Open

The ISME Journal (2012) 6, 1403–1414

www.nature.com/ismej

ORIGINAL ARTICLE

Influence of nutrients and currents on the genomic composition of microbes across an upwelling mosaic

Lisa Zeigler Allen[1,2], Eric E Allen[2,3], Jonathan H Badger[1], John P McCrow[1], Ian T Paulsen[4], Liam DH Elbourne[4], Mathangi Thiagarajan[5], Doug B Rusch[5], Kenneth H Nealson[1], Shannon J Williamson[1], J Craig Venter[1] and Andrew E Allen[1]

[1]*Microbial and Environmental Genomics, J Craig Venter Institute, San Diego, CA, USA;* [2]*Scripps Institution of Oceanography, University of California, San Diego, CA, USA;* [3]*Division of Biological Sciences, University of California, San Diego, CA, USA;* [4]*Chemistry and Biomolecular Sciences, Macquarie University, Sydney, New South Wales, Australia and* [5]*Informatics, J Craig Venter Institute, Rockville, MD, USA*

Metagenomic data sets were generated from samples collected along a coastal to open ocean transect between Southern California Bight and California Current waters during a seasonal upwelling event, providing an opportunity to examine the impact of episodic pulses of cold nutrient-rich water into surface ocean microbial communities. The data set consists of ~5.8 million predicted proteins across seven sites, from three different size classes: 0.1–0.8, 0.8–3.0 and 3.0–200.0 μm. Taxonomic and metabolic analyses suggest that sequences from the 0.1–0.8 μm size class correlated with their position along the upwelling mosaic. However, taxonomic profiles of bacteria from the larger size classes (0.8–200 μm) were less constrained by habitat and characterized by an increase in Cyanobacteria, Bacteroidetes, Flavobacteria and double-stranded DNA viral sequences. Functional annotation of transmembrane proteins indicate that sites comprised of organisms with small genomes have an enrichment of transporters with substrate specificities for amino acids, iron and cadmium, whereas organisms with larger genomes have a higher percentage of transporters for ammonium and potassium. Eukaryotic-type glutamine synthetase (GS) II proteins were identified and taxonomically classified as viral, most closely related to the GSII in Mimivirus, suggesting that marine Mimivirus-like particles may have played a role in the transfer of GSII gene functions. Additionally, a Planctomycete bloom was sampled from one upwelling site providing a rare opportunity to assess the genomic composition of a marine Planctomycete population. The significant correlations observed between genomic properties, community structure and nutrient availability provide insights into habitat-driven dynamics among oligotrophic versus upwelled marine waters adjoining each other spatially.

The ISME Journal (2012) **6,** 1403–1414; doi:10.1038/ismej.2011.201; published online 26 April 2012
Subject Category: integrated genomics and post-genomics approaches in microbial ecology
Keywords: marine; metagenomics; upwelling; California Current

Introduction

Metagenomic studies have been successful in establishing links between environmental conditions, key populations and biological mechanisms mediating biogeochemical and ecological processes (Handelsman, 2004; Venter *et al.*, 2004; Thompson *et al.*, 2005; DeLong *et al.*, 2006). Such approaches also facilitate the discovery and assignment of new metabolic capabilities to taxonomic groups, for example, bacterial proteorhodopsin (Beja *et al.*, 2000) and archaeal ammonia oxidation (Venter *et al.*, 2004).

In collaboration with the California Cooperative Oceanic Fisheries Investigations (CalCOFI), a partnership between the California Department of Fish and Game, the NOAA Fisheries Service and the SIO (Scripps Institution of Oceanography) focusing on the study of the CCE (California Current Ecosystem), samples were collected from the CC (California Current) and SCB (Southern California Bight) in July 2007. The southward CC travels offshore near Point Conception and is the eastern boundary current of the North Pacific Gyre. Wind-driven nearshore seasonal upwelling is typical in the regions around and north of Point Conception in the CCE; however in the SCB, the Southern California Countercurrent (surface) and California Countercurrent (subsurface) travel poleward along

Correspondence: AE Allen, Microbial and Environmental Genomics, J Craig Venter Institute, San Diego, CA, USA.
E-mail: aallen@jcvi.org
Received 20 September 2011; revised 21 November 2011; accepted 28 November 2011; published online 26 April 2012

the coast during summer and inject warm, low nutrient, low chl*a* (chlorophyll *a*) oligotrophic waters into the coastal environment (Hickey, 1979; Huyer, 1983; Hickey *et al.*, 2003). Coastal upwelling, in this region, occurs when north winds blow equatorward parallel to the coastline. Consequently, surface water is pushed away from the coastline causing advection of deep, cold, nutrient-rich waters. These waters contain nutrients such as nitrogen (primarily nitrate), dissolved iron (likely originating from benthic sediment remineralization; Johnson *et al.*, 1999), inorganic carbon, silica and phosphorus that accumulate in deep waters as a result of remineralization of sinking detrital matter. These conditions are associated with increased primary production (Kudela *et al.*, 2006), resulting in the generation of large phytoplankton blooms (Wilkerson *et al.*, 2006), thus changing microbial community structure in surface waters. Adding to the dynamic complexity of this region during seasons of upwelling, eddies stir the SCB waters, particularly the Santa Barbara Channel and the Santa Monica-San Pedro Basin, and can influence phytoplankton distributions, thus altering microbial and pelagic food webs (DiGiacomo and Holt, 2001).

To date, molecular information on bacterioplankton within this region have primarily been derived from taxonomic surveys involving 16S rDNA and the 16S–23S intergenic spacer region (Brown *et al.*, 2005), as well as functional gene surveys specific to particular taxonomic groups, for example, *rpoC* diversity in *Synechococcus* sp. populations (Palenik, 1994; Toledo and Palenik, 1997). Previous reports have also estimated taxonomic groups, such as *Roseobacter* sp., the cyanobacteria *Synechococcus* and *Prochlorococcus*, and the SAR11 clade to be abundant lineages within this community (Brown and Fuhrman, 2005).

To investigate further the effect of abiotic (that is, wind forcing and nutrients) and biotic influences within this dynamic coastal upwelling system, metagenomic sampling and analyses were conducted for three planktonic size fractions, 0.1–0.8, 0.8–3.0 and 3.0–200 μm from seven sites taken along hydrographic and nutrient gradients from near to offshore within the CCE. The environmental sequences generated and analyzed here are a subset of the Global Ocean Sampling (GOS) expedition (Venter *et al.*, 2004; Rusch *et al.*, 2007; Yooseph *et al.*, 2007). To date, GOS data have been comprised of Sanger-derived sequences. Other marine metagenomics studies based on 454-pyrosequencing include data sets obtained from the deep ocean (Eloe *et al.*, 2011; Quaiser *et al.*, 2011), and one from a coastal ocean environment from the North Atlantic (Gilbert *et al.*, 2010). Also, analyses in these studies are based on a single size class. Here, we report the first metagenomics study of coastal waters traversing an upwelling system from multiple size classes. The CalCOFI group collected oceanographic metadata including temperature, salinity, pH, chlorophyll fluorescence, nutrients, oxygen, primary production and phytoplankton and zooplankton biomass and biodiversity estimates (http://www.calcofi.org). These data provided an opportunity for integrated metagenomic examination of biological diversity and microbial ecological processes across an upwelling mosaic in a high productivity marine ecosystem.

Materials and methods

Metagenomic sample collection

Samples were collected on the R/V *New Horizon*, a SIO vessel, as part of the CalCOFI July 2007 annual cruise. Seven stations were sampled along the CalCOFI transect that covered near to offshore hydrographic and nutrient gradients associated with upwelling regions of the CC and SCB. These included the following sites listed by CalCOFI stations with GOS identification in parentheses 87.40 (GS257), 87.80 (GS258), 83.110 (GS259), 83.80 (GS260), 80.90 (GS262), 77.60 (GS263), 77.49 (GS264). At each of the seven stations, pre-filtration with a 200-μm nytex net, followed by serial filtration through 3.0-, 0.8- and 0.1-μm Supor 293 mm disc filters (Pall Life Sciences, Ann Arbor, MI, USA) was performed on ~200 l of surface seawater (~2 m) generating discrete microbial size fractions (Rusch *et al.*, 2007).

See Supplementary Information for details concerning DNA purification, library construction, sequencing, accession numbers, 454-sequencing post-processing—including frameshift correction (Supplementary Table S1; Supplementary Figure S11), and genome equivalent and size estimations (Supplementary Table S8).

16S rDNA prediction and protein prediction

Non-coding RNA and protein sequences were found using the JCVI metagenomics annotation pipeline, similar to the JCVI prokaryotic annotation pipeline (Tanenbaum *et al.*, 2010). Identification of non-coding RNA is accomplished through two processes: (i) tRNAScan-SE (Lowe and Eddy, 1997) and (ii) set of two increasingly stringent BLASTN (*e*-value 0.1–1e^{-4}) searches performed against the JCVI internal reduced-complexity rRNA database. The latter contains a representative sampling of known 5S, 16S, 18S and 23S rRNA sequences downloaded from GenBank. In both cases, the identified rRNA sequences were output in multi-fasta files for subsequent rRNA analysis.

Phylogenetic profiling

For taxonomic classification of predicted metagenomic peptides, the APIS (Automated Phylogenetic Inference System) (Badger *et al.*, 2006; Bowler *et al.*, 2008) pipeline, which automates the process of sequence similarity, alignment and phylogenetic inference for each protein in a given data set, was employed. Each predicted protein was compared

with an in-house curated database (phyloDB), which consists of proteins from a comprehensive set of available reference genomes (consisting of archaea, bacteria, eukarya and viruses), particular expressed sequence tag data sets from important missing phyla (for example, dinoflagellates) and selected viral metagenomes using protein BLAST (basic local alignment search tool; blastp). Briefly, for each predicted metagenomic peptide, full-length sequences with significant blastp hits (*e*-value 10^{-9}) are retrieved and then a multiple sequence alignment is generated using MUSCLE (multiple sequence comparison by log-expectation). From this alignment a neighbor-joining tree was produced using QuickTree (Howe *et al.*, 2002) to determine the phylogenetic placement of the query sequence. If the taxonomic information differs among sequences clading equally well with the query, the classification is limited to the lowest taxonomic rank where there is agreement.

Metabolic profiling via homology to COGs/KOGs
The assignment of environmental sequences to COGs (Clusters of Orthologous Groups of proteins) was performed by querying the NJ phylogenetic trees of all bacterial open reading frames (ORFs) in APIS for the nearest-neighbor reference sequence from phyloDB. Each reference protein from all genomes, expressed sequence tags and viral metagenomics data sets was BLAST against a database of annotated proteins from NCBI (ftp://ftp.ncbi.nih.gov/pub/COG/COG/ and ftp://ftp.ncbi.nih.gov/pub/COG/KOG/). If the nearest neighbor in phyloDB had a COG assignment, it was transferred to the environmental sequence.

Results and discussion

Bacterioplankton diversity across nutrient regimes
A combination of Sanger and 454-sequencing technologies were used to generate metagenomic data derived from three filter sizes (0.1, 0.8 and 3.0 μm) collected at seven stations situated across a grid of CalCOFI transect lines in the summer (July 2007) (Figure 1a; Supplementary Table S2). Figure 1 also shows Chl*a* (Figure 1b) and sea surface temperature (Figure 1c), indicators of coastal upwelling. Based on oceanographic metadata (for example, nutrients, sea surface temperature and pigments), the sites can be classified into two productivity groups, oligotrophic (<0.5 μM nitrate (NO_3^-), >17 °C) and upwelled (>1.0 μM NO_3^-, <16 °C) (Table 1). Although satellite (Figure 1b) and *in-situ* (Table 1) Chl*a* data are not in agreement, it is clear that site 77.60 (GS263) represents an upwelled sample. It is likely that this site would not be

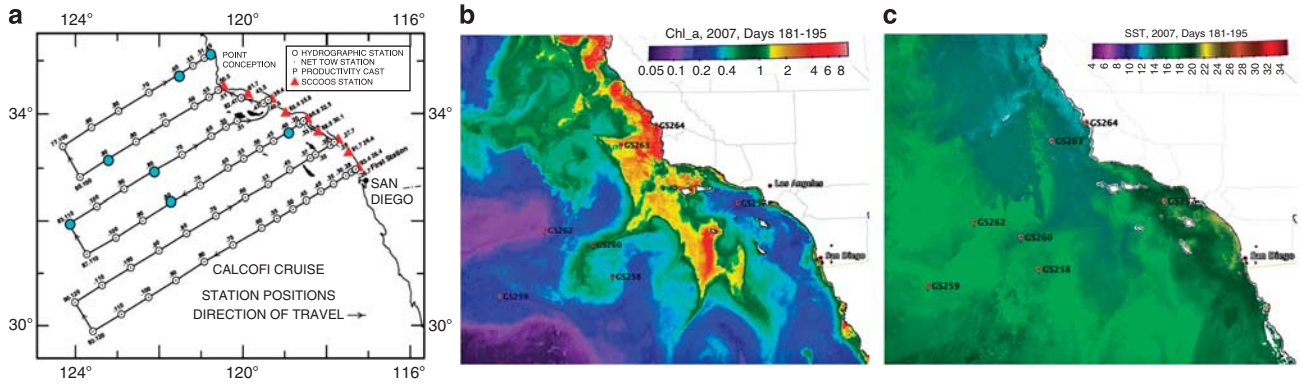

Figure 1 CalCOFI transect map. (**a**) Highlighted in blue are the sites sampled for GOS metagenomics sequencing. Two indicators for upwelling during the sampling period are shown, (**b**) chlorophyll-*a* and (**c**) sea surface temperature, both taken using satellite imagery. The GOS site identifications were superimposed onto maps.

Table 1 Oceanographic metadata collected at sea and reported by CalCOFI

GOS ID	*Station ID*	*Latitude*	*Longitude*	*Date (D/M/Y)*	*Time (PST)*	*Bottom depth (m)*	*Sample depth (m)*	*Temp. (°C)*	*Salinity*	*SiO_3 ($\mu M l^{-1}$)*	*PO_4 ($\mu M l^{-1}$)*	*NO_3 ($\mu M l^{-1}$)*	*NO_2 ($\mu M l^{-1}$)*	*Chla ($\mu g l^{-1}$)*	*Phae ($\mu g l^{-1}$)*
GS257	87.40	33 39.5 N	118 58.4 W	5/7/07	6:14:56	777	2	18.64	33.744	1.2	0.22	0	0.01	0.23	0.05
GS258	87.80	32 19.8 N	121 42.6 W	6/7/07	12:31:33	3954	2	14.71	33.511	0.4	0.42	2.7	0.16	0.89	0.01
GS259	83.110	31 54.4 N	124 10.1 W	7/7/07	15:39:03	4212	2	17.25	33.362	2.1	0.33	0.3	0.01	0.13	0.03
GS260	83.80	32 54.5 N	122 8.5 W	8/7/07	12:20:07	4186	2	14.68	33.155	2.3	0.49	2.9	0.11	1.2	0.08
GS262	80.90	33 9.1 N	123 13.4 W	11/7/07	11:00:17	4234	2	17.4	33.095	2.2	0.31	0	0.01	0.09	0.01
GS263	77.60	34 43.4 N	121 33.2 W	13/7/07	22:23:24	922	2	15.59	33.73	3.6	0.49	1.6	0.09	0.82	0.21
GS264	77.49	35 5.2 N	120 46.5 W	13/7/07	9:13:50	71	2	13.71	33.806	8.9	0.65	5.9	0.18	2.49	0.75

Abbreviations: CalCOFI, California Cooperative Oceanic Fisheries Investigations; chl*a*, chlorophyll *a*; GOS, Global Ocean Sampling; GS, glutamine synthetase.
GOS ID corresponds to the Global Ocean Sampling ID from JCVI and Station ID corresponds to the CalCOFI station ID where the sample was taken. Collection and analysis methods are described in Materials and methods.

exposed to the strong direct effects of coastal upwelling as it resides on the shelf with a water column depth of 922 m; however, 77.60 (GS263) is in a region of relatively low temperature, elevated Chl*a*, NO_3^- and silicate (SiO_3) in surface waters. Also, sites 87.80 (GS258) and 83.80 (GS260) appear to be what we operationally term as 'aged-upwelled' waters meaning they are upwelled waters that have traveled offshore and exhibit lower temperatures, relative to oligotrophic sites, and medial Chl*a* and NO_3^- levels (Kudela *et al.*, 1997). These sites appear distinct in comparison to the other upwelled sites in certain taxonomic and metabolic capacities.

In order to investigate genomic similarity, DNA sequence features were evaluated for associations samples from similar and contrasting temperature and nutrient regimes, which suggests *a priori* taxonomic and metabolic information. ORFs identified on the metagenomic reads from each site and size class were evaluated for GC composition, tri-nucleotide frequency, amino-acid usage and an estimate of the number of bacterial genome equivalents represented. The GC profile of samples 87.40 (GS257), 83.110 (GS259), 80.90 (GS262) characterized by oligotrophic waters, exhibited one peak at ~35%, which is analogous to that of *Candidatus pelagibacter* (~30%; Giovannoni *et al.*, 2005), the dominant organism present based on phylogenomic analysis (Figure 2). The remaining samples originating from upwelled sites 87.80 (GS258), 83.80 (GS260), 77.60 (GS263) and 77.49 (GS264), displayed a bimodal GC distribution with a relatively smaller peak at ~35% and a greater peak at ~45% GC. Principal component analysis of tri-nucleotide frequency and amino-acid usage (Supplementary Figure S1) also showed that sequences from oligotrophic and upwelled sites were well correlated within and between these groups compared with the two aged-upwelled samples, GS258 (87.80) and GS260 (83.80), which were intermediate. The difference in these sites is likely driven primarily by the relatively low diversity of sample GS260 (83.80), which was dominated by a Planctomycete population. Sites from oligotrophic waters also had a greater number of genomes sampled and lower estimated genome size, whereas sites from upwelling regions had lower genome equivalents and greater genome size (Supplementary Figure S2; see Supplementary Information: '*Bacterial genome estimation*'). Small genome size (genome streamlining) and low GC content is a common characteristic for organisms cultured from oligotrophic marine environments (Giovannoni *et al.*, 2005). These data, particularly evident for dominant taxa, suggest a relationship between nutrient availability, taxonomic distribution, DNA composition and estimated genome size.

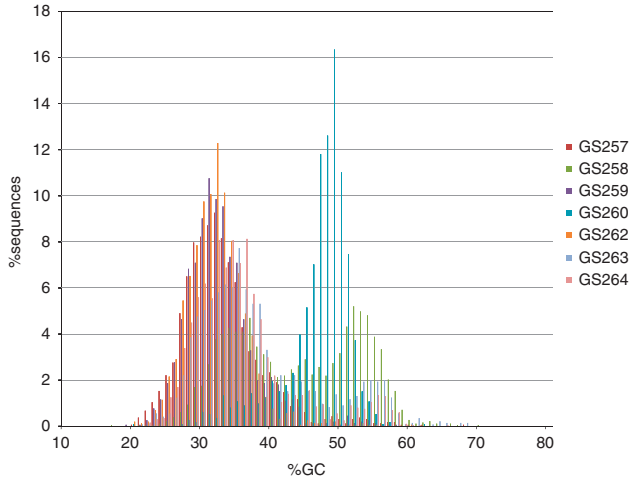

Figure 2 GC composition of bacterial protein coding DNA sequences from the 0.1-μm filter of each site sampled. Percent GC is plotted on the *y*-axis and the % of bacterial sequences on the *x*-axis.

Additionally, genome size increased in the larger size classes, which suggests that discrete bacterial populations are present and most likely comprising a unique niche compared with the 0.1-μm community rather than being an artifact arising from the filtering process (although low levels of artifactual data are possible). Our methods for determination of peptides per genome and average genome size in metagenomic data are similar to Raes *et al.* (2007) in that the approach is based on estimation of the density of genome equivalents. As such, it is an extrapolation of the much more straightforward count of genome equivalents. Accurate identification of the pool of total bacterial peptides is likely the most problematic aspect of this analysis. Here, we use all peptides identified phylogenetically as Kingdom Bacteria. Therefore, our estimates are biased by our inability to identify genes that truly belong to bacteria, but are not phylogenetically identifiable as such, thus the results are likely underestimates. Despite this, we are able to detect clear trends related to the influence of environmental variation on genome size, and taxonomic and functional composition. Importantly, as genome equivalent density is influenced by accurate identification of the total number of bacterial peptides, we use genome equivalent counts (not density) for normalization in future analyses.

Phylogenetic evaluation of all predicted proteins revealed significant differences in bacterioplankton community composition across the upwelling gradient (Figure 3; Supplementary Figure S3). Sequences from the three oligotrophic sites 87.40 (GS257), 83.110 (GS259) and 80.90 (GS262) showed taxonomic representation similar to non-upwelling sites in the GOSI data set (Rusch *et al.*, 2007) and from the Western English Channel (Gilbert *et al.*, 2010), which were typified by a dominance of α-proteobacteria and Cyanobacteria, primarily *Pelagibacter* sp. (average genome size ~1.6 Mbp) and *Prochlorococcus* sp. (average genome size ~1.5 Mbp), respectively. The remaining four sites 87.80 (GS258), 83.80 (GS260) (aged-upwelled), 77.60 (GS263) and 77.49 (GS264) (upwelled),

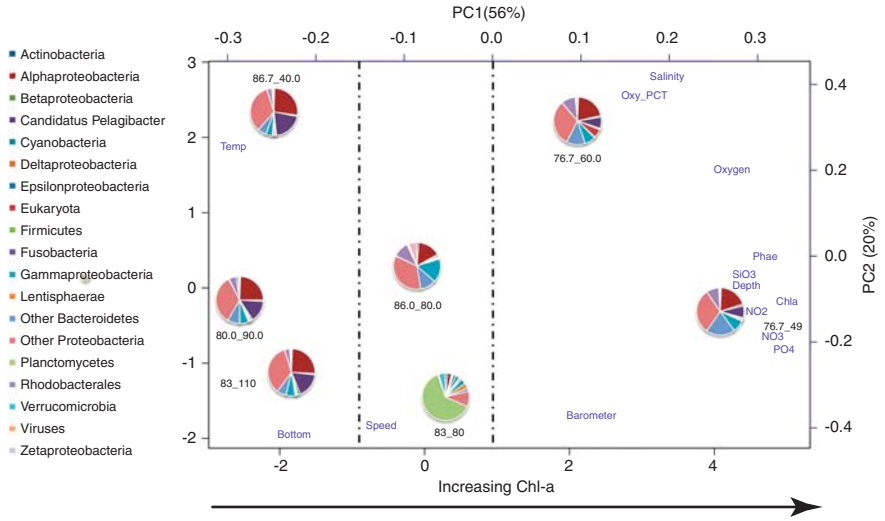

Figure 3 PCA using oceanographic metadata, including nutrients. The CalCOFI sample sites are given and a taxonomic representation of sequences from each site is shown in a pie chart.

displayed clear differences in bacterioplankton community structure. Interestingly, site 87.80 (GS258) showed an increase in Actinobacteria relative to other samples. Actinobacterial sequences were previously detected in the GOSI samples (Rusch *et al.*, 2007; Jensen and Lauro, 2008). Although these sequences appear to be phylogenetically related to known Actinobacteria, they are difficult to classify at lower taxonomic levels, indicating the potential for a novel surface ocean Actinobacterial group. Alternatively, upwelling could be a source for sediments from the benthic boundary layer and associated taxa in these waters (Johnson *et al.*, 1999). Another distinct taxonomic feature of the sample set was an increase of ORFs classified within the *Cytophage/Flavobacteria/Bacteroidetes* (CFB) group predominantly in the active upwelling sites, 77.60 (GS263) and 77.49 (GS264). Generally, Proteobacteria sequences were consistently more abundant within the oligotrophic sites, with a relative increase in α-proteobacteria (Supplementary Figures S3 and S4). In the upwelled sites, however, total percentages of sequences classified as Proteobacteria were lower and class γ-proteobacteria were relatively more abundant. Lastly, a Planctomycete bloom was evident in the 0.1-μm (and 0.8-μm) sequence data as is described below.

Community structure of 'large' microbial size classes

This study is unique to previous marine metagenomics reports in that microorganisms from 0.1 to 200 μm were investigated using filter cutoffs to reduce complexity. Based on phylogenetic placement of bacterial core genes (Figure 4; Supplementary Table S3), bacterial community structure dynamics shifted significantly ($P<0.05$) according to size class and nutrient availability. Generally, larger size class microbes appear more sensitive to shifts in nutrient availability. In addition to the 'free-living' CFB populations within the lower size class, a relative increase in sequences from the CFB group was found in the greater size class communities, consistent with their suggested lifestyle preference in association with particulate organic detritus in marine systems (Crump *et al.*, 1999). CFB enrichment on particulate detritus is likely due to their enhanced ability to degrade and assimilate high molecular weight dissolved organic matter compared with other bacterial groups (Cottrell and Kirchman, 2000). γ-Proteobacterial sequences were also more abundant on the larger filters, especially those from sites of higher nitrate concentrations; consistent with the capability of some lineages to assimilate nitrate (Allen *et al.*, 2001). These γ-proteobacterial sequences were concentrated on an internal node of the reference tree, allowing classification to the phylum level only (Figure 4a), hence distantly related to any sequenced genomes used to construct the reference tree (Supplementary Table S4). Environmental sequences within the *Roseobacter* lineage were more abundant among the α-proteobacterial sequences at the nearshore upwelled sites off Point Conception, sites 77.60 (GS263) and 77.49 (GS264). Phylogenetic data also revealed an increase of Synechococcus and Prochlorococcus populations in the 0.8–200 μm size class (Figure 4).

Additionally, the 0.8- and 3.0-μm filters contained a relatively large fraction of sequences phylogenetically classified as viruses (5.4% and 4.8%, respectively, compared with 1.9% on the 0.1-μm filter; Supplementary Table S5). These viral sequences could have originated from infection of phytoplankton (or larger-celled heterotrophic bacteria), latent infection, or from virioplankton surface associated with larger-celled organisms (DeLong *et al.*, 2006; Williamson *et al.*, 2008). Analysis of the sequence diversity of viruses based on clustering at global identity cutoffs of 70–100% on each of the filters

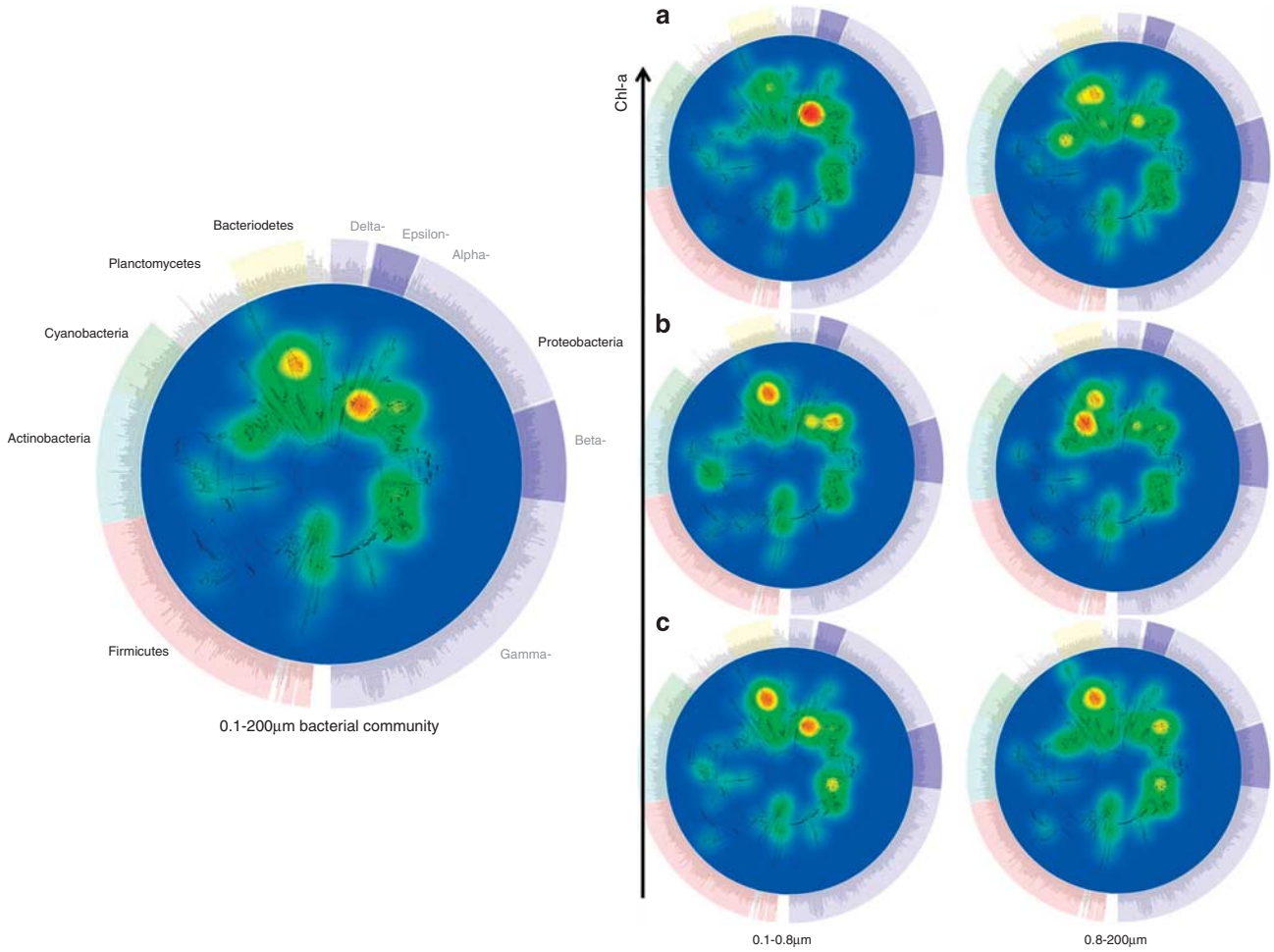

Figure 4 Taxonomic classification using core bacterial HMMs with phylogeny inferred based on placement on the reference tree of core bacterial HMMs from all sequenced bacterial genomes. Taxonomic groups of interest are highlighted as follows: blue = Proteobacteria, red = Firmicutes, light blue = Actinobacteria, green = Cyanobacteria, white = Planctomycetes, yellow = Bacteroidetes. Heatmaps were generated based on the abundance of taxonomy at each node within the reference tree for every group of samples.

revealed a higher degree of diversity in the 0.1–0.8 μm community compared with the larger (0.8–200 μm) size classes (Supplementary Figure S5), suggestive of a higher level of clonality and consistent with infection or association with dominant bacterioplankton or phytoplankton populations.

Generally, sites of active upwelling were enriched in sequences classified as diatoms in the largest size class (3.0–200 μm). The picoeukaryotic prasinophytes, *Micromonas* sp. and *Ostreococcus* sp., were relatively more abundant, when compared with other sites in the 0.8–3.0 μm size class (Supplementary Figure S6). Offshore oligotrophic sites, particularly in the case of the largest size class (3.0–200 μm), were enriched in sequences classified as dinoflagellates or other Alveolata. The >0.8 μm size class at the oligotrophic sites, 87.40 (GS257) and 83.110 (GS259), contained enrichments for pelagophytes and sequences cladding with the non-photosynthetic ciliate alveolates *Tetrahymena* sp. or *Paramecium* sp., respectively. Therefore, patterns of eukaryotic community composition display relative enrichments of different taxa associated with different size classes and nutrient availability. More specifically, within the picoeukaryotic size class, (0.8–3.0 μm), the prasinophytes were more abundant at upwelling sites whereas pelagophytes and ciliates were more abundant at oligotrophic sites. In the larger eukaryotic size class (3.0–200 μm), dinoflagellates and diatoms were relatively more abundant at oligotrophic and upwelling sites, respectively.

Genomic sampling of a Planctomycete bloom

At station 83.80 (GS260), a Planctomycete bloom event was detected by a dramatic spike in the abundance of Planctomyces genes (60.4% of total ORFs on the 0.1-μm, 26.5% on the 0.8-μm and 7% on the 3.0-μm filters), as well as, a general increase in genes from the closely related Chlamydiae/Verrucomicrobia and Lentisphaerae phyla; which likely also represent Planctomycete genes. This is significant as Planctomycetes are typically in low abundance relative to other bacterioplankton populations in marine samples (Rusch *et al*., 2007).

Planctomycetes appear to be in low abundance in all other samples analyzed in this study, ranging from 0% to 3% of the bacterial peptides classified. Why Planctomycetes bloom is unknown; however, they have been identified in similar environments, such as, Namibian and Oregon upwelling systems (Woebken *et al.*, 2007) and high densities of Planctomycete bacteria have been reported in association with diatom blooms (Morris *et al.*, 2006). This spatial association has been suggested as indicative of direct interactions, and was further exemplified by the possible gene transfer of assimilatory nitrite reductase (*nasB*) from Planctomycetes to diatoms (Bowler *et al.*, 2008). The occurrence of all of the genes required for transport and assimilation of NO_3^- in Planctomycete genomes might partially explain their apparent success in upwelled waters. Additionally, it has been proposed that Planctomycetes can derive energy through cleavage of sulfated macromolecules produced by algae (Glockner *et al.*, 2003). Silaffins are abundant cell wall proteins in diatoms and known to be associated with heavily sulfated sugars (Kroger and Poulsen, 2008). COG3119 (arylsulfatases) and KOG3867 (sulfatase) are examples of highly abundant gene families in GS260 data sets, which is consistent with Planctomycete reference genomes harboring an unusually high number of sulfatases (~41–100 genes/genome) (Glockner *et al.*, 2003; Woebken *et al.*, 2007). Although the abundance of diatom sequences in the >0.8 μm size class is relatively elevated at 83.80 (GS260) (Supplementary Figure S6), satellite imagery of regional chlorophyll in surface waters, prior and subsequent to sampling, suggest that sampling likely occurred following advection of a filament of upwelled water offshore and during a period of phytoplankton bloom decline (Supplementary Figure S7); providing further evidence of diatom–CC Planctomycete (CCP) interactions potentially supporting the high concentration of Planctomycetes at this site compared with other samples.

At the time of this study, five planctomycete genomes were available (*Rhodopirellula baltica*, *Blastpirellula marina*, *Gemmata obscuriglobus*, *Planctomyces limnophilus* and *Planctomyces maris*). To identify which features the sampled CCP population had in relation to these reference genomes, MDS analysis was performed based on the presence/absence of CCP sequences in these references (Supplementary Figure S8). These data indicated that the CCP is most closely related to the other marine planctomyces, *B. marina* and *P. maris*. Sequence assembly of the Sanger and 454 libraries for GS260 was performed and taxonomic binning used to identify CCP scaffolds and contigs. Genes on Planctomycete scaffolds and contigs not classified as Planctomycete at the phylum level and not present in the reference genomes consisted of 360 total ORFs. Over 62% of these ORFs were unable to be assigned using phylogenetic profiling (Supplementary Table S6).

SAR11 diversity

Microbes within the SAR11 cluster are ubiquitous in the world's oceans (Rappe *et al.*, 2002) and found in large numbers in surface waters (Morris *et al.*, 2002). *Pelagibacter*-like 16S rDNA sequences were mined from the metagenomic data (162 total found) to determine their diversity across the upwelling gradient (see Supplementary Information: '*SAR11 Identification*'). The majority of sequences derived from the 0.1-μm filter (Sanger or 454-Titanium; 95.6%), and primarily from the oligotrophic sites (79.6%) (Supplementary Table S7). Pplacer (Matsen *et al.*, 2010) was used to phylogenetically place metagenomic sequences onto a fixed 16S rDNA reference tree. All four of the known subgroups of SAR11 were identified within the upwelling gradient (Figure 5). Forty-six query rRNA sequences fell outside reference clades, suggesting that these sequences are evidence of new strains belonging to these subgroups or novel subgroups (Figure 5; Supplementary Table S7).

Interestingly, a few of the SAR11 16S rDNA sequences were found on the larger filters (Supplementary Table S6), as well as SAR11 non-rRNA reads that were taxonomically binned either using the phylogenomic approach or single copy core HMMs (Hidden Markov Models) (Supplementary Table S3). All 16S rDNA sequences from the larger size classes were in the subgroup 1 and 2 clades. We speculate that distinct SAR11 ecotypes, compared with the 0.1-μm ecotypes, within or similar to subgroup 1 and 2 occupy a niche in association (that is, on the cell surface) with larger bacterioplankton and phytoplankton, for example, cyanobacteria. Further, *C. pelagibacter* sp. assemblies (2082 scaffolds from Sanger and 403 contigs from 454-Titanium assemblies) were analyzed based on site characteristics. Interestingly, reads from the 0.8- and 3.0-μm filters are more likely to assemble compared with those from the 0.1-μm filters (Supplementary Figure S9). Evaluation of ORFs from *Pelagibacter* contigs that included reads from larger size classes revealed that 105 ORFs (715 total from 454 assembly) were not homologous to any predicted proteins in the three available reference genomes (*C. pelagibacter ubique* HTCC1002, *C. pelagibacter ubique* HTCC1062 and *C. pelagibacter* sp. HTCC7211) and of the 15 that showed homology to a sequence within the GenBank database, five appeared to be of viral origin.

Metabolic profiling of the CC microbial community

While certain genes or gene families are found at relatively consistent levels across genomes (for example, translation and cell division functions), others, such as transporters and most metabolic genes, often scale disproportionally with genome size. To examine the functional repertoire of microorganisms sampled within this hydrodyamically complex region, sequences were binned based on

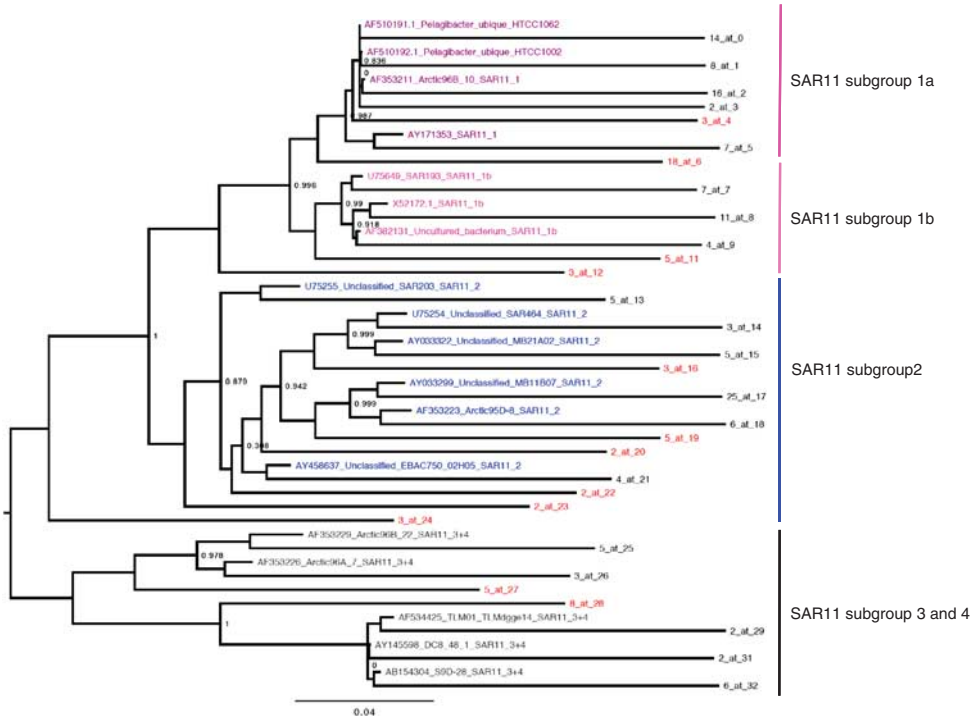

Figure 5 PhyML phylogenetic tree of 16S rDNA sequences of metagenomic sequences and SAR11 reference sequence. Each SAR11 is color-coded by subgroup: subgroup 1a (dark purple), subgroup 1b (light purple), subgroup 2 (blue), and subgroups 3 and 4 (grey). The numbers specify how many sequences were placed at each node. Red numbers indicate metagenomics sequences outside clades containing reference sequence(s).

nearest-neighbor homology to clusters of orthologous groups (COGs—bacteria; KOGs—eukaryote). In general, the number of individual COG hits per genome equivalent increased with greater size class of the sample (0.1 μm compared with >0.8 μm) and from oligotrophic to upwelled samples, which correlates with overall increases in estimated genome size. Figure 6 illustrates the relationship of the COG/KOG20 (Cuvelier *et al.*, 2010) (identified as the number of COG/KOGs per genome in the top 20% of total COG/KOGs scored) for each sample or reference genome versus genome size. Increases in COG/KOG20 are equivalent to expansions in the largest gene families, whereas decreases are suggestive of a reduced occurrence of paralogs and streamlined genomes. Although based on a limited number of samples, these results are consistent with the hypothesis that microbes in nutrient-rich waters, especially larger size class microbes, have larger genomes and an enriched repertoire of metabolic capabilities. This is reflected by a larger bin of paralogous genes.

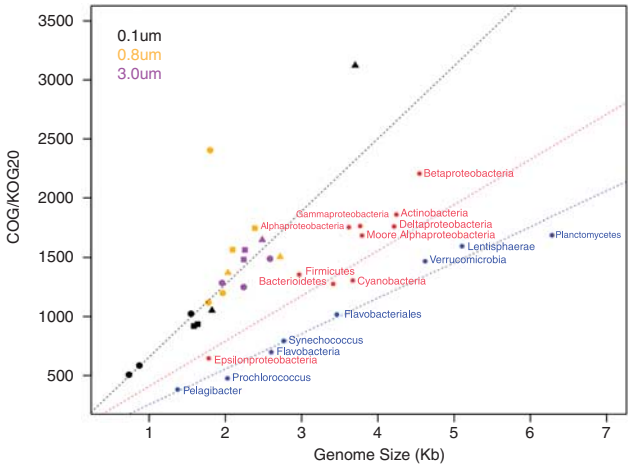

Figure 6 Metabolic profiling of samples based on COG/KOG 20. Metagenomic samples are as follows: 0.1 μm (black), 0.8 μm (orange), 3.0 μm (purple) and oligotrophic (closed circle), aged–upwelled (closed triangle), upwelled (closed square). Reference genomes were binned into groups of varying diversity levels to show that the slope of the linear regression varies with taxonomic diversity: phyla (red) and phyla with fewer than six representatives and genera (blue).

Analysis of nitrogen metabolism

Glutamine synthetases (GSI, GSII and GSIII) are key enzymes in nitrogen metabolism, specifically ammonium assimilation and glutamine biosynthesis, within all domains of life. GSI is found in diverse bacterial and archaeal lineages (Brown *et al.*, 1994), as well as, in some eukaryotic taxa. GSIII has been found in bacteria, cyanobacteria and eukaryotes (particularly diatoms and pransinophytes (green algae)). GSII has interweaving phylogenies with evidence of gene transfer events between and among bacteria and eukaryotes (Brown and Doolittle, 1997; Robertson and Tartar, 2006; Ghoshroy *et al.*, 2010).

The GSII gene family is found in all eukaryotic lineages and some bacterial lineages, however, primarily comprised of the former. Large numbers of bacterial GSII proteins were found in previous GOS data (Yooseph *et al.*, 2007). We identified two groups of GSs in the Southern California upwelling region using COG/KOG assignments; corresponding to 2763 GSI (COG0174) and 166 GSII (KOG0683) sequences. Of the 166 GSII peptides, 44 were of bacterial origin and classified as γ-proteobacteria and Actinobacteria (13.3% of the GSII peptides for each Phyla). However, 4.2% (seven total) of the GSII hits are phylogenetically classified as Mimiviridae, most likely related to Mimivirus (Raoult *et al.*, 2004) (Supplementary Figure S10). Mimivirus-like sequences have been attributed to viruses that infect marine protists and eukaryotic phytoplankton (Monier *et al.*, 2008). Within the GSII class there have been eukaryotic and bacterial genes identified, with phylogenetic placement of γ-proteobacteria sequences in a clade among green algae, thus invoking implications for non-endosymbiotic acquisition of these genes from bacteria to eukaryote (Ghoshroy *et al.*, 2010). Our finding of the Mimiviridae-like GSII sequences suggests that viruses similar to Mimivirus may have aided in the transfer of these genes among marine eukaryotic microorganisms.

Evidence for nitrogen-fixing organisms has been scarce in large-scale metagenomic data sets from the surface ocean (Johnston *et al.*, 2005). We found similar results with only one hit to nitrogenase (*nifH*), based on BLAST using a nifH reference against the complete set of 454-Titanium predicted proteins; with homology to the N-fixing marine cyanobacterium UCYN-A (Tripp *et al.*, 2010). Subsequently, we evaluated all proteins phylogenetically affiliated with UCYN-A and found a preponderance of sequences from the nearshore oligotrophic sample GS257 (87.40) and within the largest size class (77.2% in 3.0–200 μm size class and 19.4% in 0.8–3.0 μm). These data support the predicted lifestyle of UCYN-A as being associated with larger organisms (Tripp *et al.*, 2010) and indicate that there is enhanced capacity for nitrogen fixation in oligotrophic waters from organisms linked to larger size class microbes.

Transport processes

Transporters have been shown to increase as a function of bacterial genome size and represent a high percentage of total proteins per genome. For comparison, ~9% of marine bacterial genomes from Yooseph *et al.* (2010) were characterized as

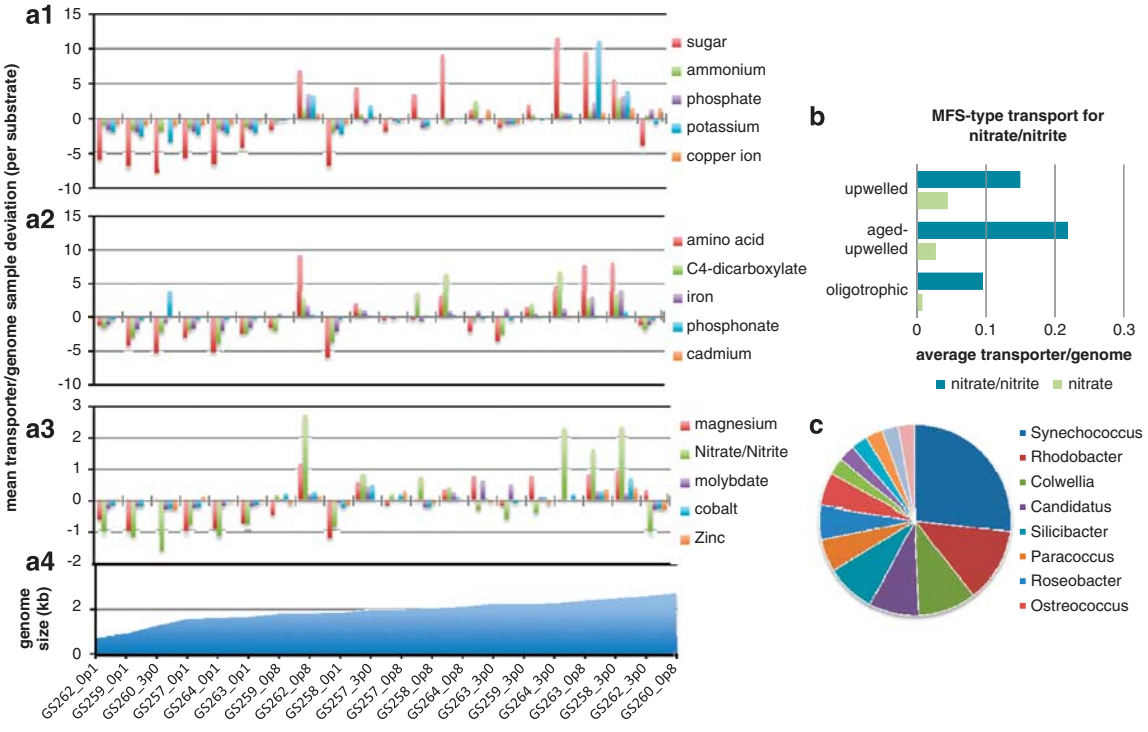

Figure 7 Transport proteins binned based on substrate plotted with increasing genome size. (**a**) The mean transporter per genome for each substrate was computed and the difference from the mean (across all sites and size classes of each transporter family based on substrate) plotted for each sample, *y*-axis. Substrates were binned into three groups based on their proportion to the total number of proteins per genome; therefore (a1) denotes enriched in large genomes (proportion increases with genome size), (a2) enriched in small genomes and (a3) unchanged (proportion does not change with change in genome size), based on data obtained from the slope of the line in Figure S11. The *x*-axis shows the samples (sites/site classes) ordered by genome size (a4). (**b**) Average transporter per genome of nitrate/nitrite substrate transporters only belonging to the Major Facilitator Superfamily (MFS). (**c**) Taxonomic classification of nitrate/nitrite MFS transport proteins in B.

transport proteins and range from 2% to 16%. All metagenomic bacterial peptides, phylogenetically classified at the kingdom level, were searched against transportDB (Ren *et al.*, 2007). A positive correlation is suggested between nutrient availability and size class, genome size and transporter abundance (Figure 7; Supplementary Figure S11). To examine whether transport protein families with known substrates were enriched or depleted in each site/size class, the proportion of transporters per genome for each of the substrates analyzed relative to the mean was estimated (Figure 7a). Additionally, transporters predicted to be elevated in nutrient-rich water, for example, nitrate/nitrite, are overrepresented in upwelled sites (Figure 7b), and within the $>0.8\,\mu m$ size classes. Major facilitator superfamily of transporters (MFS) for nitrate and nitrate/nitrite transporters taxonomy is largely dominated by *Synechoccoccus* sp., which is known to be important for nitrogen assimilation (Paerl *et al.*, 2011) and shows overlap with taxa that are enriched in upwelled sites including, γ-proteobacteria and bacteriodetes (Figure 7c). With the exception of the most nearshore upwelled site 77.49 (GS264), iron and manganese follow similar distributions to one another showing a relative enrichment in sites of low nutrient concentrations (oligotrophic) and in larger size classes, again most likely due to the abundance of Prochlorococcus-like cyanobacteria in these waters. MFS multidrug efflux transporters showed the greatest disparity in enrichment from 1.6 per genome equivalent (GS259 0.1 μm) to 9.36 per genome equivalent (GS258 3.0 μm). This may in part result from the propensity of bacteria with larger genomes to synthesize more secondary metabolites, metabolic pathways that commonly harbor MFS multidrug efflux transport proteins (Martin *et al.*, 2005). Metal transport functions, including copper and zinc, appear more prevalent in low nutrient/large genome size classes. Interestingly, cadmium was the only substrate over represented in the 0.1-μm fraction (Supplementary Figure S11).

Assembly analysis

To further investigate the relationship between sequences from sites of similar oceanographic context, assemblies were generated from 0.1 μm. Sanger sequence reads and from 0.1, 0.8 and 3.0 μm 454-Titanium sequences were used to link which sites were more likely to co-assemble into scaffolds or contigs, respectively (Supplementary Figure S9). It was clear that sites with similar oceanographic context are more likely to assemble together rather than sites of differing ocean chemistry, irrespective of spatial distance among sites. The aforementioned results from examining scaffolds/contigs taxonomically classified as *C. pelagibacter* sp. showed similar trends. These data indicate that perhaps ecotypes of organisms (even within the same genera) are forming distinct populations across the upwelling mosaic, as well as, within the same sites but among different communities (niches; size classes).

Conclusions

Delineation of spatial and temporal boundaries that define bacterioplankton communities in surface ocean environments is challenging due to the lack of clear habitat boundaries and high levels of dispersal within populations. Using community level genomics analyses we have shown that the upwelling mosaic, typical of summer months in the Southern CCE, segregates smaller bacterioplankton over small spatial scales and that community diversity is correlated with nutrient availability. Discrete bacterial communities were also observed within larger size classes. Interestingly, a clear distinction between size classes was also evident through assembly of taxonomically similar sequences between small (0.1–0.8 μm) and larger size classes (0.8–200 μm). These communities also appeared to have larger genomes, and therefore harbor enhanced metabolic capabilities; largely associated with gene family expansions that likely lead to novel functional capacity.

Acknowledgements

Support was provided by Department of Energy Office of Biological and Environmental Research DE-FC02-02ER63453 (to JCVI) and NSF-ANT-0732822 and NSF-EnGen-0722374 (to AEA). We thank Cyndi Pfannkoch for effort on library construction; Daniel Brami for assembly insight; Brian Palenik, Lihini Aluwihare and Shibu Yooseph for insightful discussion.

References

Allen AE, Booth MG, Frischer ME, Verity PG, Zehr JP, Zani S. (2001). Diversity and detection of nitrate assimilation genes in marine bacteria. *Appl Environ Microbiol* **67**: 5343.

Badger JH, Hoover TR, Brun YV, Weiner RM, Laub MT, Alexandre G *et al.* (2006). Comparative genomic evidence for a close relationship between the dimorphic prosthecate bacteria Hyphomonas neptunium and Caulobacter crescentus. *J Bacteriol* **188**: 6841.

Beja O, Aravind L, Koonin EV, Suzuki MT, Hadd A, Nguyen LP *et al.* (2000). Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science (New York, NY)* **289**: 1902.

Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A *et al.* (2008). The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature* **456**: 239.

Brown JR, Doolittle WF. (1997). Archaea and the prokaryote-to-eukaryote transition. *Microbiol Mol Biol Rev* **61**: 456.

Brown JR, Masuchi Y, Robb FT, Doolittle WF. (1994). Evolutionary relationships of bacterial and archaeal glutamine synthetase genes. *J Mol Evol* **38**: 566.

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

Brown MV, Schwalbach MS, Hewson I, Fuhrman JA. (2005). Coupling 16S-ITS rDNA clone libraries and automated ribosomal intergenic spacer analysis to show marine microbial diversity: development and application to a time series. *Environ Microbiol* **7**: 1466.

Cottrell MT, Kirchman DL. (2000). Natural assemblages of marine proteobacteria and members of the Cytophaga-Flavobacter cluster consuming low- and high-molecular-weight dissolved organic matter. *Appl Environ Microbiol* **66**: 1692.

Crump BC, Armbrust EV, Baross JA. (1999). Phylogenetic analysis of particle-attached and free-living bacterial communities in the Columbia river, its estuary, and the adjacent coastal ocean [in process citation]. *Appl Environ Microbiol* **65**: 3192.

Cuvelier ML, Allen AE, Monier A, McCrow JP, Messie M, Tringe SG *et al.* (2010). Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton. *Proc Natl Acad Sci* **107**: 14679.

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

DiGiacomo PM, Holt B. (2001). Satellite observations of small coastal ocean eddies in the Southern California Bight. *J Geophys Res Oceans* **106**: 22521.

Eloe EA, Fadrosh DW, Novotny M, Zeigler Allen L, Kim M, Lombardo MJ *et al.* (2011). Going deeper: metagenome of a hadopelagic microbial community. *PLoS One* **6**: e20388.

Ghoshroy S, Binder M, Tartar A, Robertson DL. (2010). Molecular evolution of glutamine synthetase II: phylogenetic evidence of a non-endosymbiotic gene transfer event early in plant evolution. *BMC Evol Biol* **10**: 198.

Gilbert JA, Field D, Swift P, Thomas S, Cummings D, Temperton B *et al.* (2010). The taxonomic and functional diversity of microbes at a temperate coastal site: a 'multi-omic' study of seasonal and diel temporal variation. *PLoS One* **5**: e15545.

Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D *et al.* (2005). Genome streamlining in a cosmopolitan oceanic bacterium. *Science (New York, NY)* **309**: 1242.

Glockner FO, Kube M, Bauer M, Teeling H, Lombardot T, Ludwig W *et al.* (2003). Complete genome sequence of the marine planctomycete Pirellula sp. strain 1. *Proc Natl Acad Sci USA* **100**: 8298.

Handelsman J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev* **68**: 669.

Hickey BM. (1979). The California current system–hypotheses and facts. *Prog Oceanogr* **8**: 191.

Hickey BM, Dobbins EL, Allen SE. (2003). Local and remote forcing of currents and temperature in the central Southern California Bight. *J Geophys Res* **108**: 3081.

Howe K, Bateman A, Durbin R. (2002). QuickTree: building huge neighbour-joining trees of protein sequences. *Bioinformatics* **18**: 1546.

Huyer A. (1983). Coastal upwelling in the California current system. *Prog Oceanogr* **12**: 259.

Jensen PR, Lauro FM. (2008). An assessment of actinobacterial diversity in the marine environment. *Antonie van Leeuwenhoek* **94**: 51.

Johnson KS, Chavez FP, Friederich GE. (1999). Continental-shelf sediment as a primary source of iron for coastal phytoplankton. *Nature* **398**: 697.

Johnston AWB, Li YG, Ogilvie L. (2005). Metagenomic marine nitrogen fixation – feast or famine? *Trends Microbiol* **13**: 416.

Kroger N, Poulsen N. (2008). Diatoms–from cell wall biogenesis to nanotechnology. *Ann Rev Genet* **42**: 83.

Kudela RM, Cochlan WP, Dugdale RC. (1997). Carbon and nitrogen uptake response to light by phytoplankton during an upwelling event. *J Plankton Res* **19**: 609.

Kudela RM, Garfield N, Bruland KW. (2006). Bio-optical signatures and biogeochemistry from intense upwelling and relaxation in coastal California. *Deep Sea Res Pt Ii* **53**: 2999.

Lowe TM, Eddy SR. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**: 955.

Martin JF, Casqueiro J, Liras P. (2005). Secretion systems for secondary metabolites: how producer cells send out messages of intercellular communication. *Curr Opin Microbiol* **8**: 282.

Matsen FA, Kodner RB, Armbrust EV. (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* **11**: 538.

Monier A, Larsen JB, Sandaa RA, Bratbak G, Claverie JM, Ogata H. (2008). Marine mimivirus relatives are probably large algal viruses. *Virol J* **5**: 12.

Morris RM, Longnecker K, Giovannoni SJ. (2006). Pirellula and OM43 are among the dominant lineages identified in an Oregon coast diatom bloom. *Environ Microbiol* **8**: 1361.

Morris RM, Rappe MS, Connon SA, Vergin KL, Siebold WA, Carlson CA *et al.* (2002). SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **420**: 806.

Paerl RW, Johnson KS, Welsh RM, Worden AZ, Chavez FP, Zehr JP.. (2011). Differential distributions of Synechococcus subgroups across the California Current System. *Front Microbiol* **2**: 59.

Palenik B. (1994). Cyanobacterial community structure as seen from RNA polymerase gene sequence analysis. *Appl Environ Microbiol* **60**: 3212.

Quaiser A, Zivanovic Y, Moreira D, Lopez-Garcia P. (2011). Comparative metagenomics of bathypelagic plankton and bottom sediment from the Sea of Marmara. *ISME J* **5**: 285.

Raes J, Korbel JO, Lercher MJ, von Mering C, Bork P. (2007). Prediction of effective genome size in metagenomic samples. *Genome Biol* **8**: R10.

Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H *et al.* (2004). The 1.2-megabase genome sequence of Mimivirus. *Science (New York, NY)* **306**: 1344.

Rappe MS, Connon SA, Vergin KL, Giovannoni SJ. (2002). Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature* **418**: 630.

Ren Q, Chen K, Paulsen IT. (2007). TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. *Nucleic Acids Res* **35**(Database issue): D274.

Robertson DL, Tartar A. (2006). Evolution of glutamine synthetase in heterokonts: evidence for endosymbiotic gene transfer and the early evolution of photosynthesis. *Mol Biol Evol* **23**: 1048.

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.

Tanenbaum DM, Goll J, Murphy S, Kumar P, Zafar N, Thiagarajan M *et al.* (2010). The JCVI standard operating procedure for annotating prokaryotic metagenomic shotgun sequencing data. *Stand Genomic Sci* **2**: 229.

Thompson JR, Pacocha S, Pharino C, Klepac-Ceraj V, Hunt DE, Benoit J *et al.* (2005). Genotypic diversity within a natural coastal bacterioplankton population. *Science (New York, NY)* **307**: 1311.

Toledo G, Palenik B. (1997). *Synechococcus* diversity in the California current as seen by RNA polymerase (rpoC1) gene sequences of isolated strains. *Appl Environ Microbiol* **63**: 4298.

Tripp HJ, Bench SR, Turk KA, Foster RA, Desany BA, Niazi F *et al.* (2010). Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature* **464**: 90.

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

Wilkerson FP, Lassiter AM, Dugdale RC, Marchi A, Hogue VE. (2006). The phytoplankton bloom response to wind events and upwelled nutrients during the CoOP WEST study. *Deep Sea Res Pt Ii* **53**: 3023.

Williamson SJ, Rusch DB, Yooseph S, Halpern AL, Heidelberg KB, Glass JI *et al.* (2008). The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. *PLoS One* **3**: e1456.

Woebken D, Teeling H, Wecker P, Dumitriu A, Kostadinov I, Delong EF *et al.* (2007). Fosmids of novel marine Planctomycetes from the Namibian and Oregon coast upwelling systems and their cross-comparison with planctomycete genomes. *ISME J* **1**: 419.

Yooseph S, Nealson KH, Rusch DB, McCrow JP, Dupont CL, Kim M *et al.* (2010). Genomic and functional adaptation in surface ocean planktonic prokaryotes. *Nature* **468**: 60.

Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K *et al.* (2007). The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol* **5**: e16.

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