THE ROYAL SOCIETY

Received 13 November 2003
Accepted 3 February 2004
Published online 17 May 2004

Only six kingdoms of life

Thomas Cavalier-Smith

Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK (*tom.cavalier-smith@zoo.ox.ac.uk*)

There are many more phyla of microbes than of macro-organisms, but microbial biodiversity is poorly understood because most microbes are uncultured. Phylogenetic analysis of rDNA sequences cloned after PCR amplification of DNA extracted directly from environmental samples is a powerful way of exploring our degree of ignorance of major groups. As there are only five eukaryotic kingdoms, two claims using such methods for numerous novel 'kingdom-level' lineages among anaerobic eukaryotes would be remarkable, if true. By reanalysing those data with 167 known species (not merely 8–37), I identified relatives for all 8–10 'mysterious' lineages. All probably belong to one of five already recognized phyla (Amoebozoa, Cercozoa, Apusozoa, Myzozoa, Loukozoa) within the basal kingdom Protozoa, mostly in known classes, sometimes even in known orders, families or genera. This strengthens the idea that the ancestral eukaryote was a mitochondrial aerobe. Analogous claims of novel bacterial divisions or kingdoms may reflect the weak resolution and grossly non-clock-like evolution of ribosomal rRNA, not genuine phylum-level biological disparity. Critical interpretation of environmental DNA sequences suggests that our overall picture of microbial biodiversity at phylum or division level is already rather good and comprehensive and that there are no uncharacterized kingdoms of life. However, immense lower-level diversity remains to be mapped, as does the root of the tree of life.

Keywords: kingdoms of life; Protozoa; gregarines; Amoebozoa; environmental PCR; eukaryote phylogeny

1. INTRODUCTION

All six kingdoms of life contain both unicellular and macroscopically visible organisms (Cavalier-Smith 1998). However, the higher-level classification of microbes has lagged considerably behind that of macro-organisms. Yet, in the revised six-kingdom system (Cavalier-Smith 2002*a*, 2003*a*), 34 out of the 57 living phyla consist entirely or largely of unicellular species (see table 1). With the notable exceptions of Cyanobacteria and Spirochaetes, which are sufficiently distinctive morphologically to be easily recognized microscopically, establishing the major divisions or phyla of bacteria has depended on being able to culture them, which has not yet been achieved for the majority of microbes. This has raised the question of how complete at the level of kingdoms and phyla is our inventory of the Earth's microbial biodiversity. A powerful way of exploring this question is to sequence well-conserved and phylogenetically informative genes from uncultured samples of soil or water taken directly from the environment. This approach was pioneered with bacteria using PCR primers specific for the 16S rRNA to amplify environmental DNA extracts and to generate gene libraries that could be cloned and sequenced at random (Giovannoni *et al.* 1990). Numerous lineages that are highly divergent from those known from cultures have thus been identified. A few of these are now being characterized and this process has already revealed quite novel groups of bacteria of considerable ecological, physiological and systematic importance (Morris *et al.* 2002; Sait *et al.* 2002).

Claims that the apparently most deeply diverging of such lineages, e.g. the 14 eubacterial ones found in hot springs (Hugenholtz *et al.* 1998) or korarcheota (Barns *et al.* 1996), represent undescribed divisions (phyla) or kingdoms are probably not justified. Taxonomic rank depends on phenotypic disparity (Cavalier-Smith 1998); asserting it from unidentified sequences on trees is fundamentally unsound. This is especially true for trees based on rRNA genes, which contrary to widespread assumptions are grossly non-clock-like and suffer from major systematic biases in evolutionary mode that can result in radically incorrect topologies with long branches being placed far too deeply in the tree (Philippe 2000; Cavalier-Smith 2002*b*). Large systematic biases are also found in many protein trees (see Cavalier-Smith 2002*b*; Gribaldo & Philippe 2002). Some proteins, e.g. Hsp90 (Stechmann & Cavalier-Smith 2003*a*), appear to be much more clock-like and less misleading in this respect. The gross errors possible with rRNA trees were first revealed for microsporidia, which early rRNA trees implied were the most divergent of all eukaryotes (Vossbrinck *et al.* 1987). Studies of numerous proteins and the analysis of a complete microsporidian genome have shown conclusively that microsporidia are highly derived secondarily amitochondrial fungi and not early-branching eukaryotes, and that their deeper branching position on even the best 18S rRNA trees obtained so far (Van de Peer *et al.* 2000) is a mathematical artefact (Roger 1999; Cavalier-Smith 2000*a*, 2002*b*; Williams *et al.* 2002; Keeling 2003). Accordingly microsporidia are now classified within the kingdom Fungi, not Protozoa (Cavalier-Smith 1998, 2000*b*); molecular and cytological evidence indicates that they evolved from an ancestor within the lower fungal phylum Archemycota, possibly a parasitic trichomycete (Cavalier-Smith 2000*b*; Keeling 2003). The degree of misplacement of microsporidia on the rRNA tree is so great that it far exceeds the signal supporting the positions of most other species within the tree, so the position of any long-branch clade on the tree must be treated with

Table 1. The six kingdoms of life and the 34 microbial phyla (based on Cavalier-Smith 1998, 2002*a*, 2003*a*,*b*).

empire PROKARYOTA (Cavalier-Smith 2002*b*)
- **kingdom Bacteria**
  - subkingdom Negibacteria (phyla Eobacteria, Sphingobacteria, Spirochaetae, Proteobacteria, Planctobacteria, Cyanobacteria)
  - subkingdom Unibacteria (phyla Posibacteria, Archaebacteria)

empire EUKARYOTA (Cavalier-Smith 1998)
- **kingdom Protozoa** (Cavalier-Smith 2002*a*, 2003*a*)
  - subkingdom Sarcomastigota (phyla Amoebozoa, Choanozoa)
  - subkingdom Biciliata
    - infrakingdom Rhizaria (phyla Cercozoa, Foraminifera, Radiozoa)
    - infrakingdom Excavata (phyla Loukozoa, Percolozoa, Euglenozoa, Metamonada; the latter now includes Parabasalia and Anaeromonadea; Cavalier-Smith 2003*a*,*b*)
    - infrakingdom Alveolata (phyla Myzozoa (Cavalier-Smith & Chao 2004), Ciliophora)
    - Biciliata incertae sedis: phylum Apusozoa (may be sister to Excavata); phylum Heliozoa[b]
- **kingdom Animalia** (Myxozoa and 21 other[a] phyla) (Cavalier-Smith 1998; Cavalier-Smith & Chao 2003*c*)
- **kingdom Fungi** (phyla Archemycota, Microsporidia, Ascomycota, Basidiomycota) (Cavalier-Smith 2000*b*)
- **kingdom Plantae**
  - subkingdom Biliphyta (phyla Glaucophyta, Rhodophyta)
  - subkingdom Viridaeplantae (Chlorophyta, Bryophyta[a], Tracheophyta[a])
- **kingdom Chromista**
  - subkingdom Cryptista (phylum Cryptista: cryptophytes, goniomonads, katablepharids)
  - subkingdom Chromobiota
    - infrakingdom Heterokonta (phyla Ochrophyta, Pseudofungi, Opalozoa (comprising subphyla Opalinata, Sagenista)
    - infrakingdom Haptista (phylum Haptophyta)

[a] No microbial members. All 34 phyla that contain microbes are listed.
[b] Although centrohelid Heliozoa might be Chromista, they probably belong in Protozoa (Biciliata) (Cavalier-Smith & Chao 2003*a*).

deep suspicion unless it is corroborated by independent evidence. Such evidence is not available for purely environmental sequences. Within eukaryotes both single-gene protein trees and protein trees combining data from several or many genes have revealed that most published rRNA trees have fundamentally incorrect topologies for several major long-branch clades; this is particularly true for trees published before corrections for intramolecular variation in evolutionary rates became de rigueur and those with unduly sparse taxon sampling.

Despite these problems, rRNA phylogeny remains a valuable tool for initial explorations of biodiversity. Such studies on cultured protists (unicellular eukaryotes) have contributed to major recent improvements in eukaryotic phylogeny and high-level classification (Silberman *et al.* 2002; Simpson *et al.* 2002; Cavalier-Smith 2003*a*; Cavalier-Smith & Chao 2003*a*,*b*,*c*, 2004) and through synthesis with morphological and other data to a considerable reduction in the number of protist phyla (Cavalier-Smith 2003*a*) compared with a decade ago (Cavalier-Smith 1993*a*). PCR of DNA extracts from environmental samples would *a priori* be expected to reveal many fewer deeply branching novel lineages than has been the case for bacteria. This is because most eukaryote cells are much larger and morphologically much more complex than most bacteria, so centuries of microscopical study have probably already revealed most of the major types, even of lineages that have never been cultured. This has been borne out by the first such studies of eukaryotes from aerobic habitats, where almost all the so-called 'novel' lineages discovered can be easily placed within the known phyla (López-García *et al.* 2001; Moon-van der Staay *et al.* 2001). This is true even for the alveolate protozoa, where the major groups of 'novel lineages' almost certainly belong within the phylum Myzozoa (=Miozoa) as sisters to the dinoflagellates (Cavalier-Smith & Chao 2004). Because within most phyla there are many known and well-described protists (whether of clear or obscure taxonomic position) that have not yet been cultured or had their rRNA sequenced, any study of environmental-DNA libraries is bound to come up with unidentifiable lineages that are not close to known sequences. But in eukaryotes, unlike bacteria, most of these are likely to be from groups with previously known representatives; such lineages are simply unidentified, not totally new to science.

Anaerobic aquatic habitats are ecologically important (Fenchel & Finlay 1995) but have been undersampled by culturing and ecological studies. It was formerly thought that early eukaryotes might have been anaerobic (Cavalier-Smith 1983*a*,*b*). However, all well-characterized groups of anaerobes have now been shown to have had aerobic ancestors (Embley & Hirt 1998; Roger 1999; Silberman *et al.* 2002; Simpson *et al.* 2002), making it highly probable that the ancestral eukaryote was aerobic (Cavalier-Smith 2002*a*). Nonetheless, the remote possibility remains that some little-known or entirely unknown primitively anaerobic eukaryote group might still exist, and anaerobic environments could still hold early-branching eukaryotic lineages important for understanding the origin of the nucleated (eukaryotic) cell. The recent studies of Dawson & Pace (2002) and Stoeck & Epstein (2003) of DNA extracted from marine and freshwater anoxic sites are therefore of particular interest. They found that over 90% of their sequences could easily be identified as belonging to known kingdoms and phyla, but were not able to identify relatives for 13 and nine novel sequences, respectively. Their conclusions that these new sequences represent eight and six novel kingdoms,

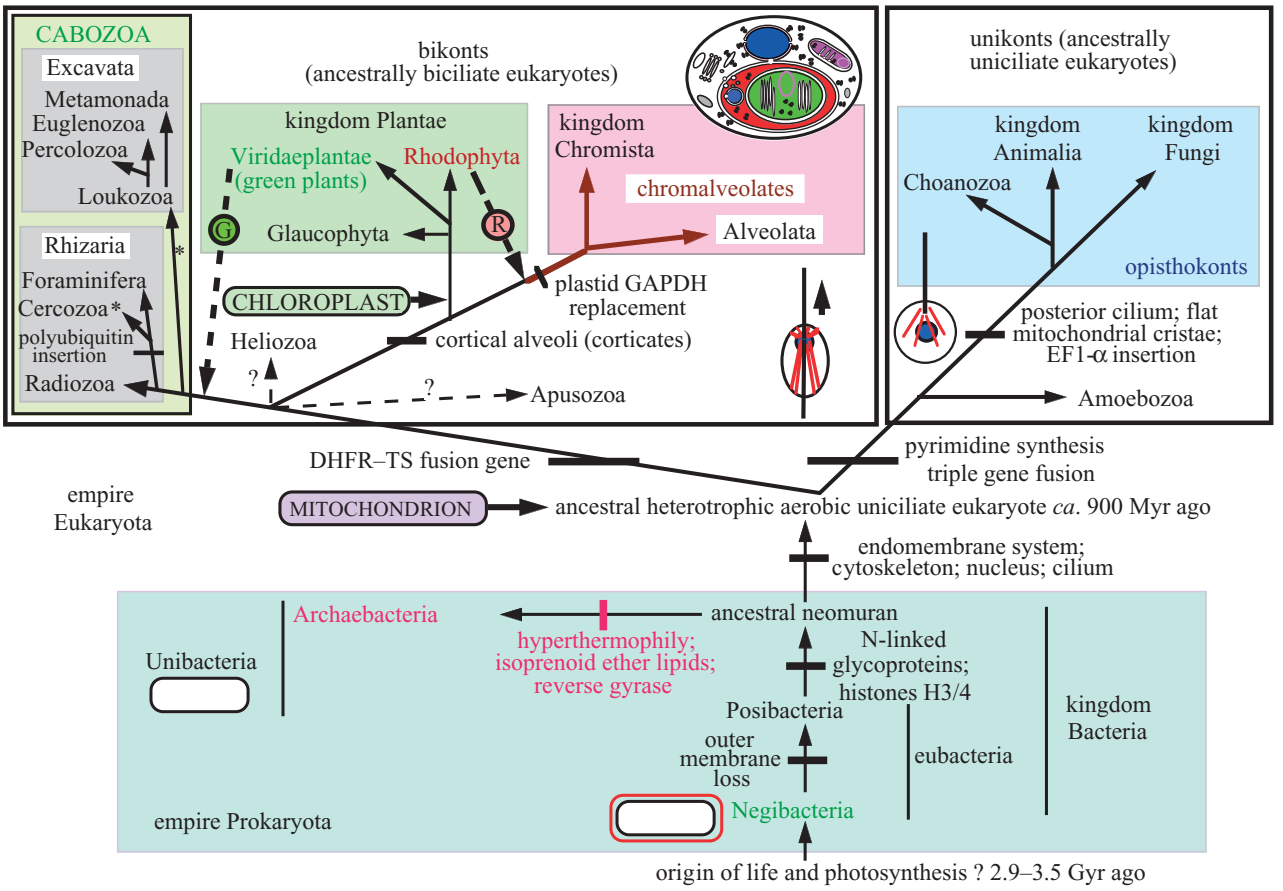

Figure 1. The tree of life based on molecular, ultrastructural and palaeontological evidence. Contrary to widespread assumptions, the root is among the eubacteria, probably within the double-enveloped Negibacteria, not between the eubacteria and archaebacteria (Cavalier-Smith 2002*b*). Establishing its precise position with confidence is one of the most challenging questions in evolutionary biology and is not yet solved; it may lie between Eobacteria and other Negibacteria (Cavalier-Smith 2002*b*); even the internal branching order of eubacterial phyla is more uncertain than for eukaryotes. The position of the eukaryotic root has been nearly as controversial, but is less hard to establish: it probably lies between the unikonts and the bikonts (Lang *et al.* 2002; Stechmann & Cavalier-Smith 2002, 2003*a*). For clarity the basal eukaryotic kingdom Protozoa is not labelled; it comprises four major groups (Alveolata, cabozoa, Amoebozoa and Choanozoa) plus the phylum Apusozoa, which may be sister to the other bikonts or, possibly more likely, sister to excavates alone (Cavalier-Smith 2003*a*,*b*), and probably Heliozoa. Heliozoa are of uncertain phylogenetic position within the bikonts, and might be chromists (Cavalier-Smith & Chao 2003*a*) not earlier branching as shown here. Symbiogenetic cell enslavement occurred four or five times: in the origin of mitochondria and chloroplasts from different Negibacteria, of chromalveolates by the enslaving of a red alga (Cavalier-Smith 1999, 2003*c*; Harper & Keeling 2003) and in the origin of the green plastids of euglenoid (excavate) and chlorarachnean (cercozoan) algae—a green algal cell (G) was enslaved either by the ancestral cabozoan (dashed arrow) or (less likely) twice independently within excavates and Cercozoa (asterisks) (Cavalier-Smith 2003*c*). The ages of eukaryotes and archaebacteria are also controversial; reasons for scepticism over substantially earlier dates than shown are explained by Cavalier-Smith (2002*b*) and Cavalier-Smith & Chao (2003*c*). The thumbnail sketches show the four major kinds of cell in the living world, plus the most complex eukaryote cells of all: the cryptophytes (upper sketch, nucleus and nucleomorph in blue, mitochondrion in purple, chloroplast in green, periplastid space in red). The middle ones show in red the contrasting ancestral microtubular cytoskeleton (ciliary roots) of unikonts (a cone of single microtubules attaching the single centriole to the nucleus) and bikonts (two bands of microtubules attached to the posterior centriole and an anterior fan of microtubules attached to the anterior centriole); cilia and plasma membrane are black and the nucleus blue. The lower ones show the single plasma membrane of unibacteria (Posibacteria plus archaebacteria), which were ancestral to eukaryotes, and the double envelope of Negibacteria, which were ancestral to mitochondria and chloroplasts (which retained the outer membrane, shown in red). Eubacteria are a paraphyletic grade comprising Negibacteria and Posibacteria.

respectively, are very surprising indeed, given that only five eukaryotic kingdoms are currently recognized (figure 1).

Have we really missed as much deep eukaryotic biodiversity as these authors claim? No! Their conclusions are invalid, not only because one cannot assess rank from divergence depth on a molecular tree, but also because the key analyses shown in their figures include far too few known eukaryotes to be able to identify their likely relatives or yield an evolutionarily reliable topology. In addition both studies were seriously flawed by a systematic misrooting caused by including the very distant bacterial outgroups. I have therefore reanalysed their data using 167, not just 37 (Dawson & Pace 2002) or only 8–35 (Stoeck & Epstein 2003), known eukaryotic taxa representing all the major groups. As recent microscopic studies have shown that the deep sea harbours many novel protists (Hausmann *et al.* 2002*a*), sometimes quite

extraordinary ones (Hausmann *et al.* 2002*b*), I have also included four environmental sequences from the deep sea not previously assigned to phyla that were more cautiously suggested as possibly representing novel kingdoms (López-García *et al.* 2003). My analysis of 193 sequences is, I think, the most taxonomically comprehensive and balanced phylogenetic study to date of eukaryotic 18S rRNA. Unsurprisingly, it does not support the idea that anaerobic habitats or the deep ocean contain uncharacterized eukaryotic kingdoms. All 26 mystery lineages discovered by the three studies can be placed within the established six-kingdom classification (table 1) in established phyla and usually also in classes within the kingdom Protozoa, except for one that proves to be an artefactual chimera of sequences from two phyla.

Dawson & Pace (2002) also interpreted their tree in terms of my former hypothesis (Cavalier-Smith 1983*a*,*b*) that the earliest eukaryotes lacked mitochondria, asserting that seven clades were 'deep branching', as did Stoeck & Epstein (2003) and another study that appeared after I completed the present analysis (Stoeck *et al.* 2003). However, recent evidence indicates that the root of the eukaryote tree lies among aerobic eukaryotes with mitochondria (Lang *et al.* 2002; Simpson & Roger 2002; Stechmann & Cavalier-Smith 2002, 2003*a*), not anaerobic ones as those authors mistakenly assumed (figure 1). Therefore, as I shall explain, the derived positions of all their sequences on my tree give added support for the now prevailing view that the last common ancestor of eukaryotes was aerobic and that mitochondria originated immediately following or even during the origin of the nucleus.

2. METHODS

The 12 unknown environmental sequences from fig. 4 of Dawson & Pace (2002), claimed to represent new kingdoms, plus another of their sequences that they did not assign to a known group (BOLA868), 10 phyletically unassigned sequences from Stoeck & Epstein (2003) and four from López-García *et al.* (2001, 2003) were aligned with 167 sequences of known organisms, representing all the major eukaryotic lineages, obtained from GenBank; wherever possible short-branch representatives were chosen to minimize phylogenetic artefacts. As the 26 environmental sequences were incomplete, the terminal regions and obviously ambiguously aligned regions were omitted from the phylogenetic analysis. Bacterial outgroups were also omitted, as these are so distant that they would simply have artefactually joined the longest eukaryotic branch (diplomonads) and caused an artefactual rooting, as happened in the trees of Dawson & Pace (2002) and Stoeck & Epstein (2003); in the former the bacterial branch was about twice as long as the *Trepomonas*/*Hexamita* part of the diplomonad one. Omitting bacteria allowed the inclusion of 1044 nucleotide positions not just 789 as included by Dawson & Pace (2002); Stoeck & Epstein (2003) did not specify how many they used. Neighbour-joining (BioNJ), weighted least-squares (power 2) distance analyses (GTR+Γ+I model: $\alpha = 0.614764$; $i = 0.120037$; parameters calculated via MODELTEST v. 3.06) and maximum parsimony were carried out using PAUP* v. 4.0b10. In addition to the tree shown in figure 2, a large number of other trees were also calculated with differing taxon samples, including the addition of longer-branch taxa such as diplomonads, retortamonads, Parabasalia, Percolozoa and Foraminifera to exclude the possibility that any of the sequences grouped with them and to test the robustness of the groupings found. Although these other trees are not shown, the generalizations about grouping are based not only on figure 2 but also on these dozens of other trees.

As rRNA trees cannot be reliably rooted using bacterial outgroups because of long-branch artefacts (Philippe 2000), an independent method is needed. Figure 1 shows that the position of the eukaryote root can be unambiguously specified by the use of two complementary gene fusions. One is the derived fusion of dihydrofolate reductase (DHFR) and thymidine synthetase (TS) first noted by Philippe *et al.* (2000) and shown by Stechmann & Cavalier-Smith (2002, 2003*a*) to characterize all groups of bikont eukaryotes; by contrast all three opisthokont groups and Amoebozoa have separate DHFR and TS genes like their bacterial ancestors (Stechmann & Cavalier-Smith 2002, 2003*a*). This fusion shows that bikonts are a clade and that the root cannot lie within them as it does on rRNA trees rooted by bacterial outgroups. The second is a fusion of the first three genes of the pyrimidine biosynthesis pathway, which is found in animals, fungi and Amoebozoa, but not in bacteria or bikonts, showing that unikonts are a clade and that the root cannot lie among them. Therefore the eukaryote root must lie precisely between the bikonts and the unikonts (Stechmann & Cavalier-Smith 2003*b*). Precisely the same position for the root is shown by the cytosolic Hsp90 tree where the branch lengths of all the eukaryote groups are relatively uniform, in marked contrast to the grossly non-clock-like rRNA, if it is rooted using the relatively closely related endoplasmic reticulum paralogue and not the much more distant bacterial outgroups (Stechmann & Cavalier-Smith 2003*b*). Concatenated trees based on mitochondrial proteins, which also have relatively uniform branch lengths (except for the aberrant Euglenozoa, which must be excluded), also place the root precisely between bikonts and unikonts (Lang *et al.* 2002) when rooted using bacterial outgroups, which are relatively much less distant than for rRNA. Ideally one should also root the rRNA tree (figure 2) between the unikonts and the bikonts. In figure 2 this is not possible, as the Amoebozoa are not cleanly separated from the bikonts, so it is rooted (only approximately correctly) between the opisthokonts and the Amoebozoa/bikonts.

3. IDENTIFYING RELATIVES OF THE PUTATIVELY ANAEROBIC MYSTERY CLADES

Figure 2 indicates that 22 out of the 24 sequences that previous authors (Dawson & Pace 2002; López-García *et al.* 2003; Stoeck & Epstein 2003) could not even place in established kingdoms form only 17 distinct groups, every one of which is related to known groups, seven belonging in a single order of eugregarine Apicomplexa. The rank of the known groups to which they can be assigned with reasonable confidence is highly variable, ranging from genus to phylum. I shall discuss them in the order shown by the numbered arrows in figure 2.

Three of these lineages belong in the protozoan phylum Amoebozoa. One (BOLA187/366) is very robustly sister to the anaerobic uniciliate amoeboflagellate '*Mastigamoeba invertens*' (the quotes signify that the strain sequenced was misidentified and is not even a *Mastigamoeba*)—it might even belong in the same genus; this clade is now treated as the class Breviatea (Cavalier-Smith *et al.* 2004). The second (BOLA868) belongs in the order Euamoebida (moderate support); most trees place it as sister to the

family Amoebidae, but in figure 2 it is sister to the Leptomyxidae instead (it was sister to the Amoebidae in the corresponding bootstrap consensus trees, both distance and parsimony, with just over 50% support); it may belong in the Amoebidae. The third lineage (LEMD267) is sister to *Filamoeba* on most trees (e.g. the bootstrapped consensus trees corresponding to figure 2; distance 62% and parsimony 84% support) or less often to *Filamoeba* plus the myxogastrids (figure 2)—it might even belong to the genus *Filamoeba* or in the family Filamoebidae, and it is certainly a member of the order Varipodida and the class Variosea (Cavalier-Smith *et al.* 2004). Thus none of these sequences represents a new kingdom, phylum, class or even order; as such a small fraction of the Amoebozoa has been sequenced, they may not even be new families, genera or species. They are simply unidentified amoebozoan sequences. It cannot safely be assumed that all 20 novel uncultured lineages of Dawson & Pace (2002) and Stoeck & Epstein (2003) are anaerobic, as aerobic cysts or other cells must sometimes enter anaerobic habitats. It is likely, however, that the BOLA187/366 clade at least is anaerobic like its sister '*M. invertens*'. The fact that this short-branch clade does not group with other Amoebozoa (it does in some trees based on longer rRNA sequences; Bolivar *et al.* 2001) may be just a consequence of the very poor resolution at the base of the ribosomal rRNA tree, probably caused by a combination of rapid early radiation and saturation effects. In figure 2 even the *Acanthamoeba* clade does not group with the rest of the Amoebozoa, although it does with slightly different taxon samples or methods.

The failure of Dawson & Pace (2002) to identify any of these sequences as Amoebozoa probably partly arose because their fig. 4 included only five Amoebozoa, not 40 as here. Their fig. 4 specifically did not include '*M. invertens*', *Filamoeba*, *Gephyramoeba*, myxogastrids or any Amoebidae or Leptomyxidae, the very groups to which their sequences are related. Furthermore, their five Amoebozoa formed three apparently unrelated clades (four counting LEMD267) rather than a single one as here and in another recent analysis with good taxon sampling (Bolivar *et al.* 2001). This emphasizes the importance of broad taxon sampling for obtaining sound trees and making sweeping conclusions about the non-affinity of environmental clones to known lineages. Their tree and those of Stoeck & Epstein (2003) were probably also distorted by including the very distant bacterial outgroups, as such extreme outgroups tend to pull the Mycetozoa, *Dictyostelium* and *Physarum* (and often also the Archamoebae) towards them and away from the other Amoebozoa, as seen in the trees of Cavalier-Smith (1993*a*), Milyutina *et al.* (2001) and Silberman *et al.* (2002) compared with the better ones of Bolivar *et al.* (2001) and figure 2. However, the distortion of fig. 4 of Dawson & Pace (2002) and the artefactually wide dispersal of the Amoebozoa were far worse than in the maximum-likelihood trees of Milyutina *et al.* (2001) that also included bacterial outgroups. The reasonably well-supported grouping together of the Archamoebae, *Vannella* and the *Hartmannella* clade (mislabelled as acanthamoebae in Dawson & Pace (2002): *Hartmannella* is not an acanthamoebid—it does not even belong in the same class; Cavalier-Smith *et al.* 2004) in Milyutina *et al.* (2001), unlike in Dawson & Pace (2002) where they appeared as three apparently unrelated clades, is also noteworthy as it suggests that their dispersal in the Dawson & Pace (2002) tree cannot be attributed solely to the inclusion of eubacterial outgroups. However, Milyutina *et al.* (2001) included 10 Amoebozoa and used 1209 alignment positions, compared with only 789 in Dawson & Pace (2002). Both factors would have improved their trees.

Lineage 4 (LEMD052/CCI78) belongs to the protozoan phylum Cercozoa and is a deep branch within the subphylum Endomyxa (Cavalier-Smith & Chao 2003*b*); on most trees it is sister to Phytomyxea with moderate bootstrap support (67% on the parsimony tree corresponding to figure 2), but in figure 2 it is sister to all other Cercozoa. The Dawson & Pace (2002) fig. 4 included only two cercozoans (not the most divergent possible), not 12 as here. This poor taxon sampling coupled with the use of only 789 not 1044 positions probably explains why LEMD052 did not group within the Cercozoa on their tree (though it was nearby). Although the bootstrap support for Cercozoa is low in figure 2 (probably because only partial sequences could be used because of the incompleteness of the environmental ones), both LEMD052 and CCI78 possess the almost unique signature deletion that characterizes all Cercozoa (Cavalier-Smith & Chao 2003*b*,*c*). In separate analyses with over 80 cercozoan sequences the LEMD052/CCI78 clade branches well within them as sister to Phytomyxea with good support. At present only one anaerobic cercozoan is known (Cavalier-Smith & Chao 2003*b*), and this novel clade does not group with it. We know from our own studies of environmental cercozoan sequences that hundreds of cercozoan sequences can be found that are not close to identified strains, and we have identified a clade from aerobic habitats in the same position on the rRNA tree as LEMD052/CCI78 (Bass & Cavalier-Smith 2004). Thus, in addition to not being a new kingdom or even phylum, this cercozoan clade may not even be anaerobic; it could be a new class or just deep-branching Phytomyxea.

It is also questionable whether sequence CCA32 (Stoeck & Epstein 2003) is from an anaerobe. Stoeck & Epstein (2003) did not claim that it represents a novel kingdom, as it grouped with strong support on their trees with the apusozoan *Ancyromonas*, but it was included here because no sequence was previously known to group robustly with *Ancyromonas*. On distance trees it is weakly sister to the zooflagellate *Diphylleia*, not included in the tree of Stoeck & Epstein (2003), but in some parsimony analyses it groups weakly with AT4-68 in an unresolved position deep among the bikonts. The taxonomic position of *Diphylleia* has itself been problematic, but it is now classified in the class Diphyllatea within the protozoan phylum Apusozoa (Cavalier-Smith 2003*a*,*b*). Figure 2 does not resolve the position of the *Diphylleia*/CCA32 clade, showing it as a very deep bikont branch not sister to any other group. However, very minor changes in the gamma correction parameters or in taxon sampling can cause it to group with Apusomonadida and *Ancyromonas*, the other Apusozoa on the tree, as well as with the breviates. This is consistent with CCA32 being a member of the class Diphyllatea of the phylum Apusozoa. Currently no anaerobic Apusozoa are known; although there is no

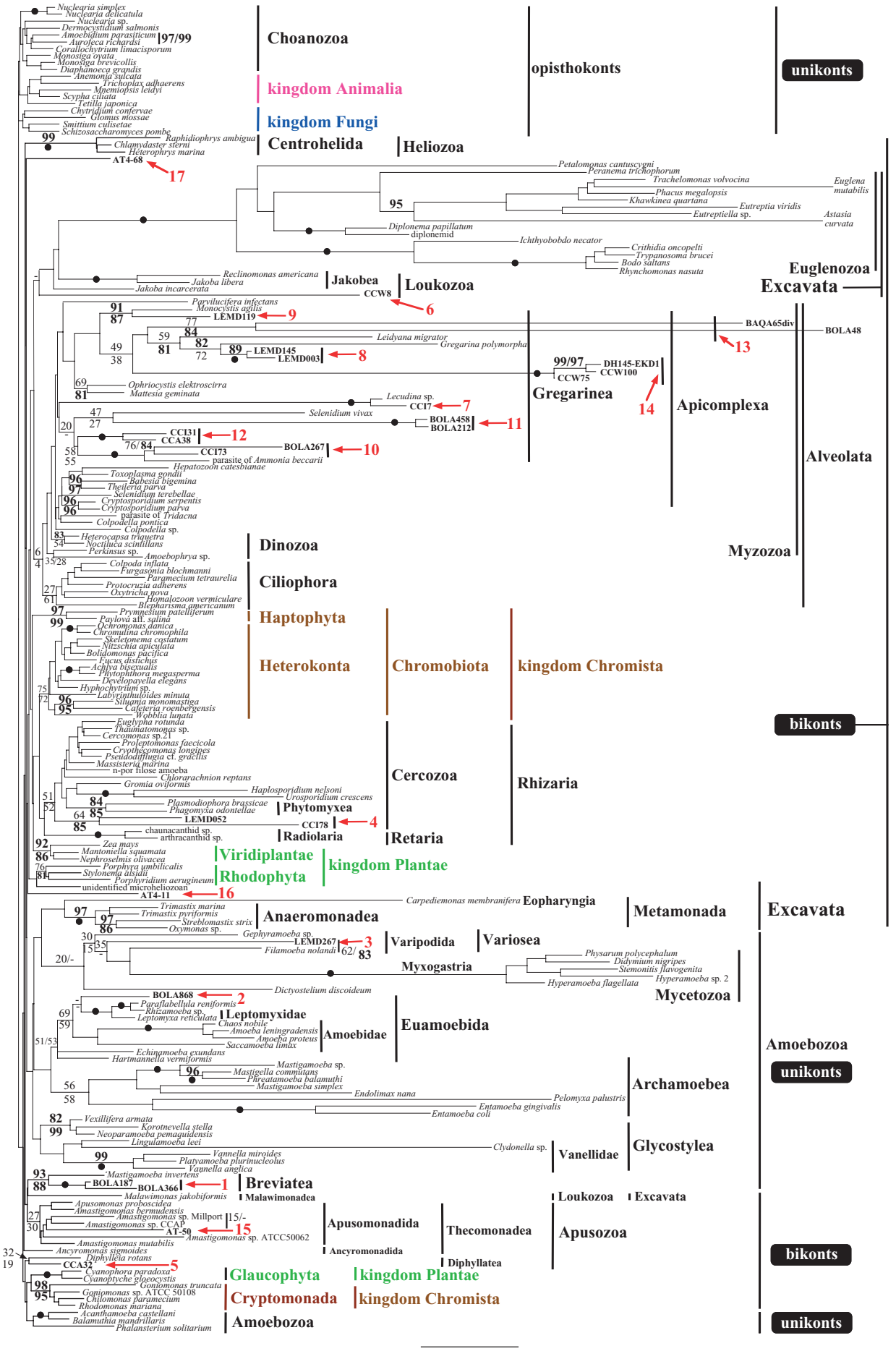

Figure 2. (*Caption opposite.*)

Figure 2. Phylogenetic analysis (BioNJ) of 193 eukaryotic 18S ribosomal RNA sequences using 1044 nucleotide positions. The 26 environmental sequences not previously assigned to established groups are in bold, and the 17 clades they form are marked by red numbered arrows. Only bootstrap percentages of 95% or more are shown except for those directly relevant to the positions of the environmental clades: distance (1000 pseudoreplicates: BioNJ) upper or left; parsimony (1000 pseudoreplicates) lower or right; as the tree is crowded, some are shown by the named clades not the bipartition points on the tree itself; bipartitions with 100% support by both methods are marked by a single black disc on the branch itself. The scale bar represents 10% sequence divergence. The great disparity of branch lengths indicates that 18S rRNA is grossly non-clock-like in its evolutionary rates; this would be several times greater than shown had not the longest-branch taxa been excluded to reduce artefacts. All taxa not included in the four derived kingdoms (Animalia, Fungi, Plantae, Chromista; shown in colour) belong in the unlabelled basal kingdom Protozoa (in black, see also table 1).

reason why some should not exist, it seems more likely that CCA32 is an aerobic flagellate (apart from breviates, anaerobic protists, unlike CCA32, have long or very long branches on rRNA trees making them very hard to place; it seems that loss of the mitochondrial genome generally causes dramatic and long-sustained acceleration of nuclear rRNA gene evolution; Cavalier-Smith 2002*b*).

Sequence ccw88 of Stoeck & Epstein (2003) turns out to be an artefactual chimera that was probably formed during the PCR process. The 3′ part of the molecule is clearly from a deep-branching heterokont as it has the characteristic heterokont signature sequence (Cavalier-Smith *et al.* 1994) and branches with or close to *Labyrinthuloides*. The 5′ part groups weakly with another of their unidentified environmental lineages, CCW8 (significantly from the same DNA sample). This chimera was omitted from the analysis shown because of the risk of its systematically distorting the tree.

Sequence CCW8 might be genuinely anaerobic. It groups as sister to the jakobid Loukozoa plus the Euglenozoa in figure 2; when Percolozoa are added to the tree they group with Euglenozoa, not CCW8. The phyletic position of CCW8 is the least clear of all sequences from anaerobic habitats; it is likely to be a member of the Loukozoa, a possibly paraphyletic phylum at the base of the excavates (Cavalier-Smith 2003*b*). The other loukozoan, *Malawimonas*, is in an isolated position in figure 2, but on most trees that also include Percolozoa it groups as sister to Metamonada, which are here in an unusual artefactual position within the Amoebozoa. The bootstrapped consensus tree for the figure 2 dataset showed *Malawimonas* as sister to Anaeromonadea with 19% support and this clade plus AT4-68 as sister to the discicristate/jakobid clade; excavates would be holophyletic on that tree except for the misplacement of *Carpediemonas* alone within the Amoebozoa. Establishing the unity and basal branching order for excavates is notoriously difficult on rRNA trees because of their tremendous rate variations (Cavalier-Smith 2003*b*). A variety of anaerobic flagellates has been observed microscopically by ecologists but not studied systematically; one such that appears to have just the right mix of characters to branch precisely as does CCW8 is flagellate 2 of Fenchel *et al.* (1995), which has a groove (like Loukozoa and Percolozoa) and two posteriad cilia (unlike these two phyla or Euglenozoa, but like the postulated missing link between Loukozoa and discicristates; Cavalier-Smith 2003*b*).

The remaining eight lineages of Dawson & Pace (2002) and Stoeck & Epstein (2003) all apparently belong in the phylum Myzozoa in the subphylum Apicomplexa, class Gregarinea and order Eugregarinida. This is most obvious for CCI7, which is so close to *Lecudina* sp. that it almost certainly belongs to that very genus. The LEMD145/003 clade is sister to *Gregarina polymorpha* with 100% support, while LEMD119 is sister to *Monocystis* with very high support. BOLA267/CCI73 has 100% support as sister to the parasite of *Ammonia*, identified by Leander *et al.* (2003) as a gregarine. BOLA458/212 groups weakly as sister to *Selenidium vivax*. Clade CCI31/CCA38 groups with moderate support with the *Ammonia* parasite clade. Two clades that consistently branch within the gregarines (BAQA65/BOLA48 and DH145 (of López-García *et al.* 2001)/CCW75/100) are, however, such long branches that their identification as gregarines is more open to question, but there is no reason from the present analysis to regard any of these sequences as representing novel kingdoms or phyla or even classes or orders.

It is not surprising that Dawson & Pace (2002) and Stoeck & Epstein (2003) failed to identify any of these sequences as gregarines, as neither included any known gregarines in their published trees, and Dawson & Pace's (2002) had only one apicomplexan. When I omit gregarines altogether from the present dataset and add Percolozoa (as in Dawson & Pace 2002), none of these sequences group within Apicomplexa, Myzozoa or (usually) even the Alveolates, but form a single large clade (or two clades) that is very distantly related and very weakly sister to Percolozoa; this long-branch artefact is corrected only by the addition of numerous gregarines. Gregarines themselves have highly variable rRNA evolutionary rates and branch lengths; coupled with the explosive basal branching of Myzozoa, this means that bootstrap support for the unity of the Myzozoa itself is very low when they are all included.

It is hard to know just how fig. 4 of Dawson & Pace (2002) was produced. The text calls it 'one typical tree' without specifying method, but implying that only one was used; the figure legend calls it 'a consensus tree' without explaining how a fully resolved consensus tree with branch lengths was calculated. The tree shows support values for different methods, three that did not allow for intersite variation and one (Bayesian) that did; for the first three there was no support at all for four bipartitions that separated most of the sequences claimed to represent separate kingdoms. The tree itself looks very different from the one presented here. It is pectinate with the known and environmental sequences in positions from top to bottom in approximate proportion to the branch lengths. Such a tree has all the hallmarks of being dominated by long-branch artefacts, and is just the kind of tree that one used to get before corrections for intersite rate variation were introduced (Cavalier-Smith 1993*a*, 1995). Correction for intersite rate variation is important to reduce long-branch artefacts, though such correction can never be totally successful for molecules that are as grotesquely non-clock-

like as rRNA. The reader can easily compare the overall eukaryote trees produced by both methods for rRNA by consulting Bolivar *et al.* (2001) and Milyutina *et al.* (2001) and for proteins by consulting Bapteste *et al.* (2002). In all cases the corrected trees are more congruent with other biological data than are the uncorrected trees. The rRNA trees shown in fig. 2*a*,*b* of Milyutina *et al.* (2001) are particularly instructive as like in Dawson & Pace (2002) and Stoeck & Epstein (2003) they were rooted by bacterial outgroups and included microsporidia and even the long-branch Foraminifera. Their uncorrected fig. 2*a* is pectinate and graded in branch lengths from top to bottom like that of Dawson & Pace (2002), but is not as bad as there is a major amoebozoan clade apart from slime moulds. Their tree is also better than that of Dawson & Pace (2002) in showing a parabasalid/diplomonad clade, now supported also by several protein trees, including cpn60 (Archibald *et al.* 2002), and shared lateral transfers (see review by Cavalier-Smith 2003*b*). Their gamma-corrected fig. 2*b* is very different. It is largely non-pectinate and non-graded with a big bang of rapidly radiating mostly short-branch lineages from which numerous long-branch sequences stem largely independently, with a general appearance similar to my present tree, even though it is rooted differently and artefactually on the parabasalid sequence. A striking difference between the two trees is in the position of the sole microsporidian: in the uncorrected tree it is well towards the base among other long-branch taxa (but not as far away from its true fungal position as in Dawson & Pace (2002)), but in the corrected tree it is between opisthokonts and Amoebozoa, i.e. in the unikont part of the tree, much closer to its true position within the Fungi. Fig. 2*b* was also biologically realistic in grouping together the short-branch radiolaria and the ultra-long-branch Foraminifera, the first molecular support for the group Retaria, established on morphological grounds (Cavalier-Smith 1999). Fig. 4 of Dawson & Pace (2002) would seem to be an uncorrected tree, which would artefactually have dispersed both named and unidentified sequences, impeding recognition of their true affinities. The frequent capacity of gamma-corrected distance trees to recover (albeit usually with weak support) clades well-substantiated by morphology but with a mixture of short- and very-long-branch taxa that usually disrupt the clade on uncorrected trees is also evident in the present figure 2: e.g. the short-branch jakobids are grouped with the long-branch Euglenozoa, metamonads are holophyletic, with the short-branch anaeromonad clade grouped with the medium-length *Carpediemonas* (if retortamonads and the ultra-long-branch diplomonads are also included they group as expected with *Carpediemonas*) and the long-branch myxogastrids are grouped with the medium-branch *Dictyostelium*.

However, figure 2 is not a perfect representation of eukaryote phylogeny. As in most previous studies, several major groups known to be holophyletic from other extensive evidence (see review by Cavalier-Smith & Chao 2003*c*) are not recovered as clades (e.g. the kingdoms Plantae and Chromista, and the chromalveolates (Chromista plus Alveolata)). Moreover, there is virtually no support for the basal-branching orders (bootstrap support for them typically in the range of 0–10%) consistent with a very rapid, virtually explosive radiation immediately following the origin of the eukaryotic cell (Cavalier-Smith & Chao 2003*c*). Nonetheless, the present tree is much more congruent with other molecular data, ultrastructure and cell biology (Cavalier-Smith 2002*b*; Cavalier-Smith & Chao 2003*c*) than any recent tree with only limited taxon sampling and/or including bacterial outgroups (Dawson & Pace 2002; Silberman *et al.* 2002) (including bacteria leads many authors to ignore parts of the molecule informative for eukaryote phylogeny). In contrast to such trees, including those of Dawson & Pace (2002), the long-branch Amoebozoa group with most of the short-branch aerobic Amoebozoa, not arbitrarily elsewhere or artefactually with the long-branch excavates, and the latter do not all artefactually cluster together. It is very difficult indeed for tree-reconstruction algorithms to cope with the extreme variations in evolutionary rate and mode of the rRNA molecule, especially in the vastly accelerated secondarily amitochondrial, and the discicristate and myxogastrid amoebozoan lineages. Nonetheless, including a large number of taxa in the tree, including as high a proportion of the molecule as is reasonable and excluding the excessively distant outgroups (bacteria) are keys to reducing these problems and obtaining a relatively undistorted phylogeny. These three features of the present tree may explain why Excavata and Amoebozoa are both much less randomly dispersed in the present tree than in taxonomically sparse bacterially rooted trees. The positions of many excavates on rRNA trees are notoriously sensitive to taxon sampling and the phylogenetic parameters, because of their exceptional disparity in rRNA evolutionary rate and mode; only rarely do all branch together (Cavalier-Smith 2002*b*, 2003*b*).

4. IDENTIFYING RELATIVES OF THE MYSTERY DEEP-OCEAN CLADES

The tree of López-García *et al.* (2003) already weakly suggested that two of their mystery sequences were sisters of *Apusomonas*. Figure 2 shows that one of them, AT-50, does indeed branch within the class Thecomonadea of the phylum Apusozoa, actually within the genus *Amastigomonas* (sister to *Amastigomonas* sp. Millport on both consensus trees), suggesting that it is simply an additional *Amastigomonas* species. From our previous work we concluded that *Amastigomonas* is probably vastly underdescribed (Cavalier-Smith & Chao 2003*c*). If AT-50 is an *Amastigomonas*, that means that we already have six dramatically different sequences for this genus in which only nine species have been described (Mylnikov 1999). Only two morphospecies (*A. mutabilis* and *A. debruynei*) were previously recorded from the deep ocean (Arndt *et al.* 2003), and their sequences are markedly different from that of AT-50. The latter need not necessarily be *Amastigomonas*, however, for there are currently no sequences available from the thecomonad order Hemimastigida. As this order is considered to have evolved from an *Amastigomonas*-like ancestor (Cavalier-Smith 2000*a*), one of these sequences might be from a hemimastigid; although hemimastigids were previously known only from terrestrial environments, two genera were recently recorded in very deep (1325–1249 m) Mediterranean sediments (Arndt *et al.* 2003). Thus, while we can be reasonably confident that AT-50 is a thecomonad

sequence, we cannot exclude the possibility that it belongs to the order Hemimastigida rather than the Apusomonadida.

On some trees AT4-11 groups as sister to the Apusomonadida (rarely within it), but it can also go within the Amoebozoa or, as in figure 2, in a deep position with no specific relative. It seems likely that it belongs either to the Apusozoa or to the Amoebozoa, as these phyla seldom appear strictly holophyletic on rRNA trees, but it might belong to a third protozoan phylum.

Sequence AT4-68 is even harder to place. It typically occupies a deep but variable position within the bikonts with no clear relatives. In some distance trees it groups with very low support as sister to the Glaucophyceae or within the excavates as sister to *Malawimonas*/Metamonada, but with parsimony it is sister to CCA32 near the base of the plant kingdom. Glaucophytes are the only one of the three major groups of the plant kingdom in which secondarily heterotrophic species have not been identified; by contrast there are non-photosynthetic green plants and red algae that have secondarily lost photosynthesis (but retained plastids, presumably for starch and/or fatty-acid synthesis; Cavalier-Smith 1993*b*). Because López-García *et al.* (2003) found no sequences from authentic photosynthetic algae in their deep-sea samples, CCA32 is probably also from a heterotroph. While it might be the first heterotrophic member of the phylum Glaucophyta and of great potential interest for understanding the early evolution of the plant kingdom if only it could be cultured, its occasional grouping with CCA32 or within the excavates makes it likely that it is simply a bikont protozoan.

5. ALL ANAEROBIC EUKARYOTES ARE PROBABLY DERIVED: PROBLEMS OF ROOTING THE TREE

In accordance with previous evidence that the last common ancestor of all eukaryotes had mitochondria capable of aerobic respiration (probably facultatively rather than obligately; Cavalier-Smith 2002*b*) and that all anaerobic eukaryotes arose secondarily by converting mitochondria into hydrogenosomes or mitosomes (Silberman *et al.* 2002; Williams *et al.* 2002; Tovar *et al.* 2003), all the putatively anaerobic lineages claimed to represent new kingdoms (Dawson & Pace 2002; Stoeck & Epstein 2003) nest well within aerobic clades in figure 2. In the past decade data from numerous proteins and the discovery of mitosomes firmly established the secondary nature of amitochondrial eukaryotes (Cavalier-Smith 2002*a*,*b*, 2003*b*; Keeling 2003; Roger 1999; Silberman *et al.* 2002; Williams *et al.* 2002). With no evidence or arguments whatsoever, Dawson & Pace (2002) superficially dismissed as 'lateral transfer' the disparate and extensive evidence for this major advance in eukaryotic phylogeny. Their tree is also topologically incorrect, as shown by the non-grouping of microsporidia with fungi from which they evolved (Cavalier-Smith 2000*b*; Keeling 2003), and by the three widely dispersed amoebozoan clades that in better analyses come together (Bolivar *et al.* 2001), as they do in figure 2; the topology in Dawson & Pace (2002) was probably distorted by long-branch bacterial outgroups. Unwise inclusion of bacteria and the drastically shortened microsporidial genes also meant that they excluded numerous phylogenetically informative sites, using only 789 nucleotide positions, compared with 1044 in my analysis.

Although their trees are technically inferior to many published ones, being topologically incorrect in major respects and rooted in entirely the wrong place, Dawson & Pace (2002) and Stoeck & Epstein (2003) assume that both features are correct when discussing so-called deep branches within them. They rooted their trees using bacterial outgroups (despite this being known to give an incorrect root among the longest-branch eukaryotes; Cavalier-Smith 2002*b*; Simpson & Roger 2002; Stechmann & Cavalier-Smith 2002). It is remarkable that eight of the lineages claimed to represent novel anaerobic kingdoms turn out to belong to a single gregarine order, as no anaerobic gregarines have so far been described. They are neither novel kingdoms nor early diverging; are they even anaerobic? The cercozoan clade and two of the three amoebozoan clades may not be anaerobic and are certainly not 'early diverging' sequences. Only one clade (BOLA187/BOLA366) has any potential to be early diverging, but it is not really novel, being sister to '*M. invertens*'. If the breviate clade ('*M. invertens*'/BOLA187/BOLA366) really belongs within a holophyletic phylum Amoebozoa, as seems most likely (Cavalier-Smith *et al.* 2004), then it would not be early diverging either. If so, as figure 1 makes clear, the rooting of the eukaryote tree between bikonts and unikonts (not within the most highly derived bikonts, as in fig. 4 of Dawson & Pace (2002)) would mean that there are also no known extant eukaryotes, whether aerobic or anaerobic, that diverged prior to the last common ancestor of animals and plants. Far from revealing 'novel' early-diverging lineages, as claimed, both studies failed to detect any novel anaerobic lineages that are early diverging (given the correct rooting of the tree), which considerably strengthens the current interpretation that there are probably no primitively amitochondrial eukaryotes. Only the breviates remain as possible candidates for such a position. '*Mastigamoeba invertens*' needs to be studied for the presence or absence of relict Hsp70 and Cpn60 chaperones from a mitochondrial ancestry and for the presence of the two fusion genes. If both sets of genes are unfused, this would provide evidence that breviates are early-diverging anaerobic eukaryotes, as postulated to exist in the Archezoa hypothesis (Cavalier-Smith 1983*a*). If the DHFR and TS genes are unfused but the three pyrimidine-biosynthesis ones are fused, this would place the breviates clearly in the Amoebozoa and unikonts; the converse would place them on the bikont side of the basal eukaryote bifurcation. While I have identified two out of the four putatively aerobic clades of López-García *et al.* (2001, 2003), the identity of the other two remains unclear. This is unsurprising as there are numerous aerobic protozoan genera of uncertain taxonomic position that have not been cultured, sequenced or studied ultrastructurally. Until we have such information we cannot tell whether they represent new orders or classes (both likely) or even phyla (unlikely, but possible). Naive interpretations of rRNA trees and protein paralogy trees have grossly misled evolutionary biology (Cavalier-Smith 2002*a*).

If breviates turn out to be secondarily anaerobic Amoebozoa and Amoebozoa prove to be holophyletic, the

rooting of the eukaryote tree between unikonts and bikonts means that there is no such thing as a deep-branching eukaryote, in the sense of one branching prior to the last common ancestor of animals and plants. The common use of the terms 'crown' and basal/deep by many rRNA sequencers reflect multiple misunderstandings (Cavalier-Smith 1999). The dichotomy between 'crown groups' and 'basal lineages' is biologically totally meaningless and simply reflects the severe long-branch exclusion artefacts and the systematic misrooting of rRNA trees when using bacterial outgroups. For this reason, and because the cladistic term 'crown' properly refers to all extant eukaryotes (Cavalier-Smith 2002*a*), the muddled and misleading phrase 'crown eukaryote' should no longer be used for any subset of extant eukaryotes.

The widespread but false assumption that rRNA is a universal molecular clock has led to very serious misinterpretations of the tree of life, especially concerning the roots of both the bacterial and eukaryotic parts of the tree (Cavalier-Smith 2002*a*). To reconstruct phylogeny satisfactorily it is fundamentally unsound to rely on a single molecule, as Dawson & Pace (2002) appear to. Trees based on large numbers of molecules (Baldauf *et al.* 2000; Bapteste *et al.* 2002), and genic, biochemical and ultrastructural data that can be treated cladistically (Stechmann & Cavalier-Smith 2002, 2003*a*; Cavalier-Smith & Chao 2003*c*), as well as palaeontology (Cavalier-Smith 2002*a*,*b*) and single-gene trees for proteins (e.g. Hsp90, which appears to be much more clock-like than rRNA; Stechmann & Cavalier-Smith 2003*b*) as well as rRNA are all taken note of in the synthesis of figure 1. Dawson & Pace (2002) and Stoeck & Epstein (2003) ignore the vast majority of such data on eukaryote evolution, including the evidence that the positions of the roots of their trees are profoundly incorrect (Simpson & Roger 2002; Stechmann & Cavalier-Smith 2002; Cavalier-Smith & Chao 2003*c*); their discussion in terms of a 'crown radiation' and 'deep branching' is basically upside down. It is the divergences among the short-branch, misnamed, 'crown groups' that are basal, whereas their so-called 'deep' branches (mostly excavates) are actually among the most derived groups. It is particularly astounding that Dawson & Pace (2002) ignore the evidence for the fungal nature of microsporidia, which was the first group to show conclusively how dramatically misleading the rRNA tree can be when naively interpreted (Embley & Hirt 1998; Roger 1999; Cavalier-Smith 2000*b*; Williams *et al.* 2002; Keeling 2003). When naively interpreted in the absence of other data, rRNA trees are the single most misleading source of information we have about the history of life. When integrated with all the other information they are very valuable.

Analogous claims of 14 novel division (phylum)/kingdom-level bacterial lineages (Hugenholtz *et al.* 1998) based on unidentified environmental sequences are probably also ill-founded; long-branch problems, many exacerbated by thermophilic bias (Cavalier-Smith 2002*b*; Gribaldo & Philippe 2002), coupled with the non-existent resolution at the base of the eubacterial rRNA tree, impede their placement in one of the seven established eubacterial phyla (table 1 in Cavalier-Smith 2002*b*). The eubacterial tree is equally poorly resolved at its base, and the problem of its rooting is even more severe than for eukaryotes. Contrary to over-confident but deeply held assumptions, we do not actually know where the base of the tree of life is, but it is much more likely to be among the Gram-negative eubacteria (Cavalier-Smith 2002*b*) than between archaebacteria and eubacteria as is often assumed. Until we know the answer to this question, any reference to 'deeply branching' or 'ancient' prokaryotic lineages (e.g. Gaucher *et al.* 2003) is potentially misleading and immensely more controversial than it is generally realized to be. While it may be that the Eobacteria are the earliest-diverging phylum (Cavalier-Smith 2002*b*), we cannot currently exclude the possibility that their apparently primitive characters are secondarily simplified and that, as in eukaryotes, there are no extant 'early diverging' lineages. Even when representatives of novel rRNA lineages are cultivated, as is increasingly being done (Leadbetter 2003), and partly characterized, ranking them as phyla purely on the basis of the degree of rRNA divergence (e.g. Zhang *et al.* 2003) is taxonomically unsound as rRNA divergence can be greatly accelerated for relatively trivial reasons and does not necessarily correlate well with biologically more important character differences.

6. CONCLUSIONS

Far from revealing novel kingdoms, the new rRNA sequences from anaerobic habitats (Dawson & Pace 2002; Stoeck & Epstein 2003) and the deep ocean (López-García *et al.* 2003) show that our understanding of eukaryote high-level diversity is actually now rather good. When properly analysed they tell us, in conjunction with numerous studies from aerobic habitats (López-García *et al.* 2001; Moon-van der Staay *et al.* 2001; Moreira & López-García 2002), that there may be very few, if any, previously unknown protist phyla—and no 'new kingdoms'—remaining to be 'discovered'. These other studies did find a very small number of sequences that are long branches and as hard to place as some of those included here, e.g. Edgcomb *et al.* (2002), who also misrooted the tree and made similar unwarranted statements about 'early branching' anaerobic eukaryotes. While it is not impossible that some of these lineages might represent new phyla, it is more probable that they also will turn out to be hard-to-place long-branch representatives of established ones. I do not wish to be misunderstood as predicting that no new phyla will be discovered. There are scores of protist genera that have not yet been studied ultrastructurally (e.g. the extraordinary *Meteora*, which waves its filopodia like semaphores (Hausmann *et al.* 2002*b*); its phylum is unknown; my hunch is that it belongs in the Cercozoa, which are unsurpassed in filopodial wonders (Cavalier-Smith & Chao 2003*b*)) and there are many others still undescribed. It is possible that some may belong in a small group sufficiently distinct from known phyla to merit one of its own; Apusozoa is one such phylum that was only recently established (Cavalier-Smith 2002*a*). There may be others, but probably rather few, and possibly none. Phyla and kingdoms are not actually things that one discovers, but conceptual entities that systematists create by deliberately grouping together known organisms. Novel organisms, molecules or lineages can be discovered, but one lesson from the present study is that molecular trees used to claim novel major eukaryote lineages should in future

include the broadest range of protist diversity as in figure 2 and not just a tiny subset of it.

Ribosomal DNA sequencing is very useful indeed for evolutionary and ecological studies, and we ourselves are using environmental rDNA sequencing to explore the immense hidden diversity of protists, but it has to be critically integrated into traditional taxonomy and not given a spurious pre-eminence.

I thank NERC for research grants, and the Canadian Institute for Advanced Research and NERC for fellowship support.

REFERENCES

Archibald, J. M., O'Kelly, C. J. & Doolittle, W. F. 2002 The chaperonin genes of jakobid and jakobid-like flagellates: implications for eukaryotic evolution. *Mol. Biol. Evol.* **19**, 422–431.

Arndt, H., Hausmann, K. & Wolf, M. 2003 Deep-sea heterotrophic nanoflagellates of the Eastern Mediterranean Sea: qualitative and quantitative aspects of their pelagic and benthic occurrence. *Mar. Ecol. Prog. Ser.* **256**, 45–56.

Baldauf, S. L., Roger, A. J., Wenk-Siefert, I. & Doolittle, W. F. 2000 A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* **290**, 972–977.

Bapteste, E. (and 10 others) 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.

Barns, S. M., Delwiche, C. F., Palmer, J. D. & Pace, N. R. 1996 Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. *Proc. Natl Acad. Sci. USA* **93**, 9188–9193.

Bass, D. & Cavalier-Smith, T. 2004 Phylum-specific environmental DNA analysis reveals remarkably high global biodiversity of Cercozoa (Protozoa). *Int. J. Syst. Evol. Microbiol.* (In the press.)

Bolivar, I., Fahrni, J. F., Smirnov, A. & Pawlowski, J. 2001 SSU rRNA-based phylogenetic position of the genera *Amoeba* and *Chaos* (Lobosea, Gymnamoebia): the origin of gymnamoebae revisited. *Mol. Biol. Evol.* **18**, 2306–2314.

Cavalier-Smith, T. 1983*a* A 6-kingdom classification and a unified phylogeny. In *Endocytobiology II* (ed. W. Schwemmler & H. E. A. Schenk), pp. 1027–1034. Berlin: de Gruyter.

Cavalier-Smith, T. 1983*b* Endosymbiotic origin of the mitochondrial envelope. In *Endocytobiology II* (ed. W. Schwemmler & H. E. A. Schenk), pp. 265–279. Berlin: de Gruyter.

Cavalier-Smith, T. 1993*a* Kingdom Protozoa and its 18 phyla. *Microbiol. Rev.* **57**, 953–994.

Cavalier-Smith, T. 1993*b* The origin, losses and gains of chloroplasts. In *Origin of plastids: symbiogenesis, prochlorophytes and the origins of chloroplasts* (ed. R. A. Lewin), pp. 291–348. New York: Chapman & Hall.

Cavalier-Smith, T. 1995 Membrane heredity, symbiogenesis, and the multiple origins of algae. In *Biodiversity and evolution* (ed. R. Arai, M. Kato & Y. Doi), pp. 75–114. Tokyo: The National Science Museum Foundation.

Cavalier-Smith, T. 1998 A revised six-kingdom system of life. *Biol. Rev. Camb. Phil. Soc.* **73**, 203–266.

Cavalier-Smith, T. 1999 Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryotic family tree. *J. Eukaryotic Microbiol.* **46**, 347–366.

Cavalier-Smith, T. 2000*a* Flagellate megaevolution: the basis for eukaryote diversification. In *The flagellates* (ed. J. R. Green & B. S. C. Leadbeater), pp. 361–390. London: Taylor & Francis.

Cavalier-Smith, T. 2000*b* What are fungi? In *The Mycota*, vol. VII part A (ed. D. J. McLaughlin, E. J. McLaughlin & P. Lemke), pp. 3–37. Berlin: Springer.

Cavalier-Smith, T. 2002*a* The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. *Int. J. Syst. Evol. Microbiol.* **52**, 297–354.

Cavalier-Smith, T. 2002*b* The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. *Int. J. Syst. Evol. Microbiol.* **52**, 7–76.

Cavalier-Smith, T. 2003*a* Protist phylogeny and the high-level classification of Protozoa. *Eur. J. Protistol.* **39**, 338–348.

Cavalier-Smith, T. 2003*b* The excavate protozoan phyla Metamonada Grassé emend. (Anaeromonadea, Parabasalia, *Carpediemonas*, Eopharyngia) and Loukozoa emend. (Jakobea, *Malawimonas*): their evolutionary affinities and new higher taxa. *Int. J. Syst. Evol. Microbiol.* **53**, 1741–1758.

Cavalier-Smith, T. 2003*c* Genomic reduction and evolution of novel genetic membranes and protein-targeting machinery in eukaryote–eukaryote chimaeras (meta-algae). *Phil. Trans. R. Soc. Lond.* B **358**, 109–134. (DOI 10.1098/rstb.2002.1194.)

Cavalier-Smith, T. & Chao, E. E. 2003*a* 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. & Chao, E. E. 2003*b* Phylogeny and classification of phylum Cercozoa (Protozoa). *Protist* **154**, 341–358.

Cavalier-Smith, T. & Chao, E. E. 2003*c* Phylogeny of Choanozoa, Apusozoa, and other Protozoa and early eukaryote megaevolution. *J. Mol. Evol.* **56**, 540–563.

Cavalier-Smith, T. & Chao, E. E. 2004 Protalveolate phylogeny and the origins of Sporozoa and dinoflagellates (phylum Myzozoa nom. nov.) *Eur. J. Protistol.* **40**, 21–48.

Cavalier-Smith, T., Allsopp, M. T. E. P. & Chao, E. E. 1994 Thraustochytrids are chromists, not Fungi: 18S rRNA signatures of Heterokonta. *Phil. Trans. R. Soc. Lond.* B **339**, 139–146.

Cavalier-Smith, T., Chao, E. E. & Oates, B. 2004 Molecular phylogeny of Amoebozoa and the evolutionary significance of the unikont *Phalansterium*. *Eur. J. Protistol.* **40**, 21–48.

Dawson, S. C. & Pace, N. R. 2002 Novel kingdom-level eukaryotic diversity in anoxic environments. *Proc. Natl Acad. Sci. USA* **99**, 8324–8329.

Edgcomb, V. P., Kysela, D. T., Teske, A., de Vera Gomez, A. & 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. & Hirt, R. P. 1998 Early branching eukaryotes? *Curr. Opin. Genet. Dev.* **8**, 624–629.

Fenchel, T. & Finlay, B. J. 1995 *Ecology and evolution in anoxic worlds*. Oxford University Press.

Fenchel, T., Bernard, C., Esteban, G., Finlay, B. J., Hansen, P. J. & Iversen, N. 1995 Microbial diversity and activity in a Danish fjord with anoxic deep water. *Ophelia* **43**, 45–100.

Gaucher, E. A., Thomson, J. M., Burgan, M. F. & Benner, S. A. 2003 Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature* **425**, 285–288.

Giovannoni, S. J., Britschgi, T. B., Moyer, C. L. & Field, K. G. 1990 Genetic diversity in Sargasso Sea bacterioplankton. *Nature* **345**, 60–63.

Gribaldo, S. & Philippe, H. 2002 Ancient phylogenetic relationships. *Theor. Popul. Biol.* **61**, 391–408.

Harper, J. T. & Keeling, P. J. 2003 Nucleus-encoded, plastid-targeted glyceraldehyde-3-phosphate dehydrogenase (GAPDH) indicates a single origin for chromalveolate plastids. *Mol. Biol. Evol.* **20**, 1730–1735.

Hausmann, K., Hülsmann, N., Polianski, I., Schade, S. & Weitere, M. 2002*a* Composition of benthic protozoan communities along a depth transect in the eastern Mediterranean Sea. *Deep Sea Res.* **49**, 1959–1970.

Hausmann, K., Weitere, M., Wolf, M. & Arndt, H. 2002*b* *Meteora sporadica* gen. nov. et sp. nov. (Protista *incertae sedis*): an extraordinary free-living protist from the Mediterranean deep sea. *Eur. J. Protistol.* **38**, 171–177.

Hugenholtz, P., Pitulle, C., Hershberger, K. L. & Pace, N. R. 1998 Novel division level bacterial diversity in a Yellowstone hot spring. *J. Bacteriol.* **180**, 366–376.

Keeling, P. J. 2003 Congruent evidence from alpha-tubulin and beta-tubulin gene phylogenies for a zygomycete origin of microsporidia. *Fungal Genet. Biol.* **38**, 298–309.

Lang, B. F., O'Kelly, C., Nerad, T., Gray, M. W. & Burger, G. 2002 The closest unicellular relatives of animals. *Curr. Biol.* **12**, 1773–1778.

Leadbetter, J. R. 2003 Cultivation of recalcitrant microbes: cells are alive, well and revealing their secrets in the 21st century laboratory. *Curr. Opin. Microbiol.* **6**, 274–281.

Leander, B. S., Clopton, R. E. & Keeling, P. J. 2003 Phylogeny of gregarines (Apicomplexa) as inferred from small-subunit rDNA and beta-tubulin. *Int. J. Syst. Evol. Microbiol.* **53**, 345–354.

López-García, P., Rodriguez-Valera, F., Pedros-Alio, C. & Moreira, D. 2001 Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* **409**, 603–607.

López-García, P., Philippe, H., Gail, F. & 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.

Milyutina, I. A., Aleshin, V. V., Mikrjukov, K. A., Kedrova, O. S. & Petrov, N. B. 2001 The unusually long small subunit ribosomal RNA gene found in amitochondriate amoeboflagellate *Pelomyxa palustris*: its rRNA predicted secondary structure and phylogenetic implication. *Gene* **272**, 131–139.

Moon-van der Staay, S. Y., De Wachter, R. & Vaulot, D. 2001 Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* **409**, 607–610.

Moreira, D. & López-García, P. 2002 The molecular ecology of microbial eukaryotes unveils a hidden world. *Trends Microbiol.* **10**, 31–38.

Morris, R. M., Rappe, M. S., Connon, S. A., Vergin, K. L., Siebold, W. A., Carlson, C. A. & Giovannoni, S. J. 2002 SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **420**, 806–810.

Mylnikov, A. P. 1999 New brackish water amoeboid flagellates of the genus *Amastigomonas* (Apusomonadida, Protozoa). *Zool. Zhurnal* **78**, 771–777. [In Russian.]

Philippe, H. 2000 Opinion: long branch attraction and protist phylogeny. *Protist* **151**, 307–316.

Philippe, H., Lopez, P., Brinkmann, H., Budin, K., Germot, A., Laurent, J., Moreira, D., Müller, M. & Le Guyader, H. 2000 Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions. *Proc. R. Soc. Lond.* B **267**, 1213–1221. (DOI 10.1098/rspb.2000.1130.)

Roger, A. J. 1999 Reconstructing early events in eukaryotic evolution. *Am. Nat.* **154**(Suppl. 4), S146–S163.

Sait, M., Hugenholtz, P. & Janssen, P. H. 2002 Cultivation of globally distributed soil bacteria from phylogenetic lineages previously only detected in cultivation-independent surveys. *Environ. Microbiol.* **4**, 654–666.

Silberman, J. D., Simpson, A. G., Kulda, J., Cepicka, I., Hampl, V., Johnson, P. J. & Roger, A. J. 2002 Retortamonad flagellates are closely related to diplomonads: implications for the history of mitochondrial function in eukaryote evolution. *Mol. Biol. Evol.* **19**, 777–786.

Simpson, A. G. & Roger, A. J. 2002 Eukaryotic evolution: getting to the root of the problem. *Curr. Biol.* **12**, R691–R693.

Simpson, A. G., Roger, A. J., Silberman, J. D., Leipe, D. D., Edgcomb, V. P., Jermiin, L. S., Patterson, D. J. & Sogin, M. L. 2002 Evolutionary history of 'early-diverging' eukaryotes: the excavate taxon *Carpediemonas* is a close relative of *Giardia*. *Mol. Biol. Evol.* **19**, 1782–1791.

Stechmann, A. & Cavalier-Smith, T. 2002 Rooting the eukaryote tree by using a derived gene fusion. *Science* **297**, 89–91.

Stechmann, A. & Cavalier-Smith, T. 2003*a* The root of the eukaryote tree pinpointed. *Curr. Biol.* **13**, R665–R666.

Stechmann, A. & Cavalier-Smith, T. 2003*b* Phylogenetic analysis of eukaryotes using heat-shock protein Hsp90. *J. Mol. Evol.* **57**, 408–419.

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

Stoeck, T., Taylor, G. T. & Epstein, S. S. 2003 Novel eukaryotes from the permanently anoxic Cariaco Basin (Caribbean Sea). *Appl. Environ. Microbiol.* **69**, 5656–5663.

Tovar, J., Leon-Avila, G., Sanchez, L. B., Sutak, R., Tachezy, J., Van Der Giezen, M., Hernandez, M., Müller, M. & Lucocq, J. M. 2003 Mitochondrial remnant organelles of *Giardia* function in iron–sulphur protein maturation. *Nature* **426**, 172–176.

Van de Peer, Y., Ali, A. B. & Meyer, A. 2000 Microsporidia: accumulating molecular evidence that a group of amitochondriate and suspectedly primitive eukaryotes are just curious fungi. *Gene* **246**, 1–8.

Vossbrinck, C. R., Maddox, J. V., Friedman, S., Debrunner-Vossbrinck, B. A. & Woese, C. R. 1987 Ribosomal RNA sequence suggests microsporidia are extremely ancient eukaryotes. *Nature* **326**, 411–414.

Williams, B. A., Hirt, R. P., Lucocq, J. M. & Embley, T. M. 2002 A mitochondrial remnant in the microsporidian *Trachipleistophora hominis*. *Nature* **418**, 865–869.

Zhang, H., Sekiguchi, Y., Hanada, S., Hugenholtz, P., Kim, H., Kamagata, Y. & Nakamura, K. 2003 *Gemmatimonas aurantiaca* gen. nov., sp. nov., a Gram-negative, aerobic, polyphosphate-accumulating micro-organism, the first cultured representative of the new bacterial phylum Gemmatimonadetes phyl. nov. *Int. J. Syst. Evol. Microbiol.* **53**, 1155–1163.

As this paper exceeds the maximum length normally permitted, the author has agreed to contribute to production costs.