environmental microbiology

Environmental Microbiology (2016) **18**(12), 4549–4563 doi:10.1111/1462-2920.13523

Oligotrophic lagoons of the South Pacific Ocean are home to a surprising number of novel eukaryotic microorganisms

Eunsoo Kim,[1*†] Ben Sprung,[2†] Solange Duhamel,[3] Christopher Filardi[4] and Mann Kyoon Shin[5]

[1] *Division of Invertebrate Zoology and Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, NY 10024, USA.*

[2] *Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA.*

[3] *Lamont-Doherty Earth Observatory, Division of Biology and Paleo Environment, Columbia University, Palisades, NY 10964, USA.*

[4] *Center for Biodiversity and Conservation, American Museum of Natural History, New York, NY 10024, USA.*

[5] *Department of Biological Sciences, University of Ulsan, Nam-Gu, Ulsan 44610, South Korea.*

Summary

The diversity of microbial eukaryotes was surveyed by environmental sequencing from tropical lagoon sites of the South Pacific, collected through the American Museum of Natural History (AMNH)'s Explore21 expedition to the Solomon Islands in September 2013. The sampled lagoons presented low nutrient concentrations typical of oligotrophic waters, but contained levels of chlorophyll *a*, a proxy for phytoplankton biomass, characteristic of meso- to eutrophic waters. Two 18S rDNA hypervariable sites, the V4 and V8–V9 regions, were amplified from the total of eight lagoon samples and sequenced on the MiSeq system. After assembly, clustering at 97% similarity, and removal of singletons and chimeras, a total of 2741 (V4) and 2606 (V8–V9) operational taxonomic units (OTUs) were identified. Taxonomic annotation of these reads, including phylogeny, was based on a combination of automated pipeline and manual inspection. About 18.4% (V4) and 13.8% (V8–V9) of the OTUs could not be assigned to any of the known eukaryotic groups. Of these, we focused on OTUs that were not divergent and possessed multiple sources of evidence for their existence. Phylogenetic analyses of these sequences revealed more than ten branches that might represent new deeply-branching lineages of microbial eukaryotes, currently without any cultured representatives or morphological information.

Introduction

Microbial eukaryotes play important roles in various ecosystems, including that of major primary producer (e.g. diatoms in coastal and upwelling areas of the ocean; Armbrust, 2009), keystone mutualist (e.g. the dinoflagellate *Symbiodinium* in coral reefs; Baker, 2003) and agent of disease (e.g. malaria parasites; Martinsen *et al*., 2008). In addition, microbial eukaryotes are critical for comprehending eukaryotic evolution and diversity: the vast majority of deep-level eukaryotic diversity is found among single-celled eukaryotic organisms (Adl *et al*., 2012). While consensus on the high-level classification of eukaryotes has not yet been reached, from a conservative point of view there are about ten major groups of eukaryotes, including the Stramenopila, Alveolata, Rhizaria (these first three constituting the 'SAR' clade; Burki *et al*., 2007), Amoebozoa, Chloroplastida, Cryptista, Excavata (the monophyly of this group remains controversial; e.g. Hampl *et al*., 2009; Katz and Grant, 2015), Haptophyta, Opisthokonta and Rhodophyta (Graham *et al*., 2009; Adl *et al*., 2012; Yabuki *et al*., 2014). In addition, there are several groups of eukaryotes that are relatively less studied and are often smaller in terms of known taxon diversity within group. These include the Ancyromonadida (Atkins *et al*., 2000; Glücksman *et al*., 2013), Apusomonadida (Karpov and Mylnikov, 1989; Cavalier-Smith and Chao, 2010), Breviatea (Cavalier-Smith *et al*., 2004; Walker *et al*., 2006; Brown *et al*., 2013), Centrohelida (Cavalier-Smith and Chao, 2003; Cavalier-Smith and von der Heyden, 2007; Cavalier-Smith and Chao, 2012), Collodictyonidae (or Diphyllatea) (Brugerolle *et al*., 2002; Zhao *et al*., 2012), Glaucophyta (Graham *et al*., 2009), Microhelida (Cavalier-Smith and Chao, 2003; Yabuki *et al*., 2012),

Received 27 July, 2016; accepted 30 August, 2016. *For correspondence. E-mail ekim1@amnh.org; Tel. +1 212 769 5510; Fax 1 212 769 5277. †Eunsoo Kim and Ben Sprung are co-first authors.

Rigifilida (Mikrjukov and Mylnikov, 2001; Yabuki *et al.*, 2013) and Telonemia (Klaveness *et al.*, 2005; Shalchian-Tabrizi *et al.*, 2006; Bråte *et al.*, 2010).

Additionally, new eukaryotic lineages continue to be discovered and characterized. One example is the Mantamonadida, a group of gliding biflagellates that was isolated in culture and first reported just several years ago (Glücksman *et al.*, 2011). *Palpitomonas* is another microbial eukaryote that was recently discovered (Yabuki *et al.*, 2010). Phylogenetic analyses based on multiple protein sequences suggest that this swimming heterotrophic biflagellate represents a major branch, together with cryptomonads and katablepharids, within the Cryptista (Yabuki *et al.*, 2014). Interestingly, since the advent of molecular sequencing tools that enable the characterization of microbial diversity directly from mixed environmental samples, some novel eukaryotic groups have been identified from sequence data before being investigated by morphology or isolated in culture. Examples include the Picozoa, a globally distributed group of heterotrophic flagellates (Not *et al.*, 2007b; Seenivasan *et al.*, 2013), and the rappemonads, a group of plastid-bearing phytoplankton, currently without any cultured representatives (Kim *et al.*, 2011).

Environmental sequencing is an excellent complementary approach to the traditional culture-based method in the study of microbial diversity. This is particularly so when organisms of interest are difficult to maintain under standard laboratory culture conditions (e.g. the picozoan *Picomonas*; Seenivasan *et al.*, 2013). A number of new rDNA sequence types have been uncovered from environmental sequencing; MAST (marine stramenopiles) and MALV (marine alveolates) groups are some of the prominent examples (López-García *et al.*, 2001; Not *et al.*, 2009; Logares *et al.*, 2012; Massana *et al.*, 2014).

In this study, the diversity of pico- and nano-sized microbial eukaryotes from oligotrophic tropical lagoon waters of the South Pacific Ocean was investigated using massively parallel sequencing technology with particular emphasis on new 18S rDNA sequences. Our work revealed a number of 18S rDNA types that do not show clear affinity to any known eukaryotic groups.

Results

Characteristics of the sampling sites

The sampling sites (Fig. 1, Supporting Information Fig. S1) presented very low nutrient concentrations, typical of open ocean surface waters (Table 1) (Conkright *et al.*, 2000; Zehr and Ward, 2002; Treguer and De La Rocha, 2013). While temperature data were not collected for the surveyed sites, nearby surface waters located further offshore were 27.5–28°C during the day. Nitrate plus nitrite (N + N) concentrations were below the detection limit of the auto-analyzer (20 nmol l^{-1}). Inorganic phosphate (P_i) concentrations were detected using the low-level MAGIC method. The Nirasa lagoon presented the lowest P_i concentrations, ranging from 25 ± 6 to 51 ± 6 nmo l^{-1} at Z2 and Z3, respectively. The New Georgia lagoon presented a gradient in P_i concentrations with values ranging from 47 ± 5 to 250 ± 65 nmol l^{-1} at Z2 and Isolated Reef respectively. Silicate concentrations were similarly low, with values ranging from 65 to 99 nmol l^{-1} and from 235 to 364 nmol l^{-1} in Nirasa and New Georgia respectively. Such low nutrient concentrations are typically found in oligotrophic waters and are expected to constrain phytoplankton growth. Yet chlorophyll *a* (Chl *a*) concentrations, a proxy for phytoplankton biomass, were relatively high in both lagoons, with values ≥ 0.8 μg l^{-1}, in the range typically found in meso- to eutrophic marine environments (Kletou and Hall-Spencer, 2012). Chlorophyll *a* concentrations were lower in Nirasa (0.798–1.666 μg l^{-1}) than in New Georgia (1.255–2.221 μg l^{-1}).

Fig. 1. Photos of the sampling sites: Nirasa (top) and New Georgia (bottom) lagoons.

18S rDNA sequencing and clustering

In order to survey the diversity of pico- and nano-sized microbial eukaryotes, eight water samples collected near

Table 1. Collection dates pH, salinity, and oxygen (mg l^{-1}), chlorophyll a (Chl *a*, μg l^{-1}), phosphate (P_i, nmol l^{-1}), silicate (Si, nmol l^{-1}) and nitrate plus nitrite (N + N) concentrations at the corresponding lagoon sites sampled in this study.

Sample	Collection date	Oxygen (mg l^{-1})	pH	Salinity (‰)	Chl a (μg l^{-1})	P_i (nmol l^{-1})	Si (nmol l^{-1})	N+N
Nirasa Z0	Sept. 19, 2013	na	na	na	na	na	na	na
Nirasa Z1	Sept. 19, 2013	5.6	8.2	35.5	0.837	32 ± 5	75	<DL
Nirasa Z2	Sept. 19, 2013	5.7	8.4	35.5	0.798	25 ± 6	99	<DL
Nirasa Z3	Sept. 19, 2013	6.3	8.4	35	1.666	51 ± 6	63	<DL
New George Z1	Sept. 22, 2013	3.5	7.9	30	1.295	na	na	na
New George Z2	Sept. 22, 2013	6.2	8.3	33.5	2.221	47 ± 5	235	<DL
New George Z3	Sept. 22, 2013	6.0	8.2	34	1.432	163 ± 19	360	<DL
New George IR	Sept. 22, 2013	5.0	8.1	34	1.255	250 ± 65	364	<DL

New Georgia and Nirasa islands were processed for DNA extraction, amplification of V4 and V8–V9 regions of 18S rDNA and sequencing on the Illumina's MiSeq platform. From each amplicon library, between 187 000 and 755 000 read pairs were obtained, tallying 3.6 and 4.5 million read pairs for the V4 and V8–V9 region respectively (Table 2). Removal of primers, quality trimming and merging read pairs preserved in both cases 83% of the reads, yielding 3.0 and 3.7 million merged reads respectively (Table 2). Of these, non-singleton reads were de-replicated, clustered at 97% sequence identity and checked for chimeric sequences, yielding 2741 (V4) and 2606 (V8–V9) final OTUs. The median cluster size was 168 for the V4 region and 61 for the V8–V9 region. The largest clusters were 183 662 (V4) and 181 426 (V8–V9) in size.

Taxonomic annotation of OTUs

The annotation of OTUs was first made by using usearch (Edgar, 2010) with the SILVA eukaryotic 18S rDNA database (Quast *et al.*, 2013), followed by manual curation. For both the V4 and V8–V9 data, the automated annotation was found to be very accurate for a sequence that had a match strength score (i.e. sequence identity) of 92 or higher. On the other hand, those sequences with a match strength score below 75 were long-branching in phylogenetic trees. Since such divergent sequences are difficult to classify due to the long-branch attraction (LBA) artifact (Philippe and Germot, 2000), these were annotated as 'not classified'. During manual inspection, inconsistencies in the taxonomic classification listing for some of the SILVA rRNA database entries were found and subsequently corrected. For instance, the microheliozoan sequence (AF534711) was ranked under the unrelated genus *Palpitomonas*, a MAST (stramenopile) sequence (KC488595) was listed as an uncultured rhizarian, and the sequence of the ancyromonad *Planomonas micra* (JF791081) was incorrectly classified under the family Apusomonadidae.

Those sequences having a match strength score between 75 and 92 were investigated by phylogeny. Sequences that were identified as alveolate, stramenopile or rhizarian were not subjected to this procedure, because preliminary assessment suggested that sequences representing these lineages were mostly accurately classified, at least in high-level classification (i.e. phylum and above). More than 50 entries were corrected during this procedure, for each of the two data sets. For example, some sequences that were obviously branching within the dinoflagellates were erroneously annotated as uncultured jakobids (Excavata). In other cases, taxonomic placement of sequences into known eukaryotic groups could not be made with great confidence, generally due to the sequence being long branching or possibly a member of a novel lineage. The latter possibility was further investigated as described in the

Table 2. Summary of 18S rDNA amplicon library data obtained from each lagoon sample.

	V4 (588...963)			V8–V9 (1435...1756)		
Sample	# read pairs	# merged	# clusters (97%)	# read pairs	# merged	# clusters (97%)
Nirasa Z0	696 515	470 354	735	295 151	202 855	713
Nirasa Z1	348 357	291 808	1230	514 403	353 111	860
Nirasa Z2	436 708	368 502	1268	695 297	484 349	1389
Nirasa Z3	382 470	322 310	1226	345 139	240 136	1290
New Georgia Z1	323 910	284 118	506	523 660	483 876	412
New Georgia Z2	622 864	568 446	247	651 079	606 246	316
New Georgia Z3	596 923	523 261	168	755 966	678 945	274
New Georgia IR	187 511	169 354	154	729 187	681 940	304
Total	3 595 258	2 998 153	2741	4 509 882	3 731 458	2606

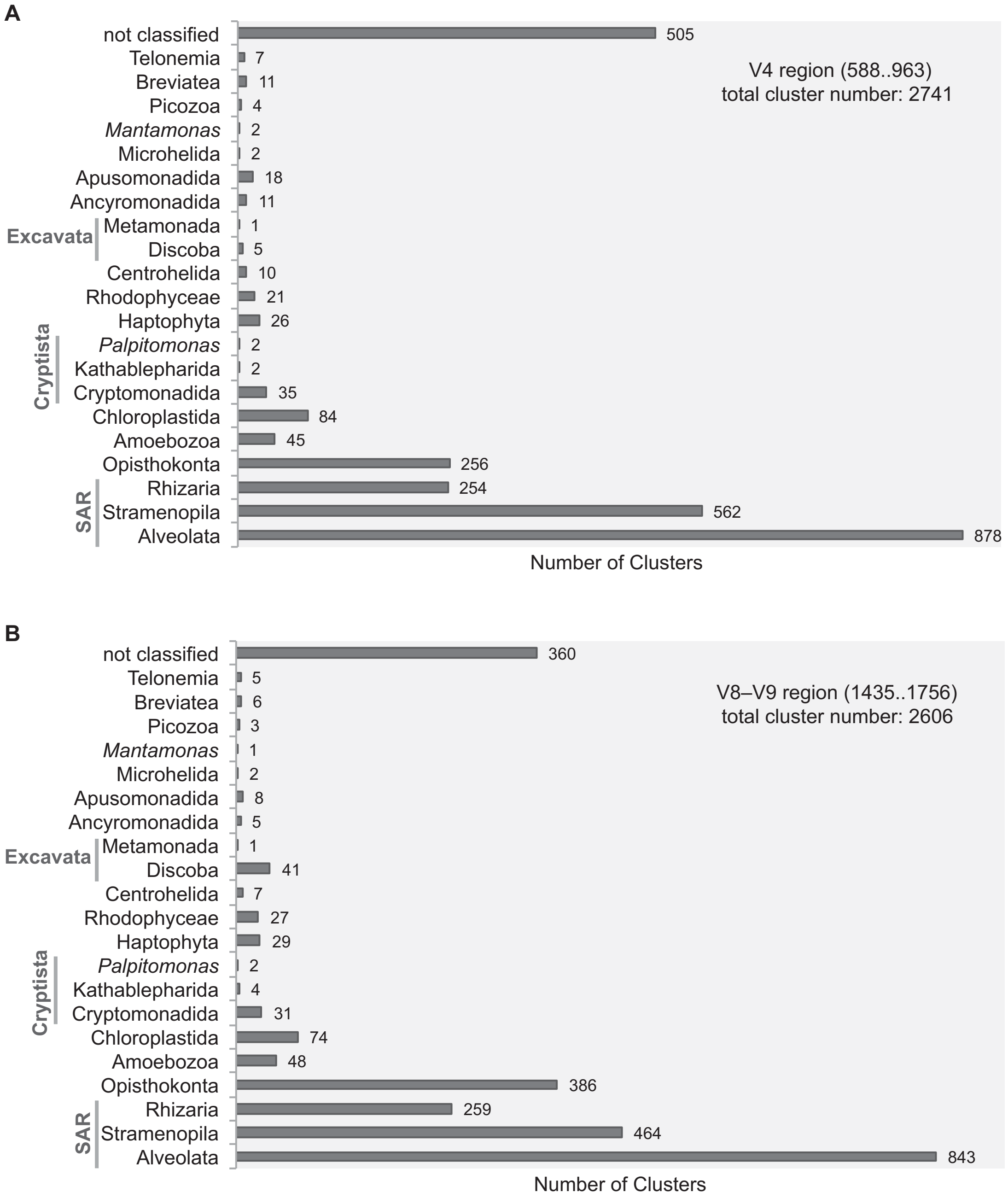


Fig. 2. High-level classification of OTU sequences for the V4 (A) and V8–V9 (B) regions.

following section. Overall, manual curation affected 4.5% (105/2311; V4) and 6.2% (146/2337; V8–V9) of the sequences having a match strength score of 75 or higher.

Taxonomic diversity of microbial eukaryotes in lagoons of the South Pacific Ocean

Of the OTUs, 81.6% (2236 out of 2741) for V4 and 86.2% (2246 out of 2606) for V8–V9 were placed to known eukaryotic groups, such as Opisthokonta, Chloroplastida and Breviatea (Fig. 2). In the lagoon samples investigated, Alveolata, which comprises dinoflagellates, ciliates and apicomplexans, as well as the paraphyletic protoalveolates (Janouškovec *et al*., 2013), was most represented, with more than 30% of the total OTUs (Fig. 2). Stramenopila, Opisthokonta and Rhizaria followed next to Aveolata; together, these four eukaryotic groups constituted 71.1% and 74.9% of the total OTU taxonomic diversity for V4 and V8–V9 data sets respectively (Fig. 2). The rest of the groups, including Amoebozoa, Centrohelida, Chloroplastida,

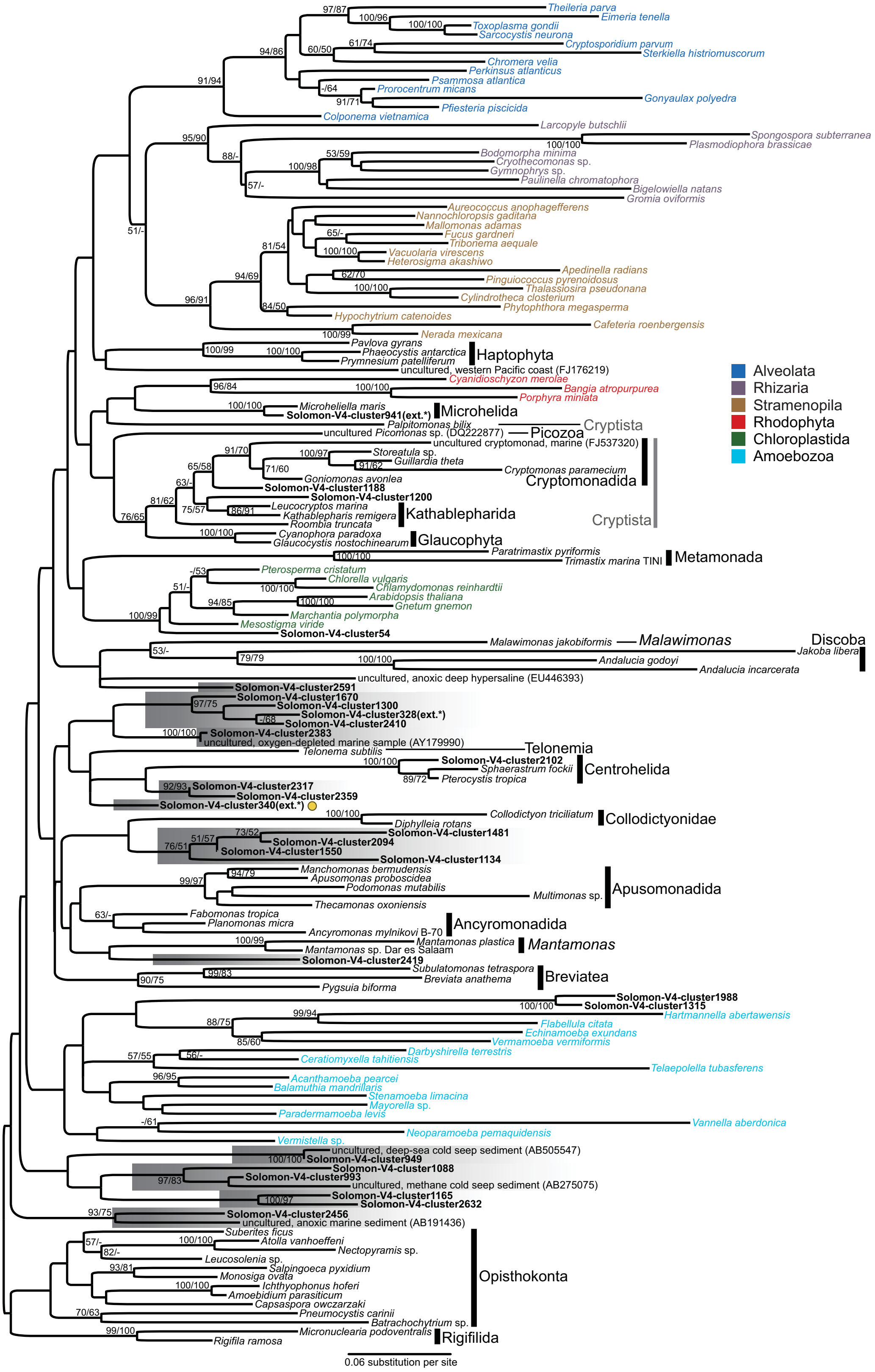

97/87
Theileria parva
100/96
Eimeria tenella
100/100
Toxoplasma gondii
Sarcocystis neurona
94/86
61/74
60/50
Cryptosporidium parvum
Sterkiella histriomuscorum
Chromera velia
Perkinsus atlanticus
Psammosa atlantica
-/64
Prorocentrum micans
91/94
91/71
Gonyaulax polyedra
Pfiesteria piscicida
Colponema vietnamica
Larcopyle butschlii
Spongospora subterranea
95/90
100/100
Plasmodiophora brassicae
88/-
53/59
Bodomorpha minima
100/98
Cryothecomonas sp.
Gymnophrys sp.
57/-
Paulinella chromatophora
Bigelowiella natans
Gromia oviformis
Aureococcus anophagefferens
Nannochloropsis gaditana
51/-
Mallomonas adamas
65/-
Fucus gardneri
81/54
Tribonema aequale
100/100
Vacuolaria virescens
Heterosigma akashiwo
94/69
Apedinella radians
62/70
Pinguiococcus pyrenoidosus
100/100
Thalassiosira pseudonana
96/91
Cylindrotheca closterium
84/50
Phytophthora megasperma
Hypochytrium catenoides
Cafeteria roenbergensis
100/99
Nerada mexicana
Pavlova gyrans
100/99
100/100
Phaeocystis antarctica
Haptophyta
Prymnesium patelliferum
uncultured, western Pacific coast (FJ176219)
Alveolata
Rhizaria
Stramenopila
Rhodophyta
Chloroplastida
Amoebozoa
Cyanidioschyzon merolae
96/84
100/100
Bangia atropurpurea
Porphyra miniata
100/100
Microheliella maris
Solomon-V4-cluster941(ext.*)
Microhelida
Palpitomonas bilix
Cryptista
uncultured Picomonas sp. (DQ222877)
Picozoa
uncultured cryptomonad, marine (FJ537320)
91/70
100/97
Storeatula sp.
91/62
Guillardia theta
65/58
71/60
Cryptomonas paramecium
Cryptomonadida
Goniomonas avonlea
63/-
Solomon-V4-cluster1188
Solomon-V4-cluster1200
81/62
Leucocryptos marina
75/57
86/91
Kathablepharis remigera
Kathablepharida
76/65
Roombia truncata
Cryptista
100/100
Cyanophora paradoxa
Glaucocystis nostochinearum
Glaucophyta
100/100
Paratrimastix pyriformis
Trimastix marina TINI
Metamonada
-/53
Pterosperma cristatum
Chlorella vulgaris
51/-
100/100
Chlamydomonas reinhardtii
Arabidopsis thaliana
94/85
100/100
Gnetum gnemon
Marchantia polymorpha
100/99
Mesostigma viride
Solomon-V4-cluster54
Malawimonas jakobiformis
Malawimonas
Discoba
53/-
Jakoba libera
79/79
100/100
Andalucia godoyi
Andalucia incarcerata
uncultured, anoxic deep hypersaline (EU446393)
Solomon-V4-cluster2591
Solomon-V4-cluster1670
97/75
Solomon-V4-cluster1300
Solomon-V4-cluster328(ext.*)
-/68
Solomon-V4-cluster2410
100/100
Solomon-V4-cluster2383
uncultured, oxygen-depleted marine sample (AY179990)
Telonemia
Telonema subtilis
100/100
Solomon-V4-cluster2102
Sphaerastrum fockii
Centrohelida
89/72
Pterocystis tropica
92/93
Solomon-V4-cluster2317
Solomon-V4-cluster2359
Solomon-V4-cluster340(ext.*)
100/100
Collodictyon triciliatum
Collodictyonidae
Diphylleia rotans
73/52
Solomon-V4-cluster1481
51/57
Solomon-V4-cluster2094
76/51
Solomon-V4-cluster1550
Solomon-V4-cluster1134
Manchomonas bermudensis
94/79
99/97
Apusomonas proboscidea
Podomonas mutabilis
Multimonas sp.
Apusomonadida
Thecamonas oxoniensis
63/-
Fabomonas tropica
Planomonas micra
Ancyromonas mylnikovi B-70
Ancyromonadida
100/99
Mantamonas plastica
Mantamonas
Mantamonas sp. Dar es Salaam
Solomon-V4-cluster2419
99/83
Subulatomonas tetraspora
Breviatea
90/75
Breviata anathema
Pygsuia biforma
Solomon-V4-cluster1988
100/100
Solomon-V4-cluster1315
99/94
Hartmannella abertawensis
88/75
Flabellula citata
Echinamoeba exundans
85/60
Vermamoeba vermiformis
57/55
56/-
Darbyshirella terrestris
Ceratiomyxella tahitiensis
Telaepolella tubasferens
96/95
Acanthamoeba pearcei
Balamuthia mandrillaris
Stenamoeba limacina
Mayorella sp.
Paradermamoeba levis
-/61
Vannella aberdonica
Neoparamoeba pemaquidensis
Vermistella sp.
uncultured, deep-sea cold seep sediment (AB505547)
100/100
Solomon-V4-cluster949
Solomon-V4-cluster1088
97/83
Solomon-V4-cluster993
uncultured, methane cold seep sediment (AB275075)
Solomon-V4-cluster1165
100/97
Solomon-V4-cluster2632
93/75
Solomon-V4-cluster2456
uncultured, anoxic marine sediment (AB191436)
Suberites ficus
57/-
100/100
Atolla vanhoeffeni
82/-
Nectopyramis sp.
Leucosolenia sp.
93/81
Salpingoeca pyxidium
Monosiga ovata
100/100
Ichthyophonus hoferi
Opisthokonta
Amoebidium parasiticum
Capsaspora owczarzaki
70/63
Pneumocystis carinii
Batrachochytrium sp.
99/100
Micronuclearia podoventralis
Rigifilida
Rigifila ramosa
0.06 substitution per site

Fig. 3. Maximum likelihood (ML) tree based on analyses of select OTU reads from the V4 data. OTUs that might represent novel eukaryotic lineages are highlighted in grey. Some OTU sequences that may represent new branches within known taxonomic groups are also included in the tree. An asterisk(*) indicates reads that were extended by additional PCR experiments as described in the text. One OTU, marked with an orange dot, has a matching entry in the BioMarKs 18S rDNA/rRNA environmental database. ML and MP bootstrap support values of 50 or higher are shown at the corresponding nodes.

Cryptista, Excavata, Haptophyta, Rhodophyta and several *incertae sedis* groups, add up to only 10.5% (V4) and 11.2% (V8–V9) of the OTU diversity.

Those OTUs that do not fall into known eukaryotic groups were identified by phylogenetic analyses (Figs. 3 and 4). To minimize the impact of artifacts that could be produced during sequence data generation or analysis, divergent (long branching) sequences, as well as those that were sampled from only one site, were excluded in final sequence matrices. Exceptions were made for those OTUs that branched robustly with other sequences. For the V4 data, 20 OTU sequences, forming 11 independent branches, represented short branches that did not associate with known clades (Fig. 3). A total of 20 OTU sequences, representing 13 branches, were identified from the V8–V9 data (Fig. 4). Of these, five OTU sequences (representing four monophyletic groups) from V4, and three from V8–V9, branched strongly with environmental sequences obtained previously from oxygen-depleted saline water basins (Alexander *et al.*, 2009; Stoeck and Epstein, 2003; Takishita *et al.*, 2005; 2007a; 2007b; 2010; Figs. 3 and 4). In addition, some of the novel OTUs – one from the V4 data and four from V8–V9 – had matches in the BioMarks and Tara Ocean databases, which include high throughput 18S rDNA/rRNA sequences from coastal and open ocean sites (Figs. 3 and 4; Supporting Information Table S1). The presence of two novel OTUs from the V4 data was further verified by PCR extension experiments using sequence-specific primers, as described by Kim *et al.* (2011). These two sequences were extended by about 350 bp towards the 5′ end. Interestingly, the majority of the novel sequence types identified herein were obtained from Nirasa sites, particularly the Z1, Z2 and Z3 sites (Fig. 5). It may be worth noting that Nirasa sites are more taxon-rich than New Georgia sites by greater than a factor of three (Table 2).

OTU distribution patterns and rarefaction analysis

When the OTUs were divided based on their frequency across the sampling sites, the following patterns were noted (Fig. 6). The OTUs that are more widely distributed, such as the dinoflagellate *Gymnodinium* sp. and the green alga *Tetraselmis* sp. in our study, are characterized by having, on average, higher match strength scores. In other words, OTUs with lower match scores tend to have limited geographical distribution. In our study, about half of all OTUs contained reads from more than one of the eight sample sites. The median cluster size increased near-monotonically with the number of sites represented, suggesting that widely distributed taxa tend to be numerically abundant.

Rarefaction curves were generated by repeated sub-sampling of the final OTU table, counting the number of distinct OTUs in each sub-sample. Curves showed asymptotic flattening at around 1.5M–2.5M reads (Fig. 7), suggesting that sequencing depth was sufficient, i.e. that few new OTUs would be found with more reads.

Discussion

Novel eukaryotic diversity

This study aims to evaluate the extent to which novel microbial eukaryotes, particularly those that branch outside of the known lineages, exist in poorly explored lagoon waters of the South Pacific Ocean by conducting massively parallel sequencing of 18S rDNA regions. We used a combination of automated and time-intensive manual procedures to accurately identify the sequenced reads.

For both the V4 and V8–V9 regions, more than 60% of the OTUs (clustered at 97%) were matched to their respective top reference entries (SILVA ver. 119) with less than 95% similarity. Such divergence corresponds to no more than 98% similarity if the full 18S rDNA region is considered; thus, these OTUs represent new taxa at least at the level of species (Caron *et al.*, 2009). Even using a conservative threshold of 90% identity, which translates to 90–94% identity across the full 18S rDNA region (Supporting Information Tables S2 and S3), 51% and 40% of the total OTUs for the V4 and V8–V9 regions, respectively, could not be matched to any reference reads. More than half of these represent novel lineages within known eukaryotic groups, leaving less than 20% of the total OTUs (18.4% for V4 and 13.8% for V8–V9) that could not be assigned to known eukaryotic groups.

Of these not-classified OTUs, the majority (93% for V4 and 86% for V8–V9) turned out to be long-branching in phylogeny. This is somewhat akin to the results in previous studies that looked at environmental sequence data generated by the Sanger method (e.g. see, Berney *et al.*, 2004; Not *et al.*, 2007a; Epstein

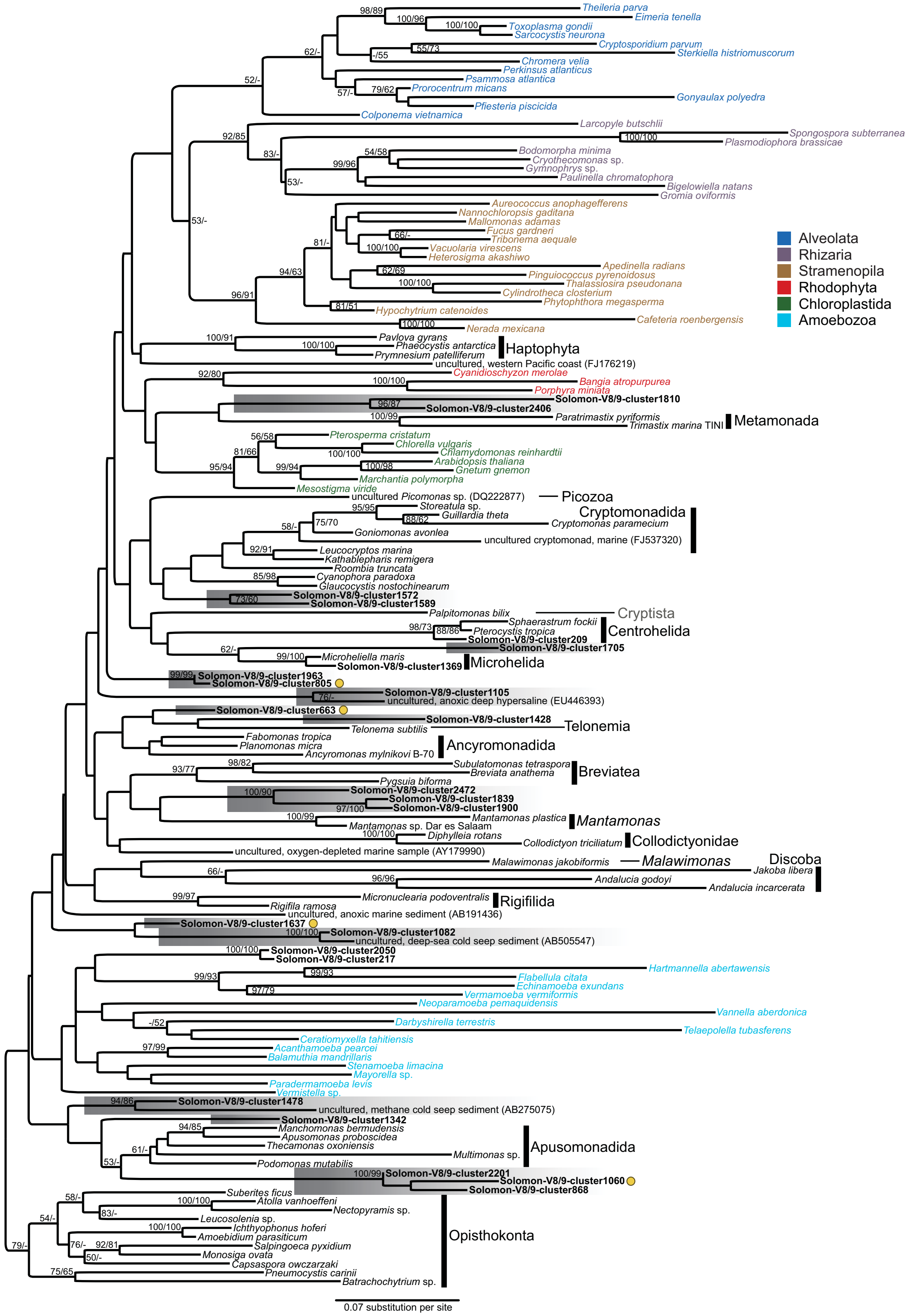
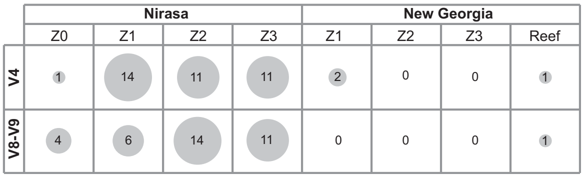


Fig. 4. Maximum likelihood tree based on analyses of select OTU reads from the V8–V9 data. OTUs that might represent novel eukaryotic lineages are highlighted in grey. OTUs that have matching entries in the Tara Ocean or BioMarKs 18S rDNA/rRNA environmental databases are marked with an orange dot. ML and MP bootstrap support values of 50 or higher are shown at the corresponding nodes.



	Nirasa				New Georgia			
	Z0	Z1	Z2	Z3	Z1	Z2	Z3	Reef
V4	1	14	11	11	2	0	0	1
V8-V9	4	6	14	11	0	0	0	1

Fig. 5. Distribution of the novel OTUs identified by phylogeny (Figs. 3 and 4) across the sampling sites. The majority of the novel reads were detected from Nirasa lagoon sites.

and López-García, 2008). For instance, in the study by Berney *et al*. (2004), out of over 400 18S sequences analyzed, five OTUs were found to be novel, of which four were long-branching. In our analyses, such fast-evolving reads were excluded from further analyses because they are prone to long-branch attraction artifacts and could wrongly appear as an independent branch from the group it actually belongs to (e.g. see, Keeling and Doolittle, 1996). After removal of divergent reads, only 1–2% of the total OTUs (36 out of 2741 for V4; 52 out of 2606 for V8–V9) remained. As far as read abundance is concerned, these novel OTUs constitute less than 0.1% of the total reads, supporting the notion that novel eukaryotic diversity is enriched in the so-called rare biosphere (de Vargas *et al*., 2015; Logares *et al*., 2015). In this study, we further focused on those OTUs that have recurring evidence for their presence in order to minimize possible artifacts. Even with our conservative approach, 11 and 13 new deep-branching groups of eukaryotes were identified from the V4 and V8–V9 datasets respectively (Figs. 4 and 5). These results suggest that our current understanding of eukaryotic diversity even at high taxonomic ranks remains limited by significant under-sampling.

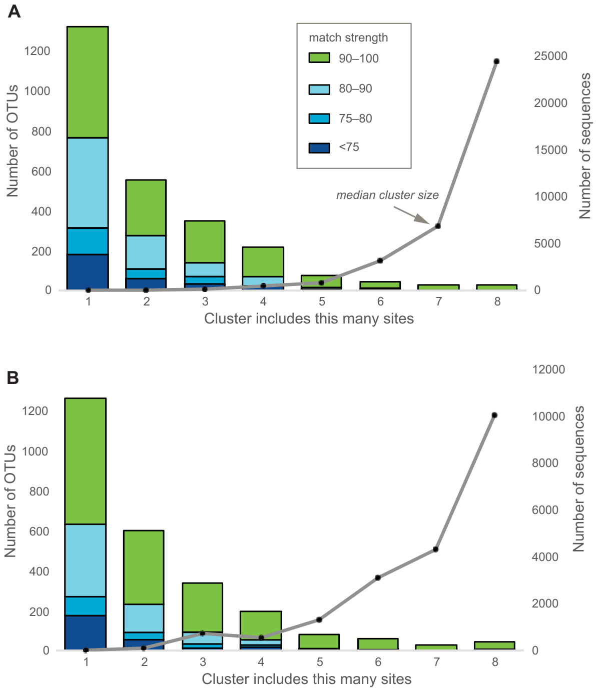


Fig. 6. Site representation and abundance of the OTUs for the V4 (A) and V8–V9 (B) data.
Those OTUs that are more widely distributed (i.e. represented in more sites) are characterized by having a larger cluster size (a proxy for abundance), a pattern also known from other environments (e.g. deep sea floor; Pawlowski *et al*., 2011) and tend to have higher match scores.

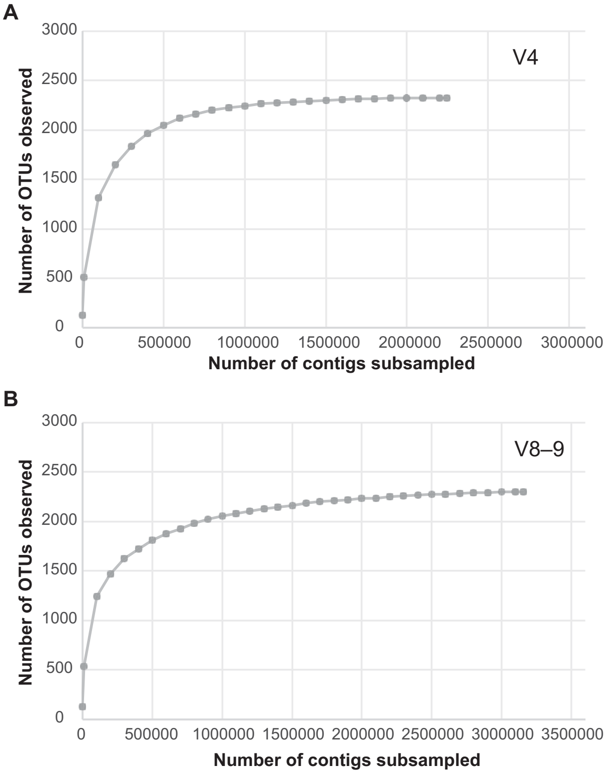


Fig. 7. Rarefaction curves for the V4 (A) and V8–V9 (B) data.
The Nirasa Z0 sample was excluded from the rarefaction analyses because this sample was size-fractionated differently from the others.

Taxonomic diversity of OTUs

The overall patterns in taxonomic diversity were similar between the V4 and V8–V9 data sets. In addition to having a comparable total OTU number (2741 vs. 2606), both sets had identical breadth in high-level taxon diversity and showed matching relative OTU richness across major eukaryotic groups (Fig. 2). This is likely because the two PCR primer sets used in this study targeted more or less the same assemblage of eukaryotic organisms, a broad range of eukaryotic diversity. The taxonomic breadth of microbial eukaryotes from all the lagoon sites combined into one dataset was high; all of the major eukaryotic groups, with particularly high representation by alveolates and stramenopiles, were reported (Fig. 2). Several *incertae sedis* taxa, such as picozoans, breviates, telonemids, mantamonads and microheliozoans, were also detected from the samples. Some, but not many, OTUs were identified as excavates; this may be due to PCR bias against excavates, which tend to have divergent 18S rDNA sequences (Simpson *et al*., 2002), or may reflect actual scarcity. For example, many members of the Metamonada are obligately anaerobic (Simpson and Roger, 2004), and thus are not expected to be found in oxygenated seawater like the surveyed equatorial lagoon regions (Table 1). Some protist groups were not detected at all from the lagoon samples. These include members of the collodictyonids, glaucophytes, malawimonads and rigifilids, all of which have thus far been known only from freshwater environments (O'Kelly and Nerad, 1999; Brugerolle *et al*., 2002; Graham *et al*., 2009; Yabuki *et al*., 2013). Therefore, the absence of OTUs corresponding to these groups is not surprising from our marine sites.

Methodological considerations

We designed PCR primers that take advantage of longer read length (up to 2 × 300 bp) available through the MiSeq system. The lengths of the amplified fragments were in the range of 283–539 bp (380 bp average) and 281–501 bp (325 bp average), after removing primer sites, for the V4 and V8–V9 regions, respectively. The V4 and V9 regions have already been shown in previous studies to be suitable as markers for the study of eukaryotic diversity (e.g. see, Amaral-Zettler *et al*., 2009; Dunthorn *et al*., 2012). Use of two variable rDNA regions instead of one was to reduce missing taxa. Even so, some taxa, particularly those with fast-evolving rDNA, may have not been amplified by either of the primer sets. There is at least one case of this: a heterotrophic protist of possible stramenopile affinity that we cultured from the Nirasa water, which has a highly divergent 18S rDNA sequence (data not shown) and is not represented in either of the data sets. In addition, those organisms that contain lineage-specific expansions within 18S rDNA also could have been missed in our study due to PCR amplification bias towards shorter fragments. Nevertheless, since our primary goal is the discovery of deeply branching, slow evolving eukaryotes of possibly novel taxonomic origins, the absence of highly divergent taxa is not expected to greatly affect our analyses.

Nirasa lagoon as a home to many previously uncharacterized eukaryotes?

Nirasa lagoon samples housed more of the novel deep-branching OTUs identified in this study than nearby New Georgia sites (Fig. 5). Both lagoon sites are similar in that they are (a) under the same equatorial, tropical climate; (b) have mangrove vegetation; (c) are covered with sediments that are pale in color and (d) are oligotrophic but with an elevated Chl *a* level (Table 1). However, the Nirasa lagoon is more isolated and contained, as it is immediately bordered by barrier reef. In contrast, the New Georgia lagoon is partially unbounded, and continues gradually into deeper ocean. Further, the Nirasa samples were slightly more saline and had less silicate than those from New Georgia (Table 1). While additional sampling and characterization of its abiotic and biotic features would be necessary to further confirm these observations, the Nirasa lagoon and perhaps other regions with similar physicochemical characteristics may be a prime location (in addition to anoxic habitats, e.g. Takishita *et al*., 2007a,b) for finding novel microorganisms that may hold important keys for understanding early eukaryotic evolution.

Perspective

High throughput sequencing technology has enabled a nearly comprehensive survey of microbial eukaryotic diversity from mixed environmental samples (Logares *et al*., 2014a; de Vargas *et al*., 2015). By applying this tool to poorly studied tropical lagoon waters of the South Pacific Ocean, we have identified more than ten 18S rDNA groups that do not show clear affinities to any known eukaryotic lineages. An obvious next step is to isolate them in cultures or characterize their morphology by fluorescence *in situ* hybridization (FISH) in order to formally describe such new groups. Flow-cytometry-based single-cell genomics (e.g. see, Yoon *et al*., 2011) would be another useful approach for investigating uncultured eukaryotes. However, our study suggests that the novel taxa inferred from the analyses, as evidenced by their low read counts, are rather scarce components of their ecosystems; if this is indeed the case, neither of the methods mentioned above are expected to be successful, due of their bias towards more abundant taxa. In fact, we suspect that inherent problems associated with rarity will be a major roadblock to our comprehensive characterization of microbial eukaryotic diversity in coming years. Another technological innovation, such as in the area of cell culturing, combined with added efforts in exploratory biodiversity research, may be necessary in order to systematically target the remaining microbial eukaryotic diversity present in nature.

Table 3. PCR Primers designed and used in this study.

Primer	5′ End	3′ End	Primer sequence (5′ → 3′)
Nex_18S_0587_F	570	587	TCG TCG GCA GCG TCA GAT GTG TAT AAG AGA CAG ***CCG CGG TAA TTC CAG CTC***
Nex_18S_0964_R	986	964	GTC TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G***GA TCC CYY AAC TTT CGT TCT TGA***
Nex_18S_1434_F	1412	1434	TCG TCG GCA GCG TCA GAT GTG TAT AAG AGA CAG ***GAG GCA ATA ACA GGT CTG TGA TG***
Nex_18S_1757_R	1777	1757	GTC TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G***CA GGT TCA CCT ACG GAA ACC T***

Sequences in bold and italics target 18S rRNA gene regions whereas those toward the 5′ end are part of the Illumina's Nextera adapters. Primer positions are relative to location within *Chlamydomonas reinhardtii* sequence (M32703).

Experimental procedures

Sampling

Samples were collected during the AMNH's Explore 21 expedition to the Solomon Islands, which took place in September 2013. Two to three liters of seawater were collected in the afternoon (14:00–16:00) by hand while walking along shallow lagoons or free-diving in inshore coral reefs adjacent to New Georgia (8° 19′ 01.13″ S, 157° 13′ 09.89″ E) and Nirasa (8° 46′ 29.92″ S, 157° 46′ 20.20″ E) Islands, both located in Western Province of Solomon Islands (Supporting Information Fig. S1). Lagoon samples were collected across three (New Georgia) to four (Nirasa) 'zones' associated with increasing depth and tidal flow regimes, and decreasing water temperature, as one moves from shallow, protected and largely stagnant interior shoreline areas to deeper lagoon centers exposed to wind, tidal currents and wave action. The New Georgia samples also included water collected around a small isolated reef edge (Supporting Information Fig. S1).

Oxygen and pH levels were measured on duplicate samples. Probes for dissolved oxygen and pH (AtlasScientific, Brooklyn, NY) were connected to Raspberry Pi computers (part#: RASPBRRY-MODB-512M). Probe calibration and sample measurements were performed using custom Python scripts. Salinity was measured using a refractometer. For chlorophyll *a* measurement, 200 ml seawater was filtered onto a 47 mm glass microfiber filter (Grade F borosilicate glass fiber; Sterlitech Corporation, Kent, WA) by gravity or under a low vacuum (< 600 mbar); the filter was folded into a cryovial and stored at −80°C until extraction and analysis. Chlorophyll extraction was performed by submerging a filter in 5 ml methanol; a few grains of sand were added to the solvent and the tube was vortexed to break open the cells attached to the filter. The filter was stored overnight at −20°C. Chlorophyll was measured using a Turner Designs TD-700 fluorometer using standard chlorophyll filters (E_x = 340–500 nm, $E_m \leq$ 650 nm). Relative fluorescence unit (RFU) values were converted to μg l^{-1} using a standard curve generated earlier.

Approximately 400 ml of seawater was stored into high-density polyethylene (HDPE) bottles at −20°C for analyses of inorganic nutrient concentrations. Note that samples from the New Georgia lagoon area were pre-filtered through a 40 μm nylon mesh strainer to remove sand particles. Nitrate + nitrite (N + N) and silicate (Si) analyses were performed on a SEAL Analytical AA3 HR AutoAnalyzer with software version 6.10 (Mequon, WI), following the Joint Global Ocean Flux Study (JGOFS) methods (Knap *et al.*, 1996), with detection limits of 20 and 30 nmol l^{-1}, for N + N and Si respectively. Phosphate (P$_i$) concentrations were determined by the magnesium-induced co-precipitation (MAGIC) method (detection limit 5 nmol l^{-1}; Karl and Tien, 1992; Thomson-Bulldis and Karl, 1998).

Amplicon library preparations and sequencing

Material for DNA extraction was collected on 47 mm polycarbonate membrane filters (Sterlitech Corporation, Kent, WA). A water sample (typically 250 ml unless the volume had to be decreased due to the presence of a large amount of particles) was first pre-filtered through a 8 μm polycarbonate membrane filter by gravity, and the eluent was filtered onto a 0.2 μm filter under a gentle vacuum (> 600 mbar). The only exception was the Nirasa Z0 sample, for which a 20 μm pre-filter was used instead. Each 0.2 μm filter was folded and placed in a cryovial and stored at −80°C. A Purelink DNA kit (Life Technologies, Carlsbad, CA) was used for DNA extraction. The membrane filter was thawed at RT and placed inside a 50 ml centrifuge tube with the microbial-mass-attached side facing inward. Lysis buffer was added to the tube, and the tube was sealed and continuously rotated vertically (end-over-end) at 55°C for 3 h. The rest of the steps followed the kit manufacturer's recommended protocol.

Two sets of PCR primers, each set targeting a region that includes either the V4 or V8–V9 regions within 18S rDNA (Table 3), were designed by examining an alignment that includes a broad spectrum of eukaryotic diversity (e.g. see, Kim *et al.*, 2011). Also included in these oligonucleotides were sequences that are part of the adapter regions for the Illumina Nextera sequencing platform (Table 3). PCR amplification was done using *TaKaRa Ex Taq* DNA polymerase (Clontech Laboratories, Mountain View, CA). A cyclic reaction consisted of (1) the initial incubation for 3 min at 95°C, (2) 25–30 cycles of [95°C for 30 s; 53°C for 30 s; 72°C for 2 min] and the final extension at 72°C for 5 min. Amplified products were cleaned up using Agencourt AMPure XP (Beckman Coulter, Pasadena, CA) following the manufacturer's protocol, except that the PCR product was mixed with an equal volume of AMPure XP solution. This reduced strength (0.5× as opposed to 0.64×) was empirically determined, and was necessary in order to remove unincorporated long PCR primers used in this study.

The cleaned PCR products were quantified using a Qubit dsDNA HS Assay kit (Life Technologies, Carlsbad, CA).

The rest of the adapter as well as index sequences – unique to each library – were added to the amplified 18S rDNA products by a subsequent PCR reaction using Nextera® index primers (Supporting Information Table S4). Approximately 50 ng of each of the 18S rDNA PCR products was used as a template in the total of 50 μl reaction volume. In addition to the index primers, two shorter, 'bridging' primers were added to facilitate the extension; these are 'Nex_Ext_P1' (5′-AAT GAT ACG GCG ACC ACC GA-3′) and 'Nex_Ext_P2' (5′-CAA GCA GAA GAC GGC ATA CGA-3′). The cyclic condition consisted of incubation at 72°C for 3 min, then at 98°C for 30 s, followed by 8 cycles of [98°C for 10 s; 63°C for 30 s; 72°C for 3 min]. Amplified products were cleaned using AMpure XP as described above, and a 1 μl aliquot was analyzed using an Agilent High Sensitivity DNA or DNA 1000 kit (Agilent Technologies, Santa Clara, CA). The amplicon libraries were sent to the New York Genome Center and Cornell Sequencing Core for sequencing on the MiSeq platform (2 × 300 bp).

Sequence assembly, clustering and annotation

Starting from raw reads, PCR primer sites were removed using cutadapt (ver. 1.2.1; Martin, 2011) and reads trimmed by quality score using Trimmomatic (ver. 0.30; Bolger *et al.*, 2014). Read pairs were then joined using FLASH (ver. 1.2.6; Magoč and Salzberg, 2011). We also used PEAR (ver. 0.9.8; Zhang *et al.*, 2014) for merging the paired reads, and no notable differences were found between the two tools. Of ~3.6 million raw paired V4 reads and ~4.5 million raw V8–V9 paired reads, ~83% (for both data sets) were successfully merged (Table 2).

Beginning with merged reads, we employed USEARCH (ver. 7.0; Edgar, 2010) in combination with the SILVA rRNA database (ver. 119; Quast *et al.*, 2013), to cluster and annotate the reads. After removing redundant sequences (i.e. dereplication), 65% (~1.9 million; V4) and 78% (~2.9 million; V8–V9) non-singleton reads were retained for OTU creation. The reads then were clustered using the UPARSE-OTU algorithm (Edgar, 2013) at 97% identity, a typical threshold value used in other studies (e.g. Massana *et al.*, 2015), generating a set of 2777 OTU sequences for V4 data and 2622 for V8–V9 data. The UPARSE algorithm performs *de novo* chimera filtering at it constructs OTUs. Next, we performed a reference-based chimera check on the OTU sequences using the uchime_ref command (Edgar *et al.*, 2011) and the SILVA database, removing 36 (1.3%) and 16 (0.6%) OTUs for V4 and V8–V9 data, respectively, for a final count of 2741 (V4) and 2606 (V8–V9) non-chimeric OTUs. The size of each cluster was measured by mapping reads back to the OTU sequences at 97% identity. For the V4 region, the clusters ranged in size from 2 to 183 662 and collectively comprised 2 827 573 reads (~94% of successfully merged reads). For the V8–V9 region, the clusters ranged in size from 2 to 181 426 and comprised 3 395 217 reads (~91% of successfully merged reads). Finally, the SILVA database was used to annotate the OTUs, using the usearch_global command of USEARCH.

This automated OTU annotation was further curated by manual entry inspection, BLAST searches against nr database and/or phylogenetic analyses. For instance, when phylogenetic placement was not clear (e.g. low bootstrap support), the annotation was revised as 'not classified'. The list of OTU sequences, its distribution and read count across the sampling sites and taxonomic annotation is provided in the Supporting Information Tables S5 and S6. The merged, dereplicated reads that were obtained in this study have been deposited to GenBank (accession numbers: KAGV00000000, KAGW00000000, KAGX00000000, KAGY00000000, KAGZ00000000, KAHA00000000, KAHB00000000, KAHC00000000, KAHD00000000, KAHE00000000, KAHF00000000, KAHG00000000, KAHH00000000, KAHI00000000, KAHJ00000000, KAHK00000000) and MG-RAST (ID: 4705419–4705434). All the custom scripts that include information on the specific parameters used in this study have been archived at Zenodo (DOI 10.5281/zenodo.56375).

Identification and phylogenetic analyses of novel sequences

Those clusters of sequences that were annotated with less than 92% match strength scores, the value we determined empirically by inspecting reads subsamples across the range of scores, were analyzed by phylogenetic methods in order to infer their evolutionary relationships to known eukaryotic groups. As our aim is to identify those sequences that may represent new deeply branching eukaryotic lineages and such that would be important for understanding early eukaryotic evolution, very fast evolving sequences were not included in the final sequence matrices due to their susceptibility to phylogenetic inference artifact (i.e. long branching attraction). The final sets of novel reads were analyzed by a hidden Markov models (HMM) based algorithm as described by Logares *et al.* (2014); all were confirmed to have 18S rDNA signatures.

Some of the novel sequences identified through the above-mentioned pipeline were further verified by PCR amplification using primers that were designed to specifically target select sequences (Supporting Information Table S7). Amplicons were gel-purified, if necessary, or otherwise cleaned up using a Wizard® SV Gel and PCR Clean-Up System (Promega Life Sciences, Wisconsin, USA) prior to cloned into the pGEM®-T Easy vector (Promega). About 3–10 colonies per cloning reaction were picked, screened and sequenced on an ABI 3730xl sequencer (Applied Biosystems, Foster City, CA). Note that many of our PCR extension attempts were unsuccessful due to false-positive amplifications of abundant taxa. Therefore, in the end, out of a dozen primer pairs tried, only two produced amplicons that correspond to their respective target regions.

Phylogenetic analyses were based on a modified, updated version of the 18S rDNA alignment used in a previous study (Kim and Archibald, 2013). OTU sequences were incorporated into the alignment by eye using Mesquite (ver. 2.75; Maddison and Maddison, 2011). Data matrices (including 1493 sites), which excluded ambiguously aligned regions, were analyzed using RAxML (ver. 8.1.11; Stamatakis, 2014) available on the CIPRES Science Gateway V. 3.3 (Miller *et al.*, 2010). The rapid bootstrap option with 1000 iterations was selected for maximum likelihood tree searching and bootstrap analysis. Maximum Parsimony bootstrap analysis was based on 1000 replicates and by using PAUP* (ver. 4.0; Swofford, 2003). The sequence alignments used in this study are included in the Supporting Information. In addition to manual alignment of sequence data, we also explored the use of an automated option – a combination of MAFFTS (with the –addfragments option; Katoh and Frith, 2012) and Gblocks (Talavera and Castresana, 2007). While both methods produced comparable phylogenies (data not shown), we opted for the manual alignment as the automated procedure produced at least some obvious cases of misalignments. Although our analyses here (Figs. 3 and 4) were based on manual alignment, our recommended strategy for analyzing a large number of OTUs is to pre-process sequences by automated alignment, followed by manual curation.

Comparison to other high-throughput amplicon data

The novel reads identified as described above were compared to the Tara Ocean (de Vargas *et al*., 2015) and BioMarKs (Logares *et al*., 2014) 18S rDNA/rRNA amplicon sequence data. For reference, the Tara Ocean data were collected from more than 300 euphotic oceanic sites across the globe; the BioMarKs data were from six European coastal offshore sites. We downloaded these data sets, converted the data to FASTA format using standard bash tools when necessary and queried our novel OTUs against the resulting fasta files using the usearch_global command of USEARCH.

Rarefaction and OTU distribution analyses

To construct rarefaction curves, we created a script that accepted the size distribution of OTUs as input, randomly subsampled at intervals of 100 000 observations and counted the number of distinct OTUs recovered at each sampling. For each interval, we used the average count of ten subsamplings. Note that the Nirasa Z0 sample was excluded from the rarefaction analyses because the sample was size-fractionated differently from the others (0.2–20 μm vs. 0.2–8 μm).

To determine how many reads in each cluster were from each study site, we used usearch's 'uc2otutab.py' script to generate an OTU table from the output of usearch_global. We also computed the median size of each cluster, binned by how many of the eight sites provided sequences to the cluster.

Acknowledgements

The work was supported by grants to EK from the Dalio Foundation and Simons Foundation (SF-382790); and to EK & SD from the National Science Foundation (OCE-1458095 & OCE-1458070). The authors thank Nigel D'souza and Naomi Shelton for assistance with chlorophyll and nutrient analyses, respectively, and Aaron Heiss for extensive editorial assistance. For field sampling, the authors thank Dr. Melchior Mataki and the Solomon Islands Ministry of Environment, Climate Change, Disaster Management and Meteorology, Rosalie Masu and the Solomon Islands Ministry of Fisheries and Marine Resources, the Western Province Government, and the customary landholders of Vonavona Lagoon and the Hele Bar Islands. They also thank anonymous reviewers who provided very helpful suggestions on an earlier version of this manuscript. The authors declare that there is no conflict of interest concerning this work.

References

Adl, S.M., Simpson, A.G., Lane, C.E., Lukeš, J., Bass, D., Bowser, S.S., *et al.* (2012) The revised classification of eukaryotes. *J Euk Microbiol* **59:** 429–514.

Alexander, E., Stock, A., Breiner, H.W., Behnke, A., Bunge, J., Yakimov, M.M., and Stoeck, T. (2009) Microbial eukaryotes in the hypersaline anoxic L'Atalante deep-sea basin. *Environ Microbiol* **11:** 360–381.

Amaral-Zettler, L.A., McCliment, E.A., Ducklow, H.W., and Huse, S.M. (2009) A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLoS One* **4:** e6372.

Armbrust, E.V. (2009) The life of diatoms in the world's oceans. *Nature* **459:** 185–192.

Atkins, M.S., McArthur, A.G., and Teske, A.P. (2000) Ancyromonadida: a new phylogenetic lineage among the protozoa closely related to the common ancestor of metazoans, fungi, and choanoflagellates (Opisthokonta). *J Mol Evol* **51:** 278–285.

Baker, A.C. (2003) Flexibility and specificity in coral-algal symbiosis: diversity, ecology, and biogeography of *Symbiodinium*. *Annu Rev Ecol Evol Syst* **34:** 661–689.

Berney, C., Fahrni, J., and Pawlowski, J. (2004) How many novel eukaryotic 'kingdoms'? Pitfalls and limitations of environmental DNA surveys. *BMC Biol* **2:** 13.

Bolger, A.M., Lohse, M., and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30:** 2114–2120.

Bråte, J., Klaveness, D., Rygh, T., Jakobsen, K.S., and Shalchian-Tabrizi, K. (2010) Telonemia-specific environmental 18S rDNA PCR reveals unknown diversity and multiple marine-freshwater colonizations. *BMC Microbiol* **10:** 168.

Brown, M.W., Sharpe, S.C., Silberman, J.D., Heiss, A.A., Lang, B.F., Simpson, A.G., and Roger, A.J. (2013) Phylogenomics demonstrates that breviate flagellates are related to opisthokonts and apusomonads. *Proc Biol Sci* **280:** 20131755.

Brugerolle, G., Bricheux, G., Philippe, H., and Coffe, G. (2002) *Collodictyon triciliatum* and *Diphylleia rotans* (= *Aulacomonas submarina*) form a new family of flagellates (Collodictyonidae) with tubular mitochondrial cristae that is phylogenetically distant from other flagellate groups. *Protist* **153:** 59–70.

Burki, F., Shalchian-Tabrizi, K., Minge, M., Skjaeveland, A., Nikolaev, S.I., Jakobsen, K.S., and Pawlowski, J. (2007) Phylogenomics reshuffles the eukaryotic supergroups. *PLoS One* **2:** e790.

Caron, D.A., Countway, P.D., Savai, P., Gast, R.J., Schnetzer, A., Moorthi, S.D., *et al.* (2009) Defining DNA-based operational taxonomic units for microbial-eukaryote ecology. *Appl Environ Microbiol* **75:** 5797–5808.

Cavalier-Smith, T., and Chao, E.E. (2003) 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. (2010) Phylogeny and evolution of apusomonadida (protozoa: apusozoa): new genera and species. *Protist* **161:** 549–576.

Cavalier-Smith, T., and Chao, E.E. (2012) *Oxnerella micra* sp. n.(Oxnerellidae fam. n.), a tiny naked centrohelid, and the diversity and evolution of heliozoa. *Protist* **163:** 574–601.

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

Cavalier-Smith, T., and von der Heyden, S. (2007) Molecular phylogeny, scale evolution and taxonomy of centrohelid heliozoa. *Mol Phylogenet Evol* **44:** 1186–1203.

Conkright, M.E., Gregg, W.W., and Levitus, S. (2000) Seasonal cycle of phosphate in the open ocean. *Deep-Sea Res I* **47:** 159–175.

Dunthorn, M., Klier, J., Bunge, J., and Stoeck, T. (2012) Comparing the hyper-variable V4 and V9 regions of the small subunit rDNA for assessment of ciliate environmental diversity. *J Euk Microbiol* **59:** 185–187.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26:** 2460–2461.

Edgar, R.C. (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods* **10:** 996–998.

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., and Knight, R. (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27:** 2194–2200.

Epstein, S., and López-García, P. (2008) "Missing" protists: a molecular prospective. *Biodivers Conserv* **17:** 261–276.

Glücksman, E., Snell, E.A., Berney, C., Chao, E.E., Bass, D., and Cavalier-Smith, T. (2011) The novel marine gliding zooflagellate genus *Mantamonas* (Mantamonadida ord. n.: Apusozoa). *Protist* **162:** 207–221.

Glücksman, E., Snell, E.A., and Cavalier-Smith, T. (2013) Phylogeny and evolution of Planomonadida (Sulcozoa): eight new species and new genera *Fabomonas* and *Nutomonas*. *Eur J Protistol* **49:** 179–200.

Graham, L., Graham, J., and Wilcox, L. (2009) *Algae*, 2nd ed. San Francisco, USA: Pearson Education.

Hampl, V., Hug, L., Leigh, J.W., Dacks, J.B., Lang, B.F., Simpson, A.G., and Roger, A.J. (2009) Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". *Proc Natl Acad Sci USA* **106:** 3859–3864.

Janouškovec, J., Tikhonenkov, D.V., Mikhailov, K.V., Simdyanov, T.G., Aleoshin, V.V., Mylnikov, A.P., and Keeling, P.J. (2013) Colponemids represent multiple ancient alveolate lineages. *Curr Biol* **23:** 2546–2552.

Karl, D.M., and Tien, G. (1992) MAGIC: a sensitive and precise method for measuring dissolved phosphorus in aquatic environments. *Limnol Oceanogr* **37:** 105–116.

Karpov, S.A., and Mylnikov, A.P. (1989) Biology and ultrastructure of colourless flagellates Apusomonadida ord. n. *Zool Zh* **68:** 5–17.

Katoh, K., and Frith, M.C. (2012) Adding unaligned sequences into an existing alignment using MAFFTS and LAST. *Bioinformatics* **28:** 3144–3146.

Katz, L.A., and Grant, J.R. (2015) Taxon-rich phylogenomic analyses resolve the eukaryotic tree of life and reveal the power of subsampling by sites. *Syst Biol* **64:** 406–415.

Keeling, P.J., and Doolittle, W.F. (1996) Alpha-tubulin from early-diverging eukaryotic lineages and the evolution of the tubulin family. *Mol Biol Evol* **13:** 1297–1305.

Kim, E., and Archibald, J.M. (2013) Ultrastructure and molecular phylogeny of the cryptomonad *Goniomonas avonlea* sp. nov. *Protist* **164:** 160–182.

Kim, E., Harrison, J.W., Sudek, S., Jones, M.D., Wilcox, H.M., Richards, T.A., *et al.* (2011) Newly identified and diverse plastid-bearing branch on the eukaryotic tree of life. *Proc Natl Acad Sci USA* **108:** 1496–1500.

Klaveness, D., Shalchian-Tabrizi, K., Thomsen, H.A., Eikrem, W., and Jakobsen, K.S. (2005) *Telonema antarcticum* sp. nov., a common marine phagotrophic flagellate. *Int J Syst Evol Microbiol* **55:** 2595–2604.

Kletou, D., and Hall-Spencer, J.M. (2012). Threats to ultraoligotrophic marine ecosystems. In *Marine Ecosystems*. Antonio Cruzado, A. (ed). InTech, pp 1–34. ISBN: 978-953-51-0176-5, DOI: 10.5772/34842. URL www.intechopen.com/books/marine-ecosystems/threats-to-ultraoligotrophic-marine-ecosystems. Rijeka, Croatia.

Knap, A.H., Michaels, A., Close, A.R., Ducklow, H., and Dickson, A.G. (1996) Protocols for the Joint Global Ocean Flux Study (JGOFS) Core Measurements. JGOFS, Reprint of the IOC Manuals and Guides No. 29, UNESCO 19. Paris, France.

Logares, R., Audic, S., Santini, S., Pernice, M.C., de Vargas, C., and Massana, R. (2012) Diversity patterns and activity of uncultured marine heterotrophic flagellates unveiled with pyrosequencing. *ISME J* **6:** 1823–1833.

Logares, R., Audic, S., Bass, D., Bittner, L., Boutte, C., Christen, R., *et al.* (2014a) Patterns of rare and abundant marine microbial eukaryotes. *Curr Biol* **24:** 813–821.

Logares, R., Sunagawa, S., Salazar, G., Cornejo-Castillo, F.M., Ferrera, I., Sarmento, H., *et al.* (2014b) Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environ Microbiol* **16:** 2659–2671.

Logares, R., Mangot, J.F., and Massana, R. (2015) Rarity in aquatic microbes: placing protists on the map. *Res Microbiol* **166:** 831–841.

López-García, P., Rodriguez-Valera, F., Pedrós-Alió, C., and Moreira, D. (2001) Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* **409:** 603–607.

Maddison, W.P. and Maddison, D.R. (2011) Mesquite: a modular system for evolutionary analysis. Version 2.75. Available at: http://mesquiteproject.org.

Magoč, T., and Salzberg, S.L. (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27:** 2957–2963.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* **17:** 10–12.

Martinsen, E.S., Perkins, S.L., and Schall, J.J. (2008) A three-genome phylogeny of malaria parasites (*Plasmodium* and closely related genera): evolution of life-history traits and host switches. *Mol Phylogenet Evol* **47:** 261–273.

Massana, R., del Campo, J., Sieracki, M.E., Audic, S., and Logares, R. (2014) Exploring the uncultured microeukaryote majority in the oceans: reevaluation of ribogroups within stramenopiles. *ISME J* **8:** 854–866.

Massana, R., Gobet, A., Audic, S., Bass, D., Bittner, L., Boutte, C., *et al.* (2015) Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. *Environ Microbiol* **17:** 4035–4049.

Mikrjukov, K.A., and Mylnikov, A.P. (2001) A study of the fine structure and the mitosis of a lamellicristate amoeba, *Micronuclearia podoventralis* gen. et sp. nov.(Nucleariidae, Rotosphaerida). *Eur J Protistol* **37:** 15–24.

Miller, M., Pfeiffer, W., Schwartz, T. (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *Gateway Computing Environments Workshop (GCE) IEEE*, pp. 1–8. 10.1109/GCE.2010.5676129. La Jolla, California, USA.

Not, F., Gausling, R., Azam, F., Heidelberg, J.F., and Worden, A.Z. (2007a) Vertical distribution of picoeukaryotic diversity in the Sargasso Sea. *Environ Microbiol* **9:** 1233–1252.

Not, F., Valentin, K., Romari, K., Lovejoy, C., Massana, R., Töbe, K., Vaulot, D., and Medlin, L.K. (2007b) Picobiliphytes: a marine picoplanktonic algal group with unknown affinities to other eukaryotes. *Science* **315:** 253–255.

Not, F., del Campo, J., Balagué, V., de Vargas, C., and Massana, R. (2009) New insights into the diversity of marine picoeukaryotes. *PLoS One* **4:** e7143.

O'Kelly, C.J., and Nerad, T.A. (1999) *Malawimonas jakobiformis* n. gen., n. sp. (Malawimonadidae n. fam.): a *Jakoba*-like heterotrophic nanoflagellate with discoidal mitochondrial cristae. *J Euk Microbiol* **46:** 522–531.

Pawlowski, J., Christen, R., Lecroq, B., Bachar, D., Shahbazkia, H.R., Amaral-Zettler, L., and Guillou, L. (2011) Eukaryotic richness in the abyss: insights from pyrotag sequencing. *PLoS One* **6:** e18169.

Philippe, H., and Germot, A. (2000) Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution. *Mol Biol Evol* **17:** 830–834.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., *et al*. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41:** D590–D596.

Seenivasan, R., Sausen, N., Medlin, L.K., and Melkonian, M. (2013) *Picomonas judraskeda* gen. et sp. nov.: the first identified member of the Picozoa phylum nov., a widespread group of picoeukaryotes, formerly known as "picobiliphytes". *PLoS One* **8:** e59565.

Shalchian-Tabrizi, K., Eikrem, W., Klaveness, D., Vaulot, D., Minge, M.A., Le Gall, F., *et al*. (2006) Telonemia, a new protist phylum with affinity to chromist lineages. *Proc Biol Sci* **273:** 1833–1842.

Simpson, A.G., and Roger, A.J. (2004) Excavata and the origin of amitochondriate eukaryotes. In *Organelles, Genomes, and Eukaryote Phylogeny: An Evolutionary Synthesis in the Age of Genomics*. Hirt, R.P., and Horner, D. S. (eds). Boca Raton, USA: CRC Press LLC, pp. 27–54.

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

Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30:** 1312–1313.

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.

Swofford, D.L. (2003) *PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods). Version 4.0*. Sunderland, Massachusetts: Sinauer Associates.

Takishita, K., Miyake, H., Kawato, M., and Maruyama, T. (2005) Genetic diversity of microbial eukaryotes in anoxic sediment around fumaroles on a submarine caldera floor based on the small-subunit rDNA phylogeny. *Extremophiles* **9:** 185–196.

Takishita, K., Tsuchiya, M., Kawato, M., Oguri, K., Kitazato, H., and Maruyama, T. (2007a) Genetic diversity of microbial eukaryotes in anoxic sediment of the saline meromictic lake Namako-ike (Japan): on the detection of anaerobic or anoxic-tolerant lineages of eukaryotes. *Protist* **158:** 51–64.

Takishita, K., Yubuki, N., Kakizoe, N., Inagaki, Y., and Maruyama, T. (2007b) Diversity of microbial eukaryotes in sediment at a deep-sea methane cold seep: surveys of ribosomal DNA libraries from raw sediment samples and two enrichment cultures. *Extremophiles* **11:** 563–576.

Takishita, K., Kakizoe, N., Yoshida, T., and Maruyama, T. (2010) Molecular evidence that phylogenetically diverged ciliates are active in microbial mats of deep-sea cold-seep sediment. *J Euk Microbiol* **57:** 76–86.

Talavera, G., and Castresana, J. (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* **56:** 564–577.

Thomson-Bulldis, A., and Karl, D. (1998) Application of a novel method for phosphorus determinations in the oligotrophic North Pacific Ocean. *Limnol Oceanogr* **43:** 1565–1577.

Treguer, P.J., and De La Rocha, C.L. (2013) The world ocean silica cycle. *Annu Rev Mar Sci* **5:** 477–501.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares., *et al*. (2015) Eukaryotic plankton diversity in the sunlit ocean. *Science* **348:** 1261605.

Walker, G., Dacks, J.B., and Martin Embley, T. (2006) Ultrastructural description of *Breviata anathema*, n. gen., n. sp., the organism previously studied as "*Mastigamoeba invertens*". *J Euk Microbiol* **53:** 65–78.

Yabuki, A., Inagaki, Y., and Ishida, K.I. (2010) *Palpitomonas bilix* gen. et sp. nov.: a novel deep-branching heterotroph possibly related to Archaeplastida or Hacrobia. *Protist* **161:** 523–538.

Yabuki, A., Chao, E.E., Ishida, K.I., and Cavalier-Smith, T. (2012) *Microheliella maris* (Microhelida ord. n.), an ultrastructurally highly distinctive new axopodial protist species and genus, and the unity of phylum Heliozoa. *Protist* **163:** 356–388.

Yabuki, A., Ishida, K.I., and Cavalier-Smith, T. (2013) *Rigifila ramosa* n. gen., n. sp., a filose apusozoan with a distinctive pellicle, is related to *Micronuclearia*. *Protist* **164:** 75–88.

Yabuki, A., Kamikawa, R., Ishikawa, S.A., Kolisko, M., Kim, E., Tanabe, A.S., *et al*. (2014) *Palpitomonas bilix* represents a basal cryptist lineage: insight into the character evolution in Cryptista. *Sci Rep* **4:** 4641.

Yoon, H.S., Price, D.C., Stepanauskas, R., Rajah, V.D., Sieracki, M.E., Wilson, W.H., *et al*. (2011) Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science* **332:** 714–717.

Zehr, J.P., and Ward, B.B. (2002) Nitrogen cycling in the ocean: new perspectives on processes and paradigms. *Appl Environ Microbiol* **68:** 1015–1024.

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30:** 614–620.

Zhao, S., Burki, F., Bråte, J., Keeling, P.J., Klaveness, D., and Shalchian-Tabrizi, K. (2012) *Collodictyon*—an ancient lineage in the tree of eukaryotes. *Mol Biol Evol* **29:** 1557–1568.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web-site

Table S1. Those novel OTUs that have significant matches in the Tara Ocean and BioMarKs 18S rDNA/rRNA databases. Five OTUs (out of forty) had matches with the similarity score of greater than 92%.

Table S2. Pairwise comparison of the *Goniomonas avonlea* 18S rDNA (J1434475) to closely related or distant sequences across the near full length, V4 or V8–V9 regions. A distance was calculated using EMBOSS Needle (Rice *et al*., 2000, Trends Genet 16:276–277).

Table S3. Pairwise comparison of the *Ancyromonas sigmoides* 18S rDNA (strain HFCC62; AY827844) to closely related or distant sequences across the near full length, V4 or V8–V9 regions.

Table S4. Nextera Index PCR primers used in this study. Index sites are indicated in grey box.

Table S5. List of OTUs from the V4 data. (see excel file attached)

Table S6. List of OTUs from the V8–V9 data. (see excel file attached)

Table S7. PCR primers used for extending two novel reads identified in this study: Solomon_V4_328 and Solomon_V4_340. The primers nu-SSU-0024-5' and nu-SSU-0033-5' are adapted from the previous study by Kim *et al*. (2006 Mol Biol Evol 23: 2455–2466).

Fig. S1. Maps showing the geographical location of the sampling sites around New Georgia and Nirasa islands. A. The islands are located in the Western Province of the Solomon Islands. B. New Georgia samples were collected first by foot from a mangrove forest area (Z1) to the deeper part of the lagoon (Z2, Z3); then by diving from boat at an isolated reef region (Isolated Reef). C. Nirasa samples were collected by foot along a gradient of temperature and water depth from Z0 (closest to land) to Z3. The maps were made using Google Earth, accessed in January of 2016.

File S1. The unmasked version of the 18S rDNA alignment used for generating the tree as shown in Fig. 3.

File S2. The masked version of the 18S rDNA alignment used for generating the tree as shown in Fig. 3.

File S3. The unmasked version of the 18S rDNA alignment used for generating the tree as shown in Fig. 4.

File S4. The masked version of the 18S rDNA alignment used for generating the tree as shown in Fig. 4.