

The molecular diversity of freshwater picoeukaryotes from an oligotrophic lake reveals diverse, distinctive and globally dispersed lineages

Thomas A. Richards,^{1,2†} Alexey A. Vepritskiy,^{3,4†}
Dilnora E. Gouliamova³ and Sandra A. Nierzwicki-
Bauer^{3,4*}

¹Department of Zoology, The Natural History Museum, Cromwell Road, London SW7 5BD, UK.

²Department of Zoology, The University of Oxford, South Parks Road, Oxford, OX1 3PS, UK.

³Darrin Fresh Water Institute and

⁴New York Center for Studies on the Origins of Life, Rensselaer Polytechnic Institute, 5060 Lakeshore Drive, Bolton Landing, NY, USA.

Summary

The recent discovery of a diverse phylogenetic assemblage of picoeukaryotes from environments such as oceans, salt marshes and acidic habitats, has expanded the debates about the extent and origin of microbial eukaryotes. However, the diversity of these eukaryote microorganisms, that overlap bacteria in size, and their environmental and biogeographical ubiquity remains poorly understood. Here we survey picoeukaryotes (microbial eukaryotes of 0.2–5 µm in size) from an oligotrophic (nutrient deficient) freshwater habitat using ribosomal RNA gene sequences. Three taxonomic groups the Heterokonta, Cryptomonads and the Alveolata dominated the detected diversity. Most sequences represented previously unsampled species, with several being unassignable to known taxonomic groups and plausibly represent new or unsampled phyla. Many freshwater phylogenetic groups identified in this study appeared unrelated to picoeukaryotic sequences identified in marine ecosystems, suggesting that aspects of eukaryote microbial diversity are specific to certain aquatic environments. Conversely, at least five phylogenetic clusters comprised sequences from freshwater and globally dispersed and often contrasting environments, supporting the concept that a number of picoeukaryotic lineages are widely distributed.

Received 26 July, 2004; revised 7 January, 2005; accepted 17 January, 2005. *For correspondence. E-mail nierzs@rpi.edu; Tel. (+1) 518 644 3541; Fax (+1) 518 644 3640. †These authors contributed equally to this work.

Introduction

An accurate record of eukaryotic microbial diversity is a key to resolving evolutionary relationships, as well as understanding a range of biological and environmental factors such as community assembly, the effects of local habitat catastrophe, environmental change, and the biological components of ecological processes. While traditional methods for sampling microbial communities have limited capacity in elucidating true environmental microbial diversity and are often biased towards species with morphologically distinctive features (Pace, 1997), environmental gene libraries of small subunit ribosomal RNA have consistently demonstrated that natural microbial diversity is far more extensive than has previously been observed (van Hannen *et al.*, 1999; Lopez-Garcia *et al.*, 2001; Moon-van der Staay *et al.*, 2001; Amaral Zettler *et al.*, 2002; Dawson and Pace, 2002; Edgcomb *et al.*, 2002; Moreira and Lopez-Garcia, 2002; Lopez-Garcia *et al.*, 2003). It has also become clear that the existing 18S rDNA catalogue is missing ecologically significant morphospecies and requires sampling from alternative environments (Lopez-Garcia *et al.*, 2001; Dawson and Pace, 2002; Stoeck and Epstein, 2003).

Freshwater environments provide diverse ecological habitats and important environmental resources. There are approximately 250 000 cubic kilometers of freshwater on earth, in the forms of lakes, inland seas and rivers, all of which potentially harbour as diverse an assemblage of eukaryotic microbes as has been discovered in open oceans (Guillou *et al.*, 1999; Diez *et al.*, 2001a,b; Lopez-Garcia *et al.*, 2001; Moon-van der Staay *et al.*, 2001). However, rRNA sequence data from freshwater are limited to environments of acidic pH (Amaral Zettler *et al.*, 2002) and freshwater microbial communities maintained in a model system and nourished by specific microbial monocultures (van Hannen *et al.*, 1999). Consequently, the diversity, distribution and natural abundance of different freshwater eukaryotic microbial taxa are largely unknown. In the present study, the diversity of eukaryotic microbes from the euphotic portion of the water column of Lake George is elucidated using molecular methods. The dimictic, oligotrophic lake is located in the Adirondack Park (North-Eastern New York, USA). It lies in a steep glaciated graben of Precambrian bedrock and is

divided into two basins by a narrow waterway section. The long and narrow shape of Lake George (51 km by 2.2 km) provides abundant opportunity for small-scale hydrological variations and nutrient fluxes. Based on morphological grounds, phytoplankton communities of the lake display compositional changes along the length of the basin and variations with depth and season (Monheimer and Baker, 1982). The data presented provides a molecular insight into the picoeukaryotic diversity in a freshwater lake, and offers an important comparison to recent surveys of marine planktonic systems (Diez *et al.*, 2001b; Lopez-Garcia *et al.*, 2001; Moon-van der Staay *et al.*, 2001).

Results and discussion

Eukaryotic microbial diversity in freshwater

One hundred and seventy 18S rDNA clones, comprising 113 restriction fragment length polymorphism (RFLP)-based phylotypes, were completely or partially sequenced. Twenty-three partially sequenced clones were excluded from further analyses as they were identical in RFLP pattern and sequence to other fully sequenced clones. Chimera detection software (Cole *et al.*, 2003) did not uncover any putative chimeras, however, manual alignment checks revealed that three sequences had identical hyper variable regions and lineage specific sequence synapomorphies to divergent lineages at different ends of the 18S rDNA clone suggesting that they were potentially chimeras. This approach has proved successful for chimera detection (Berney *et al.*, 2004). The identified three chimeras were excluded from further analyses. After BLASTN searches and parsimony phylogenetic analyses, the remaining subset of 144 nearly full-length sequences was further narrowed down to 77 sequences based on the following considerations. First, to correct for sampling artefacts, sequences exceeding 99% of sequence similarity were excluded because they could be a product of sequencing error, or represent a single species that may encode a number of variant 18S rRNA molecules, rather than true microbial diversity. Micro-heterogeneity of SSU rRNA genes from the same organism can exceed 99% (e.g. Pawlowski *et al.*, 2002) but a conservative cut-off level was arbitrarily chosen, as it allows highly related but distinct species/strains to be observed at this primary stage of analyses. Second, after the sequences were manually aligned and masked to remove areas that could not be aligned with confidence many Lake George sequences formed phylogenetic clusters with less than 1% sequence variation. These were grouped down to single sequence representatives in the final phylogenetic analyses, enabling sequence sampling to be minimized for the deployment of sophisticated phy-

logenetic methods. Ten groups of Lake George sequences appeared highly related (<1% sequence variation) to sequences from known species. These were also sampled for phylogenetic analyses according to the strategy described above. The distribution of all sampled sequences is shown in Figs 1, 2 and 6. Rarefaction curves, calculated for the complete library of all 414 clones and the corresponding 125 RFLP patterns (data not shown), did not reach a clear saturation, indicating that analysis of an increasing number of clones may have revealed further diversity and increased overlapping between the two lake basins sampled.

Sequences that grouped within the Heterokonta, Alveolata and Cryptomonads were the most diverse. Of the 77 distinct Lake George phylotypes analysed, 47 (61%) grouped with species from these taxonomic groups with 60% or greater bootstrap support in both LogDet and maximum likelihood (ML) distance analyses (Figs 1–3 and 6). This pattern of diversity was similar to that discovered in a range of different environments (Moon-van der Staay *et al.*, 2001; Berney *et al.*, 2004) implying that a relatively small number of higher taxonomic groups dominate diversity of picoeukaryotes in aquatic environments.

Microscope studies of freshwater lake ecosystems have indicated that three microbial eukaryote groups, the Heterokonta, the kathablepharids, and the Choanozoa, contribute in excess of 90% of the flagellate protozoa diversity observed (Arndt *et al.*, 2000). The Choanozoa, a group of comparatively large ($10 \geq \text{Ø} \geq 5 \mu\text{m}$) heterotrophic flagellates that thrive predominantly in salt/brackish habitats and cluster in phylogenetic trees at the base of the animal/fungi branch, were not detected in any of the Lake George samples, suggesting that these forms are either not numerous in the lake's waters or too large to be detected by the methods employed here. The kathablepharids, a group of predatory, heterotrophic biflagellates with mitochondria have been identified in microscopy studies as a significant component of freshwater microbial communities (Arndt *et al.*, 2000) and comprise three recognized genera: *Kathablepharis*, *Leucocryptos* and *Platytilomonas* (Lee *et al.*, 2000). To the best of our knowledge, no cultured strain or 18S rDNA sequence is available, thereby making it impossible to infer their contribution to the diversity reported here.

Heterokonts

Thirty (39%) of the 79 phylotypes analysed grouped with heterokont species with a bootstrap support of 60% or higher (Figs 1 and 3). A large proportion of the sequence diversity observed grouped close to the Bicosoecida, corroborating morphology-based observations, which have demonstrated that bicosoecids are abundant in freshwater environments (Arndt *et al.*, 2000). Het-

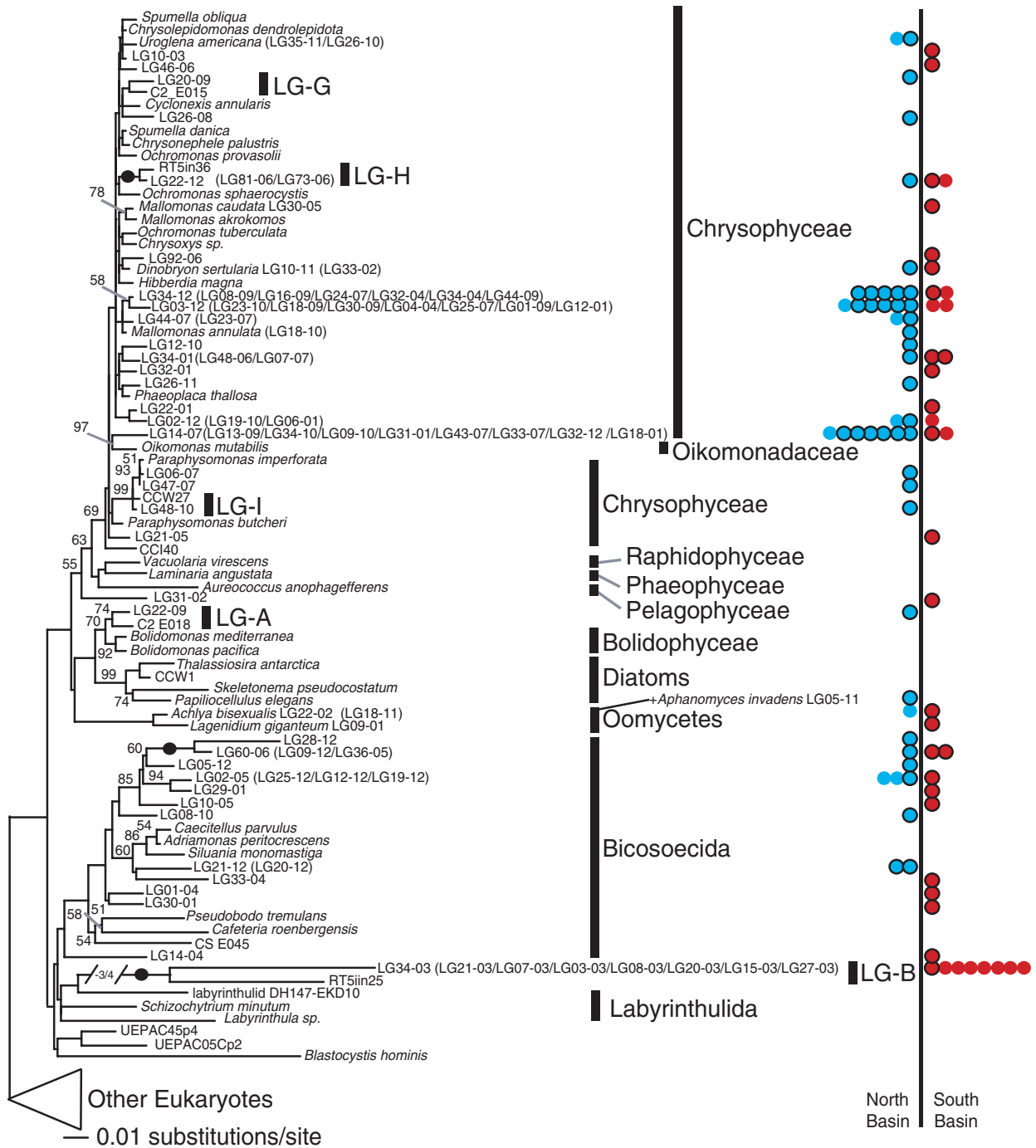


Fig. 1. Diversity of the Heterokonta picoeukaryotes sampled from Lake George. The small subunit 18S rRNA gene phylogenetic distance tree was inferred from an alignment of 207 taxa with a character sampling of 1217 nucleotides. Distance methods (topology search and 1000 bootstrap replicates) utilized a GTR+ γ +I model of sequence evolution with heuristic searches by stepwise addition using the TBR settings in PAUP. Bootstrap support values above 49% are labelled on the phylogenetic branches. A black dot on a branch indicates 100% support value. All taxa labelled with LG and a number were sampled from Lake George. Other phylotypes with alphanumeric names are from published environmental gene libraries (see Table 1). Occurrence of each LG phylotype is marked with blue and red dots on the right, representing North and South Lake George basins respectively. Circled dots indicate a clone that had a 99% or greater total sequence identity with the sequence included in the phylogeny. Non-circled dots indicate a clone that had 99% or greater sequence identity with the sequence included in the phylogeny after the alignment was masked. The Lake George sequences labelled next to known species and are grouped to the same criteria. Proportional length reduction of shortened branches is labelled.

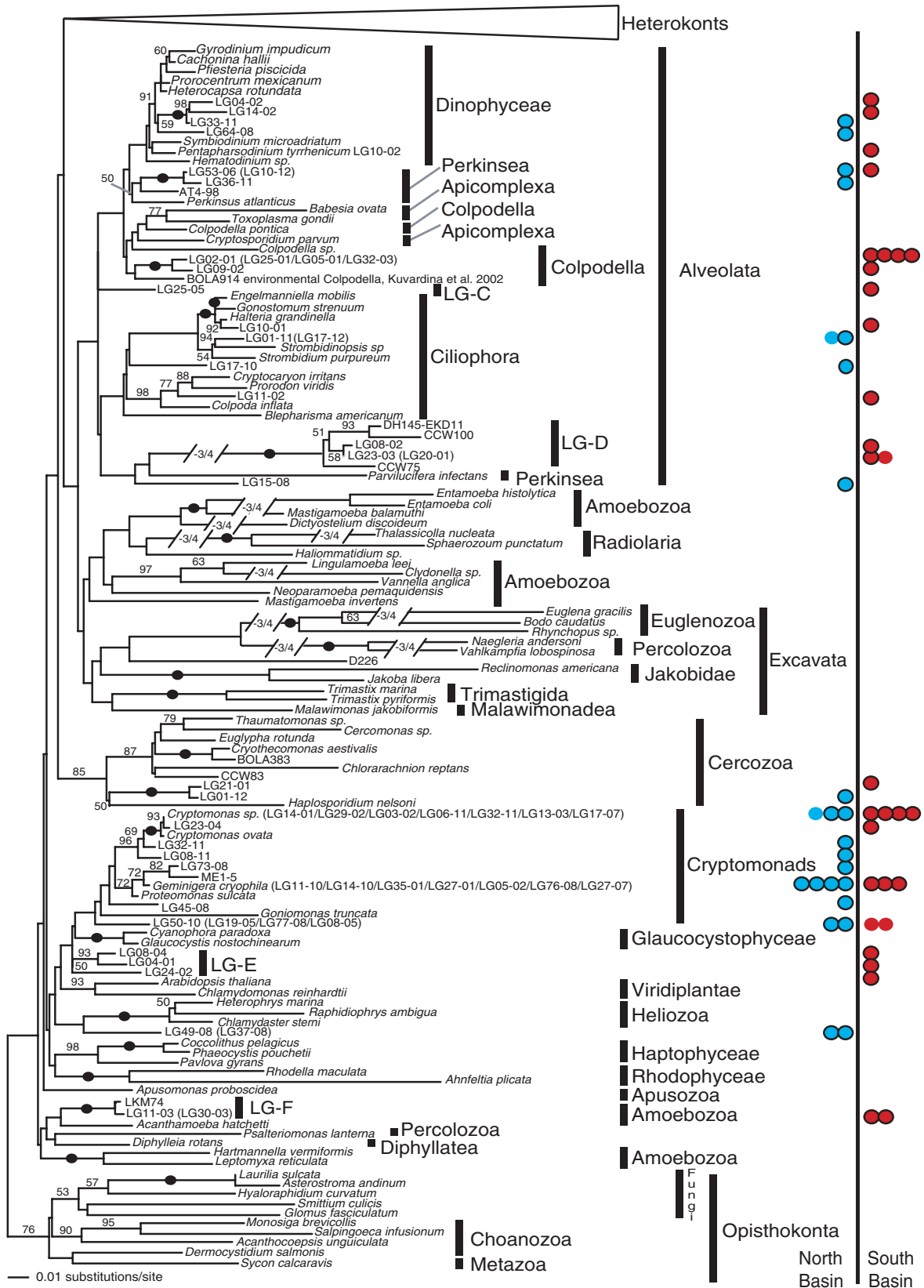


Fig. 2. Diversity of picoeukaryotes other than heterokonts sampled from Lake George. See legend to Fig. 1 for details. Reference given for BOLA9114: Kuvardina and colleagues (2002).

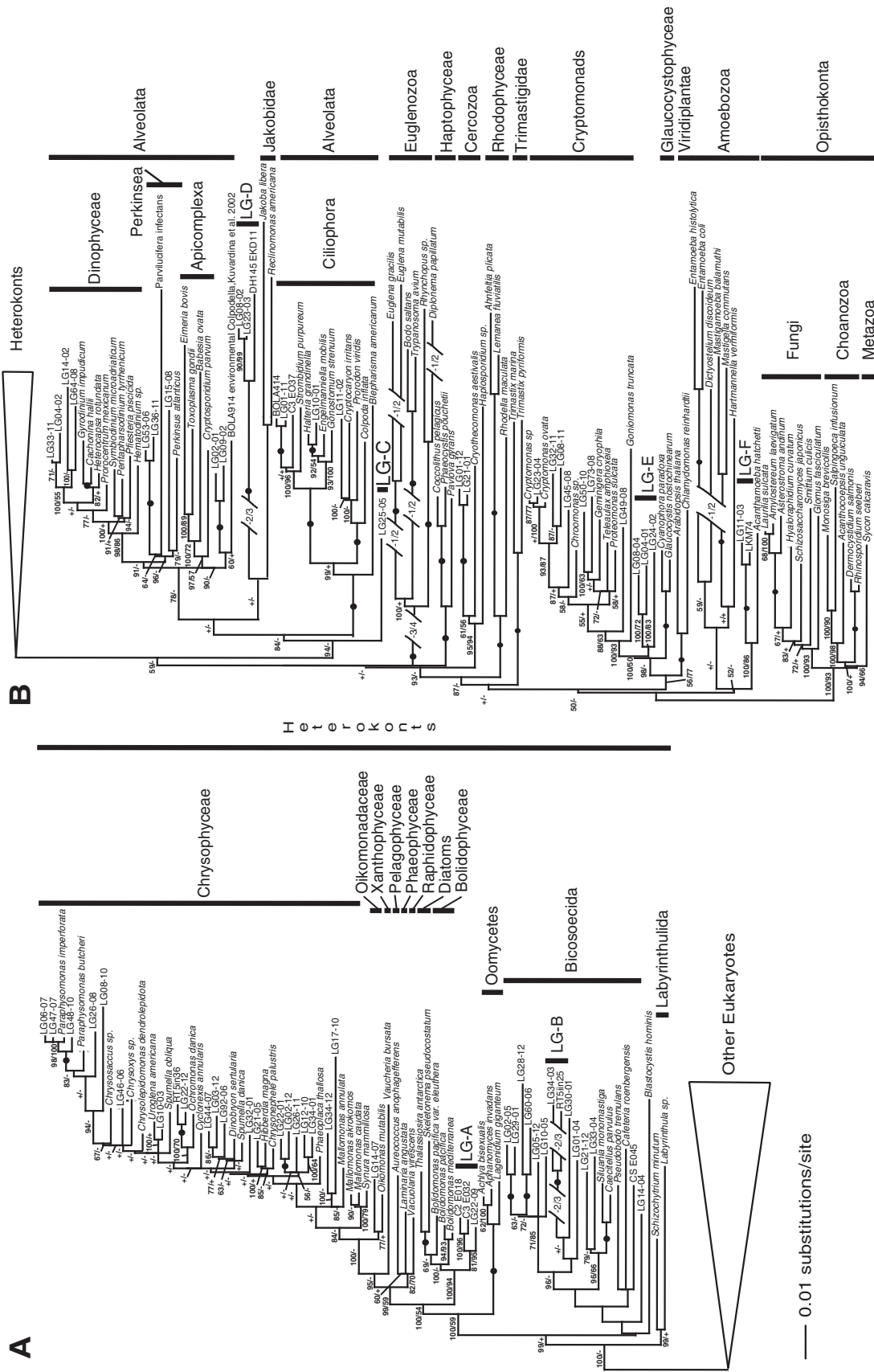


Fig. 3. Bayesian 18S rDNA phylogeny of picoeukaryotes sampled from Lake George.

A. The topology of the Heterokonta picoeukaryotes. B. The topology of the Heterokonta, Opisthokonta, and other eukaryotes. The Bayesian tree was inferred from an alignment of 184 taxa with a character sampling of 1295 nucleotides. The tree was calculated using MrBayes2 (Huelsenbeck and Ronquist, 2001) with the following settings: the ML model employed 6 substitution types, with base frequencies estimated from the data; rate variation across sites was modelled using a gamma distribution, with a proportion of sites being invariant; the MCMC search was run with 6 chains for 1000 000 generations, with trees being sampled every 100 generations (the first 5000 trees were discarded as 'burn-in'). Support for the Bayesian tree topology was evaluated using 1000 LogDet distance bootstraps. The parameters for the LogDet distance model (proportion of invariant sites) were calculated with ML tree scores in PAUP using the best parsimony tree. Tree support values are represented as X/Y, where X is the Bayesian posterior percentage probability and Y is the LogDet best bootstrap support. Values above 49% are shown. LogDet bootstrap values below 50%, but consistent with the Bayesian tree topology are labelled with +. When the topology of the LogDet bootstrap tree is incongruent, a - is labelled. A black dot indicates 100% support value in both methods. Proportional length reduction of shortened branches is labelled. The posterior probabilities are shown as percentage values. All taxa labelled with LG and a number were sampled from Lake George. Other phylogenies with alphanumeric names are from published environmental gene libraries (see Table 1).

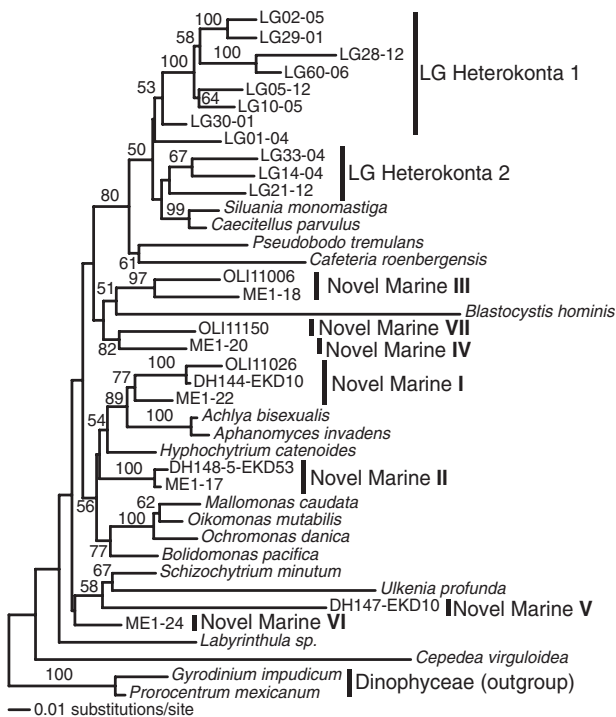


Fig. 4. Phylogenetic relationships between heterokonts sampled from Lake George and marine environments. The 18S rDNA distance tree was inferred from an alignment of 40 taxa. Character sampling comprised 1217 aligned nucleotides. The novel marine groups are specific Heterokonta groups pointed out by Massana and colleagues (2002). Distance methods (topology search and 1000 bootstrap replicates) utilized a GTR+ γ +I model of sequence evolution with heuristic searches by stepwise addition using the TBR settings in PAUP. Taxa labelled with LG and a number were sampled from Lake George. Other phylotypes with alphanumeric names are from published environmental gene libraries (see Table 1).

erotrophic heterokonts have also been shown to be abundant and diverse in marine habitats, with at least seven uncharacterized phylogenetic groups detected in coastal waters (Massana *et al.*, 2002). Our phylogenetic analysis (Fig. 4) showed that the putative heterotrophic heterokonts found in the Lake George environment did not associate with those sampled from coastal marine environments inferring that a number of freshwater and marine phylogenetic groups appear to be distinct.

The phylogenetic cluster LG-A grouped within heterokonts and formed a sister group to the Bolidophyceae, a recently discovered class of picoplankton algae, that forms a sister group to the diatoms (Guillou *et al.*, 1999). To date, only two bolidophycean species have been detected in marine environments. The Lake George sequence, LG22-09, may represent a third bolidophycean lineage, or a new sister group to the Bolidophyceae. It grouped (with moderate support) with two partial 18S rDNA sequences retrieved from a hydrothermal vent in Guaymas Basin (Gulf of California) (Edgcomb *et al.*,

2002) (see Figs 1 and 3). The discovery of a Bolidophyceae-like sequence in a freshwater ecosystem indicates that lineages related to Bolidophyceae have a wide environmental and biogeographical distribution.

The phylogenetic group LG-B includes two long-branch Heterokonta phylotypes, one of which has been detected in the acidic environment (\sim pH 2) of the river Rio Tinto in Spain (Amaral Zettler *et al.*, 2002). The placement of the group within the heterokonts was recovered in all analyses, albeit with a weak bootstrap support. Amaral Zettler *et al.* (2002) suggest that the Rio Tinto phylotype (RT5iin25) may represent a new 'stramenopile' lineage (here called heterokonts), and the data reported here agrees with these findings. The two sequences that form LG-B are divergent and occupy long branches suggesting that this putative heterokont group includes considerable sequence variation, provided their sisterhood is not an artefact of long-branch attraction (Stiller and Hall, 1999). This is supported by the discovery of eight closely related but not identical sequences in Lake George (Fig. 1). The detection of two representatives of this group in different global localities and in very different freshwater environments, suggests that they represent a globally dispersed group of unidentified protozoa. The placement of these long-branch sequences within the Heterokonta could be responsible for the low terminal support values for the heterokonts seen in these phylogenies, and as such the classification of this phylotype as a Heterokonta should be treated with caution.

Alveolates and new phylogenetic relatives

Eight Lake George phylotypes grouped with alveolate sequences with 60% or greater bootstrap support, representing a total of 10% of the 18S rDNA diversity recovered. The phylogenetic cluster LG-D formed a long branch within the alveolates and contained sequences sampled from Lake George, low oxygen salt marsh environments (Cape Cod, USA), and Antarctic marine environments (Lopez-Garcia *et al.*, 2001; Stoeck and Epstein, 2003), showing that this lineage was globally distributed and found in contrasting habitats. Published phylogenetic trees place this group as deep branching eukaryotes with no clear affinity to known higher-level taxonomic groups (Lopez-Garcia *et al.*, 2001; Stoeck and Epstein, 2003). Conversely, all our analyses (e.g. Figs 2 and 3) suggested that LG-D rather formed a long branch within the Alveolata, although with no affinity to characterized alveolate taxa. In the Bayesian analysis (Fig. 3), the group formed a weak sister relationship with Jakobidae, with the two groups clustering within alveolates with weak support. Previous phylogenetic analyses and morphological characters suggest that this position for Jakobidae is likely to be an artefact (Cavalier-Smith, 2003). When Jakobidae

sequences were excluded from the Bayesian analysis (data not shown) the LG-D group retained its placement within the alveolates, suggesting that LG-D is reliably placed within the alveolates (Cavalier-Smith, 2004). However, it is likely that the inclusion of this long-branch cluster within alveolates is resulting in a weak bootstrap support values for the monophyly of the alveolates. Consequently, the alveolate grouping of LG-D has to be treated with caution.

The Lake George sequence LG25-05 (group LG-C) showed no phylogenetic association with any known eukaryote group, and is likely to represent a previously unsampled group of protists. In the Bayesian analyses this sequence was confidently positioned between alveolates and heterokonts with 0.98 posterior probability (Fig. 3). However, this placement was not recovered in the phylogenetic analyses using distance methods (Figs 2 and 3). Instead, LG-C grouped within the alveolates with a weak bootstrap support and with no affinity to known alveolate taxa (Fig. 2). Further analyses that sampled two recently identified marine alveolate groups [marine alveolate group I and II (Lopez-Garcia *et al.*, 2001)], recovered no affiliation of LG-C with these alveolates and, in agreement with the Bayesian analyses, placed it below the alveolate radiation with a weak bootstrap support (Fig. 5). The phylogeny shown in Fig. 5 assesses whether any of the environmental alveolates detected in Lake George grouped with either of the marine alveolate group I or II (Lopez-Garcia *et al.*, 2001). Lake George sequences appeared unrelated to these unidentified marine alveolates illustrating potential differences in eukaryotic microbial diversity between marine and freshwater environments.

Cryptomonads and other eukaryote groups

Six Lake George phylotypes grouped with the cryptomonads (Figs 2 and 3) and three phylotypes grouped with the cryptomonad nucleomorphs with 60% or higher bootstrap support suggesting that 12% of the sequence diversity detected in Lake George originated from the cryptomonads (Fig. 6). The sequence LG49-08 occupies two differing phylogenetic positions in ML distance and Bayesian analyses. In the ML distance tree, LG49-08 was placed at the base of the Heliozoa (Fig. 2), although this association is weakly supported. When the Heliozoa were removed, it formed a long branch within cryptomonads, a relationship supported by 100% posterior probability and 93% LogDet distance bootstrap support (Fig. 3). A separate phylogenetic analysis with increased character sampling (29 taxa and 1335 nucleotides) and the addition of a long-branch sequence from a heliozoan-like organism isolated from a marine environment (GenBank Accession no. AF534711) (Cavalier-Smith and Chao, 2003a), recovered LG49-08 to occupy a position similar to the unnamed

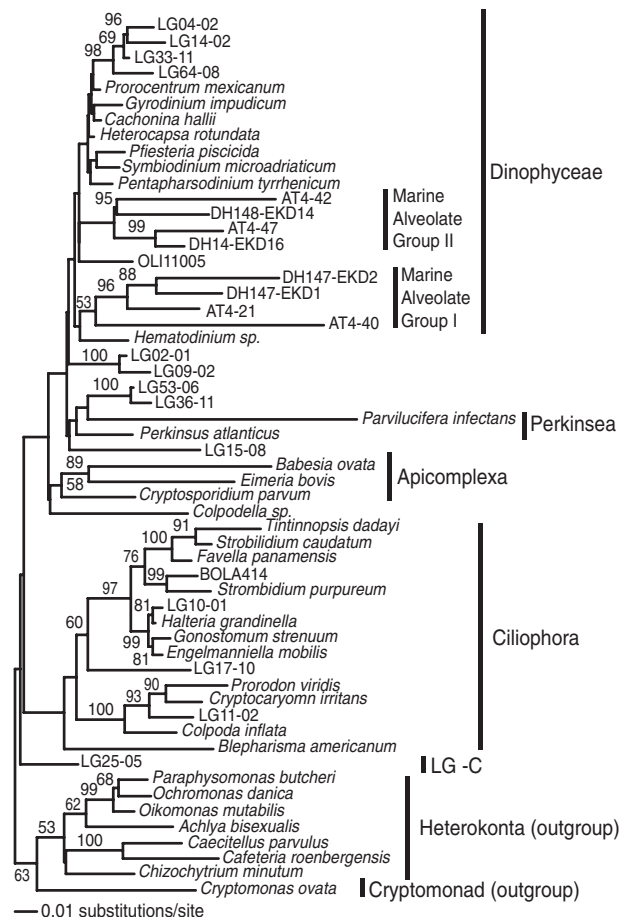


Fig. 5. Phylogenetic relationship between Alveolates sampled from Lake George and marine environments. The Marine Alveolate groups were pointed out by Lopez-Garcia and colleagues (2001). The 18S rRNA distance tree was inferred from an alignment of 56 taxa. Character sampling comprised 1217 aligned nucleotides. Distance methods (topology search and 1000 bootstrap replicates) utilized a GTR+ γ +I model of sequence evolution with heuristic searches by stepwise addition using the TBR settings in PAUP. Taxa labelled with LG and a number were sampled from Lake George. Other phylotypes with alphanumeric names are from published environmental gene libraries (see Table 1).

marine microheliozoan (Fig. 7), implying that this is a freshwater relative of this marine isolate.

Three of the 77 Lake George phylotypes clustered with sequences from the cryptomonad nucleomorph organelles. These organelles are relict nuclei and contain a remnant genome produced by the endosymbiosis of an alga and another eukaryote (Douglas *et al.*, 2001). The nucleomorph endosymbiont has undergone unique evolutionary pressures leading to the streamlining of these genomes and the evolution of highly divergent SSU gene sequences (Douglas *et al.*, 2001). Consequently, the phylogeny of the nucleomorph sequences were analysed separately (Fig. 6). The recovery of the highly variant

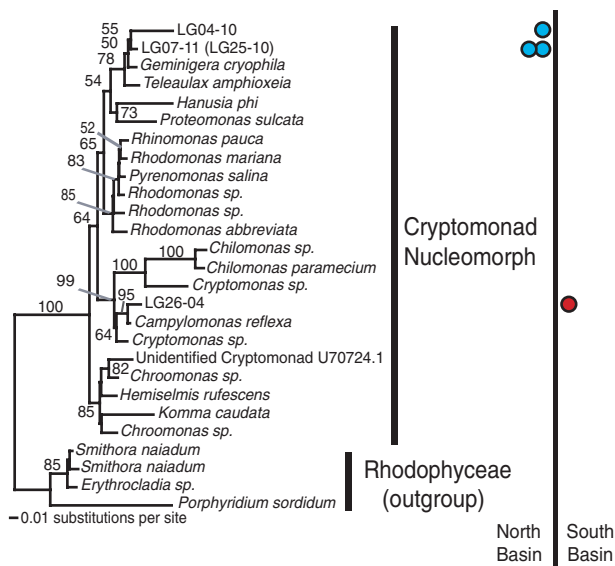


Fig. 6. SSU 18S rDNA phylogeny of cryptomonad nucleomorphs sampled from Lake George. The distance tree was inferred from an alignment of 27 taxa with a character sampling of 1270 nucleotides. Distance methods (topology search and 1000 bootstrap replicates) utilized a GTR+ γ +I model of sequence evolution with heuristic searches by stepwise addition using the TBR settings in PAUP. Taxa labelled with LG and a number were sampled from Lake George. Occurrence of each LG phylotype is marked with blue and red dots on the right, representing North and South Lake George basins respectively. Circled dots indicate a clone that had 99 or greater total sequence identity with the sequence included in the phylogeny.

nucleomorph sequences in our environmental gene library, attests to the effectiveness of the primers used to sample the picoplankton 18S rDNA diversity in Lake George.

The phylogenetic cluster LG-E did not associate with any known taxonomic group but was consistently attracted to either the cryptomonads or glaucocystophytes. Phylogenetic analyses with increased character sampling aimed to identify a consistent placement for LG-E and recovered a weakly supported sister relationship of LG-E and cryptomonads (Fig. 7). Further investigation using alternative gene data sets and morphological data are required to reliably identify the phylogenetic position of this group. It may also be important to investigate whether this group possesses a plastid organelle, as LG-E seems to be related to two taxonomic groups that include members with plastids.

Many ecologically significant and probably valid morphospecies of picoeukaryotes are missing from the current 18S rDNA database. It is therefore conceivable that a specific morphospecies which has no 18S sequence data and is abundant in the oligotrophic freshwater environments, such as the kathablepharids, may be related to LG-E or LG-D or possibly LG-C.

Bayesian analysis of the 18S phylogenetic tree

The Bayesian tree (Fig. 3) retrieves a number of phylogenetic relationships consistent with hypotheses of eukaryote evolutionary relationships. Among these are the recovery of a monophyletic Amoebozoan group (Baptiste *et al.*, 2002) (supported by 0.52 posterior probability), monophyly of the unikonts (Stechmann and Cavalier-Smith, 2003) (supported by 0.50 posterior probability), the grouping of the haplosporidia with the Cercozoa (Cavalier-Smith and Chao, 2003b) (supported by 0.95 posterior probability and 94% bootstrap support) and the sister group relationship of the Heterokonta and Alveolata (Fast *et al.*, 2001) (supported by 0.59 posterior probability). Many of these relationships are weakly supported and could be an indirect product of streamlining the data set to remove taxa that occupy unresolved positions, and which on the whole were not attracted to the Lake George phylotypes (see LG49-08 and previous discussion for the only exception). However, these considerations aside, the resulting topology has a number of relationships which are supported by compelling evidence from alternative pieces of genetic data, i.e. the monophyly of the Amoebozoa (Baptiste *et al.*, 2002), and the sisterhood of the Alveolata and the Heterokonta (Fast *et al.*, 2001). Although some of these relationships have been seen in other 18S rRNA gene analyses (e.g. Van de Peer *et al.*, 2000), the distance analyses on the same data set did not recover these

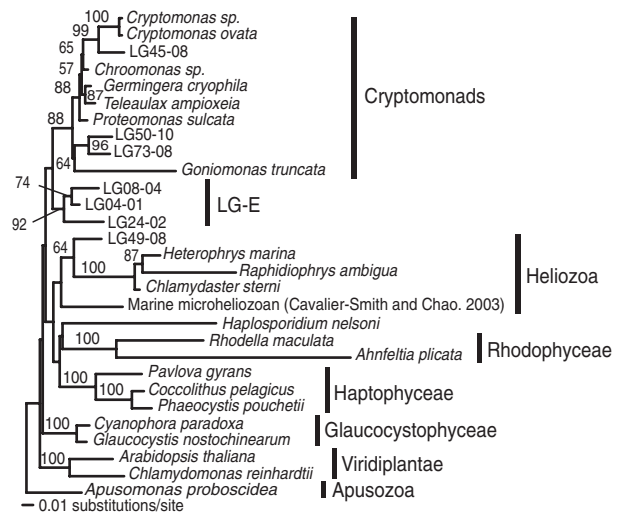


Fig. 7. Phylogenetic placements of the cluster LG-E and the phylotype LG49-08. The 18S rRNA distance tree was inferred from an alignment of 29 taxa. Character sampling comprised 1335 aligned nucleotides. Distance methods (topology search and 1000 bootstrap replicates) utilized a GTR+ γ +I model of sequence evolution with heuristic searches by stepwise addition using the TBR settings in PAUP. A black dot on a branch indicates 100% bootstrap support value. Taxa labelled with LG and a number were sampled from Lake George. Other phylotypes with alphanumeric names are from published environmental gene libraries (see Table 1).

relationships, indicating there are improvements to be gained from using the Bayesian approach.

Global distribution of picoeukaryotes

The phylogenetic cluster LG-F includes two unknown amoebozoan sequences: one from Lake George and one from a synthetic detritus environment (van Hannen *et al.*, 1999). The two sequences share 99% similarity. This is a higher level of sequence similarity than sometimes found in 18S SSU rRNA paralogues from the same species (Stothard *et al.*, 2000; Pawlowski *et al.*, 2002). It is likely that LG-F represents an amoeba species, or closely related species, living in two similar but globally disparate environments (van Hannen *et al.*, 1999). Indeed, there is evidence that some genetically identical microbial strains are ubiquitous (Bowers *et al.*, 1998), and that the small size and abundance of protozoa is conducive to random passive dispersal processes and transportation of protozoan cells by animal vectors around the globe at a faster rate than it takes strains to become genetically differentiated (Finlay *et al.*, 2001; Finlay, 2002). However, tracking microbial biogeography using DNA sequence characters is influenced by the sensitivity of the genetic marker used, and alternative markers may reveal distinct evolutionary differences. The discovery of cosmopolitan picoeukaryotic lineages in Lake George waters and other environments, as observed in the groups LG-A, LG-D, LG-F, LG-G, LG-H and LG-I (Figs 1 and 2), is indicative of a process of continual dispersal of microbial lineages around the globe and into differing environments, and is consistent, at least at the level of 18S rRNA phylogenetic lineages, with the Baas-Becking's axiom that 'everything is everywhere' (Staley and Gosink, 1999). Conversely, as figures four and five demonstrate, certain aspects of marine and freshwater biodiversity appear to be unique to particular environment types, supporting Baas-Becking's second proposition that the 'environment selects' (Staley and Gosink, 1999). However, these observations rely upon sampling saturation, which has currently not been achieved in any environment. Increased environmental gene sampling will enable the adequate testing of the speculations that smaller species tend to have wider distribution and higher dispersal efficiency, lower rate of allopatric speciation, and lower rates of local extinction.

Experimental procedures

Sampling

Picoplankton was collected in the summer of 1996 from the euphotic portion of the water column at two sites in the North basin and two sites in the South basin of Lake George (43°22-51'N and 73°21-47'W). Samples were collected at the light penetration level of 1%, mean sampling

depth being 20 m in the North and 16 m in the South. In total, 12 1-l water samples were collected. The samples were prefiltered through a nylon mesh, filtered through 5- μ m cellulose membrane filters (Millipore), and the remaining plankton collected on 0.2- μ m cellulose membrane filters (Millipore).

Environmental gene library construction and sequencing

Following enzymatic cell lyses, total community DNA was phenol-chloroform extracted and further purified using QIAamp DNA Kit (QIAGEN). SSU rDNA was selectively amplified by polymerase chain reaction (PCR) from the total community DNA pool using degenerate microbial eukaryote specific primers: 3Fphp 5'-CTGGTTGATCCTGCCAGT-3', and 1749Rphp 5'-ACCTTGTTACGACTTCWCC-3' (numbers based on rDNA of *Chlorella kessleri* SAG211-11 g, GenBank Accession no. X56105). The reaction and cycling conditions were as follows: 85 ng template DNA, 1 \times *Taq*Extender PCR buffer (Stratagene), 1.5 mM MgCl₂, 200 μ M of each dNTP (Amersham Pharmacia Biotech), 0.1 μ M of each primer, 0.8 μ g of T4 gene 32 Protein (Ambion), 1.5 U of *Taq*Extender (Stratagene), and 1.5 U of *Taq* DNA polymerase (Sigma) were combined in a total volume of 50 μ l and incubated for 4 min at 94°C, followed by 35 cycles of 45 s (s) at 94°C, 45 s at 57°C, 135 s at 72°C with an additional 12-min extension at 72°C. Amplified 18S rDNA was gel-extracted, purified with QIAquick spin columns (QIAGEN), and cloned using TOPO TA Cloning Kit (Invitrogen). Plasmid DNA was prepared using QIAprep spin columns (QIAGEN).

The clones were characterized by *Hae*III RFLP. Plasmids bearing complete inserts were used as templates to PCR-amplify cloned rDNA. The PCR and cycling conditions were essentially the same as above with some changes: 30 ng template DNA and a standard 10 \times PCR reaction buffer (Sigma) were used; T4 gene 32 Protein, *Taq*Extender were excluded, and the number of cycles reduced to 25. Twelve μ l of non-purified PCR products were digested with 3 U of the restriction endonuclease *Hae*III for 2 h. The restriction fragments were resolved by gel electrophoresis in 0.7% Agarose-1000 (Gibco)/2% Synergel (Diversified Biotech) (a functional equivalent of 5% agarose gel) in 1 \times TAE buffer, and stained with 0.5 μ g ml⁻¹ ethidium bromide. A total of 414 rDNA clones were screened and sorted by RFLP. The diversity was further evaluated by rarefaction analysis using the Analytic Rarefaction software (Analytic Rarefaction software provided by Dr Steven Holland at <http://www.uga.edu/~strata/software/Software.html>). Representative clones were sequenced entirely by automated sequencing using the ABI Model 373 A Stretch with the XL upgrade automated DNA sequencer (Applied Biosystems), standard M13 primer pair and the internal primer 3F (Bird *et al.*, 1992).

Sequence analysis

To correct for sampling artefacts, highly similar sequences, with an excess of 99% sequence similarity were excluded. These similar sequences may originate from identical genes, and minor variances (< 1%) observed could possibly be a product of sequencing error. Alternatively, highly similar

sequences might potentially originate from single organisms that possess multiple variant SSU rRNA genes. Environmental sequences were examined for potential chimerical artefacts using CHECK-CHIMERA in the Ribosomal Database Project (Cole *et al.*, 2003) and by manual alignment (Berney *et al.*, 2004). Lake George environmental 18S SSU rDNA sequences were first collated with the GenBank database using BLASTN analysis to assess superficially their phylogenetic affiliation; then aligned, using CLUSTAL X (Thompson *et al.*, 1997), with a comprehensive set of database sequences that were selected on the basis of two criteria: (i) The BLASTN best scoring sequences for each Lake George sequence; (ii) Sequences known to represent diverse, unique and intermediate phylogenetic positions on the SSU rRNA evolutionary tree were sampled based upon published analyses. These often included sequences from environmental gene library experiments (Table 1). In addition, SSU

sequences sampled from publicly available environmental gene libraries were sampled based upon BLAST similarity to sequences recovered from Lake George (Table 1). Although a taxonomically broad sample of eukaryotes was included in the analyses, a number of long-branching eukaryote groups, such as some diplomonads, some parabasalids, foraminifera, and microsporidians, were removed, because, in preliminary analyses, the Lake George sequences did not cluster with these groups. Additionally, these eukaryotic groups are characterized by high rates of gene sequence evolution, often rendering their positioning within phylogenies ambiguous (Embley and Hirt, 1998; Stiller and Hall, 1999) and the inclusion of these sequences within the alignment would reduce the number of characters that could be confidently sampled. The alignments were manually refined using the Genetic Data Environment software and masked to remove rejoins of the alignment that were ambiguous and could not be aligned

Table 1. Published environmental 18S rDNA sequences included in the phylogenetic analyses

Environmental clone	Sampled environment	Closest Lake George sequence ^a or reason for inclusion	Accession no.	Reference
CCW100	hypoxic marine waters	LG23-03 (94%)	AY180041	Stoeck and Epstein (2003)
CCW75	hypoxic marine waters	LG23-03 (96%)	AY180032	Stoeck and Epstein (2003)
CCW1	hypoxic marine waters	LG22-09 (95%)	AY180007	Stoeck and Epstein (2003)
CCW27	hypoxic marine waters	LG48-10 (97%)	AY180017	Stoeck and Epstein (2003)
CCW83	hypoxic marine waters	LG21-12 (92%)	AY180035	Stoeck and Epstein (2003)
CCI40	hypoxic waters/sediment	LG21-05 (91%)	AY179989	Stoeck and Epstein (2003)
C2_E015	hydrothermal vent	LG20-09 (96%)	AY046805	Edgcomb and colleagues (2002)
C2_E018	hydrothermal vent	LG22-09 (94%)	AY046808	Edgcomb and colleagues (2002)
C3_E032	hydrothermal vent	LG22-09 (94%)	AY046872	Edgcomb and colleagues (2002)
C3_E037	hydrothermal vent	LG01-11 (95%)	AY046865	Edgcomb and colleagues (2002)
CS_E045	hydrothermal vent	LG30-01 (91%)	AY046666	Edgcomb and colleagues (2002)
RT5iin25	acidic river waters (pH 2)	LG34-03 (92%)	AY082983	Amaral Zettler and colleagues (2002)
RT5iin36	acidic river waters (pH 2)	LG22-12 (95%)	AY082999	Amaral Zettler and colleagues (2002)
UEPAC45p4	Pacific coastal site	intermediate branch	AY129064	A.Z. Worden, unpublished
UEPAC05Cp2	Pacific coastal site	intermediate branch	AY129060	A.Z. Worden, unpublished
AT4-98	hydrothermal vent	LG53-03 (93%)	AF530536	Lopez-Garcia and colleagues (2003)
AT4-42	hydrothermal vent	Maine Alveolata II	AF530537	Lopez-Garcia and colleagues (2003)
AT4-47	hydrothermal vent	Maine Alveolata II	AF530539	Lopez-Garcia and colleagues (2003)
AT4-21	hydrothermal vent	Maine Alveolata I	AF530532	Lopez-Garcia and colleagues (2003)
AT4-40	hydrothermal vent	Maine Alveolata I	AF530540	Lopez-Garcia and colleagues (2003)
BOLA914	anoxic sediment	LG02-01 (92%)	AF372772	Dawson and Pace (2002)
BOLA414	anoxic sediment	LG01-11 (95%)	AF372793	Dawson and Pace (2002)
BOLA383	anoxic sediment	intermediate branch	AF372765	Dawson and Pace (2002)
DH145-EKD11	deep-sea Antarctic	LG23-03 (93%)	AF290065	Lopez-Garcia and colleagues (2001)
DH144-EKD10	deep-sea Antarctic	Marine Heterokonta I	AF290063	Lopez-Garcia and colleagues (2001)
DH147-EKD16	deep-sea Antarctic	Maine Alveolata II	AF290071	Lopez-Garcia and colleagues (2001)
DH147-EKD10	deep-sea Antarctic	Marine Heterokonta V	AF290070	Lopez-Garcia and colleagues (2001)
DH147-EKD2	deep-sea Antarctic	Maine Alveolata I	AF290071	Lopez-Garcia and colleagues (2001)
DH147-EKD1	deep-sea Antarctic	Maine Alveolata I	AF290040	Lopez-Garcia and colleagues (2001)
DH148-5-EKD53	deep-sea Antarctic	Marine Heterokonta II	AF290083	Lopez-Garcia and colleagues (2001)
DH148-EKD6	deep-sea Antarctic	Maine Alveolata II	AF290055	Lopez-Garcia and colleagues (2001)
D226	anoxic marine waters	intermediate branch	AY256332	Stoeck and Epstein (2003)
ME1-5	Oceanic	LG73-08 (98%)	AF363183	Diez and colleagues (2001b)
ME1-17	Oceanic	Marine Heterokonta II	AF363186	Diez and colleagues (2001b)
ME1-18	Oceanic	Marine Heterokonta III	AF363187	Diez and colleagues (2001b)
ME1-20	Oceanic	Marine Heterokonta IV	AF363189	Diez and colleagues (2001b)
ME1-22	Oceanic	Marine Heterokonta I	AF363191	Diez and colleagues (2001b)
ME1-24	Oceanic	Marine Heterokonta VI	AF363207	Diez and colleagues (2001b)
LKM74	synthetic detritus	LG11-03 (99%)	AJ130863	van Hannen and colleagues (1999)
OLI11006	oligotrophic marine waters	Marine Heterokonta III	AJ402357	Moon-van der Staay and colleagues (2001)
OLI11150	oligotrophic marine waters	Marine Heterokonta VII	AJ402355	Moon-van der Staay and colleagues (2001)
OLI11026	oligotrophic marine waters	Marine Heterokonta I	AJ402339	Moon-van der Staay and colleagues (2001)
OLI11005	oligotrophic marine waters	intermediate branch	AJ402349	Moon-van der Staay and colleagues (2001)

a. Similarity score in parentheses is given for a highest scoring segment of 500 bp or larger.

with confidence. After the alignments were masked, Lake George sequences with less than 1% sequence variance were grouped so that only one representative of each group was included in the final analyses. The final alignments used for phylogenetic analyses comprised a total of 234 taxa [207 taxa in the general eukaryote tree, Figs 1 and 2, and 27 in the cryptomonad nucleomorph tree (Fig. 6)]. Parsimony phylogenetic trees were calculated using PAUP (Swofford, 2002) with 10 random heuristic searches by stepwise addition. Using the best parsimony tree, ML tree scores were used to calculate a general time reversible rate substitution matrix, nucleotide base frequencies, proportion of invariant sites and gamma distribution (four rate categories). These model parameters were then used for ML distance analysis (GTR+ γ +I). ModelTest (Posada and Crandall, 1998) confirmed that GTR+ γ +I was the most appropriate model for the given data. The ML distance tree topology was calculated using 100 random replicate heuristic searches with stepwise addition using the tree bisection reconnection method (TBR), and the best tree was used for the phylogenetic tree topology. Support for this tree topology was assessed using 1000 ML distance bootstrap replicates with the same parameters used to obtain the tree topology but with single heuristic searches for each bootstrap replicate (Figs 1 and 2 for nuclear SSU tree and Fig. 6 for cryptomonad nucleomorph SSU phylogeny). The ML distance tree contained a number of long branches that did not group with sequences from Lake George. To further assess the validity of the conclusions derived from the ML distance tree, Bayesian analysis was performed using MrBayes2 (Huelsenbeck and Ronquist, 2001). The taxon sampling was reduced to remove some long-branch taxa and the character sampling was adjusted accordingly based upon a manual alteration of the alignment (1295 characters and 184 taxa). The MrBayes settings were as follows. The ML model employed six substitution types, with base frequencies estimated from the data. Rate variation across nucleotide sites was modelled using a gamma distribution, with a proportion of sites being invariant. The Markov chain Monte Carlo search was run with six chains for 1000 000 generations, with trees being sampled every 100 generations (the first 5500 trees were discarded as 'burn-in'). Support for the Bayesian tree topology was evaluated using 1000 LogDet distance bootstraps. The parameters for the LogDet distance model (proportion of invariant sites) were calculated with ML tree scores in PAUP using the best parsimony tree. LogDet was used here as the tree clearly suffered from long branch problems that in some cases can be caused by compositional heterogeneity (e.g. Lake, 1994). Three further phylogenetic analyses were undertaken investigating specific parts of the eukaryote phylogeny (Figs 4, 5 and 7). Maximum likelihood distance methods were used as before for both topology search and 1000 bootstrap replicates. All bootstrapping, ML-distance topology searches and Bayesian analyses were run on the BRAIN 64-processor multiclust computer.

Nucleotide sequence accession numbers

Nucleotide sequences obtained in this study have been deposited in GenBank under Accession numbers AY919677–AY919829.

Acknowledgements

James McInerney (NUI Maynooth) is acknowledged for providing the BRAIN multiclust computer. We acknowledge Sharon Danielsen and the undergraduate students Thomas Flint, Jennifer Reineke and Tommy Thomas for technical assistance and Giselle Walker for comments regarding this manuscript. This study was supported by the Helen V. Froehlich Foundation and the New York Center for Studies on the Origins of Life, a NASA Specialized Center of Research and Training. T.A.R. was supported by a BBSRC studentship and a NASA Planetary Biology Internship.

Notes Added in Proof

Highly divergent SSU rRNA gene sequences of two commonly encountered planktonic ciliates (Alveolata), *Myrionecta rubra* (GenBank Accession no. AY587129) and *Mesodinium pulex* (GenBank Accession no. AY587130), have recently been published [Johnson, M.D. *et al.* (2004) *Protist* **155**: 347–359]. The two sequences appear to be very similar to the Lake George sequences grouped in the cluster LG-D (Fig. 2). It confirms our and Cavalier-Smith's (2004) analyses that this group is a member of the Alveolata and not a deep branching eukaryotic lineage with no clear affinity to known eukaryotic taxa as previously suggested. Analyses provided by Johnson *et al.* (2004) also confirmed an alveolate placement for this lineage, as well as a pattern of cosmopolitan distribution consistent with our analyses published here.

References

- Amaral Zettler, L.A., Gomez, F., Zettler, E., Keenan, B.G., Amils, R., and Sogin, M.L. (2002) Microbiology: eukaryotic diversity in Spain's River of Fire. *Nature* **417**: 137.
- Arndt, H., Dietrich, D., Auer, B., Cleven, E.J., Gräfenhan, T., Weitere, M., and Mylnikov, A.P. (2000) Functional Diversity of Heterotrophic Flagellates in Aquatic Ecosystems. In *The Flagellates*. Leadbeater, B.S.C. and Green, J.C. (eds). London, UK: Taylor & Francis, pp. 240–268.
- Baptiste, E., Brinkmann, H., Lee, J.A., Moore, D.V., Sensen, C.W., Gordon, P., *et al.* (2002) The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. *Proc Natl Acad Sci USA* **99**: 1414–1419.
- Berney, C., Fahrni, J., and Pawlowski, J. (2004) How many novel eukaryotic 'kingdoms'? Pitfalls and limitations of environmental DNA surveys. *BMC Biol* **2**: 13.
- Bird, C.J., Rice, E.L., Murphy, C.A., and Ragan, M.A. (1992) Phylogenetic relationships in the Gracilariales (Rhodophyta) as determined by 18S rDNA sequences. *Phycologia* **31**: 510–522.
- Bowers, N., Kroll, T.T., and Pratt, J.R. (1998) Diversity and geographic distribution of riboprints from three cosmopolitan species of Colpoda Müller (Ciliophora, Colpodea). *European J Protistol* **34**: 341–347.
- Cavalier-Smith, T. (2003) The excavate protozoan phyla Metamonada Grasse emend. (Anaeromonadea, Parabasalialia, *Carpodiomonas*, Eopharyngia) and Loukozoa emend. (Jakobea, *Malawimonas*): their evolutionary affini-

- ties and new higher taxa. *Int J Syst Evol Microbiol* **53**: 1741–1758.
- Cavalier-Smith, T. (2004) Only six kingdoms of life. *Proc R Soc Lond B Biol Sci* **271**: 1251–1263.
- Cavalier-Smith, T., and Chao, E.E. (2003a) Molecular phylogeny of centrohelid heliozoa, a novel lineage of bikont eukaryotes that arose by ciliary loss. *J Mol Evol* **56**: 387–396.
- Cavalier-Smith, T., and Chao, E.E. (2003b) Phylogeny and classification of phylum Cercozoa (Protozoa). *Protist* **154**: 341–358.
- Cole, J.R., Chai, B., Marsh, T.L., Farris, R.J., Wang, Q., Kulam, S.A., *et al.* (2003) The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Res* **31**: 442–443.
- Dawson, S.C., and Pace, N.R. (2002) Novel kingdom-level eukaryotic diversity in anoxic environments. *Proc Natl Acad Sci USA* **99**: 8324–8329.
- Diez, B., Pedros-Alio, C., Marsh, T.L., and Massana, R. (2001a) Application of denaturing gradient gel electrophoresis (DGGE) to study the diversity of marine picoeukaryotic assemblages and comparison of DGGE with other molecular techniques. *Appl Environ Microbiol* **67**: 2942–2951.
- Diez, B., Pedros-Alio, C., and Massana, R. (2001b) Study of genetic diversity of eukaryotic picoplankton in different oceanic regions by small-subunit rRNA gene cloning and sequencing. *Appl Environ Microbiol* **67**: 2932–2941.
- Douglas, S., Zauner, S., Fraunholz, M., Beaton, M., Penny, S., Deng, L.T., *et al.* (2001) The highly reduced genome of an enslaved algal nucleus. *Nature* **410**: 1091–1096.
- Edgcomb, V.P., Kysela, D.T., Teske, A., de Vera Gomez, A., and Sogin, M.L. (2002) Benthic eukaryotic diversity in the Guaymas Basin hydrothermal vent environment. *Proc Natl Acad Sci USA* **99**: 7658–7662.
- Embley, T.M., and Hirt, R.P. (1998) Early branching eukaryotes? *Curr Opin Genet Dev* **8**: 624–629.
- Fast, N.M., Kissinger, J.C., Roos, D.S., and Keeling, P.J. (2001) Nuclear-encoded, plastid-targeted genes suggest a single common origin for apicomplexan and dinoflagellate plastids. *Mol Biol Evol* **18**: 418–426.
- Finlay, B.J. (2002) Global dispersal of free-living microbial eukaryote species. *Science* **296**: 1061–1063.
- Finlay, B.J., Esteban, G.F., Clarke, K.J., and Olmo, J.L. (2001) Biodiversity of terrestrial protozoa appears homogeneous across local and global spatial scales. *Protist* **152**: 355–366.
- Guillou, L., Moon-Van Der Staay, S.Y., Claustre, H., Partensky, F., and Vaulot, D. (1999) Diversity and abundance of Bolidophyceae (Heterokonta) in two oceanic regions. *Appl Environ Microbiol* **65**: 4528–4536.
- van Hannen, E.J., Mooij, W., van Agterveld, M.P., Gons, H.J., and Laanbroek, H.J. (1999) Detritus-dependent development of the microbial community in an experimental system: qualitative analysis by denaturing gradient gel electrophoresis. *Appl Environ Microbiol* **65**: 2478–2484.
- Huelsenbeck, J.P., and Ronquist, F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**: 754–755.
- Kuvaridina, O.N., Leander, B.S., Aleshin, V.V., Myl'nikov, A.P., Keeling, P.J., and Simdyanov, T.G. (2002) The phylogeny of colpodellids (Alveolata) using small subunit rRNA gene sequences suggests they are the free-living sister group to apicomplexans. *J Eukaryot Microbiol* **49**: 498–504.
- Lake, J.A. (1994) Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. *Proc Natl Acad Sci USA* **91**: 1455–1459.
- Lee, J.J., Leedale, G.F., and Bradbury, P. (2000) *An Illustrated Guide to the Protozoa*, 2nd edn. Lawrence, KS, USA: Society of Protozoologists.
- Lopez-Garcia, P., Philippe, H., Gail, F., and Moreira, D. (2003) Autochthonous eukaryotic diversity in hydrothermal sediment and experimental microcolonizers at the Mid-Atlantic Ridge. *Proc Natl Acad Sci USA* **100**: 697–702.
- Lopez-Garcia, P., Rodriguez-Valera, F., Pedros-Alio, C., and Moreira, D. (2001) Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* **409**: 603–607.
- Massana, R., Guillou, L., Diez, B., and Pedros-Alio, C. (2002) Unveiling the organisms behind novel eukaryotic ribosomal DNA sequences from the ocean. *Appl Environ Microbiol* **68**: 4554–4558.
- Monheimer, R.H., and Baker, M.D. (1982) Phytoplankton community changes in Lake George. In *The Lake George Ecosystem*, Vol. 2. Schadler, M.H. (ed.). Lake George, NY, USA: The Lake George Association, pp. 41–47.
- Moon-van der Staay, S.Y., De Wachter, R., and Vaulot, D. (2001) Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* **409**: 607–610.
- Moreira, D., and Lopez-Garcia, P. (2002) The molecular ecology of microbial eukaryotes unveils a hidden world. *Trends Microbiol* **10**: 31–38.
- Pace, N.R. (1997) A molecular view of microbial diversity and the biosphere. *Science* **276**: 734–740.
- Pawlowski, J., Fahrni, J., and Bowser, S.S. (2002) Phylogenetic analysis and genetic diversity of *Notodendrodes hyalinosphaira*. *J Foraminiferal Res* **32** (2): 173–176.
- Posada, D., and Crandall, K.A. (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**: 817–818.
- Staley, J.T., and Gosink, J.J. (1999) Poles apart: biodiversity and biogeography of sea ice bacteria. *Annu Rev Microbiol* **53**: 189–215.
- Stechmann, A., and Cavalier-Smith, T. (2003) The root of the eukaryote tree pinpointed. *Curr Biol* **13**: R665–R666.
- Stiller, J.W., and Hall, B.D. (1999) Long-branch attraction and the rDNA model of early eukaryotic evolution. *Mol Biol Evol* **16**: 1270–1279.
- Stoeck, T., and Epstein, S. (2003) Novel Eukaryotic Lineages Inferred from Small-Subunit rRNA Analyses of Oxygen-Depleted Marine Environments. *Appl Environ Microbiol* **69**: 2657–2663.
- Stothard, J.R., Brémond, P., Andriamaro, L., Loxton, N.J., Sellin, B., Sellin, E., and Rollinson, D. (2000) Molecular characterization of the freshwater snail *Lymnaea natalensis* (Gastropoda: Lymnaeidae) on Madagascar with an observation of an unusual polymorphism in ribosomal small subunit genes. *J Zool Lond* **252**: 303–315.

- Swofford, D.L. (2002) PAUP*: Phylogenetic analysis using parsimony (*and other methods). 4.0 Beta. [WWW document] URL <http://www.sinauer.com/detail.php?id=8060/>
- Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**: 4876–4882.
- Van de Peer, Y., Baldauf, S.L., Doolittle, W.F., and Meyer, A. (2000) An updated and comprehensive rRNA phylogeny of (crown) eukaryotes based on rate-calibrated evolutionary distances. *J Mol Evol* **51**: 565–576.