UC San Diego
UC San Diego Previously Published Works

Title
Metagenomic covariation along densely sampled environmental gradients in the Red Sea

Permalink
https://escholarship.org/uc/item/4tw8b3st

Journal
The ISME Journal: Multidisciplinary Journal of Microbial Ecology, 11(1)

ISSN
1751-7362

Authors
Thompson, Luke R
Williams, Gareth J
Haroon, Mohamed F
et al.

Publication Date
2017

DOI
10.1038/ismej.2016.99

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

The ISME Journal (2017) 11, 138–151
© 2017 International Society for Microbial Ecology All rights reserved 1751-7362/17
www.nature.com/ismej

ORIGINAL ARTICLE

Metagenomic covariation along densely sampled environmental gradients in the Red Sea

Luke R Thompson[1,2], Gareth J Williams[3,4], Mohamed F Haroon[1], Ahmed Shibl[1], Peter Larsen[5], Joshua Shorenstein[2], Rob Knight[2,6] and Ulrich Stingl[1]

[1]*Red Sea Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia;* [2]*Department of Pediatrics, University of California, San Diego, CA, USA;* [3]*Center for Marine Biodiversity and Conservation, Scripps Institution of Oceanography, La Jolla, CA, USA;* [4]*School of Ocean Sciences, Bangor University, Anglesey, UK;* [5]*Argonne National Laboratory, Argonne, IL, USA and* [6]*Department of Computer Science, University of California, San Diego, CA, USA*

Oceanic microbial diversity covaries with physicochemical parameters. Temperature, for example, explains approximately half of global variation in surface taxonomic abundance. It is unknown, however, whether covariation patterns hold over narrower parameter gradients and spatial scales, and extending to mesopelagic depths. We collected and sequenced 45 epipelagic and mesopelagic microbial metagenomes on a meridional transect through the eastern Red Sea. We asked which environmental parameters explain the most variation in relative abundances of taxonomic groups, gene ortholog groups, and pathways—at a spatial scale of <2000 km, along narrow but well-defined latitudinal and depth-dependent gradients. We also asked how microbes are adapted to gradients and extremes in irradiance, temperature, salinity, and nutrients, examining the responses of individual gene ortholog groups to these parameters. Functional and taxonomic metrics were equally well explained (75–79%) by environmental parameters. However, only functional and not taxonomic covariation patterns were conserved when comparing with an intruding water mass with different physicochemical properties. Temperature explained the most variation in each metric, followed by nitrate, chlorophyll, phosphate, and salinity. That nitrate explained more variation than phosphate suggested nitrogen limitation, consistent with low surface N:P ratios. Covariation of gene ortholog groups with environmental parameters revealed patterns of functional adaptation to the challenging Red Sea environment: high irradiance, temperature, salinity, and low nutrients. Nutrient-acquisition gene ortholog groups were anti-correlated with concentrations of their respective nutrient species, recapturing trends previously observed across much larger distances and environmental gradients. This dataset of metagenomic covariation along densely sampled environmental gradients includes online data exploration supplements, serving as a community resource for marine microbial ecology.
The ISME Journal (2017) **11,** 138–151; doi:10.1038/ismej.2016.99; published online 15 July 2016

Introduction

Microbial communities have a central role in energy flow and carbon and nutrient cycling in the oceans. Shotgun sequencing and analysis of microbial community DNA (metagenomics) is now an established method for understanding the microbial genomic diversity underlying these processes (DeLong *et al.*, 2006; Dinsdale *et al.*, 2008). Distribution of microbial diversity and biogeochemistry is structured in large part by environmental gradients in light, temperature, oxygen, salinity and nutrients. Oceanographic surveys spanning such environmental gradients, combining metagenomic sequencing and measurement of continuous environmental variables, are enabling quantitative understanding of microbial communities (Gianoulis *et al.*, 2009; Raes *et al.*, 2011). Global oceanographic surveys have sequenced hundreds of surface and moderately deep (epipelagic and mesopelagic) ocean microbial communities (Rusch *et al.*, 2007; Sunagawa *et al.*, 2015), cataloging the vast genomic diversity of ocean microbes; further analyses of these data have identified correlations between environmental parameters and genetic community traits (gene ortholog groups and pathways) with predictive power (Gianoulis *et al.*, 2009; Raes *et al.*, 2011; Barberán *et al.*, 2012). Local studies at individual ocean sites, meanwhile, have shown how microbial taxa and gene ortholog groups are partitioned at greater detail along the water column and between discrete ocean environments (DeLong *et al.*, 2006; Coleman

Correspondence: LR Thompson or U Stingl, Department of Pediatrics, University of California, 9500 Gilman Drive, San Diego, CA 92037, USA.
E-mail: luket@alum.mit.edu or ulistingl@gmail.com
Received 9 March 2016; revised 8 June 2016; accepted 12 June 2016; published online 15 July 2016

and Chisholm, 2010; Ghai *et al*., 2010; Thompson *et al*., 2013). Depth is a critical factor behind community structure in the open ocean (DeLong *et al*., 2006), and dense sampling is capable of capturing subtle changes in environmental parameters with sufficient replication for statistical power.

The Red Sea is an ideal oceanic site for dense sampling of metagenomes to study environment–microbe covariation. The Red Sea is a deep (>2000 m) incipient ocean with strong latitudinal and depth-dependent gradients in temperature, salinity, oxygen and nutrients (Edwards, 1987). Like the open-ocean gyres of the Atlantic and Pacific Oceans, the Red Sea is oligotrophic with surface waters dominated by the picoplankton *Prochlorococcus* and *Pelagibacter* (Ngugi and Stingl, 2012). More so than these open-ocean gyres, however, the Red Sea lies at pelagic extremes of irradiance, temperature and salinity. The Red Sea experiences a late-summer southern influx of water called the Gulf of Aden Intermediate Water (GAIW), a foreign water mass that is cooler, fresher and more nutrient-rich than the native Red Sea water mass. The Red Sea is compact enough to sample across these gradients and water masses on a single month-long expedition, sampling more densely along transects and deeper through the water column than possible on a global survey.

We undertook a high-resolution metagenomic survey of the Red Sea, conducting a multivariate community analysis of covariation between environmental parameters and metagenome-derived taxonomic and functional metrics. We followed three main lines of questioning. First, how well can both taxonomic and functional microbial diversity be explained by environmental parameters, and which environmental parameters explain the most variation? Sunagawa *et al.* (2015) showed in a recent global ocean survey that temperature could explain more variation in taxonomic abundance than any other parameter. At smaller spatial scales and narrower temperature ranges, does temperature still have the most explanatory power? Which parameters can best explain residual variation? Second, are patterns of environmental covariation conserved across co-occurring water masses? Sampling the GAIW allowed us to determine whether this co-occurring water mass follows the same organizational principles (covariation with environmental parameters) as the native Red Sea water mass, across different taxonomic and functional metrics. Third, how are microbes functionally adapted along environmental gradients of irradiance, temperature, salinity and nutrients, including extremes in these parameters? Do marine communities exhibit fine-scale genomic adapation to environmental parameters as has been observed between separate oceans? Our dataset has allowed us to address these questions, and supporting online resources will make the processed data available to the wider community for further investigations.

Materials and methods

Oceanographic sampling

Samples were collected aboard the R/V *Aegaeo* on Leg 1 of the 2011 KAUST Red Sea Expedition, 15 September–11 October 2011. At eight stations, 20 l seawater was collected from each of depths 10, 25, 50, 100, 200 and 500 m; in two cases (Stations 12 and 34) where the seafloor was shallower than 500 m, the deepest sample was taken at the seafloor. Water was collected in 10 l Niskin bottles (that is, two Niskin bottles per depth), attached to a CTD rosette. Back on deck, the seawater was filtered through a series of three 293 mm mixed cellulose esters filters (Millipore, Billerica, MA, USA) of pore sizes 5.0 μm, 1.2 μm and 0.1 μm. Filters were placed in sealed plastic bags and frozen at −20 °C. Station properties (location, depth of mixed layer, chlorophyll maximum and oxygen minimum) are described in Supplementary Table S1. Physical oceanographic measurements (pressure, temperature, conductivity, chlorophyll *a*, turbidity and dissolved oxygen) were collected on a modified SeaBird 9/11+ rosette/CTD system, described in Supplementary Methods. Nutrient measurements (nitrate+nitrite, nitrite, ammonium, phosphate and silicate) on the final 0.1 μm filtrate were carried out at the UCSB Marine Science Institute and the Woods Hole Oceanographic Institution (Supplementary Methods). Sample water properties are described in Supplementary Table S2.

DNA extraction and whole-genome shotgun sequencing of community metagenomes

Community DNA was extracted from the 0.1 μm filters (0.1–1.2 μm size fraction) using phenol–chloroform extraction, similar to Rusch *et al.* (2007) and Ngugi *et al.* (2012); the full protocol is described in Supplementary Methods. Yields of genomic DNA ranged from 200–1500 ng per sample. Whole-genome shotgun libraries were made using the Nextera DNA Library Prep Kit (Illumina, San Diego, CA, USA). Median insert size by sample ranged from 183 bp to 366 bp (Supplementary Table S3). Libraries were sequenced using Illumina HiSeq 2000 paired-end (2 × 100 bp) sequencing, filling a total of three lanes (15 samples per lane). Sequence length after adapter removal was 93 bp, and 10 million reads (for each of reads 1 and 2) per sample were generated (Supplementary Table S3). Reads were quality filtered and trimmed using PRINSEQ (Schmieder and Edwards, 2011) with parameters given in Supplementary Table S4, and final read counts and metagenome sizes are given in Supplementary Table S3. Although exact duplicates and reverse-complement exact duplicates were removed, we tested the effect of leaving in these duplicates, and it increased the number of reads retained by only 0.1–0.2%. Raw fastq files have been submitted to the NCBI BioSample database with accession numbers PRJNA289734 (BioProject) and SRR2102994–

SRR2103038 (SRA). All analyses presented here were carried out on the quality filtered and trimmed reads. Both reads 1 and 2 were analyzed initially; however, unless otherwise indicated, only the results of read 1 are presented here because of a high degree of redundancy between results of reads 1 and 2. Genomic assemblies were built from each sample; these assemblies were used to calculate insert sizes of metagenomic libraries (Supplementary Table S3) but provided limited value for quantitation of taxa and gene ortholog groups. The assemblies did, however, yield contigs belonging to uncharacterized clades, which are the subject of a separate study (Haroon *et al.*, 2016).

Calculation of metagenomic 'response variables' from metagenomic reads

Data tables of merged environmental metadata and response variables are provided in Supplementary Information. Scripts used in the preparation of this manuscript are available at https://github.com/cuttlefishh/papers in the directory red-sea-spatial-series.

Taxonomic composition. The 45 metagenomes were analyzed at the read level for the relative abundance of taxonomic groups using CLARK. CLARK (full mode) (Ounit *et al.*, 2015) and CLARK-S (spaced mode) (Ounit and Lonardi, 2015) were used to classify paired metagenomic reads at species and genus level, respectively, based on a k-mer approach against the NCBI RefSeq database (Release 74). CLARK was run using default parameters but with the high-confidence option, which reports only results with high confidence (assignments with confidence score $\geqslant 0.75$ and gamma value $\geqslant 0.03$), as suggested by the developers. For both species-level and genus-level CLARK results, the column Proportion_All(%) (relative normalized abundance such that each sample sums to 100%) was exported and merged with sample metadata (environmental parameters) using the Python package Pandas. Hierarchically-clustered heatmaps were generated using MetaPhlAn2 utilities (Truong *et al.*, 2015).

In order to specifically capture the diversity within the *Pelagibacter* and *Prochlorococcus* groups in the Red Sea, we used GraftM (https://github.com/geronimp/graftM), which classifies reads based on HMM profiles in concert with a reference phylogeny. HMM profiles of *Pelagibacter* 16S rRNA gene and *Prochlorococcus rpoC1* were generated from forward reads using HMMer v3.1b1 (Eddy, 2011). Reference phylogenies were constructed using MEGA6 (Tamura *et al.*, 2013) from ClustalW alignments (Larkin *et al.*, 2007) of publicly available *Pelagibacter* 16S rRNA gene sequences (Luo *et al.*, 2015) and *Prochlorococcus rpoC1* (DNA-directed RNA polymerase subunit gamma) genes (Shibl *et al.*, 2016). Phylogenies were estimated by maximum-likelihood using the GTR+I+G model of nucleotide evolution, chosen with the Perl script ProteinModelSelection.pl that comes with RAxML (Stamatakis, 2014). GraftM was run with default parameters based on the the built GraftM packages, which are available here: https://github.com/fauziharoon/graftm-packages. Counts were fourth-root transformed.

Gene ortholog group and pathway relative abundance. The 45 metagenomes were analyzed for the relative abundance of gene ortholog groups (KEGG orthologs or KOs) and biochemical pathways (KEGG pathways) using HUMAnN v0.99 (Abubucker *et al.*, 2012) with KEGG release 66.0. First, because the focus of this study was prokaryote genomes, and to increase search speed, reads were recruited to only the prokaryotic fraction of the KEGG genome database, containing all (as of the KEGG release) 1377 prokaryotic genomes (proteomes translated from open reading frames) using a translated search with USEARCH v7.0.1001 (Edgar, 2010) with options –ublast, –accel 0.8 and –evalue 1e–5. The fraction of reads mapped to the KEGG genome database averaged 26.2% (range: 16.2–42.5%) across 45 samples (Supplementary Table S6). Using these results, HUMAnN was run in both standard mode (all taxa merged) and in 'per-organism' mode (option: c_fOrg = True). KO counts and the KEGG pathway counts were normalized to counts per million counts, that is, the number of reads mapped to the KO (or pathway) divided by the sum of all reads mapped in that sample times 1e6, such that all values for a given sample sum to 1 million. Note that KO counts were not normalized to gene size (for example, average length of each KO in KEGG) because this was unnecessary: comparisons of KO relative abundances were to environmental parameters and not to each other, and the multivariate community models used are insensitive to absolute magnitudes.

Statistical analyses

We utilized multivariate statistical techniques to relate an array of environmental parameters to metagenomic response variables: taxon relative abundance, KO relative abundance and pathway relative abundance. All analyses were completed using R v3.1.1 (www.r-project.org) and PERMANOVA+ (Anderson *et al.*, 2008).

Exploratory analyses. Pearson correlations between pairwise combinations of environmental parameters were calculated and displayed as a heatmap. Similarity profile analysis (SIMPROF) was used to identify significant groupings within the KO relative abundance response matrix using the *clustsig* package (http://cran.r-project.org/web/packages/clustsig/index.html). Partitioning around medoids was used to partition the KOs by relative abundance using the *cluster* package v1.15.2 (Kaufman and Rousseeuw, 2005) with Kullback–

Leibler distances (Kullback and Leibler, 1951); 12 clusters were chosen based on minimization of the gap statistic.

Explaining variability using environmental parameters. To quantify the spatial variation (both horizontally and vertically) in the response variable matrices explained by the co-occurring gradients in our environmental parameters, we used a multivariate distance-based linear model (DistLM) (McArdle and Anderson, 2001). Eight environmental parameters were considered: temperature, salinity, dissolved oxygen, chlorophyll, turbidity, nitrate, phosphate and silicate. These parameters were normalized and fitted in a conditional manner to each response variable matrix using step-wise selection and 9999 permutations of the residuals under a reduced model. Model selection was based on Akaike's information criterion with a second-order bias correction applied (AICc) (Hurvich and Tsai, 1989). The best-fit model (the one that balanced performance with parsimony) was then visualized using distance-based redundancy analysis (McArdle and Anderson, 2001) in order to identify the directionality of the correlations between the response variable matrix and the environmental parameters. Variation explained by all parameters combined was calculated by forcing all parameters to be included in the final model.

Visualization of metagenome–environment relationships. Pairwise relationships between environmental parameters and KO relative abundance plus other metagenomic response variables were visualized using scatter plots, available using the Bokeh-based HTML files in Supplementary Information. Environmental parameters and metagenomic response variables were visualized in the 3D volume of the Red Sea using the ỳili Toolbox, also available in Supplementary Information. KOs having strong correlations with environmental parameters were visualized with canonical correspondence analysis (CCA) using the vegan 2.3-1 package, implemented according to Legendre and Legendre (2012). For clarity, only KOs with abundance in the top half and variance in the top tenth of all KOs were visualized by CCA (Supplementary Methods).

Results and discussion

Overview of Red Sea metagenomic dataset and analysis
To measure covariation of microbial diversity with oceanic gradients, we sampled a north–south transect of the Red Sea at eight stations (Supplementary Table S1), sampling six depths from the surface to 500 m (Figure 1a), totaling 45 samples. Concurrent with microbial sampling we measured temperature, salinity, dissolved oxygen, chlorophyll *a*, turbidity, nitrate, phosphate and silicate (values in Supplementary Table S2, covariance matrix in Figure 2). The microbial size fraction (0.1–1.2 μm) was sequenced at 10M reads per sample with 93-bp paired reads (Supplementary Table S3). From the metagenomic reads, we calculated five metagenomic response variables: genus-level taxon relative abundance, species-level taxon relative abundance, gene ortholog group (KEGG Orthology or KO) relative abundance, the KEGG pathway coverage and the KEGG pathway relative abundance. Of the 1738 taxa, 5775 KOs and 162 pathways detected in the metagenomes, many exhibited ecologically meaningful correlations with environmental parameters. As an example, the inverse relationship between phosphate concentration and relative abundance of phosphate-acquisition gene *pstS* (K02040) is shown in Figure 1. Samples generally grouped by depth, as indicated by hierarchical clustering of samples based on all taxa (Figure 3) and KOs (Supplementary Figure S1), and by abundance patterns of individual taxa and KOs (Figures 1b and 4; Supplementary Information).

The acquired set of metagenomic response variables and environmental parameters allowed us to assess the predictive power of environmental parameters at multiple levels of microbial genotype. We tested how much variation in genus-level taxonomy, KO relative abundance and pathway relative abundance could be explained using a small number of environmental parameters. Distance-based multivariate linear models (DistLM) and redundancy analysis were used, balancing parsimony and performance (using AICc) to derive an optimal model for explaining variation in each response variable Figure 5). We acknowledge that the analyses presented here, by necessity, are constrained by the databases available for assigning taxonomy and KOs and the available mappings of KOs to pathways.

Variation in metagenomic diversity metrics explained by environmental parameters
We first asked which environmental parameters explained the most variation in both taxonomic and functional diversity metrics, and we looked for differences in total variation explained. Environmental parameters explained similar amounts of variation in the various metrics used (Figure 5a). Total variation explained using all available environmental parameters was only marginally higher for KO relative abundance (79.0%) than for pathway relative abundance (77.0%) and genus-level taxon relative abundance (75.1%). Variation explained was similar even at greater phylogenetic resolution within two important marine microbial groups, the autotroph *Prochlorococcus* and the heterotroph *Pelagibacter* (SAR11 clade), which are the two most abundant genera across our dataset (Figure 3). At ecotype-level taxonomy (*Prochlorococcus* 'ecotypes' and *Pelagibacter* 'subclades') and genus-level KO abundance, the percent variation explained was similar to the community as a whole (Figure 5b).

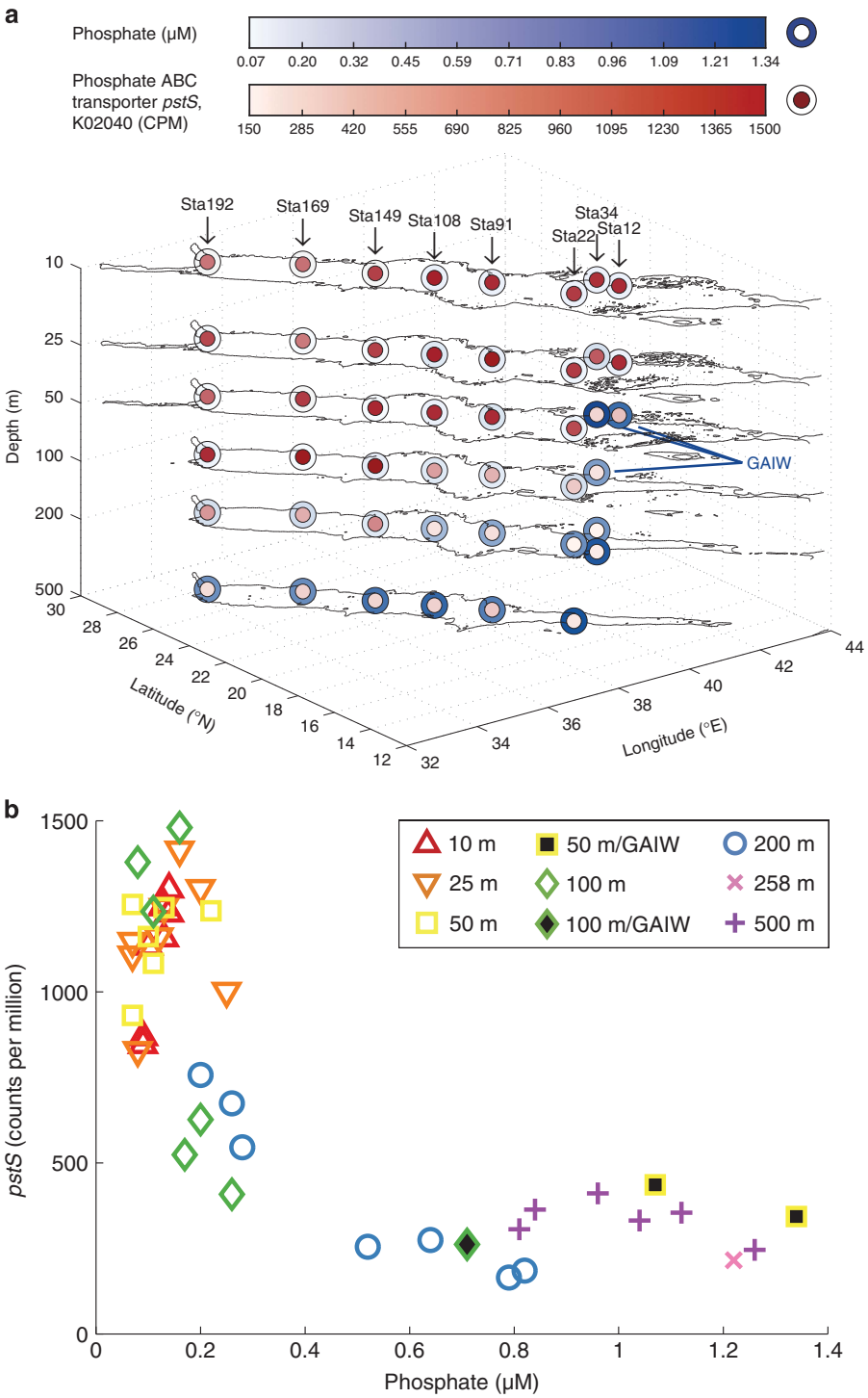

Figure 1 Covariation of gene ortholog group abundance and environmental parameters in the water column. (**a**) 3D contour map of the Red Sea, with outlines (isobaths) showing boundaries of the Red Sea at sampling depths, and samples colored by phosphate concentration (outer circle) and relative abundance of gene ortholog group (KO) for phosphate ABC transporter *pstS* (inner circle). (**b**) Scatter plot of KO relative abundance versus phosphate concentration. Samples taken within the foreign water mass GAIW are indicated. KO relative abundance is given in units of counts per million of total KO counts in each sample (that is, all KOs sum to 1 million in each sample).

Overall, environmental parameters explained more variation in our dataset than in other microbial ecosystems where this has been tested. For example, in a similarly sized dataset on reef-associated microbes, the best parameter explained only 15% of taxonomic variation and 18% of metabolic variation (Kelly *et al.*, 2014). Variations in water column microbial communities appear easier to predict. In an English Channel time-series, day length explained over 65% of variance in taxonomic diversity (Gilbert *et al.*, 2011). The better performance of water column data could be because the

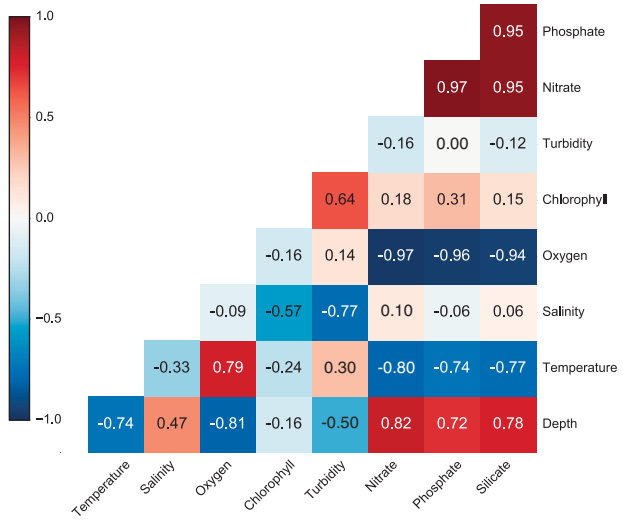

Figure 2 Pearson correlations between environmental parameters shown as a colored covariance matrix. A Pearson's *r* value of 1 (red) indicates a total positive correlation, a value of −1 (blue) indicates a total negative correlation, and a value of 0 (white) indicates no correlation.

open ocean is not patchy but well mixed and stably stratified by depth into layers, especially in the Red Sea in late summer. Relative to open-water samples, the increased complexity of the response matrix in reef-associated samples resulting from microhabitats and higher diversity likely reduces model performance.

Temperature explained the most variation in each of the response variables; this was followed in each case by nitrate (second) and then chlorophyll (third) for the functional response variables and salinity (third) for genus-level taxonomy (Figure 5a; Supplementary Figure S2). Although nitrate and phosphate ($r=0.97$) and silicate ($r=0.95$ with nitrate and phosphate) were very highly correlated (Figure 2), nitrate was consistently ranked higher (explaining more variation) than phosphate in the optimal model, and silicate was not implicated (Figure 5a). Across the whole dataset, temperature explained more variation than oxygen in every response variable. Although temperatue and oxygen were correlated ($r=0.79$), oxygen was never part of the optimal model. Temperature has been identified as a key predictor of microbial diversity in the ocean by other studies (Johnson *et al.*, 2006; Sunagawa *et al.*, 2015). Specifically, Sunagawa *et al.* (2015) showed that temperature is a better predictor of taxonomic composition than is oxygen. Here we show that the same is true for gene functional composition (KOs): the absence of oxygen in any optimal model suggests that temperature is a stronger predictor (and possibly also driver) of microbial diversity than oxygen.

Nitrate (measured as nitrate+nitrite) and phosphate both formed part of the optimal model for each functional response variable, with nitrate always explaining slightly more variation than phosphate. This finding hints at the relative selective pressures these two key nutrients exert. The idea that limitation of a given nutrient leads to the gain of genes for uptake and assimilation of that nutrient is supported by numerous studies (Coleman and Chisholm, 2010; Kelly *et al.*, 2013; Thompson *et al.*, 2013). Here, we extend that idea to the quantitative explanatory power of the nutrient's concentration for predicting KO relative abundance. The predictive power of nitrate relative to phosphate in our genetic results may indicate that nitrogen (N) is relatively more limiting than phosphorus (P) in the Red Sea. Limited data exist on this topic, but N:P ratios of 0.3–5 (well below the Redfield ratio of 16, the atomic ratio of N to P in phytoplankton (Redfield, 1958)) in the Gulf of Aqaba (Lindell *et al.*, 2005) and a high frequency of N-acquisition genes in a Red Sea surface metagenome relative to the Atlantic ocean (Thompson *et al.*, 2013) suggest N limitation; however, in the northern Gulf of Aqaba, a P-stress response and lack of N-stress *ntcA* response in Red Sea cyanobacteria supports the opposite conclusion (Post, 2005). Nevertheless, our own nutrient measurements from this cruise show that the N:P ratio (calculated here as the ratio of nitrate+nitrite to phosphate) in surface waters was 2, whereas a prototypical ratio of 16 was observed in deeper waters (Supplementary Figure S3), possibly due to remineralization of N from phytoplankton at depth. Regarding nitrate, it is interesting that for *Pelagibacter* KO relative abundance, nitrate (59.1%) not temperature explained the most variation. N limitation has strong effects on the transcriptional response of *Pelagibacter* in culture, with genes for assimilation of organic sources of N up-regulated under N stress (Smith *et al.*, 2013). However, none of the differentially expressed genes identified by Smith *et al.* (2013) covaried strongly with nitrate in our dataset (reads from corresponding KOs assigned to *Pelagibacter*). Thus, the nature of a potential selective force of this putative N limitation on *Pelagibacter* gene content remains a mystery.

Chlorophyll has a non-monotonic relationship with depth, unlike the other environmental parameters analyzed here, which are either low at the surface and high at depth (salinity, phosphate, nitrate, silicate) or high at the surface and low at depth (temperature, oxygen, turbidity). Chlorophyll peaks below the surface mixed layer at the deep chlorophyll maximum (100 m in the Red Sea, Supplementary Table S1), due to a confluence of sunlight from above, nutrients from below, and the tendency of deeper phytoplankton to possess higher chlorophyll per cell. Because chlorophyll is effectively orthogonal to other environmental parameters, it should not be unexpected that it has significant explanatory power, and that chlorophyll and temperature (a key depth-dependent parameter) together could explain much of the genetic variation.

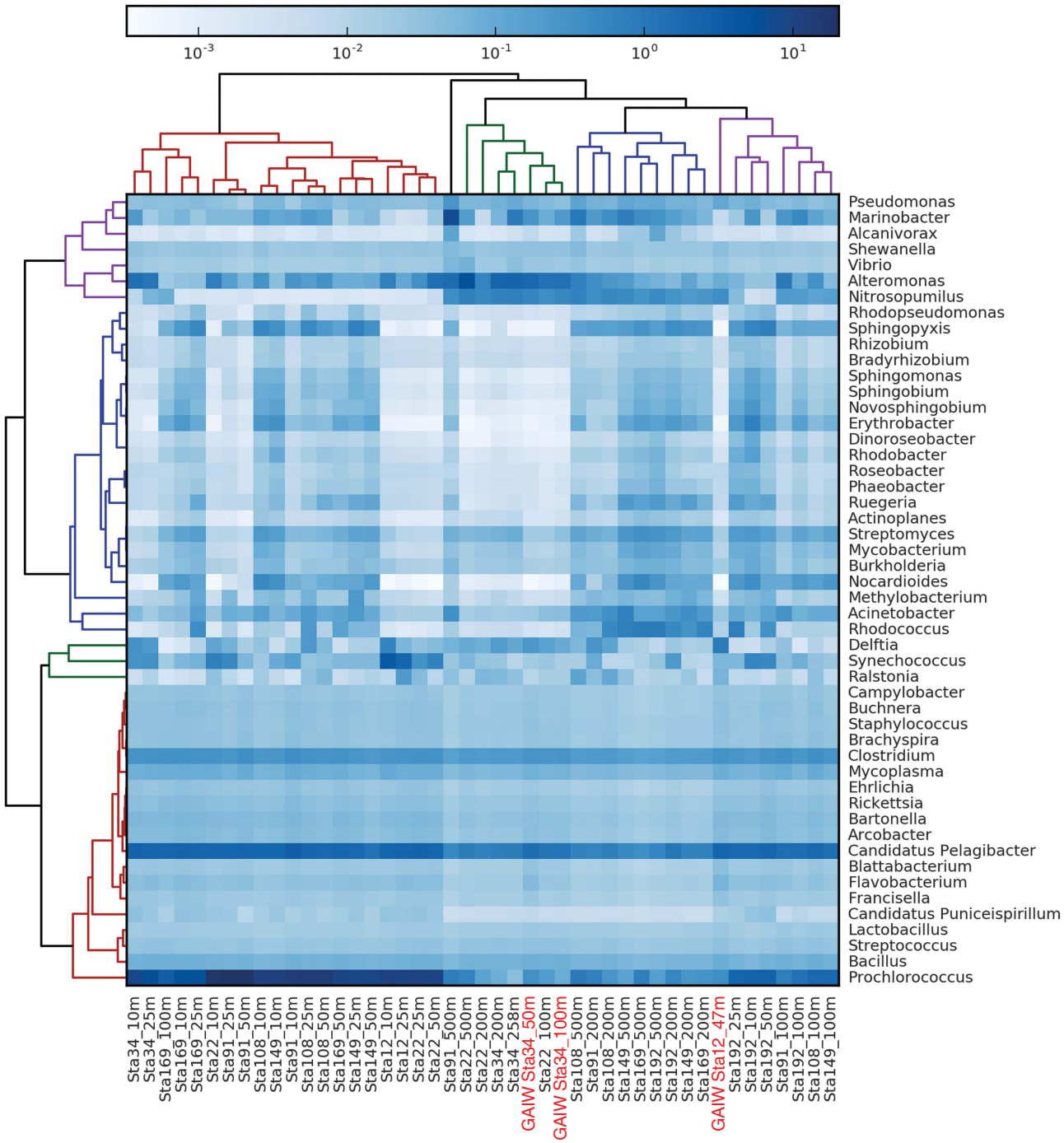

Figure 3 Relative abundance of genera across metagenomes displayed as a hierarchically-clustered heatmap, with clustering of samples by Bray–Curtis distance (top dendrogram) and clustering of taxa by correlation between samples (left dendrogram), with branch colors indicating major clusters. GAIW sample labels are colored red. The top 50 most abundant genera are shown. Relative abundances of all 683 genera detected for each sample sum to 100. Genus-level taxonomy was calculated based on k-mer frequency in comparison with the NCBI RefSeq database (see Materials and methods section).

Comparison with a foreign water mass

We next asked whether the ability to predict metagenomic response variation from environmental parameters was sufficiently robust to extend to alternate water masses. Fortuitously, the Red Sea experiences a water influx each summer from the Indian Ocean, called the GAIW, which was captured in three of our samples. The GAIW brings cooler, less saline, oxygen-rich, nutrient-rich water from the Gulf of Aden (Churchill *et al*., 2014). The three GAIW samples were clearly distinct from their neighboring samples in the temperature–salinity (T–S) profile (Supplementary Figure S4A) and Red Sea water column (Supplementary Figure S4B). The properties

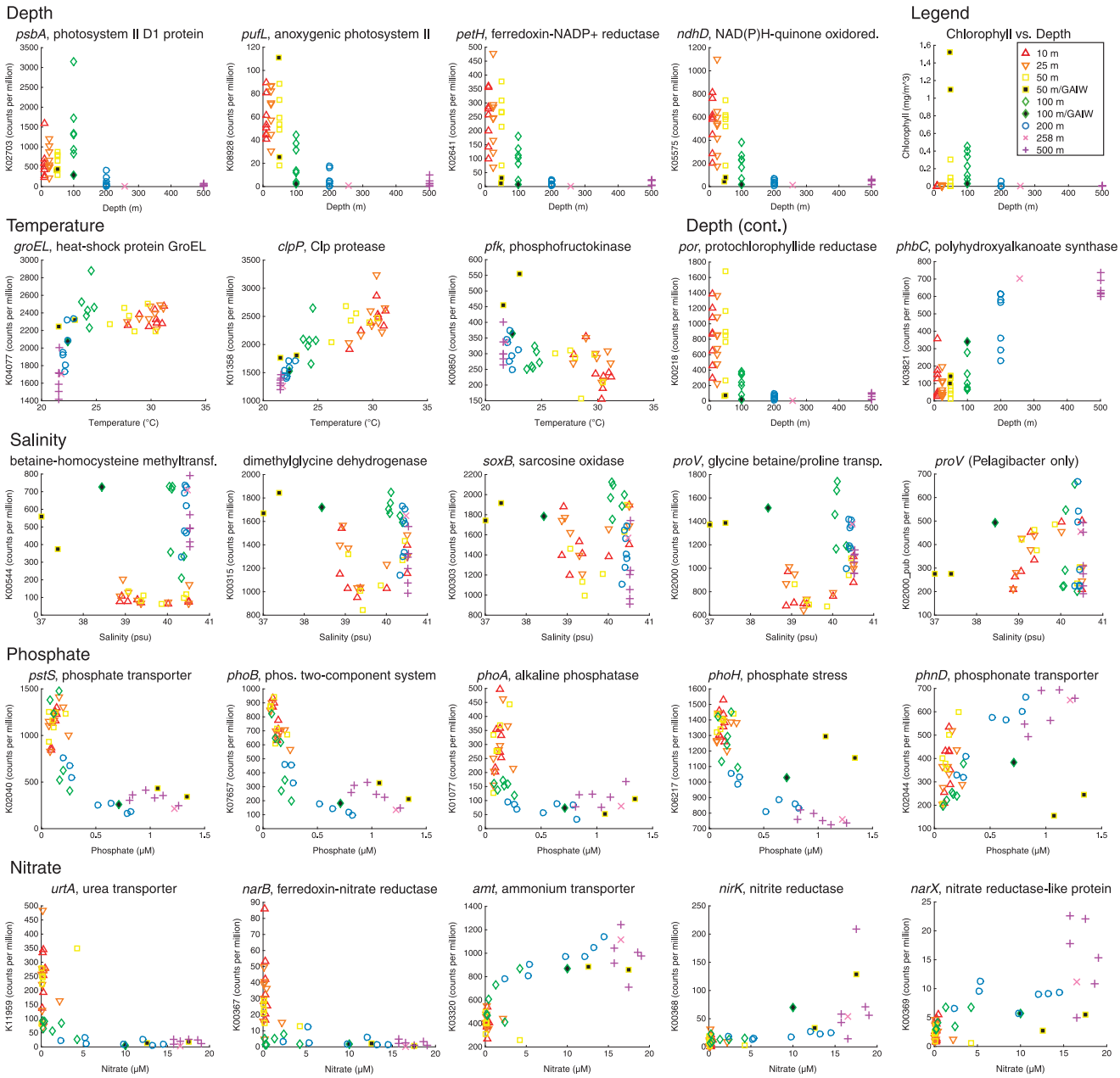

Figure 4 Covariation of select KOs with environmental parameters. KO relative abundance is given in units of counts per million of total KO counts in each sample (that is, all KOs sum to 1 million in each sample).

of GAIW samples resembled those of deeper samples in the native Red Sea water mass; the GAIW samples, which were from 50 m to 100 m depth, had markedly different environmental parameters from non-GAIW samples from 50m to 100 m. We were curious if our multivariate community models would be able to highlight any differences between response variables in model performance across different water masses.

Considering the distance-based redundancy analysis (Figure 5a; Supplementary Figure S2), all of the functional metrics placed the GAIW samples amidst the native Red Sea samples, though clustering with deeper samples, owing to the lower temperature and higher nutrients of the GAIW samples. The taxonomic metrics, however, placed the GAIW samples either far apart from the other samples (genus level) or with much deeper samples than expected even based on physicochemical properties (species-level), driven by the high nitrate and low temperature and salinity of the GAIW samples relative to the non-GAIW samples (Figure 5a). These results suggest that environmental covariation patterns of taxonomy are less conserved across water masses (that is, different combinations of environmental parameters) than are environmental covariation patterns of functional metrics.

Supporting the idea that functional covariation with environmental parameters is conserved across

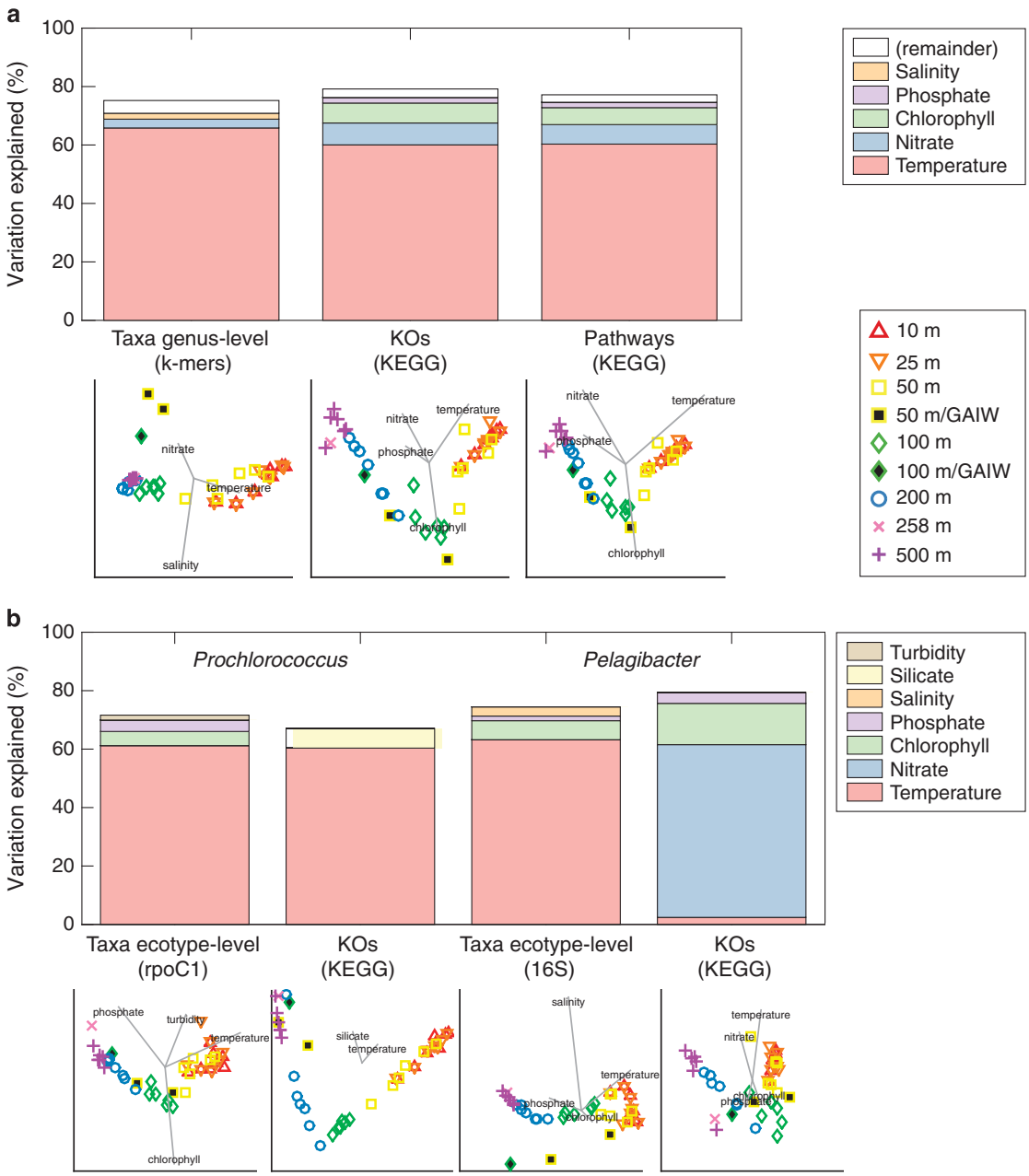

Figure 5 Maximization of linear relations between environmental parameters and metagenomic response variables using a distance-based multivariate linear model and distance-based redundancy analysis for (**a**) the whole dataset and (**b**) genera *Prochlorococcus* and *Pelagibacter* (see Materials and methods section). Percent variation explained by each parameter is shown as a bar graph. The optimal model using AICc to balance performance and parsimony is shown for both (**a**) and (**b**); also shown for (**a**) is the remainder of variation explained by other environmental parameters unused in the optimal model. The distance-based redundancy analysis ordination of the optimal model is shown along distance-based redundancy analysis axes 1 and 2, with stations colored by depth and water mass (GAIW in black).

different water masses, we note anecdotally that for most of the individual environment–KO relationships examined below (Figure 4), GAIW samples followed a similar pattern to the non-GAIW majority. One notable exception was salinity, with the salinity of GAIW much lower than anything in the native Red Sea water mass and the covariation of KOs with salinity very different for GAIW samples compared to non-GAIW samples.

Environmental covariation patterns of individual KOs

We finally turned our attention to the covariation patterns of individual KOs, which partition along the three-dimensional water column in ecologically meaningful ways. Which KOs have the strongest covariation with environmental parameters? Can previously observed patterns between oceans also be observed along gradients within a single sea? Which KOs are implicated in the adaptive response

of microbes to the low nutrients and high irradiance, temperature and salinity of the Red Sea?

We used CCA to identify and visualize correlations between KOs and environmental parameters, with KOs organized by the metabolic pathway (Figure 6; Supplementary Table S9). We note that all KOs were included in the distance-based linear model above, whereas a subset of the most differentially represented and abundant KOs are shown in the CCA (methods); most of the KOs discussed below are visualized in Figure 6. In addition, KOs were ranked by total abundance across all samples (Supplementary Table S7), and KO abundance patterns were clustered using partitioning around medoids into 12 clusters (Supplementary Table S8).

Nutrient acquisition and energy metabolism contributed most of the functional covariation with environmental parameters (Figure 6). The patterns documented here were not exclusively depth-dependent but also captured subtle covariation with gradients along isobaths. The full set of environmental parameters and metagenomic response variables can be visualized interactively in the 3D volume of the Red Sea using web-based tools with files in Supplementary Information. Visualization examples showing the temperature–salinity profile, and temperature in the 3D volume of the Red Sea, are provided in Supplementary Figure S4.

Depth is a spatial parameter that is not 'felt' by microorganisms, except as it relates to pressure, but nevertheless structures virtually all environmental parameters in the water column. Light attenuates with depth, and thus photosynthesis is mostly confined to the upper 200 m of the water column. As expected, KOs for photosynthesis were most abundant in shallow waters and gradually less abundant in deeper waters. This was true for both oxygenic (*psbA*/K02703) and anoxygenic (*pufL*/K08928) photosynthesis, photosynthetic electron transport (*petH*/K02641, *ndhD*/K05575) and pigment biosynthesis (*por*/K00218) (Figure 4). Some heterotrophic bacteria accumulate carbon-rich polymers called polyhydroxyalkanoates when organic carbon is readily available but growth is limited by nutrients (Stubbe *et al.*, 2005). We observed that polyhydroxyalkanoate synthase (*phbC*/K03821) was more abundant in mesopelagic samples than epipelagic samples, consistent with relatively more heterotrophy than phototrophy at depth.

Temperature covaries with depth (warmer at surface, colder at depth), but in the Red Sea southern surface waters are warmer than northern surface

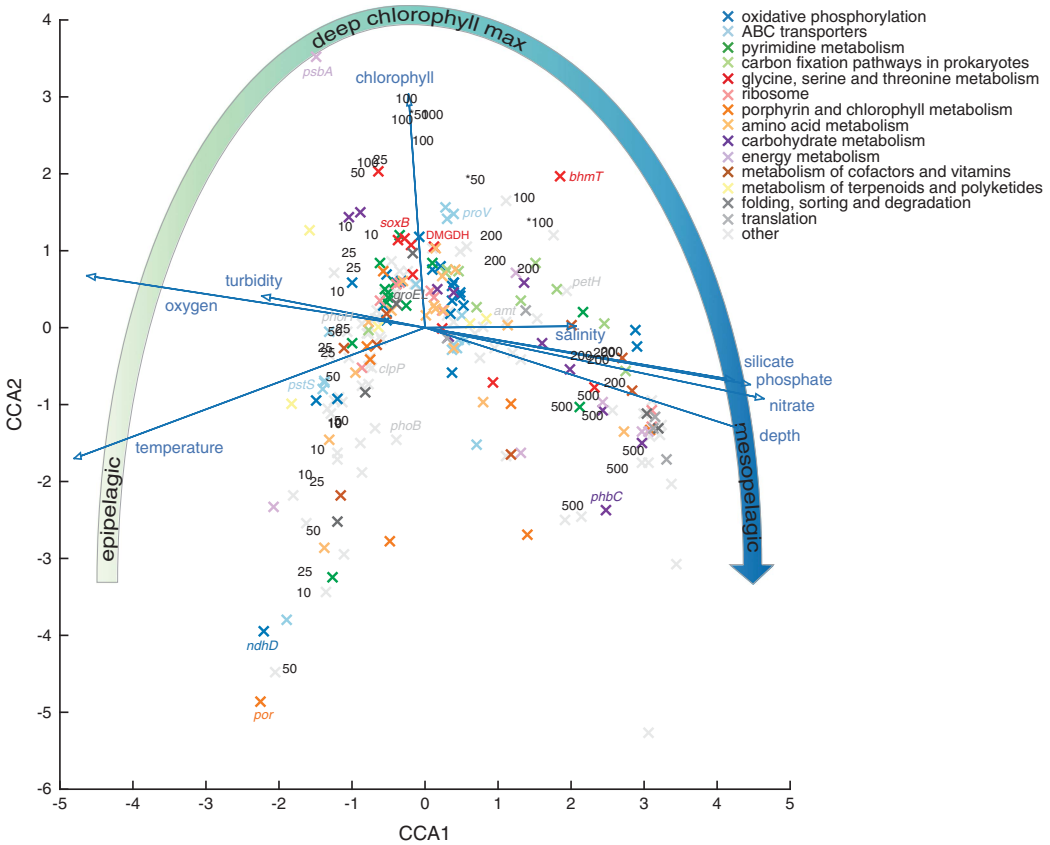

Figure 6 Canonical correspondence analysis of KO relative abundance with environmental parameters. Samples are shown as black numerals indicating depth in meters (GAIW samples marked with asterisk), environmental parameters as dark blue arrows, and KOs colored by the KEGG pathway. For clarity, only KOs were displayed that were found in all samples, with a total count of at least one per thousand counts over all samples, and variance in the top 10% (see Materials and methods section). The large arrow indicates the trend of sample position from surface (epipelagic), to deep chlorophyll maximum, to deep (mesopelagic).

waters, and the GAIW is cooler than surrounding depths in the native Red Sea water mass. We observed that KOs for chaperonins, including heat-shock proteins GroEL/ES (K04077, K04078) and proteases, including Clp protease (*clpP*/K01358), had greater relative abundance in warm (24–32 °C) samples than in cooler (21–23 °C) samples (Figure 4). Both GroEL/ES (Zeilstra-Ryalls *et al.*, 1991) and Clp protease (Zybailov *et al.*, 2009) have important roles in protein folding, which is sensitive to high temperature. The increase in *groEL* relative abundance leveled off above 23 °C, whereas the increase in *clpP* relative abundance increased along the full temperature range from 21 °C to 32 °C. In an opposite trend, glycolysis was relatively more abundant in colder samples than warmer samples, as exemplified by phosphofructokinase (*pfk*/K00850). This is likely related to the relative increase in heterotrophy at depth, as deeper waters tend to be cooler. The most cold, eutrophic samples from the GAIW have the highest relative abundance of *pfk* by far, indicating relatively more heterotrophy in this foreign water mass.

Salinity in the Red Sea is higher at depth and in northern surface waters and lower in southern surface waters and the GAIW. Saline-rich waters of the the Mediterranean and Red Seas were previously shown to have high relative abundance of genes for degradation of osmolytes, in particular recruiting to *Pelagibacter* (Thompson *et al.*, 2013). We put forth a hypothesis that high salinity leads to high production of osmolytes by algae and other organisms, a valuable organic carbon and nutrient source for *Pelagibacter* (Sun *et al.*, 2011), and therefore there is selective pressure to encode osmolyte-degrading enzymes (Thompson *et al.*, 2013). Across the 45 Red Sea metagenomes, KOs for glycine betaine (GBT) transport and degradation (Sun *et al.*, 2011—glycine betaine/proline transporter (*proV*/K02000), betaine-homocysteine *S*-methyltransferase (*bhmT*/K00544), dimethylglycine dehydrogenase (DMGDH/K00315) and sarcosine oxidase (*soxB*/K00303)—were correlated with high chlorophyll and with high or moderate salinity (Figure 6). The shape of covariation of these four KOs was not as clearly dependent on salinity as we expected (Figure 4). As suggested by the CCA plot (Figure 6), both salinity and chlorophyll help explain the relative abundance of GBT transport and degradation KOs. The legend at top-right of Figure 4 indicates chlorophyll *a* fluorescence of the samples as a function of depth. Among the GBT-utilization KOs, samples with either high chlorophyll (green and yellow-black) or high salinity (blue and purple) tended to have the highest abundance. Thus this multifaceted trend is completely consistent with the hypothesis of phototroph (chlorophyll) production of osmolytes in high-salinity waters as a source of reduced carbon and nutrients for heterotrophic bacteria. Regarding the phototrophs responsible for producing GBT, we note that although *Prochlorococcus* are thought to use glucosylglycerate and sucrose as their main osmolytes, some low-light *Prochlorococcus* strains and *Synechococcus* strains are thought to accumulate GBT as well (Scanlan *et al.*, 2009), and these low-light strains are more abundant in the high-chlorophyll samples. Interestingly, the KO patterns with reads assigned to *Pelagibacter* specifically (for example, *proV*/K02000)—one- to two-thirds of the recruited reads for these salinity-related KOs—were similar to the overall KO patterns but more dependent on salinity than chlorophyll (Figure 4).

Phosphate and nitrate are both low in Red Sea surface waters but higher at depth and in the GAIW (for example, phosphate shown in Figure 1a). Several studies have shown that genes for nutrient acquisition are enriched in waters limited for those nutrients, for example, phosphate acquisition in the low-phosphate Mediterranean and Sargasso Seas (Coleman and Chisholm, 2010; Kelly *et al.*, 2013; Thompson *et al.*, 2013). Across the gradients of the Red Sea, numerous KOs for nutrient transport and assimilation were differentially distributed between nutrient-poor (surface and non-GAIW) and nutrient-rich (deep and GAIW) samples. Although depth was a major factor underlying the covariation observed here, we also detected more subtle differences along gradients within isobaths, as well as more striking differences between GAIW and non-GAIW samples at the same depth. This is to our knowledge the first demonstration of differential abundance patterns of nutrient-acquisition genes on such a small scale, not between disparate oceans but across environmental gradients within a single sea.

Phosphate-acquisition and phosphate-response KOs were enriched in low-phosphate samples (Figure 4), including phosphate ABC transporter (*pstS*/K02040), phosphate two-component system PhoBR (K07657, K07636), alkaline phosphatase (*phoA*/K01077) and phosphate stress-response protein PhoH (K06217). Trends were observed even within isothermal samples binned in two-degree increments, both for cooler isotherms with a wide range of phosphate concentrations, and for warmer isotherms with a narrow and low range of phosphate concentrations, for example, *phoB*/K07657 (Figure S5). Phosphonate-acquisition genes, in an opposite pattern, were enriched in high-phosphate (and low-chlorophyll) samples, as exemplified by the phosphonate ABC transporter (*phnD*/K02044). Phosphonate utilization genes (*phn*) are abundant in proteobacteria such as *Pelagibacter* (Villarreal-Chiu *et al.*, 2012), and are enriched in deeper waters of the Sargasso Sea (Martinez *et al.*, 2010) and generally in low-P waters (Coleman and Chisholm, 2010; Feingersch *et al.*, 2010). Although phosphonate-acquisition genes are found in some *Prochlorococcus* in the environment (Feingersch *et al.*, 2012), genomes and transcriptional responses of cultured strains (Martiny *et al.*, 2006) suggest that inorganic phosphate is the major P source for *Prochlorococcus*. Therefore, in addition to ecotype-level genome

variability tuned to ambient concentrations of phosphate and phosphonate (Martiny *et al.*, 2006), different distributions of phosphate- and phosphonate-acquisition genes along the water column are likely also due to genus-level differences in taxonomic composition (and therefore gene content) along the water column, for example, phosphate-utilizing *Prochlorococcus* in the epipelagic and phosphonate-utilizing *Pelagibacter* in the mesopelagic. Indeed, many of the low nutrient-associated KOs such as phosphate and urea transporters had very similar abundance patterns to KOs typical of a phototrophic bacterium like *Prochlorococcus*: photosystems and photosynthetic electron transport, chlorophyll binding proteins, the Calvin cycle, and transport and chelation of metal cofactors essential for photosynthesis (partitioning around medoids cluster 8, Supplementary Table S8).

Nitrogen-acquisition KOs were differentially distributed with respect to nitrate concentration and, like with phosphorus, also followed one of two opposite patterns (Figure 4), which were also observed within isotherms (Supplementary Figure S5). KOs for urea transport (*urtA*/K11959) and assimilatory ferredoxin-nitrate reductase (*narB*/K00367) were enriched in low-nitrate relative to high-nitrate samples. Conversely, KOs for ammonium transport (*amt*/K03320), nitrite reductase (*nirK*/K00368) and nitrate reductase-like protein (*narX*/K00369) were enriched in high-nitrate relative to low-nitrate samples, with the shift to high abundance occurring at 5 μM for *amt* and 15 μM for *nirK* and *narX*. Opposite of *narB*, *narX* was most abundant in the mesopelagic, where oxygen was low (1 ml/l at 500 m); this is consistent with a putative nitrate reductase fusion gene (*narX*) being upregulated under anaerobic conditions in *Mycobacterium* (Hutter and Dick, 1999). Our measurements of N species besides nitrate were either below detection (nitrite) or unreliable (ammonium), but using global averages, nitrite and ammonium peak around the chlorophyll maximum and nitracline (where nitrate increases most rapidly) and then decrease through the deep epipelagic and mesopelagic (Gruber, 2008); urea is generally low and patchy through the water column (Remsen, 1971). Abundance patterns of several N-acquisition KOs thus appear to follow the 'low nutrient–high KO' paradigm: nitrate reductase was abundant where nitrate was low (surface), and ammonium transport and nitrite reductase were abundant where ammonium and nitrite were low (mesopelagic). Urea transport, if it follows the same paradigm, indicates that urea in the surface of the Red Sea is very low relative to in deeper waters.

Conclusions

We have analyzed a 3D array of marine metagenomes across environmental gradients in the Red Sea, showing that three-quarters of taxonomic and functional variation could be explained by temperature, nitrate and chlorophyll. Covariation patterns with environmental parameters were largely conserved across water masses, notably more so for gene orthologs and pathways than for taxonomic groups. Individual patterns of KO covariation with environmental parameters revealed protein folding functions highly correlated with temperature, osmolyte degradation functions correlated with salinity and chlorophyll, and acquisition functions of nitrogen and phosphorus species anti-correlated with concentrations of their respective species. Subtle trends shown here across isobathic and isothermal gradients have hitherto been observed only between distant and disparate oceans. It is expected that this high-resolution marine metagenomic map of the Red Sea, accessible using interactive visualization tools, will serve as an important resource for marine microbiology and modeling.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We thank chief scientist Amy Bower, co-chief scientist Yasser Abualnaja, Leah Trafford, Dan McCorkle and other scientists from the Woods Hole Oceanographic Institution, the captain and crew of the R/V *Aegaeo* and the Hellenic Center for Marine Research and Red Sea Research Center Director James Luyten for their help on the 2011 KAUST (King Abdullah University of Science and Technology) Red Sea Expedition. Assistance with DNA extraction was provided by Matt Cahill, David Ngugi and Francisco Acosta Espinosa. Bioinformatics assistance was provided by Mamoon Rashid and James Morton. Statistics assistance was provided by Mikyoung Jun, Myoungji Lee, Yoan Eynaud and James Morton. We thank Jon Sanders, Jenan Kharbush and Lihini Aluwihare for helpful comments on the manuscript. We also thank colleagues who suggested KOs hypothesized to have interesting ecological patterns: Paul Berube, Yue Guan, Laura Villanueva, Francisco Rodríguez-Valera, Nathan Ahlgren, Zhenfeng Liu, Francy Jiménez and Ulrike Pfreundt. This work was funded in part by a postdoctoral fellowship to LRT from the Saudi Basic Industries Corporation (SABIC).

Data Deposition

Sequence data have been submitted to the NCBI BioSample database with accession numbers PRJNA289734 (BioProject) and SRR2102994–SRR2103038 (SRA).

Author contributions

LRT planned the study, organized the cruise, collected samples, curated physical and chemical data, extracted DNA, processed sequence data, generated graphics and tables and wrote the paper. GJW planned and executed statistical analyses and

wrote the paper. MFH tested and ran taxonomic analyses and wrote the paper. AS collected samples and extracted DNA. PL ran metabolite prediction analysis. JS generated interactive visualizations. RK provided analytical input and wrote the paper. US planned the study, organized the cruise and wrote the paper.

References

Abubucker SS, Segata NN, Goll JJ, Schubert AAM, Izard JJ, Cantarel BBL *et al.* (2012). Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput Biol* **8**: e1002358–e1002358.

Anderson MJ. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecol* **26**: 32–46.

Anderson MJ, Gorley RN, Clarke KR. (2008). *PERMANOVA+ for PRIMER: Guide to Software and Statistical Methods*. PRIMER-E Ltd: Plymouth, UK.

Barberán A, Fernández-Guerra A, Bohannan MBJ, Casamayor EO. (2012). Exploration of community traits as ecological markers in microbial metagenomes. *Mol Ecol* **21**: 1909–1917.

Churchill JH, Bower AS, McCorkle DC, Abualnaja Y. (2014). The transport of nutrient-rich Indian ocean water through the Red Sea and into coastal reef systems. *J Mar Res* **72**: 165–181.

Coleman ML, Chisholm SW. (2010). Ecosystem-specific selection pressures revealed through comparative population genomics. *Proc Natl Acad Sci USA* **107**: 18634–18639.

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

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM *et al.* (2008). Functional metagenomic profiling of nine biomes. *Nature* **452**: 629–632.

Eddy SR. (2011). Accelerated profile HMM searches. *PLoS Comput Biol* **7**: e1002195.

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

Edwards FJ. (1987). Climate and oceanography. In: Edwards AJ, Head SM. (eds). *Key Environments: Red Sea*. Pergamon: Oxford, pp 45–68.

Feingersch R, Suzuki MT, Shmoish M, Sharon I, Sabehi G, Partensky F *et al.* (2010). Microbial community genomics in eastern Mediterranean Sea surface waters. *ISME J* **4**: 78–87.

Feingersch RR, Philosof AA, Mejuch TT, Glaser FF, Alalouf OO, Shoham YY *et al.* (2012). Potential for phosphite and phosphonate utilization by *Prochlorococcus*. *ISME J* **6**: 827–834.

Ghai R, Martin-Cuadrado A-B, Molto AG, Heredia IG, Cabrera R, Martin J *et al.* (2010). Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing. *ISME J* **4**: 1154–1166.

Gianoulis TA, Raes J, Patel PV, Bjornson R, Korbel JO, Letunic I *et al.* (2009). Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc Natl Acad Sci USA* **106**: 1374–1379.

Gilbert JA, Steele JA, Caporaso JG, Steinbrück L, Reeder J, Temperton B *et al.* (2011). Defining seasonal marine microbial community dynamics. *ISME J* **6**: 298–308.

Gruber N. (2008). The marine nitrogen cycle: overview and challenges. In: Capone DG, Bronk DA *et al.* (eds). *Nitrogen in the Marine Environment*. Elsevier: Oxford.

Haroon MF, Thompson LR, Parks DH, HugenholtzP, Stingl U. (2016). A catalogue of 136 microbial draft genomes from Red Sea metagenomes. *Sci Data*; e-pub ahead of print 5 July 2016; doi:10.1038/sdata.2016.50.

Hurvich CM, Tsai C-L. (1989). Regression and time series model selection in small samples. *Biometrika* **76**: 297–307.

Hutter B, Dick T. (1999). Up-regulation of narX, encoding a putative fused nitrate reductase in anaerobic dormant *Mycobacterium bovis* BCG. *FEMS Microbiol Lett* **178**: 63–69.

Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward SEM, Chisholm SW. (2006). Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science* **311**: 1737–1740.

Kaufman L, Rousseeuw PJ. (2005). *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ: Wiley-Interscience.

Kelly L, Ding H, Huang KH, Osburne MS, Chisholm SW. (2013). Genetic diversity in cultured and wild marine cyanomyoviruses reveals phosphorus stress as a strong selective agent. *ISME J* **7**: 1827–1841.

Kelly LW, Williams GJ, Barott KL, Carlson CA, Dinsdale EA, Edwards RA *et al.* (2014). Local genomic adaptation of coral reef-associated microbiomes to gradients of natural variability and anthropogenic stressors. *Proc Natl Acad Sci USA* **111**: 10227–10232.

Kullback S, Leibler RA. (1951). On information and sufficiency. *Ann Math Statist* **22**: 79–86.

Larkin MA *et al.* (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947–2948.

Legendre P, Legendre L. (2012). *Numerical Ecology*. Amsterdam: Elsevier.

Lindell D, Penno S, Al-Qutob M, David E, Rivlin T, Lazar B *et al.* (2005). Expression of the nitrogen stress response gene ntcA reveals nitrogen-sufficient Synechococcus populations in the oligotrophic northern Red Sea. *Limnol Oceanogr* **50**: 1932–1944.

Luo H, Thompson LR, Stingl U, Hughes AL. (2015). Selection maintains low genomic GC content in marine SAR11 lineages. *Mol Biol Evol* **32**: 2738–2748.

Martinez A, Tyson GW, DeLong EF. (2010). Widespread known and novel phosphonate utilization pathways in marine bacteria revealed by functional screening and metagenomic analyses. *Environ Microbiol* **12**: 222–238.

Martiny AC, Coleman ML, Chisholm SW. (2006). Phosphate acquisition genes in *Prochlorococcus* ecotypes: Evidence for genome-wide adaptation. *Proc Natl Acad Sci USA* **103**: 12552–12557.

McArdle BH, Anderson MJ. (2001). Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology* **82**: 290–297.

Ngugi DK, Antunes A, Brune A, Stingl U. (2012). Biogeography of pelagic bacterioplankton across an antagonistic temperature-salinity gradient in the Red Sea. *Mol Ecol* **21**: 388–405.

Ngugi DK, Stingl U. (2012). Combined analyses of the ITS loci and the corresponding 16 S rRNA genes reveal high micro- and macrodiversity of SAR11 populations in the Red Sea. *PLoS ONE* **7**: e50274.

Ounit R, Lonardi S. (2015). Higher classification accuracy of short metagenomic reads by discriminative spaced k-mers. In: *Algorithms in Bioinformatics. 15th International Workshop, WABI 2015*. Springer: Berlin, pp 286–295.

Ounit R, Wanamaker S, Close TJ, Lonardi S. (2015). CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics* **16**: 236.

Post AF. (2005). Nutrient limitation of marine cyanobacteria. In: Huisman J, Matthijs HCP *et al.* (eds). *Harmful Cyanobacteria*. Springer: Dordrecht, pp 87–107.

Raes J, Letunic I, Yamada T, Jensen LJ, Bork P. (2011). Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. *Mol Syst Biol* **7**: 473–473.

Redfield AC. (1958). The biological control of chemical factors in the environment. *Am Scientist* **46**: 205–221.

Remsen CC. (1971). The distribution of urea in coastal and oceanic waters. *Limnol Oceanogr* **16**: 732–740.

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.

Scanlan DJ, Ostrowski M, Mazard S, Dufresne A, Garczarek L, Hess WR *et al.* (2009). Ecological genomics of marine picocyanobacteria. *Microbiol Mol Biol Rev* **73**: 249–299.

Schmieder R, Edwards R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**: 863–864.

Shibl AA, Haroon MF, Ngugi DK, Thompson LR, Stingl U. (2016). Distribution of Prochlorococcus Ecotypes in the Red Sea Basin Based on Analyses of rpoC1 Sequences. *Front Mar Sci*; e-pub ahead of print 24 June 2016; doi:10.3389/fmars.2016.00104.

Smith DP, Thrash JC, Nicora CD, Lipton MS, Burnum-Johnson KE, Carini P *et al.* (2013). Proteomic and transcriptomic analyses of 'Candidatus Pelagibacter ubique' describe the first PII-independent response to nitrogen limitation in a free-living Alphaproteobacterium. *mBio* **4**: e00133–12.

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

Stubbe J, Tian J, He A, Sinskey AJ, Lawrence AG, Liu P. (2005). Nontemplate-dependent polymerization processes: polyhydroxyalkanoate synthases as a paradigm. *Annu Rev Biochem* **74**: 433–480.

Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G *et al.* (2015). Ocean plankton. Structure and function of the global ocean microbiome. *Science* **348**: 1261359–1261359.

Sun J, Steindler L, Thrash JC, Halsey KH, Smith DP, Carter AE *et al.* (2011). One carbon metabolism in SAR11 pelagic marine bacteria. *PLoS ONE* **6**: e23973.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol* **30**: 2725–2729.

Thompson LR, Field C, Romanuk T, Ngugi D, Siam R, El Dorry H *et al.* (2013). Patterns of ecological specialization among microbial populations in the Red Sea and diverse oligotrophic marine environments. *Ecol Evol* **3**: 1780–1797.

Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E *et al.* (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods* **12**: 902–903.

Villarreal-Chiu JF, Quinn JP, McGrath JW. (2012). The genes and enzymes of phosphonate metabolism by bacteria, and their distribution in the marine environment. *Front Microbiol* **3**: 1–13.

Zeilstra-Ryalls J, Fayet O, Georgopoulos C. (1991). The universally conserved GroE (Hsp60) chaperonins. *Annu Rev Microbiol* **45**: 301–325.

Zybailov B, Friso G, Kim J, Rudella A, Rodriguez VR, Asakura Y *et al.* (2009). Large scale comparative proteomics of a chloroplast Clp protease mutant reveals folding stress, altered protein homeostasis, and feedback regulation of metabolism. *Mol Cell Proteomics* **8**: 1789–1810.

Supplementary Information accompanies this paper on *The ISME Journal* website (http://www.nature.com/ismej)