nature microbiology

LETTERS

DOI: 10.1038/s41564-017-0008-3

Environmental drivers of a microbial genomic transition zone in the ocean's interior

Daniel R. Mende[1], Jessica A. Bryant[1,2], Frank O. Aylward[1,4], John M. Eppley[1], Torben Nielsen[1,3], David M. Karl[1] and Edward F. DeLong[1,2*]

The core properties of microbial genomes, including GC content and genome size, are known to vary widely among different bacteria and archaea[1,2]. Several hypotheses have been proposed to explain this genomic variability, but the fundamental drivers that shape bacterial and archaeal genomic properties remain uncertain[3–7]. Here, we report the existence of a sharp genomic transition zone below the photic zone, where bacterial and archaeal genomes and proteomes undergo a community-wide punctuated shift. Across a narrow range of increasing depth of just tens of metres, diverse microbial clades trend towards larger genome size, higher genomic GC content, and proteins with higher nitrogen but lower carbon content. These community-wide changes in genome features appear to be driven by gradients in the surrounding environmental energy and nutrient fields. Collectively, our data support hypotheses invoking nutrient limitation as a central driver in the evolution of core bacterial and archaeal genomic and proteomic properties.

Major physicochemical features that distinguish the ocean's twilight zone (mesopelagic zone, 200–1,000 m) from well-lit surface layers include lower light, temperature and oxygen levels, together with higher hydrostatic pressure and macronutrient concentrations. Although some differences between surface and deeper-water microbes are now predictable[8–11], the specific nature of genomic variability along the vertical depth continuum is less clear.

To gain a better understanding of the evolutionary and ecological trends that shape the ocean microbiome, we conducted a time-resolved deep metagenomic survey of bacterioplankton from the ocean's surface to a depth of 1,000 m in the North Pacific Subtropical Gyre at Station ALOHA (Supplementary Fig. 1 and Supplementary Table 1)[12]. Taxa and genes clustered primarily by depth, consistent with previous reports of stratified microbial community depth distributions[8,9,13] (Fig. 1 and Supplementary Fig. 2). Our analyses also revealed a marked vertical transition, where all samples above the deep chlorophyll maximum (DCM) clustered together (Fig. 1a,c and Supplementary Tables 2 and 3). Below the DCM, samples formed three distinct clusters: all 125 m samples, all 200 m samples and all those from 500 m and deeper.

The microbial taxonomic diversity also reflected a sharp discontinuity below the euphotic zone (below depths having <1% of surface 475 nm blue light[14]). The largest differences were found between the 75–125 m, 125–200 m and 200–500 m depth intervals (Fig. 1d). Similar trends in community richness were evident, with surface waters containing the lowest number of unique metagenomic operational taxonomic units (mOTUs, which are near-species-level sequence clusters of single-copy, universally conserved protein-coding genes. We used COG0012 in our analyses, since it has been demonstrated to be a robust universally conserved phylogenetic marker for near-species-level determinations[15,16]). A peak in taxon richness was observed at 125 m and 200 m, followed by a drop in richness at depths of >200 m (Fig. 1e). These shifts in microbial diversity coincided with changes in numerous physical and biogeochemical parameters (Fig. 1b)[17].

Community transitions in the 75–200 m depth interval were accompanied by major changes in the genomic and proteomic properties of resident microbes across the genomic transition zone (GTZ). Notably, across the DCM there was an abrupt change in the aggregate microbiome GC content. Collectively, genes found above the DCM had a lower average GC content than those from below the DCM, with the transition occurring at ~35% GC content (Fig. 2a). The increasing GC content shift below the DCM was reflected in every gene in our universal single-copy gene set (Supplementary Fig. 3). This trend was also accompanied by a distinct bias towards lower GC codons and lower codon diversity in surface water microbiomes (Supplementary Fig. 4). Additional changes across the GTZ included significant shifts towards larger average bacterial and archaeal genome sizes (Supplementary Fig. 5a) and intergenic spacer regions (Supplementary Fig. 5b) below the DCM. (Because genome size estimates could be confounded by viruses and eukaryotes, we removed any sequences identified as such from this analysis.)

The GC content shift occurred within a wide variety of disparate clades, including *Roseobacter*, Actinobacteria, *Prochlorococcus* and Thaumarchaeota (Fig. 2b and Supplementary Fig. 6). Although GC content trends were similar among most clades, some taxon-specific variability was observed. For example, the elevated GC content at depth in SAR11 populations spanned a much smaller range than was seen in other clades (Fig. 2b and Supplementary Figure 6). One bacterial clade (SAR324) showed an opposite GC trend below the GTZ. In aggregate, however, our data indicate that most pelagic, surface-dwelling bacterial clades have low-GC-content genomes, consistent with earlier reports from cultivar and single-cell genomes[18–21]. Furthermore, members of those same diverse bacterial clades show elevated GC content below the DCM.

Previous studies have found that low-GC genomes tend to encode proteomes with lower nitrogen but higher carbon contents, while high-GC genomes exhibit the opposite trend[22–24]. Grzymski and Dussaq[24] reported apparent differences in GC content between coastal and open ocean surface-water bacterioplankon in samples

[1]Daniel K. Inouye Center for Microbial Oceanography: Research and Education (C-MORE), University of Hawaii, Honolulu, HI 96822, USA. [2]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge 02139 MA, USA. [3]DOE Joint Genome Institute, Walnut Creek, CA 94598, USA. Present address: [4]Department of Biological Sciences, Virginia Tech, Blacksburg, VA 24061, USA. Daniel R. Mende, Jessica A. Bryant and Frank O. Aylward contributed equally to this work. *e-mail: edelong@hawaii.edu

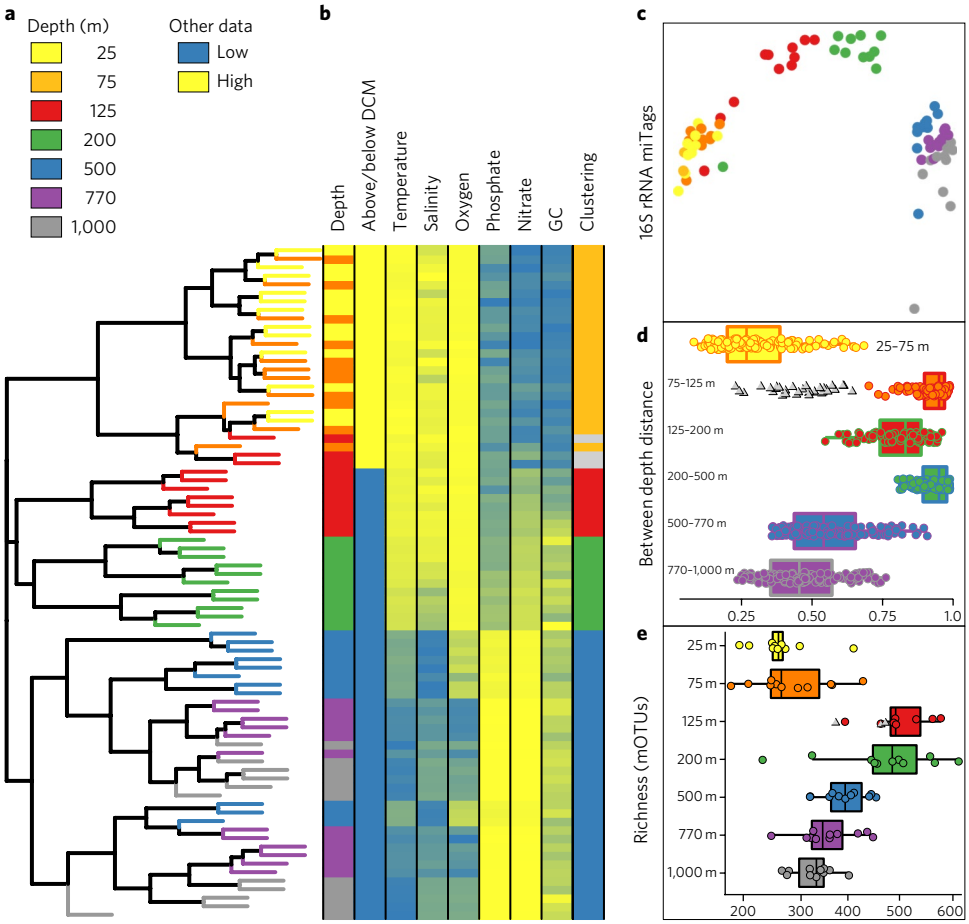

Fig. 1 | Quantitative relationships of microbiome genes and taxa and as a function of depth, time and environmental variables at Station ALOHA. **a**, Dendrogram displaying the Bray–Curtis distances between sample mOTU abundance profiles across the time series. **b**, Environmental data represented by a heatmap, ranging from blue (low) to yellow (high). **c**, Non-metric multidimensional scaling (NMDS) plot of small subunit ribosomal RNA miTag OTU abundance profiles. **d**, Bray–Curtis distances of mOTU abundances between samples at adjacent depths. Grey triangles at 125 m represent comparisons between 75 m samples and 125 m samples located above the DCM. Whiskers (error bars) show the lowest datum still within the 1.5 interquartile range (IQR) of the lower quartile, and the highest datum still within 1.5 IQR of the upper quartile. **e**, mOTU richness coloured according to depth as in **a**, except for the grey triangles at 125 m, which represent samples located above the DCM. Whiskers are defined as in **d**. All figures are based on 79 samples.

separated by hundreds to thousands of kilometres. In surface waters, predicted protein nitrogen content was negatively correlated with distance from land, and appeared positively correlated with 10 m depth annual global nitrate concentrations near sampling regions[24]. In our depth profiles, GC changes observed in microbial genes across just tens of metres in the GTZ corresponded to parallel shifts in predicted protein elemental composition (Fig. 2c,d). Specifically, the average number of nitrogen or carbon atoms per amino-acid-residue side chain (N-ARSC and C-ARSC) shifted in opposing directions across the GTZ. Bacteria and archaea below the DCM had high N-ARSC and low C-ARSC values, while the opposite trend was observed in surface water (Fig. 2c,d). Analyses we conducted on Tara Ocean Project depth profiles[25] revealed similar GC, N-ARSC and C-ARSC depth trends in all Atlantic and Pacific open ocean samples examined (Supplementary Fig. 7). Our analyses suggest that the GTZ and its associated features are common throughout the open ocean.

To identify changes in community composition that co-occur with whole-genome transitions across the GTZ, we performed a weighted gene correlation network analysis using mOTU abundances. This analysis revealed six primary sets of correlated mOTUs (termed modules) encompassing the majority of all identified taxa (Fig. 3a). Modules 1 and 6 contained mOTUs that dominated surface and deep mesopelagic waters, respectively, while the other modules showed well-defined abundance peaks between 125 and 500 m (Fig. 3a). Module taxonomic composition captured well-known distributions of high-light and low-light *Prochlorococcus*[20,26] and an increased abundance in Thaumarchaota at 125 m and deeper (Fig. 3b). Overall, the module variation trended similarly to patterns of whole-community GC content and N-ARSC, reflecting fundamental shifts in aggregate microbiome properties across the GTZ.

Nitrogen concentrations at a depth of 125 m varied considerably over the course of the time series. In tandem, microbial community composition, aggregate microbiome GC content, N-ARSC and C-ARSC varied with changing inorganic nitrogen at 125 m (Fig. 3c and Supplementary Fig. 8). Inorganic nitrogen availability was highly correlated with Module 3 in the mOTU network analyses, consisting mainly of Thaumarchaeota, SAR11 and SAR324 (Rho = 0.9; Fig. 3d). Inorganic nitrogen concentration at 125 m also correlated with community proteomic nitrogen content (N-ARSC), accounting for a large fraction of N-ARSC variation (Rho = 0.66) and suggesting a direct influence of ambient nitrate availability on community genomic properties (Fig. 3d). The abundance of key genes in nitrogen metabolic pathways also changed across the GTZ. High-affinity inorganic nitrogen transporters and nitrilases were over-represented in nitrogen-depleted surface

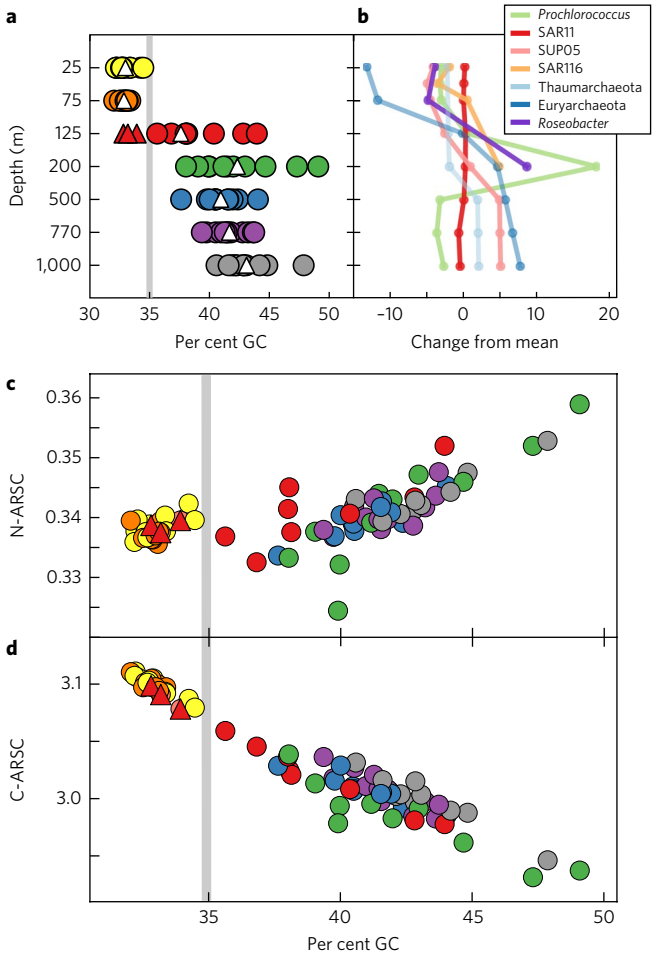

Fig. 2 | Microbiome GC content, N-ARSC and C-ARSC versus depth at Station ALOHA. a, Weighted average GC content of all assembled genes in bulk microbial communities. Red triangles indicate 125 m samples collected during periods when the DCM was located below a depth of 125 m. The vertical grey line highlights the partitioning of GC values in samples located above (left) and below (right) the DCM. **b**, Average difference between the GC content of mOTU genes that map to select taxa at a given depth, and the overall GC mean across all samples. **c**,**d**, Weighted average of N-ARSC (**c**) and C-ARSC (**d**) values of all Station ALOHA genes as a function of their corresponding GC content across all samples. Samples to the left of the vertical grey line were collected above, and to the right were collected below the DCM, respectively. Sample points are coloured by depth of origin as in Fig. 1. All figures are based on 79 samples.

waters (Supplementary Fig. 9), while ammonia and nitrite oxidation pathway genes were elevated in higher-nitrate, deeper waters (Supplementary Fig. 10). Collectively, our observations implicate nitrogen availability as a key driver of microbial community and genomic property shifts observed across the GTZ.

Elucidating the ecological and evolutionary factors that shape fundamental microbial genome properties is a central theme in comparative microbial genomics. Curiously, some microbes with strikingly similar global genomic properties (such as genome size or GC content), differ dramatically in their physiologies, life histories, population sizes and habitats[1–3,20]. For example, some open ocean surface-water cyanobacterial genomes share remarkably similar characteristics (low GC content, reduced genome size) with obligate bacterial symbionts of aphids[1–3,20]. Several different hypotheses have been proposed as drivers of such genomic evolutionary trends. These include Muller's ratchet[27], adaptation to nutrient limitation[24,28], high mutation rate[29] and the 'black queen hypothesis'[30]. A recent review of the various forces that might drive such trends concluded that nutrient limitation[24,28] was currently the most parsimonious explanation[3]. The dynamic co-variation of nitrogen availability and the microbial genomic and proteomic properties we observed strongly support this hypothesis.

Bacteria and archaea employ a variety of strategies to minimize cellular demand for limiting nutrients. Our observations reveal that across a very narrow depth stratum, microbial communities have optimized their genomic and proteomic elemental stoichiometry in response to prevailing environmental conditions. Similar macromolecular adjustments probably occur in other environmental contexts, which may provide further insight into both universal as well as unique adaptive features that help shape the structure, function and evolution of microbial genomes in the wild.

Methods

Samples were collected at approximately monthly intervals over a 1.5 year sampling period and metagenomic DNA was extracted, sequenced, individually assembled and annotated using custom workflows (Supplementary Fig. 1 and Supplementary Table 1). Individual genes from all samples were consolidated into a non-redundant gene catalogue consisting of 8.9 million genes. The Station ALOHA gene catalogue was then used to explore the properties, variability and distributional patterns of genomic properties and gene functions of Station ALOHA oligotrophic ocean microbiomes. Taxonomic distributions over time and space were investigated using sequence clusters of universally conserved, single-copy protein-coding marker genes referred to as mOTUs as previously described[15].

All samples were processed using the following procedure. Volumes of seawater (20 l) were collected and subsequently pre-filtered with a 1.6 µm, 42.5 mm Whatman GFA filter (cat. no. 1820-042). The filtrate was collected on 0.22 µm Sterivex GV filter for DNA (cat no. SVGV01015, Millipore). Cells were stored in 2 ml sucrose storage buffer (40 mM EDTA, 50 mM Tris (pH 8.3), 0.75 M sucrose) at −80 °C. For cell lysis, 2 mg ml^{-1} of lysozyme was added and cells were incubated at 37 °C for 30 min. Final concentrations of 1% SDS and 0.75 mg ml^{-1} proteinase K were subsequently added and the solution was incubated for 2 h at 55 °C. Final DNA purification was performed using the FujiFilm Quick Gene instrument with the QuickGene DNA Tissue Kit (cat. no. DT-L, Life Science). Libraries were created using the Illumina TruSeq LT Nano kit set A (PN: FC-121-4001). Sequencing data were generated using Illumina MiSeq and NextSeq 500 systems (Supplementary Table 1).

Metagenomic sequencing data were generated from 83 samples obtained between August 2010 and December 2011 at seven depths between 25 m and 1,000 m at Station ALOHA on 11 HOT cruises of the Hawaii Ocean Time-series (HOT). Only four samples from 45 m were sequenced, so they were excluded from analyses comparing different depths. Physicochemical data for all cruises are available in Supplementary Table 2 and on the Hawaii Ocean Time-series Data Organization and Graphical System (HOT-DOGS) website (http://hahana.soest.hawaii.edu/hot/hot-dogs/).

DNA samples for metagenomic sequencing and physicochemical data were obtained from separate hydrocasts. Data from different casts were matched using their potential density to account for internal waves (the inertial period of oscillation, which has a ~31 h period at Station ALOHA). In short, we interpolated measurements for all physicochemical data shown in Supplementary Table 2. DNA samples obtained in the mixed layer (depth calculated using HOT-DOGS applet) were matched to physicochemical data using the sampling depth. Samples from below the mixed layer were matched using their potential density (instead of depth) in order to account for the inertial oscillations.

Similarly, sample collection for metagenomes and chlorophyll determinations were performed on separate hydrocasts. To determine the position of the metagenomic samples relative to the DCM, we used the seawater potential density measurements taken during DNA sampling casts and calculated the corresponding density and depth in the chlorophyll cast. Next, we determined whether the metagenomic sample was located above or below the chlorophyll maximum. The results are summarized in Supplementary Table 3.

Meteorological data, including wind speed, precipitation and solar irradiance, were measured by the Upper Oceans Processes Group at the Woods Hole Oceanographic Institution with the WHOTS buoy located at Station ALOHA, and were retrieved from http://uop.whoi.edu/projects/WHOTS/whots.html (accessed 2 February 2016). Measurements during the 5 and 30 days leading up to sampling, all made at regular intervals, were averaged for subsequent analyses.

Assembly. Raw sequencing data were quality filtered using MIRA v. 4.9.5_2 with the qc and pec options and standard parameters to retain a high confidence region (HCR) for every read. This step also included the removal of contamination by phiX[31]. MIRA was also used to assemble the sequencing data of each sequencing run into contigs using the standard workflow for accurate

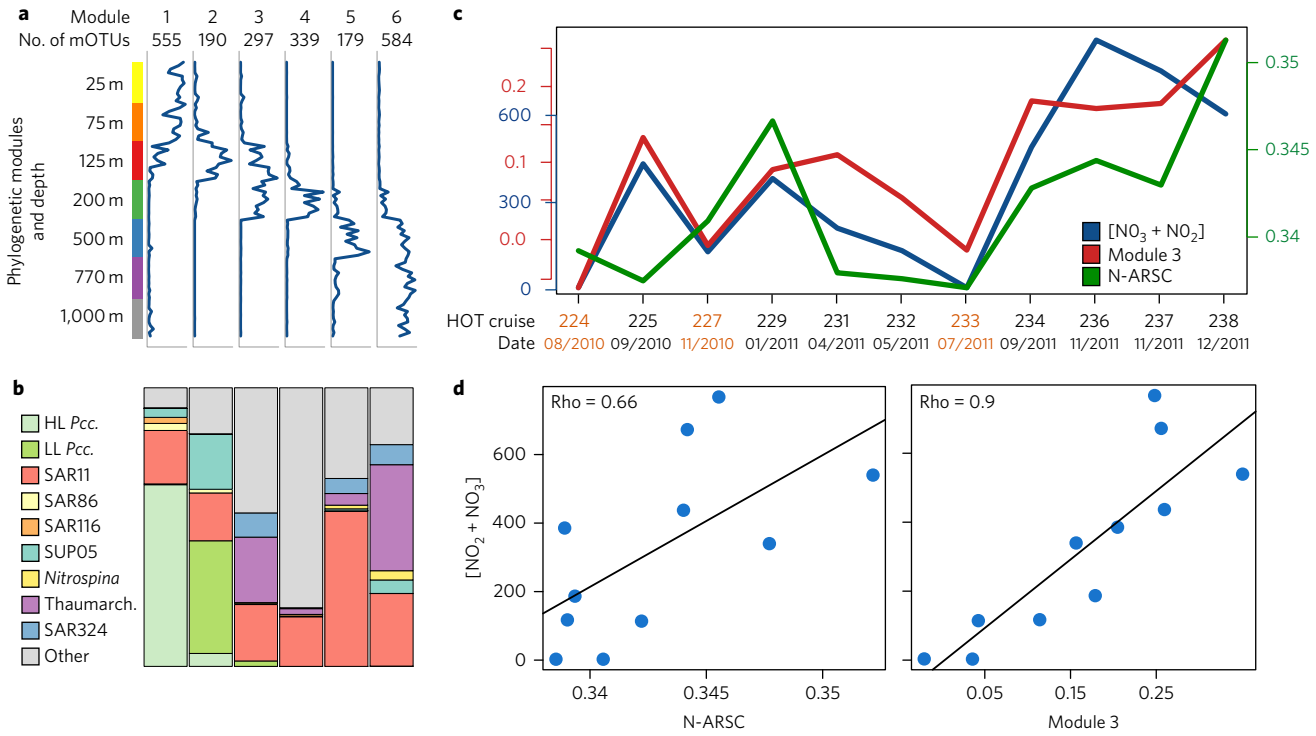

Fig. 3 | Distribution of taxon modules through the GTZ, and with nitrogen concentrations over time at 125 m. a, Eigengenes, or first-principle components, representing the relative abundance of the six most abundant mOTU modules identified in weighted network analyses. **b**, Prominent microbial clades represented by the mOTUs in the six modules. Only mOTUs that could be classified are shown. HL *Pcc.*, high-light *Prochlorococcus*; LL *Pcc.*, low-light *Prochlorococcus*. **c**, Trend lines for [NO_2^- + NO_3^-] concentrations, N-ARSC and mOTU module 3 abundance in the 125 m samples, showing broad congruence between the three. Samples coloured orange on the *x* axis correspond to 125 m samples occurring above the DCM. **d**, Regression plots of both N-ARSC and the module 3 eigengene versus [NO_2^-+NO_3^-], demonstrating correspondence between these variables. All correlations reported as Spearman's rho. Results in **a** and **b** are based on 79 samples, and in **c** and **d** are based on 11 samples.

de novo genome assembly. The assembly quality statistics are summarized in Supplementary Table 1.

ALOHA gene catalogue. Genes were predicted from the assembled contigs using Prodigal[32], and only genes that were predicted to be complete were retained. This yielded a total of 39,436,252 protein-coding genes. We generated a non-redundant gene catalogue by clustering this set of genes using CD-HIT (95% nucleotide identity and 90% overlap of the shorter gene), resulting in the ALOHA gene catalogue, which encompasses 8,966,703 non-redundant gene clusters, each with a single representative sequence used for downstream annotation[33,34].

Functional and taxonomic annotation of the gene catalogue. We annotated the non-redundant gene catalogue using multiple databases. For taxonomic annotations we used an augmented version of RefSeq release 75 (ref. [35]), which was amended by a number of high-quality single-cell amplified genomes (SAGs) from marine environments. Functional annotations were generated using the KEGG[35,36] and eggNOG databases[37].

For both RefSeq and KEGG annotations, all genes were aligned to the respective databases using LAST version 756 (ref. [38]), with scoring parameters '-b 1 -x 15 -y 7 -z 25'. Each gene was assigned to the most specific taxon common to all RefSeq hits scoring within 1% of the best hit. Genes were assigned to the KEGG orthologous group or groups represented by all KEGG genes scoring within 5% of the best hit. The HMM-based eggnog-mapper tool[39] was used to obtain eggNOG annotations.

Clade designations. Reference genomes in NCBI were classified to well-known clades of marine microbial groups to assist with phylogenetic annotations of genes in the ALOHA catalogue. Some previous studies have provided clade-level classifications of sequenced genomes[11], but the affiliation of many publicly available genomes was not specified. To resolve these taxon affiliations in reference genomes, a concatenated phylogeny was constructed from a custom reference genome data set composed of 480 genomes found to be highly represented in the gene catalogue homology search annotations. We extracted the complete set of 40 universal, single-copy marker genes[40,41] from these genomes using fetchMGs[16]. These genes were then used to build a phylogenetic tree using the standard_fasttree workflow in ete3 (ref. [41]) and FastTree[42]. Initial designations were provided via well-known genomes and previous designations, and subsequent classifications were propagated to genomes within the same monophyletic clade of the tree. Genes in the ALOHA gene catalogue that were taxonomically annotated to one of the clade-designated genomes were assigned to the respective clades.

mOTUs. mOTUs are near-species-level sequence clusters of a set of near-universal, single-copy, protein-coding genes[15,16]. These genes encompass multiple orthologous groups, each of which has a different speed of evolution and hence is clustered at an individualy optimized, near-species-level sequence identity cutoff. We established a customized version of the mOTUs using the data set accompanying this publication in combination with genes from the Tara Expedition[25] and a custom database of select marine genomes. The fetchMGs tool[15] was used to extract the universal, single-copy genes from all data sets. The nucleotide identities between all pairs of orthologous sequences were computed using vsearch (version v1.9.3) and only alignments with more than 20 aligned bases were retained[43]. The resulting distance matrix was used to generate an average linkage hierarchical clustering and the mOTU near-species-level clusters were extracted from this using optimized cutoffs[43]. Linking of different orthologous groups was not possible in this data set, probably due to the high degree of co-abundance between different phylogenetic groups, and we therefore focused our analyses of mOTU richness and beta diversity on individual mOTUs based on one gene, namely COG0012 (a ribosome-associated GTPase) (Fig. 1a,d,e and Fig. 3). This gene was first noted to be a universal protein in bacteria and archaea by Galperin and Koonin[44], and a recent survey confirmed that COG0012 was present in >99% of 25,000 microbial genomes in single copy. For genome size estimates we used all ten universal, single-copy mOTU genes (see section below). The mOTU abundances were calculated as described below (see section 'Mapping and abundance estimations').

mOTUs were designated to clades and other taxonomic levels (such as phyla and genera) using LAST v. 756 (ref. [38]) alignments to the custom version of RefSeq release 75 (ref. [35]), described above. The best alignment for each gene sequence of a mOTU was used to calculate a majority taxonomic assignment for the mOTU at all taxonomic levels.

mOTU richness. COG0012 mOTU richness was calculated from read mapping count data (for this purpose only uniquely mapping inserts were used). For

each sample, read counts were downsampled to the lowest total read count of all samples using the function rarefy of the R package Vegan[45]. COG0012 mOTU richness was calculated using the function specnumber, both from the R package Vegan. To further validate these patterns we performed the same analysis using another universal single-copy protein, COG0533, a metal-dependent protease with predicted chaperone activity. The results obtained with COG0533 were nearly identical to those using COG0012, and confirm all of the conclusions derived from the results using COG0012.

miTags and SSU rRNA GC. We assembled a SSU rRNA OTU database based on the miTag approach described by Logares et al.[46], by using usearch v. 8.1 (ref. [47]) to cluster all nearly full-length sequences in the SILVA SSU Ref NR99 database (release 123)[48,49], and then compiled a set of genes that shared less than 97% sequence similarity to one another. SSU rRNA gene fragments were then extracted from our quality-filtered unassembled Illumina data sets using riboFrame[50]. Extracted fragments were aligned to our custom SSU rRNA OTU database using bowtie2 with the parameters '--local', '--very-sensitive' and '-k 100'[51]. This returned up to 100 distinct alignments, which was sufficient to identify fragments that hit equally well to multiple database sequences. Fragments were assigned to their top database OTU provided alignment lengths were greater than 70 bps and had a sequence identity of at least 97%. Reads with the top alignment score to multiple database sequences (ambiguous matches) were assigned to OTUs in equal proportions to that of unambiguous matches to each OTU in the same sample. This approach could in theory produce OTUs encompassing SSU rRNA fragments with 94% or greater sequence identities (roughly considered the family level). However, 99% of our SSU fragments had 99% or greater sequence identity to their assigned database OTU.

Reads aligning to multiple database sequences equally well were assigned to OTUs in equal proportions to unambiguous matches to each OTU in the same sample. A total of 6,990 SSU rRNA fragments assigned to OTUs were then randomly chosen from each sample. OTU relative abundances for the subsampled data set were used to generate an NMDS plot to visualize Bray-Curtis distances with the metaMDS command in the R package Vegan[45].

Mapping and abundance estimations. Quality-trimmed sequencing reads were aligned to the ALOHA gene catalogue using BWA (mem algorithm, standard parameters)[52]. Results were filtered using a 95% identity cutoff and a minimum alignment length of 45 using msamtools[53]. For alignments that did not encompass a complete read, a more stringent minimum alignment length of 60 bp was applied. Alignment quality was assessed using BWA alignment scores. If both reads of an insert could be aligned to the same reference, a summed alignment score for the insert was calculated. The highest scoring alignment for each insert was kept for abundance counting. Inserts with multiple highest scoring alignments were flagged as multiple mappers. To estimate the abundance of each gene, we first counted all unique alignments to each of the genes (alignments not flagged as multiple mappers). In a second step, all multiple mappers were distributed among the different genes according to the abundance profiles of the unique alignments. Gene coverage was calculated by calculating the total number of bases mapping to a gene and then dividing this number by the length of the gene. To calculate an average-per-genome-copy number, all coverages were divided by the average coverage of all 10 universal, single-copy genes (mOTUs) found in the same sample.

Weighted gene co-abundance network analysis. Abundance profiles for the COG0012 mOTUs were used to create a weighted gene correlation network using methods similar to those previously described[13,54,55]. Total reads mapped to mOTUs was used as the abundance criterion, and counts were normalized using the variance stabilizing transformation implemented in DESeq2 (ref. [56]). A soft-threshold of 3 was chosen based on the scale-free network criterion[57], and modules of co-abundant mOTUs were constructed using the blockwiseModules command (parameters: minModuleSize = 2, mergeCutHeight = 0.25). Module eigengenes, or first-principle components, were constructed using the moduleEigengenes command. Networks were visualized using igraph.

Hierarchical clustering cladogram. COG0012 mOTU abundances were calculated as coverage as described, and then normalized so the total sum equalled 1. The abundances were then used to calculate Bray-Curtis distances between all pairs of samples using the R package Vegan[45]. The distances were used to compute a complete linkage hierarchical clustering. The clustering was annotated with information about the sampling environment and the microbial communities (Fig. 1b).

Genome size estimation. We utilized all 10 mOTU genes (near single-copy, near universal marker genes)[15] to estimate genome sizes. For this purpose, we normalized the total gene abundance in each sample so that the average abundance of the 10 mOTU genes was 1. We then subtracted the abundance of genes annotated as viral and eukaryotic. Even though these genes are usually found in low abundance in our samples, they could potentially influence the results. After these calculations, the total gene abundance represents an estimate for the average number of genes per genome. The results are displayed in Supplementary Fig. 5a.

Gene characteristic calculations for GC, codon usage and encoded protein elemental composition (N-ARSC and C-ARSC). GC content was calculated using sequence utilities within biopython[58]. The effective codon number, which ranged from 20, where only one codon from each synonymous codon set is used, to 61, where all synonymous codons for each amino acid are used at even frequencies, was then calculated for each gene[59]. To quantify preferences for codon GC content, we developed a degenerate codon ranking scheme from 0 to 1, with 0 and 1 indicating the codon(s) with highest or lowest G+C content, respectively, relative to all codons encoding each amino acid. Codon rankings for each amino acid with degeneracies in the codon table were averaged across each assembled gene. N-ARSC and C-ARSC values for each gene were then calculated by translating DNA sequences into amino acids (NCBI codon table 11, as used by Prodigal), tallying the number of nitrogen and carbon atoms in the side chains of encoded amino acids residues and dividing N and C counts by the amino-acid length of the gene. GCs, effective codon number, codon GC ranks, N-ARSC and C-ARSC were averaged for all genes assembled within a sample. Abundance-weighted average gene GC, N-ARSC and C-ARSC values of all gene representatives from the non-redundant reference gene set (Figs. 2a,c,d and 4) and average GC for gene-catalogue representatives mapping to select clades (Supplementary Fig. 6) were calculated using gene coverage estimates as described above. The average of all assembled genes mapping to select clades (Supplementary Fig. 6), and all assembled mOTU sequences from each mOTU set (Supplementary Fig. 4), as well as all assembled COG0012 mOTUs that mapped to select clades (Fig. 2b), were also calculated.

GC and encoded protein elemental composition (N-ARSC and C-ARSC) in the Tara data set. Tara Ocean's reference gene catalogue, gene abundance data and sample environmental data were downloaded from the Tara Ocean's Project Companion website (http://ocean-microbiome.embl.de/companion.html, accessed 15 October 2016). The Tara Ocean's reference gene catalogue contains both full and partial genes. We selected full-length genes from the catalogue by screening each sequence for start and stop codons and calculated the GC, N-ARSC and C-ARSC of full-length sequences. Next, we calculated the weighted-average GC, N-ARSC and C-ARSC values for each sample using the sample abundances of the full-length genes reported by Tara. To compare these values among different Tara sampling depths, we selected 13 open ocean samples for which metagenomic data for all three depths (surface, DCM and mesopelagic) were available (stations 38, 64, 72, 76, 78, 98, 100, 102, 112, 132, 133, 138 and 142).

Functional analysis of nitrogen metabolism. All functions annotated to the KEGG pathway for 'Nitrogen Metabolism' (map00910)[36] were extracted from the KEGG orthologue abundance matrix. From these, the average abundance per depth as well as the correlation with nitrate + nitrite concentrations was calculated. KEGG orthologues that were either highly abundant (>0.1 average gene copies per genome) or highly correlated or anticorrelated with nitrate + nitrite concentrations (Spearman correlation, Bonferroni corrected P value of <0.01) were flagged as central genes in nitrogen metabolism at Station ALOHA. From these key functions in the nitrate + nitrite-poor surface and the nitrate + nitrite-rich deeper water layers, metabolic reconstructions were compiled by first collecting those functions that were highly abundant in 25 m and 75 m samples and/or highly anticorrelated with nitrate + nitrite. These KEGG orthologues were then mapped onto the KEGG pathway for 'Nitrogen Metabolism' (Supplementary Fig. 9). The same procedure was performed for deeper, nitrogen-rich samples (Supplementary Fig. 10).

Code availability. All custom code used in this study has been deposited on github.com. General scripts are available from https://github.com/pangenomics/GTZ. Scripts for the calculation of C-ARSC and N-ARSC values are available at https://github.com/JessAwBryant/gene-characteristics. Scripts for taxonomic and KEGG annotations are available on https://github.com/jmeppley/py-metagenomics.

Data availability. Data supporting the findings of this study that are not included in the manuscript are included in the Supplementary Figures and Supplementary Tables 1–4. Sequence data are available from the NCBI short read archive (SRA) under Bioproject no. PRJNA352737, as indicated in Supplementary Table 4. All other data products associated with this study are available from the corresponding author upon request.

Received: 29 March 2017; Accepted: 4 July 2017;

References

1. Ochman, H. & Davalos, L. M. The nature and dynamics of bacterial genomes. *Science* **311**, 1730–1733 (2006).
2. McCutcheon, J. P. & Moran, N. A. Extreme genome reduction in symbiotic bacteria. *Nat. Rev. Microbiol.* **10**, 13–26 (2011).
3. Batut, B., Knibbe, C., Marais, G. & Daubin, V. Reductive genome evolution at both ends of the bacterial population size spectrum. *Nat. Rev. Microbiol.* **12**, 841–850 (2014).

4. Lynch, M. & Conery, J. S. The origins of genome complexity. *Science* **302**, 1401–1404 (2003).
5. Daubin, V. & Moran, N. A. Comment on 'The origins of genome complexity'. *Science* **306**, 978 (2004).
6. Giovannoni, S. J. Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**, 1242–1245 (2005).
7. Viklund, J., Ettema, T. J. G. & Andersson, S. G. E. Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade. *Mol. Biol. Evol.* **29**, 599–615 (2012).
8. DeLong, E. F. et al. Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**, 496–503 (2006).
9. Konstantinidis, K. T., Braff, J., Karl, D. M. & DeLong, E. F. Comparative metagenomic analysis of a microbial community residing at a depth of 4,000 meters at station ALOHA in the North Pacific subtropical gyre. *Appl. Environ. Microbiol.* **75**, 5345–5355 (2009).
10. Mizuno, C. M., Ghai, R., Saghaï, A., López-García, P. & Rodriguez-Valera, F. Genomes of abundant and widespread viruses from the deep ocean. *mBio* **7**, e00805–16 (2016).
11. Swan, B. K. et al. Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. *Science* **333**, 1296–1300 (2011).
12. Karl, D. M. & Lukas, R. The Hawaii Ocean Time-series (HOT) program: background, rationale and field implementation. *Deep Sea Res. Part II* **43**, 129–156 (1996).
13. Bryant, J. A. et al. Wind and sunlight shape microbial diversity in surface waters of the North Pacific Subtropical Gyre. *ISME J.* **10**, 1308–1322 (2016).
14. Laws, E. A., Letelier, R. M. & Karl, D. M. Estimating the compensation irradiance in the ocean: the importance of accounting for non-photosynthetic uptake of inorganic carbon. *Deep Sea Res. Part I* **93**, 35–40 (2014).
15. Sunagawa, S. et al. Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* **10**, 1196–1199 (2013).
16. Mende, D. R., Sunagawa, S., Zeller, G. & Bork, P. Accurate and universal delineation of prokaryotic species. *Nat. Methods* **10**, 881–884 (2013).
17. Letelier, R. M., Karl, D. M., Abbott, M. R. & Bidigare, R. R. Light driven seasonal patterns of chlorophyll and nitrate in the lower euphotic zone of the North Pacific Subtropical Gyre. *Limnol. Oceanogr.* **49**, 508–519 (2004).
18. Swan, B. K. et al. Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc. Natl Acad. Sci. USA* **110**, 11463–11468 (2013).
19. Giovannoni, S. J., Cameron Thrash, J. & Temperton, B. Implications of streamlining theory for microbial ecology. *ISME J.* **8**, 1553–1565 (2014).
20. Biller, S. J., Berube, P. M., Lindell, D. & Chisholm, S. W. *Prochlorococcus*: the structure and function of collective diversity. *Nat. Rev. Microbiol.* **13**, 13–27 (2015).
21. Dupont, C. L. et al. Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J.* **6**, 1186–1199 (2012).
22. Bragg, J. G. & Hyder, C. L. Nitrogen versus carbon use in prokaryotic genomes and proteomes. *Proc. Biol. Sci.* **271**(Suppl. 5), S374–S377 (2004).
23. Baudouin-Cornu, P., Schuerer, K., Marlière, P. & Thomas, D. Intimate evolution of proteins. Proteome atomic content correlates with genome base composition. *J. Biol. Chem.* **279**, 5421–5428 (2004).
24. Grzymski, J. J. & Dussaq, A. M. The significance of nitrogen cost minimization in proteomes of marine microorganisms. *ISME J.* **6**, 71–80 (2012).
25. Sunagawa, S. et al. Ocean plankton. Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
26. Rocap, G. et al. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424**, 1042–1047 (2003).
27. Moran, N. A., McCutcheon, J. P. & Nakabachi, A. Genomics and evolution of heritable bacterial symbionts. *Annu. Rev. Genet.* **42**, 165–190 (2008).
28. Baudouin-Cornu, P., Surdin-Kerjan, Y., Marlière, P. & Thomas, D. Molecular evolution of protein atomic composition. *Science* **293**, 297–300 (2001).
29. Partensky, F. & Garczarek, L. *Prochlorococcus*: advantages and limits of minimalism. *Ann. Rev. Mar. Sci.* **2**, 305–331 (2010).
30. Morris, J. J., Lenski, R. E. & Zinser, E. R. The black queen hypothesis: evolution of dependencies through adaptive gene loss. *mBio* **3**, e00036–12 (2012).
31. Chevreux, B., Wetter, T. & Suhai, S. Genome sequence assembly using trace signals and additional sequence information. *GCB* **99**, 45–46 (1999).
32. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
33. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
34. Li, W. & Godzik, A. CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
35. Tatusova, T., Ciufo, S., Fedorov, B., O'Neill, K. & Tolstoy, I. RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res.* **42**, D553–D559 (2014).
36. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2016).
37. Huerta-Cepas, J. et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **44**, D286–D293 (2016).
38. Kiełbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21**, 487–493 (2011).
39. Huerta-Cepas, J. et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.* https://doi.org/10.1093/molbev/msx148 (2017).
40. Ciccarelli, F. D. et al. Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283–1287 (2006).
41. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
42. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
43. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).
44. Galperin, M. Y. & Koonin, E. V. Who's your neighbor? New computational approaches for functional genomics. *Nat. Biotechnol.* **18**, 609–613 (2000).
45. Dixon, P. VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* **14**, 927 (2003).
46. Logares, R. et al. 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 (2014).
47. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
48. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
49. Yilmaz, P. et al. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res.* **42**, D643–D648 (2013).
50. Ramazzotti, M., Berná, L., Donati, C. & Cavalieri, D. riboFrame: an improved method for microbial taxonomy profiling from non-targeted metagenomics. *Front. Genet.* **6**, 329 (2015).
51. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
52. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
53. Arumugam, M., Harrington, E. D., Foerstner, K. U., Raes, J. & Bork, P. SmashCommunity: a metagenomic annotation and analysis tool. *Bioinformatics* **26**, 2977–2978 (2010).
54. Aylward, F. O. et al. Microbial community transcriptional networks are conserved in three domains at ocean basin scales. *Proc. Natl Acad. Sci. USA* **112**, 5443–5448 (2015).
55. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
56. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
57. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, 17 (2005).
58. Cock, P. J. A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
59. Wright, F. The 'effective number of codons' used in a gene. *Gene* **87**, 23–29 (1990).

Acknowledgements

The authors thank the captain and crew of the R/V *Kilo Moana*, and the Hawaii Ocean Time-series marine operations team, for their expert assistance with sample collection and oceanographic data acquisition and analyses at sea. The authors also thank T. Palden, A. Romano and P. Den Uyl for their able assistance in DNA library preparation and DNA sequencing, and B. Barone and L. Fujieki for expert advice and assistance in accessing and displaying HOT oceanographic data sets. The authors also thank S. Sunagawa and G. Zeller for advice and assistance with mOTU analyses. This research was supported by the Simons Foundation (SCOPE award ID 329108 to E.F.D. and D.M.K.), the Gordon and Betty Moore Foundation (through grants GBMF 3777 to E.F.D. and GBMF3794 to D.M.K.) and the National Science Foundation for support of the HOT programme (including the most recent OCE1260164), as well as support to D.R.M. from EMBO (ALTF 721-2015) and the European Commission (LTFCOFUND2013, GA-2013-609409) and support to J.A.B. through the US EPA Science to Achieve Results Fellowship. This work is a contribution of the Simons Collaboration on Ocean Processes and Ecology, and the Center for Microbial Oceanography: Research and Education.

Author contributions

E.F.D. and D.M.K. designed the overall study, sample and data collection, and analyses. T.N., J.M.E., D.R.M., J.A.B. and F.O.A. performed bioinformatics analyses with input from E.F.D. The manuscript was written by E.F.D., D.R.M., J.A.B. and F.O.A.

Competing interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available for this paper at doi:10.1038/s41564-017-0008-3.

Correspondence and requests for materials should be addressed to E.F.D.