown is a Bayesian consensus of 20,000 sampled trees (see text). The posterior probability of each resolved node is indicated. Branches are drawn to scale, except those marked with an asterisk (*), which were reduced by half for clarity.

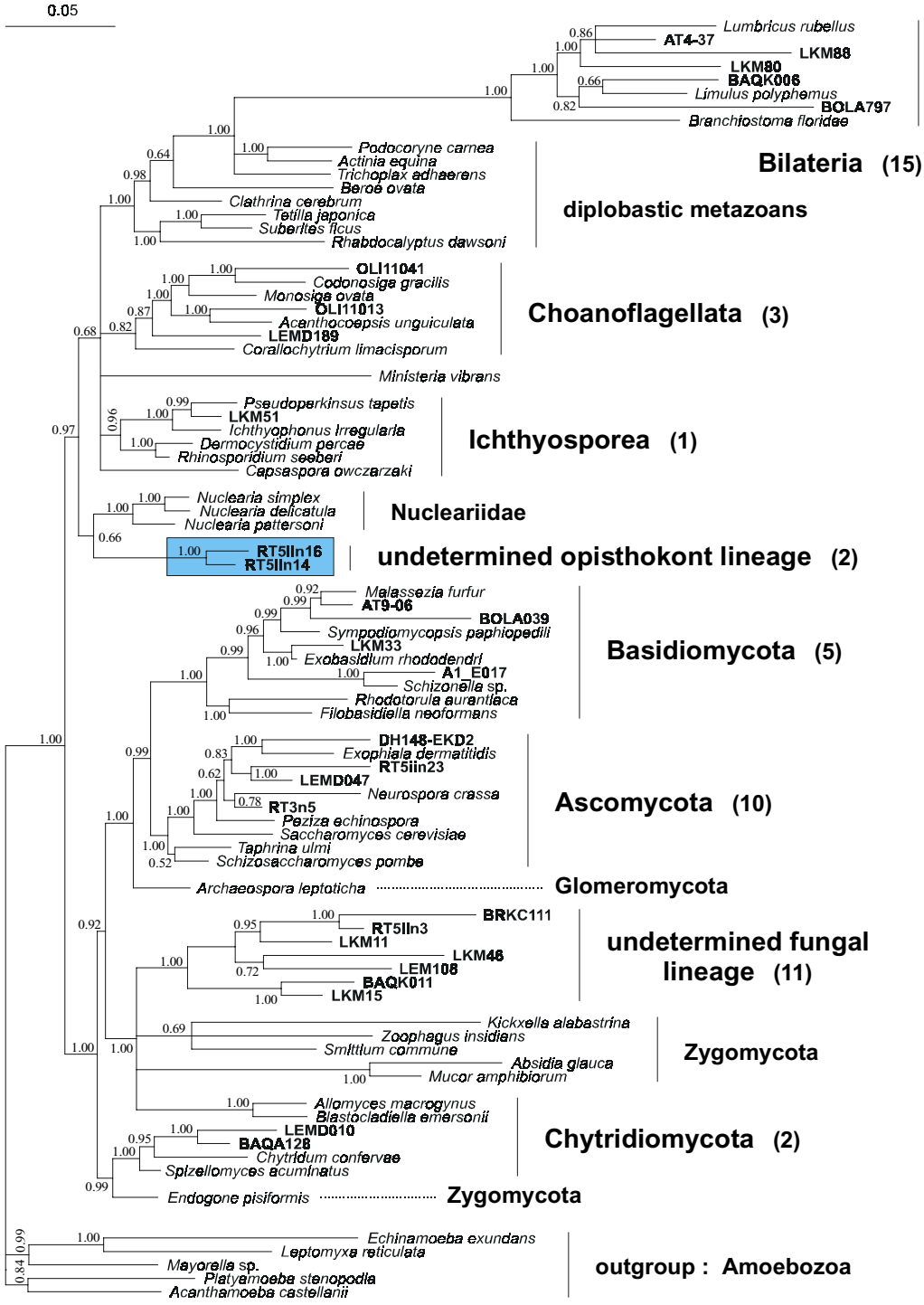

Figure 4

Bayesian phylogeny of opisthokonts based on the analysis of 80 complete or nearly complete SSU rRNA gene sequences (1,395 positions), including 28 selected environmental phylotypes (indicated in bold). The number of phylotypes belonging to each opisthokont lineage is indicated in brackets next to the clade name. An as yet undetermined lineage is highlighted in blue. The tree is rooted with five amoebozoan sequences. The GTR + G model of evolution was used, and the topology shown is a Bayesian consensus of 20,000 sampled trees (see text). The posterior probability of each resolved node is indicated. All branches are drawn to scale.

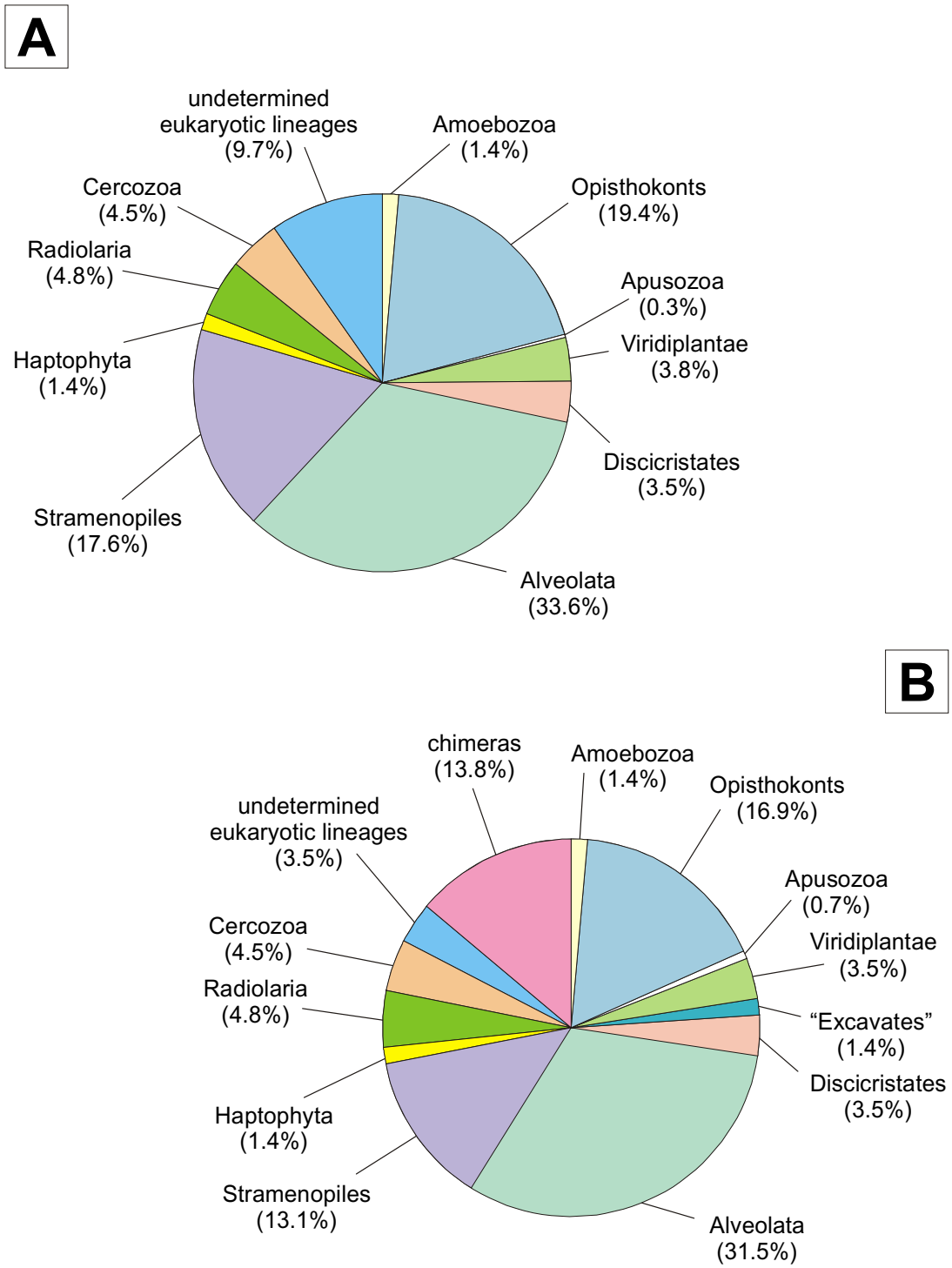

Figure 5
Identification of the 289 published phylotypes we re-analysed. **(A)** As determined by their authors and **(B)** after our re-analysis, highlighting the relative proportion of previously undetected chimeras and the reduced number of phylotypes of undetermined taxonomic position, compared to the proportion of phylotypes belonging to each defined higher-level eukaryotic group. The phylotypes related to, respectively, *Mastigamoeba invertens*, *Jakoba incarcerata*, and the *Carpediemonas* + *Retortamonas* + diplomonads lineage were grouped together as 'Excavates'.

Table 1: Summary of our re-analysis of 28 published phylotypes proposed to represent novel high-level eukaryotic diversity

Phylotype	GenBank accession number	Taxonomic status proposed after our re-analysis	Reference
Previously undetected chimeric sequences			
CS_E042	AY046663	1/2 = CS_E028 (Dinoflagellata)/2/2 = CS_E022 (Jakobidae)	Edgcomb *et al.* 2002 [7]
LEMD145	AF372805	1/2 = undet. Ascomycota/2/2 = LEMD003 (Gregarinia)	Dawson & Pace 2002 [5]
LEMD119	AF372777	1/2 = undet. Apicomplexa/2/2 = LEMD003 (Gregarinia)	Dawson & Pace 2002 [5]
Misplaced fast-evolving phylotypes			
LEMD267	AF372778	Lobosea (Amoebozoa)	Dawson & Pace 2002 [5]
C1_E016	AY046618	Dinoflagellata (Alveolata)	Edgcomb *et al.* 2002 [7]
C3_E014	AY046873	Apicomplexa (Alveolata)	Edgcomb *et al.* 2002 [7]
C1_E017	AY046619	Apicomplexa (Alveolata)	Edgcomb *et al.* 2002 [7]
C2_E016	AY046806	Apicomplexa (Alveolata)	Edgcomb *et al.* 2002 [7]
BOLA267	AF372774	Apicomplexa (Alveolata)	Dawson & Pace 2002 [5]
LEMD134	AF372806	Apicomplexa (Alveolata)	Dawson & Pace 2002 [5]
LEMD003	AF372797	Apicomplexa (Alveolata)	Dawson & Pace 2002 [5]
CS_E036	AY046668	Labyrinthulata (Stramenopiles)	Edgcomb *et al.* 2002 [7]
LEMD052	AF372744	Cercozoa (Rhizaria)	Dawson & Pace 2002 [5]
Phylotypes identified with an increasing molecular sampling of described organisms			
AT4-11	AF530526	Apusomonadidae (Apusozoa)	López-García *et al.* 2003 [8]
BOLA187	AF372745	'Mastigamoeba invertens group'	Dawson & Pace 2002 [5]
BOLA366	AF372746	'Mastigamoeba invertens group'	Dawson & Pace 2002 [5]
CS_E022	AY046649	Jakobidae (Excavates)	Edgcomb *et al.* 2002 [7]
C1_E027	AY046628	'Retortamonas/Carpediemonas group' (Excavates)	Edgcomb *et al.* 2002 [7]
Phylotypes that passed our checking procedure:			
DH145-EKD11	AF290065	possibly novel high-level lineage	López-García *et al.* 2001 [3]
DH148-5-EKD18	AF290084	possibly novel high-level lineage	López-García *et al.* 2001 [3]
C3_E012	AY046842	possibly novel high-level lineage	Edgcomb *et al.* 2002 [7]
C2_E026	AY046816	possibly novel high-level lineage	Edgcomb *et al.* 2002 [7]
CS_R003	AY046643	possibly novel high-level lineage	Edgcomb *et al.* 2002 [7]
BOLA212	AF372767	possibly novel high-level lineage	Dawson & Pace 2002 [5]
BOLA458	AF372771	possibly novel high-level lineage	Dawson & Pace 2002 [5]
BOLA048	AF372821	possibly novel high-level lineage	Dawson & Pace 2002 [5]
BAQA065	AF372825	possibly novel high-level lineage	Dawson & Pace 2002 [5]
AT4-68	AF530543	possibly novel high-level lineage	López-García *et al.* 2003 [8]

SSU rRNA gene sequences, inferred from 1,325 unambiguously aligned positions, using the GTR + G model of evolution (alpha = 0.38). Figure 4 presents the result of a Bayesian analysis of 75 opisthokont SSU rRNA gene sequences, inferred from 1,395 unambiguously aligned positions, using the same model (alpha = 0.37). Remarkably, 10 of the 25 non-chimeric phylotypes that could not be attributed to known lineages of eukaryotes are now robustly identified as fast-evolving members of different well-known groups (mainly alveolates), and five other phylotypes can be linked to recently published sequences of various small eukaryotic lineages (Figures 2 and 3). Figure 5 summarizes the proportion of phylotypes belonging to each of the higher-level eukaryotic groups identified in EES as previously published (Figure 5A) and after our re-analysis (Figure 5B).

Discussion

Our study, based on a detailed visual screening of an exhaustive alignment of SSU rRNA gene sequences of eukaryotes and the phylogenetic re-analysis of previously published environmental sequences using Bayesian methods, shows that at least 18 of the 28 previously published phylotypes proposed to represent novel high-level eukaryotic diversity were misidentified (Table 1). Three main sources of errors are responsible for this situation.

Undetected chimeric sequences

When performing PCR amplifications of SSU rRNA genes on total environmental DNA extracts, chimeric sequences are easily formed because highly conserved regions of ribosomal genes can anneal even between sequences from distantly related organisms. As a result, chimeras can represent a relatively large proportion of environmental sequences [12,13]. In our samples, at least 10 out of the 58 phylotypes we obtained (about 17%) could be identi-

fied as chimeras (see Additional file 2). Comparison with previous studies are difficult, because although most authors checked for the presence of chimeric sequences in their data, some do not indicate precisely how many clones were sequenced and how many of them were chimeras. However, we found at least 40 undetected chimeras among previously published environmental sequences, of which three of the phylotypes are considered as novel high-level taxa (Figure 5, Table 1). The fact that chimeras represent about 14% of the 289 phylotypes we re-analysed is of concern, given that chimeric sequences are a source of artifactual diversity and can bias phylogenetic reconstructions [14].

These results support the idea that the methods generally used for the identification of chimeras might be misleading [13]. In order to detect potential chimeric sequences, programs such as CHECK_CHIMERA [15] can be used. However, the efficiency of these programs depends largely on the completeness of the databases. When a chimera is composed of two parts for which no closely related sequences are available, then either part will have low similarity to all sequences in the database. Furthermore, the asymmetric composition of some chimeric sequences – that is, chimeras in which the putative breakpoint is very close to one extremity of the sequence – can limit drastically the available signal for their detection. Of the 10 chimeras we detected in our sequences, only five could be unambiguously identified as such using CHECK_CHIMERA. A thorough visual checking of all putative new phylotypes for the presence of specific sequence tags might thus prove a more efficient way to detect chimeras.

Long-branch attraction artifacts

The high heterogeneity of the rates of substitution in the SSU rRNA gene sequences of eukaryotes is a second source of errors for an accurate evaluation of the diversity in EES. López-García *et al*. [8] showed that the two undetermined phylotypes BOLA267 [5] and C3_E014 [7] belong to some as yet undescribed, fast-evolving, apicomplexan lineage. We decided to screen the 23 remaining non-chimeric, undetermined phylotypes by eye to look for rare sequence signatures that would also support their inclusion in already known eukaryotic supergroups. At least eight sequences displayed such signatures, suggesting that they are not representatives of novel high-level, taxonomic diversity, but fast-evolving members of well-known groups, such as lobose amoebae or apicomplexan alveolates. These results were strongly confirmed by our Bayesian analyses, which correctly placed all 8 sequences (Figures 2 and 3; Table 1).

Because of the well-known long-branch attraction (LBA) phenomenon [16], fast-evolving sequences tend to be artifactually attracted to each other in phylogenetic trees [17]. In the case of eukaryotic phylogenies, this is especially problematic when prokaryotic sequences are used to root the trees, because distant outgroup sequences act as long branches [18]. The resulting topologies often merely correspond to a 'sequential attachment of longer and longer branches in the absence of any evolutionary signal' [19]. A possible solution to this problem is to avoid the use of prokaryotic outgroup sequences in eukaryotic phylogenies. We recently used this approach to demonstrate the relationship between Foraminifera and Cercozoa, in spite of the extreme SSU rRNA gene divergence between both groups [20]. Besides, the rapid accumulation of SSU rRNA sequences in the databases constantly diminishes the risks of LBA artifacts in phylogenetic analyses. With an increasing taxonomic sampling, the chances of finding slowly evolving taxa closely related to the fast-evolving ones increase. This can hopefully help to position correctly the long branches in the trees. However, the problem of the position of the root persists in the case of phylogenetic analyses performed without obvious outgroup sequences because it can still be argued that the root lies along the stem-branch of one of the apparently fast-evolving lineages. Other evidence, such as rare genomic rearrangements, should hopefully help resolving this problem. Following recent hypotheses on the position of the root of the eukaryotic tree [21,22], we decided to root our eukaryotic phylogeny between unikonts (opisthokonts + Amoebozoa) and bikonts. Under these conditions, we are confident that the topology obtained best reflects the true phylogenetic signal present in the sequences, and that LBA artifacts are minimized.

Incomplete molecular sampling of described eukaryotes

The incompleteness of molecular databases for known eukaryotes is a third source of misinterpretation of the results of EES. The identification of molecular phylotypes as novel eukaryotic groups is correct only if we can be sure that these phylotypes do not belong to some described, but as yet unsequenced eukaryotes. Unfortunately, the proportion of eukaryotic taxa for which no molecular data exist is still relatively high. To our knowledge, SSU rRNA data are available only for about 35 of the 170 genera considered as amoebae and flagellates of uncertain affinities in a recent taxonomic review of protists [23]. Besides, no molecular data exist yet for some higher-level, morphologically well-defined taxa, such as the Hemimastigophora, the testate, lobose Arcellinida, or for some members of the polyphyletic heliozoans, and for many families of the testate, filose amoebae and the so-called 'ramicristate amoebae' (gymnamoebae) [23], whose monophyly is uncertain.

There are several examples that show how the putatively novel eukaryotic phyla disappear with publication of new

sequences. When Edgcomb *et al*. [7] published their results, no SSU rRNA sequences from jakobid flagellates were available, and data on other so-called excavate taxa were scarce. Re-analysis of their sequences in light of the results published by Simpson *et al*. [24] reveals that two of the phylotypes that previously did not show any close resemblance to molecularly described groups of eukaryotes (CS_E022 and C1_E027) turned out to be related to *Jakoba incarcerata* and the clade comprising *Carpediemonas* + *Retortamonas* + diplomonads, respectively (Figure 2). Similarly, the phylotype AT4-11 [9] is related to the apusomonad *Amastigomonas* (Figure 2). The same applies to the potential new diversity within known eukaryotic supergroups. Dawson and Pace [5] obtained three sequences forming a novel alveolate lineage near apicomplexans (BOLA176, BOLA553, and BOLA914). In light of the results published by Kuvardina *et al*. [25], Leander *et al*. [26] showed that it corresponds to the family Colpodellidae. Likewise, we recently obtained the first SSU rRNA gene sequence from the heliozoan-like *Sticholonche zanclea* [27]. This sequence turned out to be closely related to Acantharea and Polycystinea (data not shown), and it corresponds to the previously published environmental 'radiolarian' phylotypes DH145-KW16 [28] and CS_E043 [7]. Obtaining molecular data on a comprehensive sampling of described protists is thus of prime importance to avoid over-interpretation of the diversity revealed by EES.

Other pitfalls of eukaryotic environmental DNA surveys

The correct identification of higher-level phyla is only one of the problems related to EES. Another obvious problem is the accurateness of the diversity inferred from EES data. Whether molecular surveys correctly represent the eukaryotic diversity in a given biotope is of crucial importance for inferring accurate ecological conclusions from the samples. Foraminifera are a good example of an important taxonomic group that is absent in all environmental surveys reported so far, although they are present in both planktonic and benthic marine biotopes, as well as in freshwater biotopes, including the small river we sampled [29]. This may be due to the extreme divergence of foraminiferan ribosomal genes, which cannot be directly amplified with most known universal primers [30], although other explanations, such as a low abundance of Foraminifera in the samples, cannot be discarded. Apart from sequence divergence, the wide range of possible lengths for eukaryotic SSU rRNA genes can also be an important limiting factor during PCR amplifications or cloning. Even with appropriate primers, it is doubtful that complete SSU rRNA gene sequences of more than 3,000 nucleotides, such as those of most foraminiferans, many euglenozoans, and some amoebozoans, would amplify or be cloned in the presence of competing sequences of normal length. Finally, in case of some eukaryotes (lobose amoebae, actinophryid heliozoans) it is difficult to obtain good PCR amplifications even from cultured organisms (A. Smirnov, personal communication). It is not surprising that these eukaryotes are rarely found in EES.

In our EES of the River Seymaz, we decided to amplify only the second half of the SSU rRNA gene, which is generally more conserved both in sequence length and primary structure, in the hope of avoiding negative competition against SSU rRNA gene sequences of unusual length or high divergence. The analysis of our data reveals the presence of many common eukaryotic groups, including ciliates, cercozoans, chlorophytes, diatoms, and fungi, which are expected to be present in a freshwater environment like the small river, the Seymaz (Figure 1, Additional file 2). However, some groups of common protists that were repeatedly observed microscopically in the same area over previous years, such as lobose amoebae and euglenozoans (R. Peck, personal communication), were widely under-represented or even absent from the sequences we obtained (Figure 1, Additional file 2). We suspect that this discrepancy might also apply to previously published EES of marine biotopes. The use of several combinations of universal and/or specific primers, coupled with the use of a range of different PCR conditions, might allow a more realistic qualitative sample of the diversity of organisms present in a given biotope, although this would never be guaranteed.

Identifying novel eukaryotic lineages

After carefully re-analyzing most available near full-length environmental eukaryotic sequences, we found that the number of supposedly novel higher-level phylotypes that cannot be included in defined eukaryotic supergroups is much smaller than enthusiastically proclaimed by the authors of some previous studies [5,7] (see Figure 5 and Table 1). Among 28 phylotypes, three were identified as chimeras, 10 were misplaced fast-evolving sequences, and five were identified after new molecular data on described eukaryotes became available. Among the remaining 10 candidates, three phylotypes (DH148-5-EKD18, CS_R003, BOLA048) from three different EES form a strongly supported clade with two of the phylotypes we obtained in our samples (Sey010 and Sey017), suggesting that they belong to a group of organisms present in all types of environment (Figures 1 and 2). Another candidate cluster that passed our checking procedure comprises the phylotypes C3_E012 and C2_E026 [7] and the BOL2 cluster [5] (Figure 2). Finally, three isolated phylotypes from previous EES might also represent novel high-level diversity: AT4-68 [8], DH145-EKD11 [3], and BAQA065 [5] (Figure 2).

Although these phylotypes passed our checking, it is premature to claim that they truly represent novel eukaryotic kingdoms. First, we cannot exclude the possibility that the

three phylotypes represented by single sequences are amplification artifacts, especially in the case of the extremely divergent sequence BAQA065, or as yet undetected chimeras. It is also uncertain what the real nature of the two clusters of undetermined phylotypes is. The fact that sequences belonging to these two clades were independently found in several different EES indicates that they probably represent real taxonomic lineages. With an increasing taxonomic sampling and/or the development of better phylogenetic tools, it might ultimately be possible to link these clusters with already known groups of eukaryotes. In the tree shown in Figure 2, all fast-evolving phylotypes of unknown affiliation are grouped in a clade that also includes the jakobid flagellates and the discicristates (Heterolobosea + Euglenozoa). Because these sequences are fast evolving (especially the clade present in our samples), we cannot exclude the possibility that their grouping in the tree is the result of LBA artifacts that even Bayesian analyses and a large taxa sampling could not avoid. Supposing that this is not the case, however, these phylotypes might belong to the recently proposed supergroup of excavates [31]. Alternatively, these sequences might belong to some extremely fast-evolving apicomplexan parasites, as suggested by some distance analyses performed on a larger dataset (data not shown), and as proposed by Cavalier-Smith [32] in a similar, simultaneous study. No clear sequence signature could be detected to support the inclusion of any of these phylotypes in the apicomplexan alveolates. However, such signatures are secondarily absent in some of the fastest evolving sequences of gregarines known to date. Finally, it is possible that these sequences represent as yet unrecognized nucleomorphs, which are generally characterized by rapid rates of substitution [33].

Whatever hypothesis is correct, the influence of LBA will be difficult to disprove convincingly in the case of such fast-evolving sequences. Therefore, the only way to ascertain the nature of these putative novel high-level taxa is to identify them in environmental samples using such approaches as the fluorescent *in situ* hybridization. This technique was successfully used by Massana *et al*. [10] to identify representatives of two novel lineages of stramenopiles. One of these lineages was shown to be an important component of the total stock of bacterial grazers in a coastal environment [10]. Similarly, the novel eukaryotic lineages that might be revealed with this approach might turn out to be quantitatively and/or ecologically important members of the biotopes to which they belong.

How large is the novel eukaryotic mega-diversity?

The fact that most of the new phylotypes discovered in EES can be attributed to already known supergroups of eukaryotes is not surprising given the new view of eukaryotic evolution that is emerging from recent analyses of multigenic databases [34,35]. Following this view, most of the eukaryotic diversity can be distributed into eight 'supergroups' [36], with a limited number of possibly independent, smaller, high-level lineages such as apusozoans or centroheliozoans [37-39]. Most of the taxa that were traditionally considered early diverging branches of the eukaryotic tree [40] are now seen as highly derived members of groups belonging to the so-called crown of eukaryotes [41]. It seems, therefore, that the eukaryotes, in terms of cytological innovations and fundamental body plans, are much less diverse than previously thought; the opposite view that emerged at the dawn of the molecular systematics era was strongly biased by LBA artifacts. Furthermore, the whole diversity of eukaryotes may even be reduced to a single basal bifurcation between unikonts and bikonts [22]. However, the existence among extant eukaryotes of truly ancient lineages predating the unikont/bikont divergence cannot be excluded. The detection of such early diverging organisms, if they exist, might prove difficult and necessitate different molecular approaches, such as the use of randomly modified eukaryotic primers. In this respect, the use of cultivation-independent identification of eukaryotes by PCR amplification of SSU rRNA gene sequences should not be neglected, provided that results of such EES are correctly interpreted, and the pitfalls discussed in our study are circumvented.

Conclusions

Environmental DNA surveys undoubtedly contribute to unraveling many novel eukaryotic lineages. In view of our results, however, there is no clear evidence for a spectacular increase of the diversity at a megaevolutionary level. This is in agreement with the recent view of eukaryotic evolution, proposing that most of the known diversity of eukaryotes can be attributed to a relatively small number of 'supergroups'. After re-analysis of previously published data, we found only five candidate lineages of possibly novel high-level eukaryotic taxa, four of which are typically fast evolving. Only two of these lineages comprise several phylotypes that were found independently in different studies, suggesting that they represent real taxonomic lineages. To ascertain their taxonomic status, however, the organisms themselves must now be identified.

Methods

Sediment was sampled in the small river, the Seymaz (Geneva, Switzerland), in May and June, 2002. Total DNA extractions were performed following a protocol modified from Zhou *et al*. [42], as detailed in Holzmann *et al*. [29]. A fragment of about one half of the SSU rRNA gene was amplified by PCR with the universal primers s12.2 (5'-GATYAGATACCGTCGTAGTC-3') and sB (5'-TGATCCT-TCTGCAGGTTCACCTAC-3'). PCR amplifications, purifi-

cations, cloning and sequencing were done as described elsewhere [43].

SSU rRNA gene sequences were aligned manually with the Genetic Data Environment software, version 2.2 [44], following a secondary structure model [45]. Chimeras were identified by visual screening of the alignment in search of contradictory sequence signatures, and confirmed by partial treeing analysis [12,46]. PAUP* [47] was used for minimum evolution analyses using the GTR model of substitution [48,49], and taking into account a gamma-shaped distribution of the rates of substitution among sites, with eight rate categories. Maximum likelihood-corrected estimates of the distances were used, and parameters were estimated from the dataset. Bayesian analyses were performed with MrBayes, version 3.0b4 [50], using the GTR + G model, as above. For each dataset, the chains were run for 2,500,000 generations, and 25,000 trees were sampled. The first 5,000 sampled trees, corresponding to the initial phase before the chains reach stationarity (burn-in), were discarded. The reliability of internal branches was assessed using the posterior probabilities (PP) calculated with MrBayes. Alternatively, the bootstrap method [51] was used with 10,000 replicates for minimum evolution analyses, as described above. The 48 non-chimeric phylotypes reported in this paper have been deposited in the EMBL/GenBank database under accession numbers AY605183 to AY605230.

List of abbreviations used

EES, eukaryotic environmental DNA survey; LBA, long-branch attraction; SSU rRNA, small-subunit ribosomal RNA

Authors' contributions

CB constructed and screened the sequence alignments, carried out all phylogenetic analyses, drafted the manuscript and prepared all figures and tables. JF performed total DNA extractions and carried out PCR amplifications, purifications, cloning and sequencing. JP supervised the study and participated in the preparation of the manuscript. All authors read and approved the final manuscript.

Additional material

Additional File 1

Supplementary Figure 1. Illustration of the methods we used for the detection of chimeric sequences.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1741-7007-2-13-S1.pdf]

Additional File 2

Supplementary Table 1. Identification of the 81 environmental eukaryotic sequences we obtained from our samples of the small river, the Seymaz (Geneva, Switzerland). Two phylotypes of undetermined taxonomic position are highlighted in blue.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1741-7007-2-13-S2.xls]

Additional File 3

Supplementary Table 2. Identification of the 403 published, environmental eukaryotic sequences we re-analysed. Previously undetected chimeras are highlighted in pink. Phylotypes previously considered as novel eukaryotic lineages, which are in fact fast-evolving members of known groups are highlighted in orange. Phylotypes that could be identified thanks to an increasing taxon sampling are highlighted in green. Remaining phylotypes of undetermined taxonomic position are highlighted in blue.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1741-7007-2-13-S3.xls]

Acknowledgements

The authors wish to thank Louisette Zaninetti, Alexey Smirnov, Robert Peck, and Juan Montoya for helpful discussion. This work was supported by the Swiss NSF grant 3100-064073 and 3100A0-100415.

References

1. Barns SM, Delwiche CF, Palmer JD, Pace NR: **Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences.** *Proc Natl Acad Sci USA* 1996, **93:**9188-9193.
2. Hugenholtz P, Goebel BM, Pace NR: **Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity.** *J Bacteriol* 1998, **180:**4765-4774.
3. López-García P, Rodríguez-Valera F, Pedrós-Alió C, Moreira D: **Unexpected diversity of small Eukaryotes in deep-sea Antarctic plankton.** *Nature* 2001, **409:**603-607.
4. Moon-van der Staay SY, De Wachter R, Vaulot D: **Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity.** *Nature* 2001, **409:**607-610.
5. Dawson SC, Pace NR: **Novel kingdom-level eukaryotic diversity in anoxic environments.** *Proc Natl Acad Sci USA* 2002, **99:**8324-8329.
6. Amaral Zettler LA, Gómez F, Zettler E, Keenan BG, Amils R, Sogin ML: **Eukaryotic diversity in Spain's River of Fire.** *Nature* 2002, **417:**137.
7. Edgcomb VP, Kysela DT, Teske A, de Vera Gomez A, Sogin ML: **Benthic eukaryotic diversity in the Guaymas Basin hydrothermal vent environment.** *Proc Natl Acad Sci USA* 2002, **99:**7658-7662.
8. López-García P, Philippe H, Gail F, Moreira D: **Autochthonus eukaryotic diversity in hydrothermal sediment and experimental microcolonizers at the Mid-Atlantic ridge.** *Proc Natl Acad Sci USA* 2003, **100:**697-702.
9. van Hannen EJ, Mooij W, van Agterveld MP, Gons HJ, Laanbroek HJ: **Detritus-dependent development of the microbial community in an experimental system: qualitative analysis by denaturing gradient gel electrophoresis.** *Appl Environ Microbiol* 1999, **65:**2478-2484.
10. Massana R, Guillou L, Díez B, Pedrós-Alió C: **Unveiling the organisms behind novel eukaryotic ribosomal DNA sequences from the ocean.** *Appl Environ Microbiol* 2002, **68:**4554-4558.
11. Moreira D, López-García P: **The molecular ecology of microbial Eukaryotes unveils a hidden world.** *Trends Microbiol* 2002, **10:**31-38.
12. Hugenholtzt P, Huber T: **Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases.** *Int J Syst Evol Microbiol* 2003, **53:**289-293.

13. Robison-Cox JF, Bateson MM, Ward DM: **Evaluation of nearest-neighbor methods for detection of chimeric small-subunit rRNA sequences.** *Appl Environ Microbiol* 1995, **61:**1240-1245.
14. Liesack W, Weyland H, Stackebrandt E: **Potential risks of gene amplification by PCR as determined by 16S rDNA analysis of a mixed-culture of strict barophilic bacteria.** *Microb Ecol* 1991, **21:**191-198.
15. Cole JR, Chai B, Marsh TL, Farris RJ, Wang Q, Kulam SA, Chandra S, McGarrell DM, Schmidt TM, Garrity GM, Tiedje JM: **The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy.** *Nucleic Acids Res* 2003, **31:**442-443.
16. Felsenstein J: **Cases in which parsimony or compatibility methods will be positively misleading.** *Syst Zool* 1978, **27:**401-410.
17. Philippe H: **Opinion: Long branch attraction and protist phylogeny.** *Protist* 2000, **151:**307-316.
18. Wheeler WC: **Nucleic acid sequence phylogeny and random outgroups.** *Cladistics* 1990, **6:**363-368.
19. Stiller JW, Hall BD: **Long-branch attraction and the rDNA model of early eukaryotic evolution.** *Mol Biol Evol* 1999, **16:**1270-1279.
20. Berney C, Pawlowski J: **Revised small subunit rRNA analysis provides further evidence that Foraminifera are related to Cercozoa.** *J Mol Evol* 2003, **57(Suppl 1):**120-127.
21. Stechmann A, Cavalier-Smith T: **Rooting the Eukaryote tree by using a derived gene fusion.** *Science* 2002, **297:**89-91.
22. Stechmann A, Cavalier-Smith T: **The root of the Eukaryote tree pinpointed.** *Curr Biology* 2003, **13:**R665-R666.
23. Lee JJ, Leedale GF, Bradbury P: *An Illustrated Guide to the Protozoa* 2nd edition. Lawrence, Kansas: Society of Protozoologists; 2000.
24. Simpson AGB, Roger AJ, Silberman JD, Leipe DD, Edgcomb VP, Jermiin LS, Patterson DJ, Sogin ML: **Evolutionary history of "early-diverging" Eukaryotes: the excavate taxon *Carpediemonas* is a close relative of *Giardia*.** *Mol Biol Evol* 2002, **19:**1782-1791.
25. Kuvardina ON, Leander BS, Aleshin VV, Myl'nikov AP, Keeling PJ, Simdyanov TG: **The phylogeny of Colpodellids (Alveolata) using small subunit rRNA gene sequences suggests they are the free-living sister group to Apicomplexans.** *J Eukaryot Microbiol* 2002, **49:**498-504.
26. Leander BS, Kuvardina ON, Aleshin VV, Myl'nikov AP, Keeling PJ: **Molecular phylogeny and surface morphology of *Colpodella edax* (Alveolata): insights into the phagotrophic ancestry of Apicomplexans.** *J Eukaryot Microbiol* 2003, **50:**334-40.
27. Nikolaev SI, Berney C, Fahrni JF, Boliver I, Polet S, Mylnikov AP, Aleshin VV, Petrov NB, Pawlowski J: **The twilight of Heliozoa and rise of Rhizaria, a new supergroup of amoeboid Eukaryotes.** *Proc Natl Acad Sci USA* 2004, **101:**8066-8071.
28. López-García P, Rodríguez-Valera F, Moreira D: **Toward the monophyly of Haeckel's Radiolaria: 18S rRNA environmental data support the sisterhood of Polycystinea and Acantharea.** *Mol Biol Evol* 2002, **19:**118-121.
29. Holzmann M, Habura A, Giles H, Bowser SS, Pawlowski J: **Freshwater Foraminiferans revealed by analysis of environmental DNA samples.** *J Eukaryot Microbiol* 2002, **50:**135-139.
30. Pawlowski J: **Introduction to the molecular systematics of Foraminifera.** *Micropaleontology* 2000, **Suppl 1:**1-112.
31. Cavalier-Smith T: **The phagotrophic origin of Eukaryotes and phylogenetic classification of Protozoa.** *Int J Syst Evol Microbiol* 2002, **52:**297-354.
32. Cavalier-Smith T: **Only six kingdoms of life.** *Proc R Soc Lon B Biol Sci* 2004 in press.
33. Van de Peer Y, Rensing SA, Maier UG, De Wachter R: **Substitution rate calibration of small subunit ribosomal RNA identifies chlorarachniophyte endosymbionts as remnants of green algae.** *Proc Natl Acad Sci USA* 1996, **93:**7732-7736.
34. Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF: **A kingdom-level phylogeny of Eukaryotes based on combined protein data.** *Science* 2000, **290:**972-977.
35. Bapteste E, Brinkmann H, Lee JA, Moore DV, Sensen CW, Gordon P, Duruflé L, Gasterlan T, Lopez P, Müller M, Philippe H: **The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*.** *Proc Natl Acad Sci USA* 2002, **99:**1414-1419.
36. Baldauf SL: **The deep roots of Eukaryotes.** *Science* 2003, **300:**1703-1706.
37. Atkins MS, McArthur AG, Teske AP: **Ancyromonadida: a new phylogenetic lineage among the Protozoa closely related to the common ancestor of Metazoans, Fungi, and Choanoflagellates (Opisthokonta).** *J Mol Evol* 2000, **51:**278-285.
38. Cavalier-Smith T, Chao EEY: **Phylogeny of Choanozoa, Apusozoa, and other Protozoa and early Eukaryote megaevolution.** *J Mol Evol* 2003, **56:**540-563.
39. Cavalier-Smith T, Chao EEY: **Molecular phylogeny of centrohelid Heliozoa, a novel lineage of bikont Eukaryotes that arose by ciliary loss.** *J Mol Evol* 2003, **56:**387-396.
40. Sogin ML: **Early evolution and the origin of Eukaryotes.** *Curr Opin Genet Dev* 1991, **1:**457-463.
41. Philippe H, Germot A: **Phylogeny of Eukaryotes based on ribosomal RNA: long-branch attraction and models of sequences evolution.** *Mol Biol Evol* 2000, **17:**830-834.
42. Zhou J, Bruns AM, Tiedje JM: **DNA recovery from soils of diverse composition.** *Appl Environ Microbiol* 1996, **62:**316-322.
43. Fahrni JF, Bolivar I, Berney C, Nassonova E, Smirnov A, Pawlowski J: **Phylogeny of lobose amoebae based on actin and small-subunit ribosomal RNA genes.** *Mol Biol Evol* 2003, **20:**1881-1886.
44. Larsen N, Olsen GJ, Maidak BL, McCaughey MJ, Overbeek R, Macke TJ, Marsh TL, Woese CR: **The ribosomal database project.** *Nucleic Acids Res* 1993, **21:**3021-3023.
45. Wuyts J, De Rijk P, Van de Peer Y, Pison G, Rousseeuw P, De Wachter R: **Comparative analysis of more than 3000 sequences reveals the existence of two pseudoknots in area V4 of eukaryotic small subunit ribosomal RNA.** *Nucleic Acids Res* 2000, **28:**4698-4708.
46. Kopczynski ED, Bateson MM, Ward DM: **Recognition of chimeric small-subunit ribosomal DNAs composed of genes from uncultivated microorganisms.** *Appl Environ Microbiol* 1994, **60:**746-748.
47. Swofford DL: *PAUP*, phylogenetic analyses using parsimony (* and other methods)* Sunderland, Massachussets: Sinauer Associates; 1998.
48. Lanave C, Preparata G, Saccone C, Serio G: **A new method for calculating evolutionary substitution rates.** *J Mol Evol* 1984, **20:**86-93.
49. Rodriguez F, Oliver JL, Marin A, Medina JR: **The general stochastic model of nucleotide substitution.** *J Theor Biol* 1990, **142:**485-501.
50. Huelsenbeck JP, Ronquist F: **MrBayes: Bayesian inference of phylogenetic trees.** *Bioinformatics* 2001, **17:**754-755.
51. Felsenstein J: **Confidence limits on phylogenies: an approach using the bootstrap.** *Evolution* 1985, **39:**783-791.