npg

ORIGINAL ARTICLE

Microbial community structure in the North Pacific ocean

Mark V Brown[1,5], Gayle K Philip[1], John A Bunge[2], Matthew C Smith[3], Andrew Bissett[4], Federico M Lauro[5], Jed A Fuhrman[6] and Stuart P Donachie[1,7]

[1]NASA Astrobiology Institute, University of Hawaii, Honolulu, HI, USA; [2]Department of Statistical Science, Cornell University, Ithaca, NY, USA; [3]Department of Marine Sciences, University of Puerto Rico, Mayagüez, PR, USA; [4]Max Planck Institute for Marine Microbiology, Bremen, Germany; [5]Department of Biotechnology and Biomolecular Science, University of New South Wales, Sydney, New South Wales, Australia; [6]Department of Biological Sciences, University of Southern California, Los Angeles, CA, USA and [7]Department of Microbiology, University of Hawaii, Honolulu, HI, USA

We report a ribosomal tag pyrosequencing study of the phylogenetic diversity of *Archaea*, *Bacteria* and *Eucarya* over a depth profile at the Hawaii Ocean Time-Series Station, ALOHA. The V9 region of the SSU rRNA gene was amplified from samples representing the epi- (10 m), meso- (800 m) and bathy- (4400 m) pelagia. The primers used are expected to amplify representatives of $\sim$80% of known phylogenetic diversity across all three domains. Comparisons of unique sequences revealed a remarkably low degree of overlap between communities at each depth. The 444 147 sequence tags analyzed represented 62 975 unique sequences. Of these, 3707 (5.9%) occurred at two depths, and only 298 (0.5%) were observed at all three depths. At this level of phylogenetic resolution, *Bacteria* diversity decreased with depth but was still equivalent to that reported elsewhere for different soil types. *Archaea* diversity was highest in the two deeper samples. *Eucarya* observations and richness estimates are almost one order of magnitude higher than any previous marine microbial *Eucarya* richness estimates. The associations of many *Eucarya* sequences with putative parasitic organisms may have significant impacts on our understanding of the mechanisms controlling host population density and diversity, and point to a more significant role for microbial *Eucarya* in carbon flux through the microbial loop. We posit that the majority of sequences detected from the deep sea that have closest matches to sequences from non-pelagic sources are indeed native to the marine environment, and are possibly responsible for key metabolic processes in global biogeochemical cycles.

Introduction

In the ecology of macroorganisms, ecosystem multifunctionality has been related to species diversity, with some degree of redundancy 'built in' to community structure (Hector and Bagchi, 2007). However, the high abundance of microorganisms has previously precluded sampling at a scale considered robust enough to reliably estimate community richness. Hence, ecological theories pertaining to the impact of microbial diversity on ecosystem function are currently poorly informed. The recent development of ribosomal tag pyrose-

quencing methods has allowed the collection of more detailed organismal abundance data for the analysis of community composition and diversity. These methods are capable of detecting both the most abundant community members and those less-abundant components that are seldom, if ever, retrieved by traditional cultivation or cloning and sequencing methods (Pedrós-Alió, 2006). This is an important development in microbial ecology, because, although abundant and easily detectable organisms may have an obvious impact on energy harvesting and nutrient cycling, the role of less-abundant organisms is not easily understood. These rare organisms may have the following important environmental roles: they may act as keystone species in complex microbially mediated cycles, serve as a reservoir of genetic, and functional diversity and/or buffer ecosystems against species loss or environmental change. As technology has improved, the roles of these rare organisms have

been further appreciated; for example, Montoya et al. (2004) showed the importance of numerically unimportant unicellular diazatrophs in nitrogen fixation in the oligotrophic ocean, whereas Musat et al. (2008) showed that, despite representing only 0.3% of the total community, Chromatium okenii was potentially responsible for 40% and 70% of the total ammonium and carbon uptake, respectively, in the lake system sampled.

Adequate baseline data are required to set the framework against which to gauge community shifts in response to global climate change and other anthropogenic perturbation. Although Eucarya–Bacteria–Archaea interactions are common in the ocean and undoubtedly have a major role in determining community structure (Fuhrman and Hagstrom, 2008), these groups have rarely been analyzed together in environmental surveys. There is a growing recognition of microbial eukaryotes as primary producers, bacterial grazers and parasites in the ocean (Massana and Pedrós-Alió, 2008). Determining how the community structure of each domain changes over spatial and temporal scales is key to developing a more holistic overview of marine ecosystem dynamics. Recent pyrosequencing surveys of Bacteria and Archaea in deep-sea pelagic and benthic communities have highlighted an 'unexpected' diversity in these biomes (Sogin et al., 2006; Huber et al., 2007), whereas metagenomic analyses have revealed a remarkable functional diversity (Venter et al., 2004; DeLong et al., 2006). A recent meta-analysis of marine pyrosequencing and metagenomic data concluded that planktonic bacterial communities in the upper ocean were less diverse than those around hydrothermal vents in the deep sea and soil communities (Quince et al., 2008). However, ribosomal tag sequencing analysis has not been performed over a depth profile at any location, and no pyrosequencing or metagenomic survey to our knowledge has focused on all three domains in the marine environment. To address this paucity of data, we examined the community structure of Bacteria, Archaea and micro-Eucarya along a depth gradient at Station ALOHA, an open-ocean oligotrophic site in the sub-tropical North Pacific. The depths sampled represent the epi-, meso- and bathypelagia, each of which has distinct physical and biochemical characteristics, and potentially different scales of spatial and temporal variability (Hewson et al., 2006a, 2006b). Differences in the Bacteria and Archaea functional gene content have been observed across this depth gradient (DeLong et al., 2006); however, in that study, samples from different depths were collected at different times and may thus be temporally confounded (Hewson et al., 2006b). How phylogenetic structure changes throughout the water column and what factors influence community diversity are still debated. Here, we deconstruct the composition of each domain at different levels of phylogenetic resolution, analyzing how taxonomy

and community structure change over this depth profile. We also compare phylogenetic data with those from previous marine bacterioplankton surveys, and compare the diversity of marine Bacteria with that found in soils using similar survey methods.

Materials and methods

Water body parameters
Information on water body characteristics, such as counts of microbial cells, was taken from the HOT-DOGS website (http://hahana.soest.hawaii.edu/hot/hot-dogs/interface.html) of the Hawaii Ocean Time-Series (HOT).

Sampling, assay preparation and sequencing
Seawater (4 l) from depths of 10, 800 and 4400 m was collected from the HOT Station ALOHA (22.45 °N, 158 °W) during HOT cruise 169 (May 16–20, 2005) and was sequentially filtered through GFA glass fiber filters (Whatman, Clifton, NJ, USA) before passage through 0.22 μm HV Durapore (Millipore, Bedford, MA, USA) filters. Sample preparation and DNA extraction protocols have been reported previously (Brown and Donachie, 2007) and are summarized in the Supplementary information. PCR primer sequences used to amplify the ribosomal sequence tags (Table 1) used slightly modified common 16S/18S rDNA specific primers 1392F and 1492R (Lane, 1991), amended with 454 Life Sciences A and B adapters. A full discussion of primer specificity is supplied in the Supplementary materials. We calculated that primers would match to >80% of known phylogenetic diversity across all three domains. DNA extracted from both the Whatman and Durapore filters for the same depths was pooled, and ~5 ng was used as template in PCRs. Each PCR reaction (100 μl volumes) contained 5U AmpliTaq Gold (Applied Biosystems, Carlsbad, CA, USA), 1 × PCR Amplitaq Gold buffer (Applied Biosystems), 2.5 mM $MgCl_2$, 400 μM of each deoxynucleotide and 400 nM of each primer. Reaction mixtures were held at 94 °C for 10 min, followed by 24 amplification cycles of 94 °C for 40 s, 55 °C for 40 s and 72 °C for 40 s, with a final step at 72 °C

Table 1 Primer sequences used in this study

Primer	Sequence
Forward	gcctccctcgcgccatcag-xxx-TACACACCGCCCGT xxx depth codes: AAA = 10 m, CTA = 800 m, TTT = 4400 m
Reverse	gccttgccagcccgctcag-RGMAACCTTGTACGACTT

Uppercase characters represent the rDNA priming sequence. Lower case characters represent 454 Life Sciences Adapter A (forward primer) and B (reverse primer) and xxx represents the 3-bp key code. (R = A or G, M = A or C).

for 2 min. Amplification products from three separate reactions for each sample were pooled and purified in a MinElute PCR purification kit (Qiagen, Valencia, CA, USA).

Pyrosequencing data that were returned from 454 Life Sciences were subject to a systematic check to remove low-quality reads. We eliminated (i) sequences that did not perfectly match the 3 bp key code and primer sequence at the beginning of the read, (ii) sequences that did not perfectly match at least the first 10 bp of the distal primer, (iii) sequences that contained any undetermined nucleotides (N) and (iv) sequence reads of < 50 bp after the removal of both primers. Use of these quality control measures greatly increases the resultant read accuracy in pyrosequencing assays (Huse et al., 2007). Quality sequences were compared by BLAST (Altschul et al., 1990) (e-value cutoff of 10^{-5}) with the GenBank nr database, and the domain of origin was determined from the best high-scoring pair. A further extensive phylogenetic classification was carried out as described in the Supplementary materials. BLAST phylotypes were analyzed to identify those statistically more likely to be found at one depth (overrepresented), following the re-sampling method of Rodriguez-Brito et al. (2006).

Community comparison
We clustered and compared samples using the nonalignment clustering algorithms in the CD-HIT package (downloaded from http://www.bioinformatics.org/cd-hit/) (Li and Godzik, 2006). *Archaea*, *Bacteria* and *Eucarya* databases, containing all sequences from each depth, were compared across all similarity thresholds between 0.8 and 1 using the program CD-HIT-EST-2D (downloaded from http://www.bioinformatics.org/cd-hit/), with word sizes set according to the manual (http://cd-hit.org/). This is a quantitative two-way comparison (for example, the percentage of all sequences from 800 m that cluster with sequences from 10 m at a given percentage sequence similarity is determined, followed by the reverse), and thus this method takes into account sequence abundance as well as identity. A detailed description of the derivation of parametric and nonparametric diversity estimators, based on frequency count data from CD-HIT clustering, and choice of models used is available in the Supplementary materials.

Results

Hydrographic considerations
Compartments of the water column investigated correspond to the epi- (10 m), meso- (800 m) and bathypelagia (4400 m) in the Sub-Tropical North Pacific Gyre, and differ dramatically in their physical and chemical characteristics (Supplementary Figure S1). The availability of photo-

synthetically active radiation drives the growth of photosynthetic organisms, resulting in high chlor-opigment concentrations in surface waters (Supplementary Figure S1). Although epipelagia is truly oligotrophic, as defined by levels of inorganic nutrients, both the meso- and bathypelagia are characterized by high inorganic and low organic nutrient concentrations (Supplementary Table S1). The mesopelagia sample was taken at the oxygen minimum zone (Supplementary Figure S1), where pressure was approximately one-fifth of that at the point sampled in the bathypelagia.

Flow cytometry counts placed numbers of *Prochlorococcus* and heterotrophic *Bacteria* at the higher end of historical data covering 102 cruises (Supplementary Figures S2a and b), whereas *Synechococcus* and *Eucarya* abundances were at the lower end (Supplementary Figures S2c, d). Cell counts indicate that cyanobacteria (*Prochlorococcus* and *Synechococcus*) accounted for 30% of total prokaryote abundance at 5 m compared with an average of 25% during the 102 cruise data set.

Analysis of tag sequences
In total, 444 147 sequences passed our quality control methods, of which 66 990 were unique at a given depth (Table 2). Unique sequences and abundance data obtained from this study are available at http://www.ifa.hawaii.edu/UHNAI/Gayle/. A comparison of unique sequences at each depth revealed remarkably little overlap; 3707 sequences (5.9%) were detected at two depths, whereas only 298 (0.5%) were detected at all three depths. Total unique sequences from all depths numbered 62 975 and represented all domains (Table 2). A total of 6652 sequences (1.5% of those analyzed), including 3300 unique sequences, could not be assigned to any one domain by BLAST and were removed from further analyses.

Community structure
Although rarefaction analyses indicate an undersampling of the community in all three domains (Supplementary Figure S3), our large sequencing

Table 2 Domain level distribution of unique and total sequence abundances retrieved from samples in the epi- (10 m), meso- (800 m) and bathypelagic (4 400 m) zones of the oligotrophic open ocean at Station ALOHA

	Depth (m)	Total	Bacteria	Archaea	Eucarya	No Hit
Unique	10	26 064	12 802	433	10 433	2396
	800	23 708	10 102	884	12 130	592
	4400	17 208	8658	853	7385	312
All	10	202 092	141 931	3821	51 737	4603
	800	137 926	60 500	9761	66 312	1353
	4400	104 129	55 684	8935	38 814	696

effort provided a substantial subsample from which to generate parametric (maximum likelihood) and nonparametric (ACE1) richness estimates and lower bounds (ACE and Chao1) at different levels of phylogenetic resolution (Supplementary Table S2). Richness estimates within the *Archaea* were lowest of the three domains, but higher at 800 m and 4400 m than at the surface (Figure 1a). Rank abundance and k-dominance plots for *Archaea* at all depths were similar, indicating that this domain displayed the least evenness in community composition, with sequence abundances of the top 10 unique tags contributing >60% to the total sequence count at all depths (Supplementary Figures S4, S5). We used a two-way clustering algorithm to cluster tags from two different depths at a variety of phylogenetic resolutions. This takes into account both tag identity and tag abundance and provides a measure of the proportion of the total number of tag sequences from one depth that have a taxonomic equivalent (at a specified cutoff) at the other depth. There was a greater taxonomic overlap between the *Archaea* at 800 and 4400 m than either had with the surface sample (Figure 2a). In fact, at 98% tag sequence identity, over 90% of the unique *Archaea* tags from the 800 m sample had a taxo-

nomic equivalent in the 4400 m sample, and vice versa. However, even at 80% tag sequence identity, ~5% of sequences from the 800 and 4400 m samples have no 'relative' in the surface sample.

Within the *Bacteria* there was a transition in the trends in richness depending on phylogenetic resolution. In general, richness at the highest phylogenetic resolution, that is, 'species/strain' level (100% and 99% sequence identity), decreased with depth (Supplementary Table S2, Figure 1b). However, higher order phylogenetic richness estimates (for example, 95–80% identity) increased with depth (Figure 1b). Effectively, there was greater richness at the 'phylum/class' level with increasing depth. *Bacteria* composition at 10 m was less evenly distributed and showed greater dominance than at 800 or 4400 m (Supplementary Figures S4, S5). Two-way clustering of tag databases indicated that 800 and 4400 m samples hosted a similar overall taxonomic composition, although the relative abundances of taxa varied considerably (Figure 2b). Furthermore, although the 800 and 4400 m samples contained much of the phylogenetic diversity found at 10 m, the converse was not true. This may reflect an impact of sinking material on the composition of *Bacteria* in deeper waters, or alternatively for the
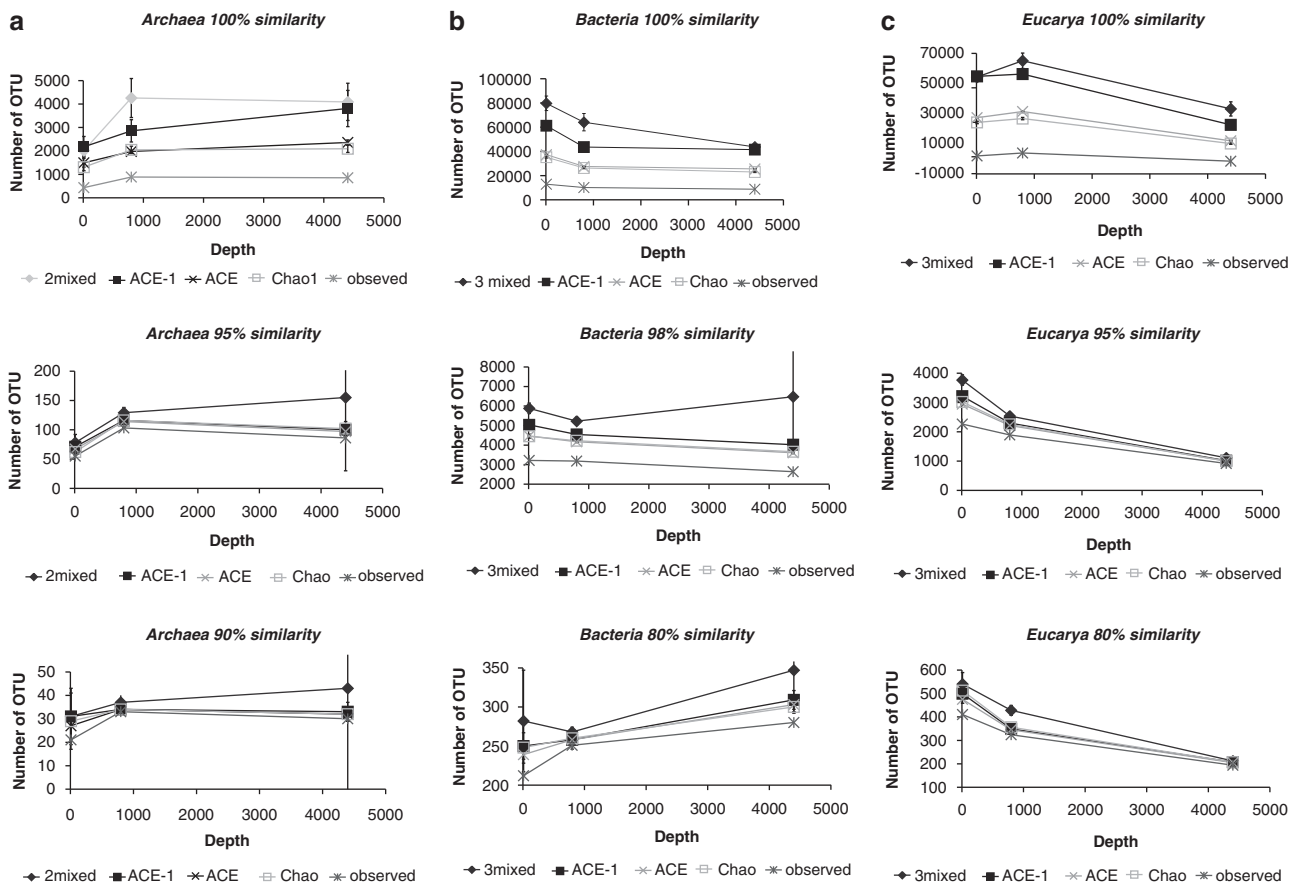

Figure 1 Observed and estimated richness of *Archaea*, *Bacteria* and *Eucarya*, calculated at a variety of phylogenetic levels. Left hand column (**a**) Archaea, middle column (**b**) Bacteria, right hand column (**c**) Eucarya.

1378

presence of many closely related, but less abundant, deep phylotypes of common surface taxa. Comparison of *Bacteria* richness estimates indicates that, at the highest phylogenetic resolution, the samples analyzed here contain taxonomic diversity equivalent to that reported in North and South American
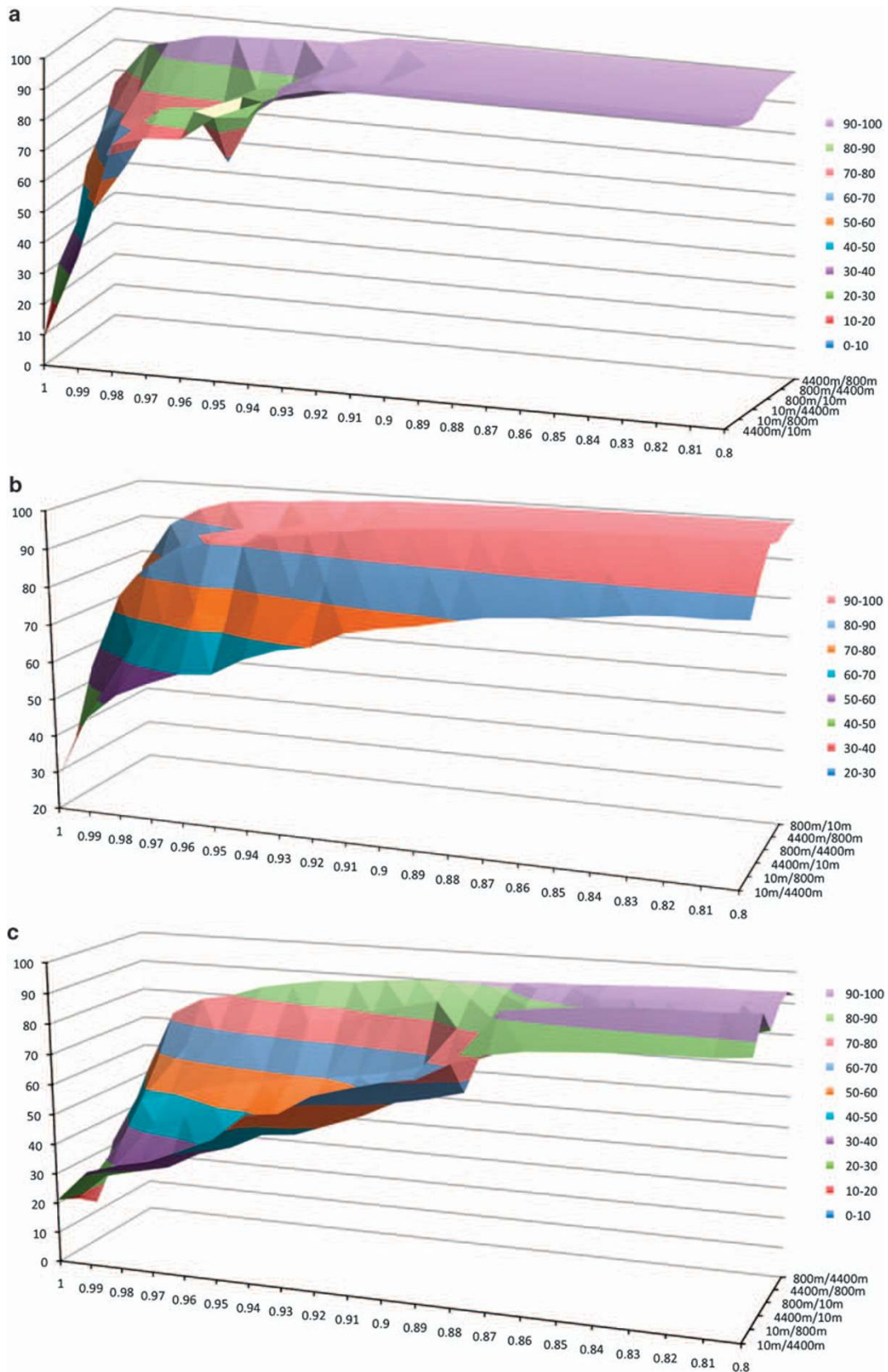
soils similarly analyzed by V9 pyrosequencing (Supplementary Table S3). At all other levels of phylogenetic resolution, however, soil hosts greater microbial diversity.

Eucarya unique sequence richness peaked at 800 m (Figure 1c). However, at all other phylogenetic levels, richness decreased with depth (Figure 1c). Rank abundance and k-dominance plots were similar at all depths and indicated that *Eucarya* had the most even population distributions (Supplementary Figure S4). When tags from two depths were clustered together, the communities showed little overlap (Figure 2c) compared with that observed in the *Bacteria* and *Archaea*. *Eucarya* taxonomic assemblages thus changed significantly with depth.

Taxonomic transitions: Archaea

A total of 12 *Archaea* phylotypes were observed at all depths, whereas 49% of phylotypes were significantly overrepresented at a given depth. Overrepresentation may indicate that a phylotype is better adapted to conditions at a given depth than at the other depths analyzed. The dominant *Archaea* clade at 10 m was the *Euryarchaeota* Marine Group II (84.5% of Archaea tag sequence abundance at 10 m, 30.4% at 800 m and 37% at 4400 m) (Figure 3a, numerical abundances are provided in Supplementary Table S6). Abundance of *Euryarchaeota* Marine group III sequences was similar throughout the water column, whereas that of the *Euryarchaeota* Methanopyri increased with depth. The *Crenarchaeota* Marine group I, which includes the ammonia-oxidizing Candidatus Nitrosopumilus, the unclassified *Crenarchaeota* and the Cenarchaeum, accounted for 55.7% and 47.1% of *Archaea* tag sequences at 800 and 4440 m, respectively.

Taxonomic transitions: Bacteria

Of the 2978 *Bacteria* phylotypes observed from common high-scoring pair analysis, 320 were observed at all depths, whereas 19.8% were significantly overrepresented at a given depth. Tag sequences representing organisms from the *Alpha*- and *Gamma Proteobacteria* were abundant at all depths (Figure 3b, Supplementary Table S6), but at finer taxonomic resolution, the phylogenetic composition within these classes varied dramatically (Supplementary Figures S6, S7). Sequence tags that affiliated with the *Alpha Proteobacteria*, *Cyanobacteria*, *Bacteroidetes* and unclassified Greengenes OTU181, all decreased in relative abundance from 10 m with increasing depth (Figure 3b). Those sequences that increased in abundance at 800 or 4400 m included the *Verrucomicrobia*, *Actinobacteria*, *Chloroflexi*, *Planktomycetaceae*, *Beta Proteobacteria*, *Acidobacteria* and *Firmicutes*. Major groups of *Bacteria* that were not observed at 800 and/or 4400 m, but were observed at 10 m, include Candidate divisions BRC-1, OD1, OD3 and WS3, *Aquifaceae*, *Deferribacteres*, *Deinococcus-Thermus*, *Fibrobacteres ktedobacter*, *Spirochaetes*, *Thermobaculum* and the genus *Mariprofundus* in the *Proteobacteria*.

Taxonomic transitions: Eucarya

Of the 1313 *Eucarya* phylotypes observed by common high-scoring pair analysis, 230 were detected in all three samples, whereas 34.2% were significantly overrepresented at a given depth. Ribosomal tag sequences representing members of the *Alveolata* were the most abundant microbial eukaryote. However, the major subgroups within the *Alveolata* displayed different patterns of abundance with depth (Figure 3c, Supplementary Table S6); for example, *Dinophyceae* tag sequence abundance dropped markedly in the bathypelagia, whereas that of the marine Alveolate groups I and II (including the *Amoebophyra*) increased in the meso- and bathypelagia. An analysis of previously defined marine Alveolate groups I and II subclusters revealed that, although most were present at all depths, some were clearly more abundant in either surface or deeper waters (Supplementary Figure S8). Tag sequences associated with several radiolarian classes increased in relative abundance with depth, including the *Acantharea* (*Taxipodida*), RAD-III and the *Sticholonchidae*. Conversely, *Polycystinea*

Figure 2 (**a**) A two-way comparison of the percentage abundance of the *Archaea* sequence tags from each depth that cluster with tags from each of the other depths at different levels of phylogenetic similarity. The z axis represents the two tag sequence databases that are being compared. The x axis represents the level of tag sequence identity (1 (identical) to 0.8). The y axis represents the relative percentage of the tag sequences occurring in the second database from the z axis, which are phylogenetically 'accounted for' in the first database from the z axis. This analysis takes into account both tag identity and tag abundance and provides a measure of the proportion of the total number of tag sequences from one depth that has a taxonomic equivalent (at a specified cutoff) at the other depth compared. For example, consider a cluster of sequences (98% identical) at 10 m that accounts for 1.5% of the total number of sequences at that depth, and for which there is a taxonomically equivalent cluster (at 98% sequence identity level) at 800 m, which accounts for 0.01% of the total tag sequences at 800 m. This is reported in the two-way analysis as 1.5% of the tag sequences from 10 m being taxonomically 'accounted for' at 800 m, but only 0.01% of the tag sequences from 800 m being 'accounted for' at 10 m. (**b**) A two-way comparison of the percentage abundance of the *Bacteria* sequence tags from each depth that cluster with tags from each of the other depths at different levels of phylogenetic similarity. The z axis represents the two tag sequence databases being compared. The x axis represents the level of tag sequence identity (1 (identical) to 0.8). The y axis represents the relative percentage of the tag sequences occurring in the second database from the z axis, which are phylogenetically 'accounted for' in the first database from the z axis. (**c**) A two-way comparison of the percentage abundance of the *Eucarya* sequence tags from each depth that cluster with tags from each of other depths at different levels of phylogenetic similarity. The z axis represents the two tag sequence databases being compared. The x axis represents the level of tag sequence identity (1 (identical) to 0.8). The y axis represents the relative percentage of the tag sequences occurring in the second database from the z axis, which are phylogenetically 'accounted for' in the first database from the z axis.
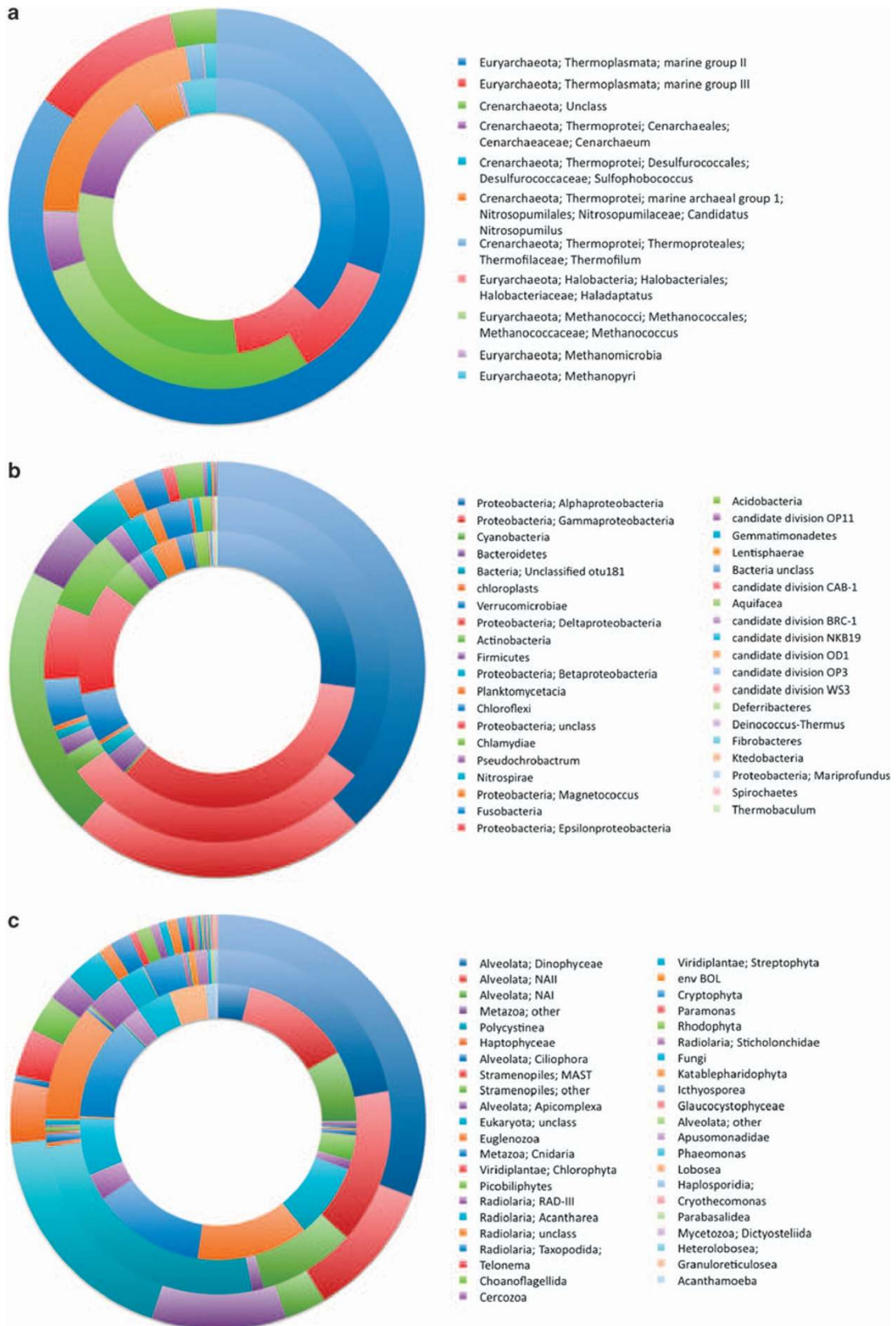
Figure 3 (**a**) The abundance of major *Archaea* phylogenetic lineages observed at 10 m (outside ring), 800 m (middle ring) and 4 400 m (inside ring) depths. (**b**) The abundance of major *Bacteria* phylogenetic lineages observed at 10 m (outside ring), 800 m (middle ring) and 4 400 m (inside ring) depths. (**c**) The abundance of major *Eucarya* phylogenetic lineages observed at 10 m (outside ring), 800 m (middle ring) and 4 400 m (inside ring) depths.

(*Radiolaria*) sequences, which were abundant at 10 m (16% of relative abundance) and peaked at 800 m (26%), were almost absent (0.2%) from 4400 m. Tag sequences affiliating with the marine *Stramenopiles* (MAST) and *Haptophyceae*, and with the novel *Picobiliphyte* cluster, decreased in abundance with depth.

Comparison with other marine samples

UPGMA clustering (see Supplementary materials for methods) of relative abundances of phylum level *Bacteria* lineages identified from 16S rDNA clone libraries from seven depths at Station ALOHA (DeLong *et al.*, 2006) and from 16S rDNA tag sequencing of deep Labrador Seawater (Sogin *et al.*, 2006), and 16S rDNA fragments parsed from metagenomic analysis of 43 surface ocean samples (Rusch *et al.*, 2007) show that a higher order community structure within deep-sea assemblages is consistently different and forms a separate cluster

from surface assemblages (Figure 4). A canonical analysis of principal coordinates analysis (see Supplementary materials for methods) further highlights the separation between surface and deep communities. Correlations of environmental variables indicated depth as the major contributing factor to this separation. The correlation of *Bacteria* lineages to canonical axes (Figure 4) highlights the major contributors to deep community separation and includes relative abundances of the *Delta Proteobacteria*, *Chloroflexi*, *Gemmatomonadetes*, *Acidobacteria*, *Verrucomicrobia* and *Nitrospirae*. The 10-m sample clusters within the major surface group, adjacent to other tropical open-ocean surface samples, which also included others from the photic zone at Station ALOHA.

Origin and novelty of rare and abundant phylotypes

The abundance of sequences from all domains related to those identified in deep-sea environments increased with the increasing depth. However, over
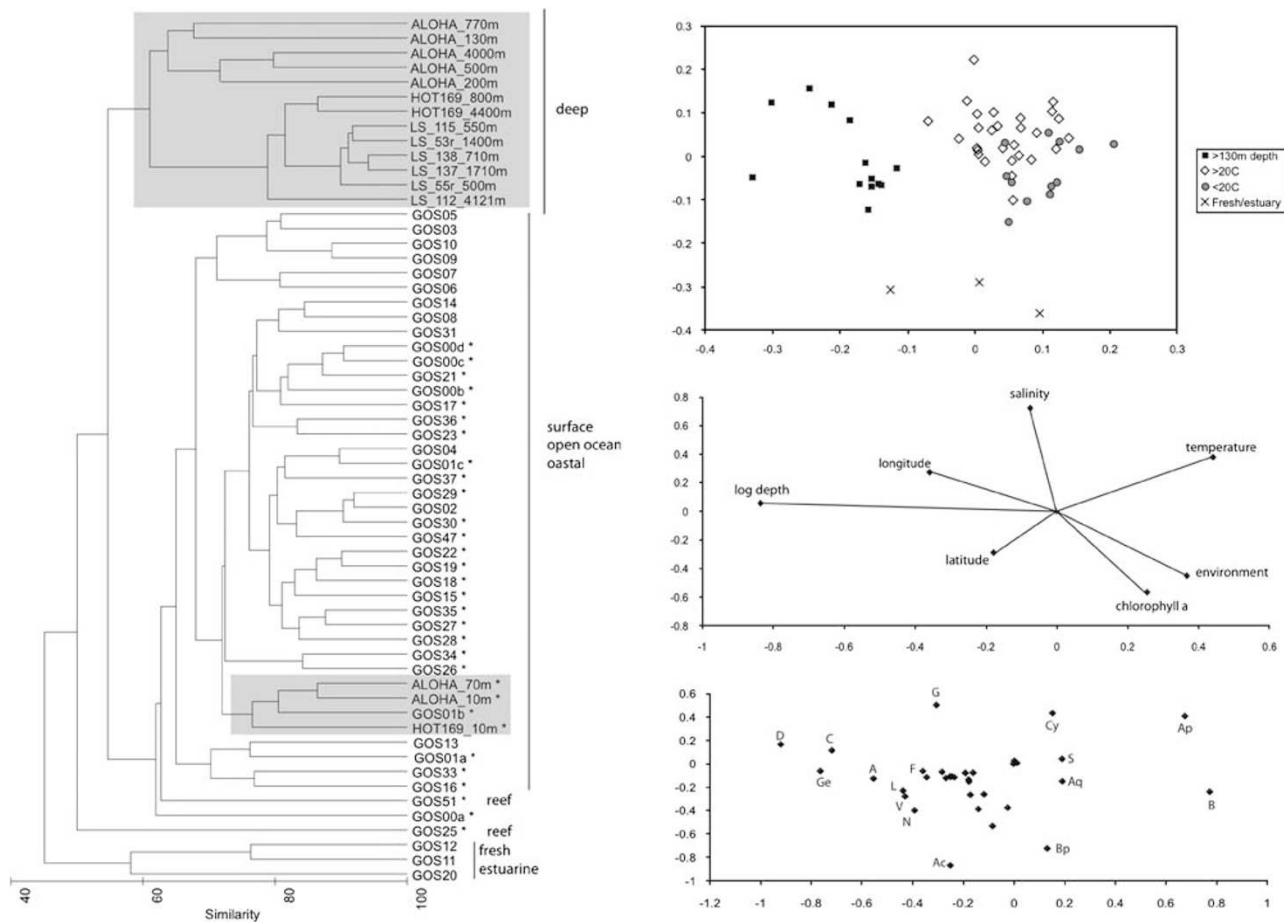

Figure 4 UPGMA clustering and canonical analysis of principal coordinates analysis of a higher order *Bacteria* community structure indicate a consistent shift from surface to deep waters. Samples in the analysis include those from the Global Ocean Survey (GOS, Rusch *et al.*, 2007), Station ALOHA (ALOHA, Delong *et al.*, 2006 and HOT169, this study) and the Labrador Sea (LS, Sogin *et al.*, 2006). Asterisk (*) indicates samples from water >20 °C. *Bacteria* lineages affecting sample distribution in CAP analysis are coded as follows: D, *deltaProteobacteria*; G, Gemmatimonadetes; C, Chloroflexi; A, *Acidobacteria*; L, Lentisphaera; V, Verrucomicrobia; N, Nitrospirae; F, Firmicutes; G, *gammaProteobacteria*; Ap, *alphaProteobacteria*; B, Bacteroidetes; Cy, Cyanobacteria; S, Spirochetes; Bp, *betaProteobacteria*; Ac, Actinobacteria.
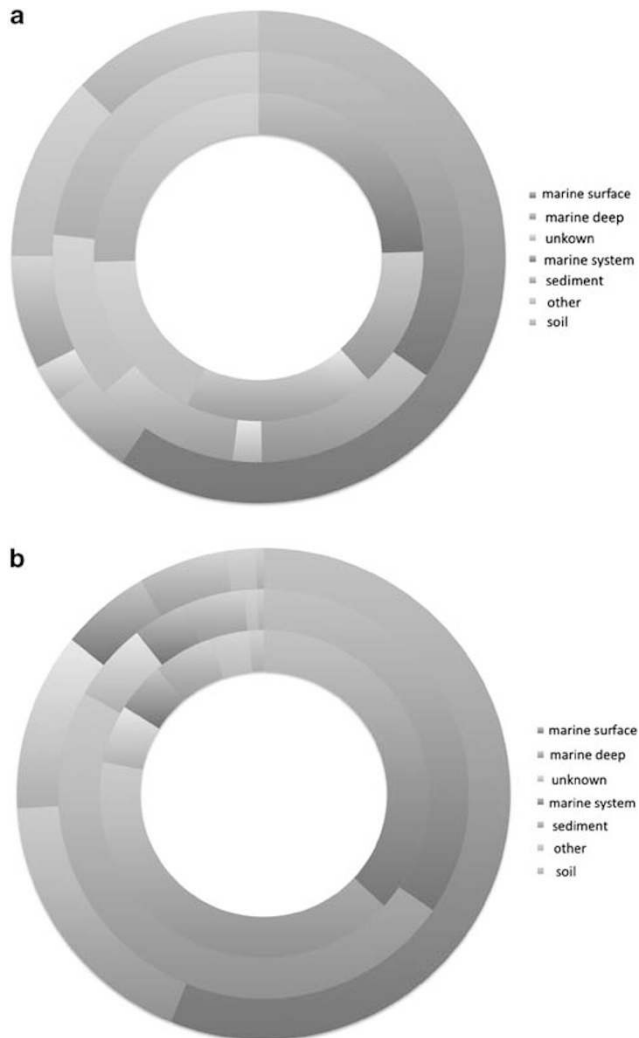
Figure 5 (**a**) The environment of origin of the highest-scoring pairs in the Genbank database for rare *Bacteria* tag sequences occurring at 10 m (outside ring), 800 m (middle ring) and 4400 m (inside ring) depths. (**b**) The environment of origin of the highest-scoring pairs in the Genbank database for *Eucarya* tag sequences occurring at 10 m (outside ring), 800 m (middle ring) and 4400 m (inside ring) depths.

half the rare *Bacteria* phylotypes observed at 800 and 4400 m, and over a quarter of those at 10 m, returned best matches to sequences in GenBank from non-pelagic sources, including soil and sediment (Figure 5a). Approximately one-quarter of the rare *Eucarya* phylotypes were similarly affiliated (Figure 5b). Rare phylotypes in all domains (rare *Archaea* considered as phylotypes with <10 sequences) were more likely to have a lower sequence similarity score (that is, novel) when compared with the Genbank database than were abundant ones (Figure 6). *Bacteria* phylotypes from 800 and 4400 m displayed greater novelty than those from 10 m, but the converse was true of *Archaea* phylotypes. *Archaea* sequences from 10 m displayed a peak in novelty from 95–97% sequence identity, which accounts for 49% of total abundances, providing

evidence of a relatively abundant surface-associated *Archaea* group (~genera level) not reported in previous molecular surveys. *Eucarya* phylotypes were as likely to be novel, regardless of the depth from which they were sampled. On an average, 32% of abundant and 44% of rare *Eucarya* phylotypes displayed <95% identity with their best BLAST matches in the GenBank database, compared with 5% and 11% for *Archaea* and 3% and 13% for *Bacteria*. Furthermore, 62% of sequences that did not significantly match GenBank sequences (and so were removed from the above analysis) were >90 bp in length; this, when over 99% of all sequences identified as *Bacteria* or *Archaea* were 90 bp or less, making it more likely that sequences not matching any others in GenBank, if ribosomal in origin, are eukaryotic.

Discussion

Novel technologies continue to expand our understanding of microbial diversity and community structure. Metagenomic analysis (Venter *et al.*, 2004; DeLong *et al.*, 2006) and ribosomal tag pyrosequencing strategies (Sogin *et al.*, 2006; Huber *et al.*, 2007) have previously identified 'unexpectedly' high *Bacteria* and *Archaea* phylogenetic and functional diversity. Here, we extend this observation to the microbial *Eucarya*. Even at the conservative 95% tag sequence identity (~97% SSU rRNA similarity), our OTU observations and richness estimates are almost one order of magnitude higher than any previously reported (Supplementary Table S2) (Jeon *et al.*, 2006; Zuendorf *et al.*, 2006; Countway *et al.*, 2007). The difference between our *Eucarya* richness estimates for unique and clustered sequences indicates that *Eucarya* display equivalent levels of microdiversity to the *Bacteria* community (Acinas *et al.*, 2004; Brown and Fuhrman, 2005). For many clades, this microdiversity may represent population, rather than intra-genomic, diversity, given the fact that not a single polymorphism was found in a tandem repeat array of the rDNA operon from a fosmid representing an organism within the Novel Avleolate group I, cluster I (NA1.1) (Massana *et al.*, 2008). The Novel Alveolate groups I and II (Groisillier *et al.*, 2006) collectively contributed to 12.4, 23.6 and 21.2% of tag sequence abundance at 10, 800 and 4400 m, respectively. Recently, sequences clustering within these clades have been identified as representing endoparasitic dinoflagellates related to the genera *Amoebophyra* (Kim *et al.*, 2008) and *Duboscquella* (Harada *et al.*, 2007). Infection by both genera is mediated by the production of dispersal spores and infections and is mostly lethal (Kim *et al.*, 2008). If the diversity within NAI and NAII does indeed represent organisms displaying a potentially lethal parasitic lifestyle, this may have significant impacts on our concepts of the mechanisms controlling host population density
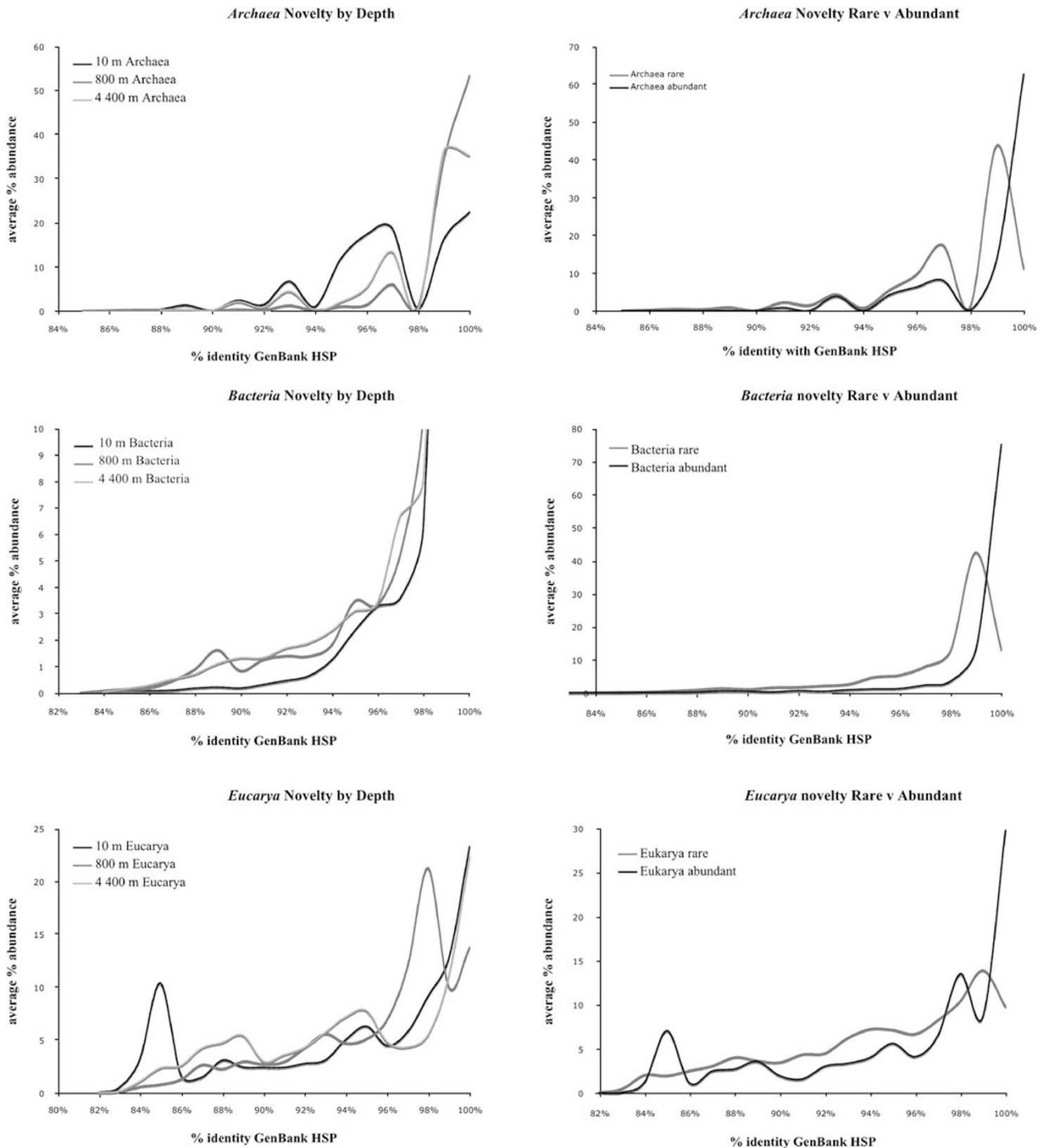
Figure 6 The average percentage identity of ribosomal tag sequences to GenBank HSPs for three depths (10, 800 and 4400 m) at Station ALOHA.

and diversity. The various subgroups within NAI and NAII varied with depth (Supplementary Figure S8). These patterns of (putative) parasitic abundance and diversity may be associated with either host or parasite characteristics. Members of the *Amoebophyra* and *Dubosquella* are known to parasitize a range of organisms including dinoflagellates, radiolarians, ciliates, acanthareans and sticholonche

(Kim *et al.*, 2008), all of which displayed different patterns of diversity and abundance in our depth gradient. Multiple species of *Amoebophyra* have been reported to infect a single host species (Salomon *et al.*, 2003), but strong host specificity may drive the genetic diversification of the host species. Conversely, lethal infections may reduce the dominance of superior competitors, allowing the

1384

coexistence of many competing species, analogous to the viral 'kill the winner' hypothesis, previously shown to alter community structure and increase diversity in bacterial populations (Fuhrman and Schwalbach, 2003). These effects may go some way toward explaining the diversity in eukaryotic populations. Although some of the observed host diversity may be environmentally neutral, it may also act to boost a population's biological robustness (which enables a system to maintain its functionalities against external and internal perturbation) (Kitano, 2004) by confounding parasitic attack. A similar mechanism has been proposed for bacterial microdiversity (Brown and Fuhrman, 2005). Furthermore, the mass release of spores from an infected or dead host would have an effect in returning significant amounts of carbon from sinking material to the immediate environment, enhancing carbon flow through the microbial loop, as has been recently implied for freshwater systems (Lefèvre et al., 2008).

Besides the persistent presence of NAI and NAII, some general phylogenetic patterns have emerged from the few molecular surveys of *Eucarya* carried out at various locations (Epstein and López-García, 2008) across depth profiles. Both Countway et al. (2007) and Not et al. (2007) reported shifts in community structure when comparing euphotic zones with deeper waters in the North Atlantic and Sargasso Sea, respectively. As with this analysis, Countway et al. (2007) observed decreasing OTU abundance with depth (although only one of three shallow versus deep comparisons showed statistically significant differences). A potential explanation for this is that decreased resource availability (Karner et al., 2001) leads to reduced niche partitioning in deeper waters. Further analysis is required to determine whether these changes are congruent across ocean basins. Much of the apparent *Eucarya* diversity is yet to be adequately described, as indicated by the degree of novelty shown by *Eucarya* tag sequences compared with those of the *Bacteria* and *Archaea* (Figure 6).

The 800 and 4400 m samples were more closely related phylogenetically than was either to the 10 m sample for all domains, despite the five times greater physical distance between these sites. A number of factors likely contribute to this. Physicochemical conditions such as oxycline, chemocline and thermocline may provide a strong barrier to genetic exchange/community homogenization between the surface and the two deeper sample sites. Furthermore, secondary adaptation to pressure may require relatively little genetic change for low-temperature-adapted *Bacteria* (Lauro and Bartlett, 2008), allowing organisms adapted to the temperature regime at 800 m to populate the higher pressure environment at 4400 m, which is only a few degrees cooler. This is consistent with the hypothesis that, in the absence of sunlight, temperature is a major ecological/taxonomic stratifying factor in the marine water column, overriding pressure at least at depths of up to 4000 m (Martín-Cuadrado et al., 2007). Indeed, UPGMA cluster analysis of phylum and class level *Bacteria* lineages from our study and others indicates a consistent shift in the higher order taxonomic structure in meso- and bathypelagia when compared with that in coastal and open-ocean surface samples (Figure 4). Geography and temporal variation may have an important role in the makeup of the deep-sea community structure, as evidenced by the clustering patterns of samples taken from the Labrador Sea (Sogin et al., 2006) and Station ALOHA in different years (this study; DeLong et al. (2006)). This observation agrees with studies that used other molecular methods to describe spatial heterogeneity in the bacterioplankton community composition in samples from marine epi-, meso- and bathypelagia (Lee and Fuhrman, 1991; Hewson et al., 2006a). However, some of the variation may be due to differences in methodology. The greater disparity in phylogenetic composition within the *Eucarya* between depths may mean that the physiological adaptations required for *Eucarya* survival in the oxygen-depleted deep sea may not be so easily expressed and may require greater genomic (and hence phylogenetic) differentiation.

A recent meta-analysis (Quince et al., 2008) comparing (meta)genomic data from marine surface samples with tag sequence data from different soils and deep-sea sediment habitats concluded that deep-sea diversity, at least in sediments near hydrothermal vents, and soil diversity significantly exceed diversity in the ocean surface. Our results, although from a single oligotrophic open-ocean site, but generated using consistent methodology across the depth profile, suggest that species/strain level diversity within the *Bacteria* is greater at the surface than at the depth in pelagia, whereas the reverse is true for a higher order phylum/class level. A confounding factor not considered by Quince et al. (2008) is the variability inherent in the different SSU rDNA helix regions examined. A comparison of the V6 and V9 regions excised from near full-length *Bacteria* 16S sequences indicates that the V6 region examined by Sogin et al. (2006) and by Huber et al. (2007) is approximately twice as variable as the V9 region used in this study and by Roesch et al. (2007) to examine soil diversity ($y = 2.28$, data not shown). Hence, comparisons using a simple 97% tag sequence identity cutoff when considering surveys using different variable regions should be avoided, and careful consideration should be paid to relative rates of divergence. Roesch et al. (2007) suggested that for the V9 region, 100% sequence similarity best equates to species level. Comparison of ACE and Chao1 statistics of our samples with those of the soil samples analyzed by Roesch et al. (2007) (re-analyzed by Quince et al. (2008)) reveals that, at this phylogenetic level, microbial diversity in the oligotrophic open ocean is equivalent to that in three agricultural soils and

one boreal forest soil (Supplementary Table S3). At all other phylogenetic levels, however, the soil samples host two- to fourfold greater richness, clearly far from the orders of magnitude previously speculated to be the difference between aquatic and soil/sediment microbial diversity (Torsvik et al., 2002). Indeed, the difference may simply be a function of spatially constrained ecosystem processes in soil as opposed to those occurring throughout the integrated 4500 m water column in the open ocean. An interesting observation of our work, and that of Roesch et al. (2007), is that some samples seem to contain higher than anticipated phylum level diversity; for example, a Canadian boreal soil and our 4400 m sample. This may relate directly to a greater variety in functional pathways. At the surface, a higher Bacteria and Eucarya species/strain diversity probably reflects greater niche availability. The availability of photosynthetically active radiation provides an opportunity for mixotrophy, and drives photoautotrophy, the net result being increased concentrations, diversity and availability of dissolved organic matter in the euphotic zone. Such an enhanced dissolved organic matter pool may support many related heterotrophic lineages. Below 100 m, DOC concentrations decrease rapidly and dissolved organic matter in the deep ocean largely comprises the refractory pool that is not easily broken down by the microbial community (Carlson et al., 1994; Witte et al., 2003; Hopkinson and Vallino, 2005). It may be that those microbes that are more abundant in deep waters, yet are not part of shallow-water communities, have important roles in the degradation of the 'refractory' dissolved organic matter pool; the alternative metabolic pathways required may well explain the higher phylum level diversity observed in the deep sea. In this respect, the closest matches in the GenBank of many phylotypes rare in the meso- and bathypelagia were from non-planktonic habitats. Some of these matches may be due to allochthonous microbes, such as those from surfaces, digestive tracts, fecal pellets and so on, of higher organisms, or from suspended sediments and coastal erosion, aeolian processes or other processes. Indeed, many Bacteria recovered from deep-sea environments show no evidence of a physiological adaptation to in situ environmental conditions (Yayanos, 1986; Lauro et al., 2004), and survive, but do not grow, under prevailing low-temperature high-pressure conditions (Lauro et al., 2004). However, it is likely that the majority of organisms represented by rare tag sequences are, indeed, native to the ocean, but remains poorly represented in databases due to a lack of breadth and scale in sampling efforts. If this is indeed the case, microbes of the rare biosphere, which have the closest known homology to organisms from soil and sediment, may be responsible for key metabolic processes in global biogeochemical cycles, similar to those carried out by these related organisms in other environments.

Acknowledgements

References

Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, Distel DL et al. (2004). Fine-scale phylogenetic architecture of a complex bacterial community. Nature 430: 551–554.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. J Mol Biol 215: 403–410.

Brown MV, Donachie SP. (2007). Evidence for tropical endemicity in the marine delta-Proetobacterial Marine Group B/SAR324 bacterioplankton clade. Aquat Microb Ecol 46: 107–115.

Brown MV, Fuhrman JA. (2005). Marine bacterial microdiversity as revealed by internal transcribed spacer analysis. Aquat Microb Ecol 41: 15–23.

Carlson CA, Ducklow HW, Michaels AF. (1994). Annual flux of dissolved organic carbon from the euphotic zone in the northwestern Sargasso Sea. Nature 371: 405–408.

Countway PD, Gast RJ, Dennett MR, Savai P, Rose JM, Caron DA. (2007). Distinct protistan assemblages characerize the euphotic zone and deep sea (2500 m) of the western North Atlantic (Sargasso Sea and Gulf Stream). Environ Microbiol 9: 1219–1232.

DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard NU et al. (2006). Community genomics among stratified microbial assemblages in the ocean's interior. Science 311: 496–503.

Epstein S, López-García P. (2008). Missing protists: A molecular prospective. Biodiversity Conserv 17: 261–276.

Fuhrman JA, Hagstrom A. (2008). Bacterial and archaeal community structure and its patterns. In: Kirchman, DL (eds). Microbial Ecology of the Oceans, 2nd edn. John Wiley & Sons, Inc., Hoboken, NJ, USA. pp 45–90.

Fuhrman JA, Schwalbach M. (2003). Viral influence on aquatic bacterial communities. Biol Bull 204: 192–195.

Groisillier A, Massana R, Valentin K, Vaulo D, Guiillou L. (2006). Genetic diversity and habitats of two enigmatic marine alveolate lineages. Aquat Microb Ecol 42: 277–291.

Harada A, Ohtsuka S, Horiguchi T. (2007). Species of the parasitic genus Duboscquella are members of

the enigmatic Marine Alveolate Group I. *Protist* **158**: 337–347.

Hector A, Bagchi R. (2007). Biodiversity and ecosystem multifunctionality. *Nature* **448**: 188–190.

Hewson I, Steele JA, Capone DG, Fuhrman JA. (2006a). Remarkable heterogeneity in meso- and bathypelagic bacterioplankton assemblage composition. *Limnol Oceanogr* **51**: 1274–1283.

Hewson I, Steele JA, Capone DG, Fuhrman JA. (2006b). Temporal and spatial scales of variation in bacterioplankton assemblages of oligotrophic surface waters. *Mar Ecol Prog Ser* **311**: 67–76.

Hopkinson CS, Vallino JJ. (2005). Efficient export of carbon to the deep ocean through dissolved organic matter. *Nature* **433**: 142–145.

Huber JA, Mark Welch DB, Morrison HG, Huse SM, Neal PR, Butterfield DA *et al.* (2007). Microbial population structures in the deep marine biosphere. *Science* **318**: 97–100.

Huse SM, Huber JA, Morrrison HG, Sogin ML, Mark Welch D. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* **8**: R143.

Jeon SO, Bunge J, Stoeck T, Barger KJA, Hong SH, Epstein SS. (2006). Synthetic statistical approach reveals a high degree of richness of microbial eukaryotes in an anoxic water column. *Appl Environ Microbiol* **72**: 6578–6583.

Karner MB, DeLong EF, Karl DM. (2001). Archaeal dominance in the mesopelagic zone of the Pacific Ocean. *Nature* **409**: 507–510.

Kim S, Park, MG, Kim KY, Kim CH, Yih W *et al.* (2008). Genetic diversity of parasitic dinoflagellates in the genus amoebophrya and its relationship to parasite biology and biogeography. *J Eukaryot Microbiol* **55**: 1–8.

Kitano H. (2004). Biological robustness. *Nature Rev Gen* **5**: 826–837.

Lane DJ. (1991). 16S/23S rRNA sequencing. In: Stackebrandt E and Goodfellow M (eds). *Nucleic Acid Techniques in Bacterial Systematics*. John Wiley & Sons: New York. pp 115–176.

Lauro FM, Bartlett DH. (2008). Prokaryotic lifestyles in deep sea habitats. *Extremophiles* **12**: 15–25.

Lauro FM, Bertoloni G, Obraztsova A, Kato C, Tebo BM, Bartlett DH. (2004). Pressure effects on Clostridium strains isolated from a cold deep-sea environment. *Extremophiles* **8**: 169–173.

Lee SH, Fuhrman JA. (1991). Spatial and temporal variation of natural bacterioplankton assemblages studied by total genomic DNA cross-hybridization. *Limnol Oceanogr* **36**: 1277–1287.

Lefèvre E, Roussel B, Amblard C, Sime-Ngando T. (2008). The molecular diversity of freshwater picoeukaryotes reveals high occurrence of putative parasitoids in the plankton. *PLOS One* **3**: e2324.

Li WZ, Godzik A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**: 1658–1659.

Martín-Cuadrado A, López-García P, Alba J, Moreira D, Monticelli L, Strittmatter A *et al.* (2007). Metagenomics of the deep mediterranean, a warm bathypelagic habitat. *PLOS One* **2**: e914.

Massana R, Pedrós-Alió C. (2008). Unveiling new microbial eukaryotes in the surface ocean. *Curr Opin Microbiol* **11**: 213–218.

Massana R, Karniol B, Pommier T, Bodaker I, Beja O. (2008). Metagenomic retrieval of a ribosomal DNA repeat array from an uncultured marine alveolate. *Environ Microbiol* **10**: 1335–1343.

Montoya JP, Holl CM, Zehr JP, Hansen A, Villareal TA, Capone DG. (2004). High rates of N2 fixation by unicellular diazotrophs in the oligotrophic Pacific Ocean. *Nature* **430**: 1027–1032.

Musat N, Halm H, Winterholler B, Hoppe P, Peduzzi S, Hillion F *et al.* (2008). A single-cell view on the ecophysiology of anaerobic phototrophic bacteria. *Proc Natl Acad Sci USA* **105**: 17861–17866.

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

Pedrós-Alió C. (2006). Marine microbial diversity: can it be determined? *Trends Microbiol* **14**: 257–263.

Quince C, Curtis TP, Sloan WT. (2008). The rational exploration of microbial diversity. *ISMEJ* **2**: 997–1006.

Rodriguez-Brito B, Rohwer F, Edwards RA. (2006). An application of statistics to comparative metagenomics. *BMC Bioinformatics* **7**: 162.

Roesch LFW, Fulthorpe RR, Riva A, Casella G, Hadwin AKM, Kent AD *et al.* (2007). Pyrosequencing enumerates and contrasts soil microbial diversity. *ISMEJ* **1**: 283–290.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al.* (2007). The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLOS Biol* **5**: e77.

Salomon PS, Janson S, Granéli E. (2003). Multiple species of the dinophagous dinoflagellate genus Amoebophrya infect the same host species. *Environ Microbiol* **5**: 1046–1052.

Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR *et al.* (2006). Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc Natl Acad Sci USA* **103**: 12115–12120.

Torsvik V, Øvreås L, Thingstad TF. (2002). Prokaryotic diversity—magnitude, dynamics, and controlling factors. *Science* **296**: 1064–1066.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA *et al.* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66–74.

Witte U, Wenzhofer F, Sommer S, Boetius A, Heinz P, Aberle N *et al.* (2003). *In situ* experimental evidence of the fate of a phytodetritus pulse at the abyssal sea floor. *Nature* **424**: 763–766.

Yayanos AA. (1986). Evolutional and ecological implications of the properties of deep-sea barophilic bacteria. *Proc Natl Acad Sci USA* **83**: 9542–9546.

Zuendorf A, Bunge J, Behnke A, Barger KJA, Stoeck T. (2006). Diversity estimates of microeukaryotes below the chemocline of the anoxic Mariager Fjord, Denmark. *Fems Microbiol Ecol* **58**: 476–491.