npg

## ORIGINAL ARTICLE

# Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction

Robert M Morris[1], Brook L Nunn[2], Christian Frazar[1], David R Goodlett[2], Ying S Ting[1,2], Gabrielle Rocap[1]

[1]School of Oceanography, Center for Environmental Genomics, University of Washington, Seattle, WA, USA and [2]Department of Medicinal Chemistry, University of Washington, Seattle, WA, USA

**Bacteria and Archaea play critical roles in marine energy fluxes and nutrient cycles by incorporating and redistributing dissolved organic matter and inorganic nutrients in the oceans. How these microorganisms do this work at the level of the expressed protein is known only from a few studies of targeted lineages. We used comparative membrane metaproteomics to identify functional responses of communities to different nutrient concentrations on an oceanic scale. Comparative analyses of microbial membrane fractions revealed shifts in nutrient utilization and energy transduction along an environmental gradient in South Atlantic surface waters, from a low-nutrient gyre to a highly productive coastal upwelling region. The dominant membrane proteins identified (19%) were TonB-dependent transporters (TBDTs), which are known to utilize a proton motive force to transport nutrients across the outer membrane of Gram-negative bacteria. The ocean-wide importance of TonB-dependent nutrient acquisition in marine bacteria was unsuspected. Diverse light-harvesting rhodopsins were detected in membrane proteomes from every sample. Proteomic evidence of both TBDTs and rhodopsins in the same lineages suggest that phototrophic bacterioplankton have the potential to use energy from light to fuel transport activities. We also identified viral proteins in every sample and archaeal ammonia monooxygenase proteins in the upwelling region, suggesting that Archaea are important nitrifiers in nutrient-rich surface waters.**

## Introduction

Bacteria and Archaea dominate biomass in the world's oceans and have key roles in marine biogeochemical cycles. As most of these marine microorganisms have not been cultured, it has been difficult to directly link organisms with the biogeochemical processes they mediate. Recently, metagenomic studies have provided insights into the vast diversity and metabolic potential of microorganisms in the environment (Tyson *et al.*, 2004; Venter *et al.*, 2004; Coleman *et al.*, 2006; DeLong *et al.*, 2006; Rusch *et al.*, 2007; Yooseph *et al.*, 2007). Functional approaches, such as transcriptomics and proteomics, are now needed to link this genomic diversity with activity. Metaproteomic approaches that detect expressed proteins and thus verify *in situ* microbial functions in a given habitat have been applied to inhabitants of low-diversity acid-mine drainage,

wastewater sludge and anaerobic reductive-dechlorinating microbial communities (Ram *et al.*, 2005; Lo *et al.*, 2007; Morris *et al.*, 2007; Wilmes *et al.*, 2008), as well as more complex systems such as the oceans and human microbiome (Giovannoni *et al.*, 2005; Chen *et al.*, 2008; Sowell *et al.*, 2009). Successful metaproteomic studies have succeeded by taking advantage of rich metagenomic data sets, by targeting specific lineages or proteins, and by rapid advances in MS/MS-based proteomic technologies (for a review see VerBerkmoes *et al.*, 2009).

In the first MS/MS-based metaproteomic study of marine microorganisms, *in situ* expression of a light-dependent proton pump (proteorhodopsin) from the alpha-proteobacterial SAR11 clade was detected in an enriched microbial membrane fraction from the Oregon coast (Giovannoni *et al.*, 2005), confirming that these dominant heterotrophic microbes express proteins that enable them to use light to generate a proton motive force (PMF). A subsequent metaproteomic study of nutrient-poor seawater in the North Atlantic gyre targeted all proteins expressed by SAR11 cells and the marine cyanobacteria *Prochlorococcus* and *Synechococcus* (Sowell *et al.*, 2009). This whole-cell approach maximized the

674

number of proteins detected, but as membrane proteins such as proteorhodopsin are not easily solvated and are generally less abundant (Stapels et al., 2004), proteorhodopsins were not identified. Transport proteins dominated the Sargasso Sea microbial metaproteome, including a preponderance of inorganic and organic phosphate transporters. Although the potential for these energy transduction and nutrient transport processes had been recognized in genome sequence data, functional approaches such as these were required to move from prediction to verified in situ function. The challenge is to use protein-centric approaches that move beyond targeted lineages and resolve functional differences between proteins expressed by complex communities of genetically diverse and largely uncultured microorganisms in the oceans.

In this study, we apply a protein-centric comparative metaproteomics approach on an oceanic scale. We targeted membrane proteins because they are involved in nutrient transport and energy transduction. We evaluated a suite of 10 surface seawater samples collected along a natural gradient in nutrient concentrations that spanned the low-nutrient South Atlantic gyre (open ocean) and high-nutrient Benguela upwelling region (coastal). Existing marine metagenomic data were used to identify the potential range of taxonomic origins and functions of expressed membrane proteins. Rather than focus on the proteomes of sequenced lineages, we analyzed functional differences expressed by whole communities. Although we restricted our analyses to the membrane fractions, we detected an open ocean to coastal shift in bacterial community structure, nutrient utilization and energy transduction, as well as viral and archaeal activities.

## Materials and methods

### Cell concentrations
Microorganisms were concentrated from 10 large volume (100–200 l) South Atlantic surface seawater samples (5–8 m) collected in November and December 2007. We targeted the $< 0.8\,\mu m$ size fraction to obtain the same bacterioplankton community used for the global ocean survey (GOS) shotgun sequencing (Rusch et al., 2007). This was to increase the number proteins identified by eliminating proteins from organisms for which sequence data were less likely to be available. All concentrations were performed in a 50 l polystyrene reservoir using a Pellicon 2 cassette tangential flow filtration system equipped with one 30 kD Biomax Polyethersulfone cassette (Millipore Corporation, Billerica, MA, USA). Seawater was continuously added to the concentration reservoir until cell densities reached $\sim 10^8$ cells ml$^{-1}$ (Supplementary Figure 1). In all, 100–200 ml of cell concentrate was obtained in 1.5–3 h. Concentrated cells were flash frozen in liquid nitrogen and stored at $-80\,°C$ until further proces-

sing at the University of Washington. Cell counts and recovery rates (average, $26 \pm 10\%$) were determined by staining cells from whole water and concentrated seawater samples with the nucleic acid stain 4′,6-diamidino-2-phenylindole (DAPI) and by fluorescence in situ hybridization as previously described (Morris et al., 2002). Cells were imaged using a Nikon 80i microscope equipped with a CoolSNAP HQ² camera (Photometrics, Tucson, AZ, USA) and NIS Elements Basic Research software (Nikon Instruments, Melville, NY, USA). More than 500 cells were counted from each slide (15 frames).

### Cell fractionations
Membrane fractions were enriched using methods previously determined to reduce soluble proteins and provide deeper coverage of membrane proteins (Morris et al., 2006; Fung et al., 2007). A cell pellet was prepared from each tangential flow filtration concentrated sample by centrifugation at $4\,°C$ for 60 min ($17\,000\,g$). The supernatant was discarded and cell pellets were resuspended in 3 ml of 20 mM Tris buffer pH 7.4. Crude extracts were prepared by passing the cells through a French pressure mini cell at $8000$ lb inch$^{-2}$ two times and subsequent centrifugation at $4\,°C$ for 30 min ($18\,000\,g$). Crude pellet and soluble fractions were separated and pellets were rinsed with 100 ml 20 mM Tris buffer pH 7.4. Highly enriched membrane material for proteomic analyses was obtained from soluble cell fractions by centrifugation at $4\,°C$ for an additional 60 min ($104\,000\,g$). Membrane-enriched material was isolated and rinsed with 100 ml 20 mM Tris buffer pH 7.4. Crude pellets, membrane-enriched pellets and soluble cell fractions were stored at $-80\,°C$ until further processing. Membrane-enriched and soluble cell fractions from the same cells were used for community protein and DNA analyses, respectively.

### Mass spectrometry
Proteomic analyses were performed on the membrane-enriched cell fraction (Nunn et al., 2009). Crude cell pellets and soluble fractions were not analyzed, as cytoplasmic proteins were not the focus of this study. The membrane-enriched pellet from each station was digested for proteomic analyses in a microfuge tube to reduce protein loss; pellets were first solubilized in urea (100 ml, 6 M) and Tris–HCl (6.6 ml, 1.5 M, pH 8.8). Disulfide bonds were reduced with tris(2-carboxyethyl) phosphine (2.5 ml, 200 mM) for 1 h ($37\,°C$) and then alkylated with iodoacetamide (IAM: 20 ml, 200 mM) in the dark for 1 h ($25\,°C$). Excess IAM was neutralized with dithiothreitol (20 ml, 200 mM: 1 h, $25\,°C$). Ammonium bicarbonate (800 ml, 25 mM) was added to dilute the urea before the addition of MeOH (200 ml), and sequence grade trypsin (Promega, Madison, WI, USA) at $\sim 50$:1 substrate:enzyme (w/w).

Trypsin digestions were vortexed and incubated 8 h at 37 °C. Additional trypsin was added at the same ratio to ensure full digestion of proteins and incubated for 18 additional hours at 37 °C. Samples were then taken to near dryness in a speedvac. Just before MS analysis, trypsin digestions were desalted using a micro-spin C18 column (Nest Group, Southborough, MA, USA) following the manufacturer's guidelines. Each sample consisted of ~1 μg total protein, allowing injections of ~200 ng for each of the 5 LC-MS/MS analyses. Samples were introduced into the mass spectrometer by reverse-phase chromatography using a 12-cm long, 75 μm i.d. fused silica capillary column packed in-house with C18 particles (Magic C18AQ, 100A, 5 μ: Michrom Bioresources, Inc., Auburn, CA, USA) fitted with a 2-cm long, 100 μm i.d. precolumn (Magic C18AQ, 200A, 5 μ; Michrom). Peptides were eluted using an acidified (formic acid, 0.1% v/v) water–acetonitrile gradient (5–35% acetonitrile in 60 min, full run-time 90 min). Mass spectrometry was performed on an LTQ-Orbitrap hybrid mass spectrometer (Thermo Fisher, San Jose, CA, USA). Data-dependent scans were completed by precursor ion selection in the fourier transform (FT)-based analyzer (Orbitrap) followed by collision induced dissociation (CID) in the linear ion trap (LTQ). Both the orbitrap and LTQ mass analyzers were calibrated with a vendor-recommended calibration solution; mass accuracies of <1 p.p.m. and <100 p.p.m., respectively, were confirmed using methionine–arginine–phenylalanine–alanine standards. To increase peptide identifications and protein sequence coverage one full $m/z$ analysis (400–2000) was followed by four gas-phase fractionations (350–520, 515–690, 685–970, 965–2000) for each sample. Gas-phase fractionations are ideal in situations in which minimal protein is available because it provides the user with a fractionation method that requires no additional sample handling, which inherently incurs protein loss. Rather than liquid bench-top fractionations, fractionations are performed in the mass spectrometer when ions are in the gas phase. Repeat injections of the sample are introduced into the LC-MS/MS; however, collision-induced dissociation is performed on ions selected across narrow $m/z$ ranges (Nunn et al., 2006; Scherl et al., 2008), allowing the mass spectrometer to increase its duty cycle and collection of MS2 spectra from that $m/z$ range.

*Protein identification*
All mass spectral results for this paper were interpreted and searched with an in-house copy of SEQUEST (PVM v.27 20070905) (Eng et al., 1994); SEQUEST is a correlative data-interpretation software package that matches observed spectra to theoretical spectra generated from predicted peptide sequences (for review see Nunn and Timperman, 2007). Tandem mass spectra were interrogated using the GOS Combined Assembly Protein (P) database (Rusch et al., 2007; Yooseph et al., 2007) available through the Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA) downloaded on 29 September 2008. The database contained >6 000 000 proteins predicted from microbial environmental genomic data through clustering and hidden Markov model approaches. Data searches were completed with no enzyme specificity and modifications of cysteine residues by 57 Da (resulting from the iodoacetamide modification) and methionine by 15 999 Da (oxidation) were allowed. Dynamic exclusion was set to 20 s. Peptides were only accepted if both termini were tryptic. Minimum ProteinProphet and PeptideProphet thresholds were set differently to ensure that peptide matches to CID spectra were very strong, but included proteins identified by one peptide (90% and 99%, respectively). Using this high stringency cutoff, we report peptides identified only once in a sample. It was not practical, however, to calculate false discovery rates from a search database that contained >6 million forward proteins.

*Protein quantification*
Relative protein abundance was assessed at the peptide level using a semiquantitative method: peptide spectral counting (Chen et al., 2008; Ryu et al., 2008). Peptide spectral counts were determined by the number of times a peptide that correlates with a given protein was selected for CID, including all repeated selections of the same peptide. As peptides elute off the column, those that are from abundant proteins are selected more often for CID using a data-dependent acquisition strategy. Thus, a protein's spectral count value was the sum of all identified peptide tandem mass spectra acquired for that protein. Spectral counts were determined from the four gas-phase fractionations and subsequently summed by consensus annotation (below).

*Consensus annotation*
As conserved amino-acid sequences in homologous proteins or conserved protein domains can be identified by the same peptide or peptides, taxonomic assignments and protein annotations were determined by consensus. All GOS proteins identified by interrogating peptide tandem mass spectra were searched against CAMERA's Non-Identical Peptide Sequences database using the Basic Local Alignment Search Tool (BLAST) available through CAMERA. BLAST expect scores >10⁻⁴ were annotated as 'unknown'. The best BLAST hit for all GOS protein sequences identified by the same peptide or peptides were first evaluated to determine the level of consensus among all taxonomic identifications. Taxonomic identities and protein function assignments were made at the most specific level that applied to all of the GOS protein annotations. For example, if all best BLAST hits of GOS proteins

identified by the same peptide or peptides were from different bacterial lineages, then the taxonomic assignment was Bacteria. However, if all best BLAST hits of GOS proteins identified by the same peptide or peptides were from *Proteobacteria*, then the taxonomic assignment was *Proteobacteria.* The process was repeated for all proteins. An analogous approach was used to determine the level of consensus among functional annotations. We assigned consensus taxonomic and functional annotations to 42 377 putative GOS protein sequences (see examples in Supplementary Figure 2).

*Clone library construction*
Bacterial 16S rRNA gene clone libraries were constructed from the soluble cell fractions of the concentrated bacterioplankton cell lysates prepared for proteomic analyses. At the pressures used, the French press shears DNA to fragments with an average length of 500–600 bp, thus partial rather than full-length 16S rRNA sequences were cloned. Community genomic DNA was extracted from 200 μl of cell lysate using a DNeasy Blood and Tissue kit (QIAGEN, Germantown, MD, USA). DNA was extracted according to the manufacturer's instructions. Ribosomal RNA genes were amplified from community genomic DNA for cloning by PCR with *Taq* polymerase (Fermentas, Hanover, MD, USA) and variations of commonly used bacterial primers, 8F and 519R. Briefly, amplifications were performed in a C1000 thermal cycler (Bio-Rad Laboratories, Hercules, CA, USA) using the following conditions: 35 cycles, annealing at 55 °C for 1 min, elongation at 72 °C for 2 min and denaturation at 94 °C for 30 s. A single band of the predicted length was observed by agarose gel electrophoresis, excised and purified using a DNeasy minelute kit (QIAGEN) according to the manufacturer's instructions. Clone libraries were constructed using the resulting mixed-template amplicons and the pGEM-T-Easy vector (Promega) following the manufacturer's instructions. Clone sequences were obtained from transformations by plating, rolling-circle amplification and cycle sequencing at the High-Throughput Genomics Unit (University of Washington, Seattle, WA, USA). Clones from each station were assigned library and station prefixes and numbered sequentially from 1 to 96 (GenBank Accession: GU460426-GU461274). Cloned 16S rRNA gene sequences were aligned in the ARB software package (Ludwig *et al.*, 2004) using a custom database that contained 151 952 sequences from cultured organisms and environmental gene clone libraries. Unambiguously aligned nucleotide sequences were added to a custom tree using the parsimony insertion tool for phylogenetic identification available in ARB. Taxonomic assignments were determined by phylogenetic inference.

*Ordination*
Nonmetric multidimensional scaling (NMS) searches for the best positions of *n* entities (samples) on *k* dimensions (axes) that minimize stress of the final configuration. Calculations are on the basis of a distance matrix generated from the main matrix, and stress is measured as the departure from monotonicity between distance in the original space and distance in ordination space. The software package PC-ORD (MjM Software Design, Gleneden Beach, OR, USA) was used for NMS analyses of bacterial 16S rDNA clone and peptide tandem mass spectra count data using Sorensen distance, a random starting configuration, and autopilot slow and thorough options. The main matrices were transformed into relative units by dividing 16S rDNA clones or spectral counts by the total for their corresponding sample. Normal analyses were performed independently on each main matrix, which contained stations in rows and either clone lineages or spectral counts in columns. Dimensionality was determined by assigning categorical variables to South Atlantic western gyre, eastern gyre and coastal locations. The final stress obtained for each two-dimensional solution was < 4.36 with instability near zero. The probability of a similar stress obtained by chance (0.0196) was determined for 40 real and 50 randomized Monte Carlo runs for each ordination. Axes were rotated to maximize orthogonality (> 98%). To evaluate environmental correlations with ordination axes in the clone library joint plot, a second data matrix contained stations in rows and environmental variables (Table 1) in columns. To evaluate lineage correlations with ordination axes in the tandem mass spectra joint plots, a different second data matrix contained stations in rows and lineage variables (transpose of Supplementary Table 1) in columns. Second matrices were transformed into relative units by dividing environmental variables (columns) or lineage counts (rows) by column and row totals, respectively.

# Results

*Shifts in microbial community structure*
We compared microbial communities from stratified nutrient-poor surface waters of the South Atlantic gyre with those from the highly productive Benguela upwelling region off the West Coast of Africa (Figure 1a). A total of 849 16S rRNA genes (85 ± 7 clones from each sample) were sequenced to evaluate bacterial diversity along an oceanic gradient in nutrient concentrations across the South Atlantic (Table 1, Supplementary Table 1). In all, 31 bacterioplankton lineages were identified in 16S rDNA clone libraries. Taxonomic assignments were determined by sequence alignment and phylogenetic inference using a custom ARB database. Genes belonging to the uncultured monophyletic marine *Actinobacteria* clade were recovered in high numbers from both open ocean and coastal seawater samples and were the most abundant category of

**Table 1** Ancillary data from South Atlantic sample sites

| Cruise KN192-5 | Date (dd/mm/yy) | Latitude | Longitude | DAPI (cells ml$^{-1}$) | Temperature (°C) | Salinity (PSU) | DCM (m) | $PO_4$ (μM) | $SiO_4$ (μM) | $NO_2$ (μM) | $NO_3$ (μM) | $NH_4$ (μM) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ST3 | 20/11/07 | 11°30.0′S | 25°00.0′W | 6.3E+05 | 25.0 | 36.9 | 102 | 0.136 | 0.81 | 0.048 | 0.212 | 0.034 |
| ST9 | 25/11/07 | 12°30.0′S | 9°59.9′W | 8.6E+05 | 22.7 | 36.5 | 73 | 0.167 | 0.88 | 0.014 | 0.066 | 0.005 |
| ST13 | 29/11/07 | 13°28.2′S | 0°01.1′W | 8.7E+05 | 21.0 | 36.3 | 64 | 0.242 | 0.69 | 0 | 0.162 | 0 |
| ST19 | 05/12/07 | 14°45.0′S | 12°12.0′E | 1.9E+06 | 21.1 | 35.8 | 15 | 1.022 | 3.58 | 0.24 | 11.96 | 0.203 |
| ST20 | 06/12/07 | 17°29.9′S | 11°15.0′E | 2.4E+06 | 16.1 | 35.4 | 28 | 1.431 | 3.12 | 0.786 | 21.114 | 0.175 |
| ST21 | 07/12/07 | 20°00.0′S | 11°59.9′E | 2.9E+06 | 16.3 | 35.2 | 14 | 1.124 | 1.84 | 0.405 | 12.905 | 0.189 |
| ST23 | 09/12/07 | 25°00.0′S | 14°30.0′E | 2.9E+06 | 13.1 | 34.9 | 9 | 1.792 | 31.48 | 0.31 | 25.61 | 0.4 |
| ST24 | 09/12/07 | 25°00.0′S | 13°30.0′E | 6.6E+05 | 16.6 | 35.0 | 23 | 0.851 | 1.44 | 0.371 | 7.809 | 1.268 |
| ST25 | 10/12/07 | 25°00.0′S | 12°30.0′E | 2.0E+06 | 18.4 | 35.2 | 46 | 0.431 | 1.09 | 0.093 | 2.087 | 0.054 |
| ST27 | 11/12/07 | 24°59.9′S | 10°00.0′E | 1.3E+06 | 19.2 | 35.5 | 62 | 0.231 | 1.59 | 0.002 | 0.288 | 0.081 |

Abbreviation: DAPI, 4′,6-diamidino-2-phenylindole; DCM, depth of the deep chlorophyll maximum; PSU, practical salinity unit.
Data include the nine variables (DAPI through $NH_4$) in the environmental data matrix used for the joint plot in (Figure 1b). Technicon AutoAnalyzer II (Bad Vilbel, FRG) components were used to measure phosphate and ammonium; and Alpkem RFA 300 (Clackamas, OR, USA) components were used for silicic acid, and nitrate plus nitrite and nitrite at the Oregon State University. Nitrate concentrations were determined by subtracting nitrite values from N+N values.
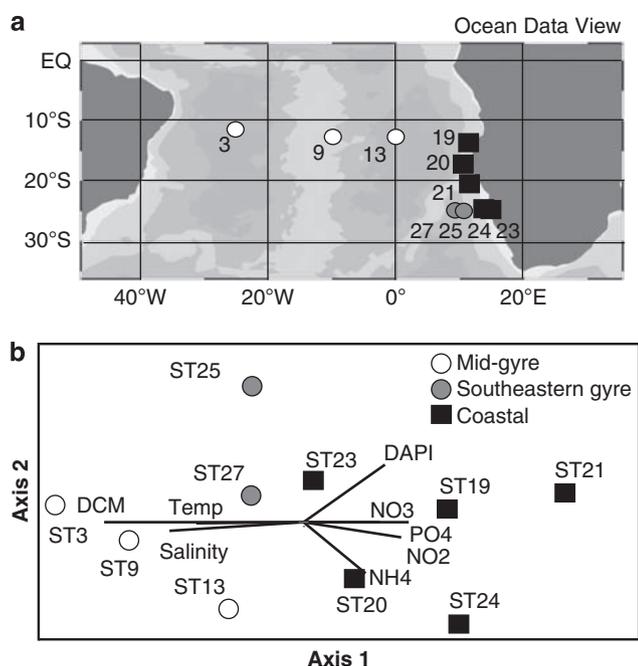


**Figure 1** South Atlantic sample distributions. (**a**) Sample locations of surface seawater collected from five open ocean stations, three mid-gyre and two southeastern gyre, and five coastal stations. (**b**) Sample ordination in a nonmetric multidimensional scaling (NMS) joint plot of cloned 16S rRNA genes. Line lengths along ordination axes in the clone joint plot indicate correlation strengths of measured environmental variables with ordination axes (Table 1).

16S rRNA genes obtained in this study (184 of 849). We also identified 90 *Prochlorococcus* clones (11% of all clones) and 177 SAR11 clones (21% of all clones). The *Prochlorococcus* clones we identified derived only from open ocean sites, whereas SAR11 clones were identified at every site.

NMS was used to examine patterns in community structure across the sample set (Figure 1b, Supplementary Table 1). NMS analyses indicate that microbial community structure shifted along the open ocean to coastal gradient in nutrient concentrations. A community shift along this gradient explained 83% of the variability in cloned lineages (Figure 1b, Axis 1). Clones of *Prochlorococcus* were highly correlated with open ocean stations along ordination Axis 1 (−0.79), defined by higher temperature, salinity and depth of the deep chlorophyll maximum (DCM) and lower concentrations of nutrients ($NO_3$, $NO_2$, $NH_4$, $PO_4$). Clones from lineages typically associated with coastal seawater samples, such as *Bacteroidetes* and *Roseobacter*, were highly correlated with coastal sites along ordination Axis 1 (0.92 and 0.63, respectively), defined by higher concentrations of nutrients ($NO_3$, $NO_2$, $NH_4$, $PO_4$) and lower temperature, salinity and DCM depths. Thus, lineage distributions reflected physical and chemical conditions typical of open ocean and coastal ecosystems.

*Membrane metaproteomics*
We used MS/MS analysis to examine the membrane metaproteome from the same 10 samples in the South Atlantic. Cytoplasmic proteins were not the focus of this study. We identified 5389 peptide spectra (average $539 \pm 232$ per sample) from 2273 distinct proteins ($428 \pm 158$ per sample). We assigned 939 unique annotations using a consensus annotation approach developed for this study (Supplementary Figure 2, Supplementary Tables 2 and 3). Protein identifications were determined at the level of both taxonomic and functional consensus. Protein abundance estimates were determined by summing spectral counts for repeated selections of the same peptide spectra and different peptide spectra with same consensus annotation (Figure 2). The most frequently identified proteins in the South Atlantic microbial membrane metaproteome (all stations) were those related to transport, proteins of unknown function
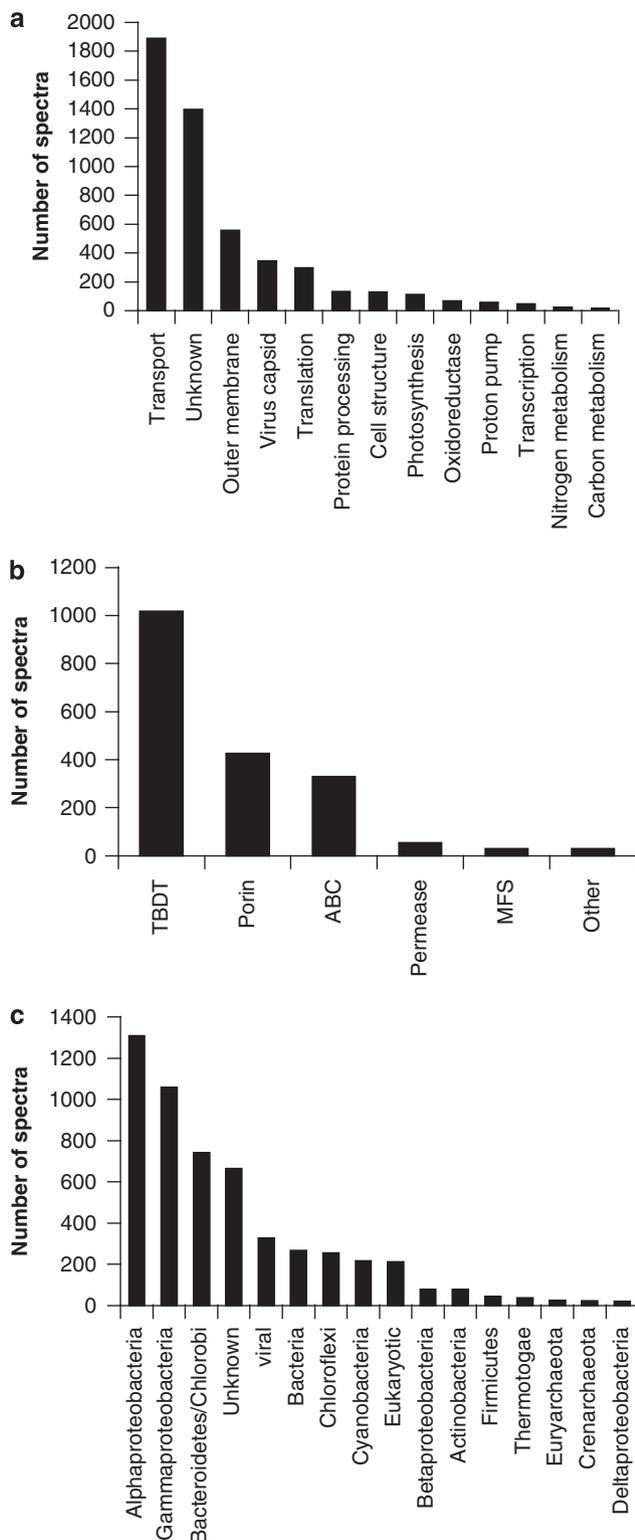
678



**Figure 2** Abundance and distribution of proteins identified in the South Atlantic. (**a**) MS/MS spectra pooled by functional annotation category. (**b**) Transport-related MS/MS spectra pooled by transporter type. (**c**) MS/MS spectra pooled by lineage annotation category.

and uncharacterized outer membrane proteins (Figure 2a). Viral proteins were identified at every site in the South Atlantic, confirming transcript data

that viral infection is ubiquitous in marine microbial communities (Frias-Lopez *et al.*, 2008).

Over 50% (1017 of 1891) of the transport-related peptide tandem mass spectra were from diverse TonB-dependent transporters (TBDTs) (Figure 2b). TBDT systems are known to utilize energy from the cytoplasmic PMF to transport nutrients across the outer membrane of Gram-negative bacteria. Microbial genome sequence data have revealed that TBDT paralogs are enriched in some marine bacterial genomes (Giovannoni and Stingl, 2007). However, the ocean-scale importance of TBDTs was unsuspected. Porins with unknown transport functions and ATP-dependent ABC transporters ranked second and third in transport protein abundance (427 and 330, respectively), followed by permeases and major facilitator superfamily proteins (Figure 2b).

Lineage-specific protein abundance estimates were consistent with abundance estimates determined by 16S rDNA clone library analyses (Figure 2c, Table 1). Proteins most closely related to alpha-proteobacteria, gamma-proteobacteria and Bacteroidetes/Chlorobi were among the most abundant taxa identified. However, we identified far fewer *Actinobacteria* proteins than expected from 16S rDNA clone library data. Proteins of unknown origin ranked fourth in overall abundance (Figure 2c) and some proteins could only be identified at higher taxonomic levels, such as Bacteria. Many of these were highly conserved ribosomal proteins with identical peptides present in sequences from different bacterial phyla (Supplementary Table 3).

*Comparative metaproteomics*
To identify patterns in membrane protein abundance across the 10 sampling sites, we again used NMS analysis. First, we analyzed the 2273 proteins identified from MS/MS spectral counts to examine community function independently of potential biases introduced by the consensus annotation approach (Figure 3a). Relationships between samples based on spectral counts were similar to those based on 16S rDNA abundances (Figure 3a, Figure 1b). Furthermore, 16S rDNA lineage correlations with ordination axes in the NMS joint plot suggest that community structure and function covaried with physical and chemical conditions in the South Atlantic (Figure 3a, Table 1). Sample variability along the open ocean to coastal gradient in membrane proteins explained 45% of the variability in the tandem mass spectra data matrix (Figure 3a, Axis 1A). However, functional annotations were required to compare the types of membrane proteins expressed by bacterioplankton at open ocean and coastal sites.

A nearly identical trend was obtained in a second NMS analysis using the 939 proteins with unique annotations (Figure 3b, Supplementary Table 2). This approach allowed us to compare taxonomic
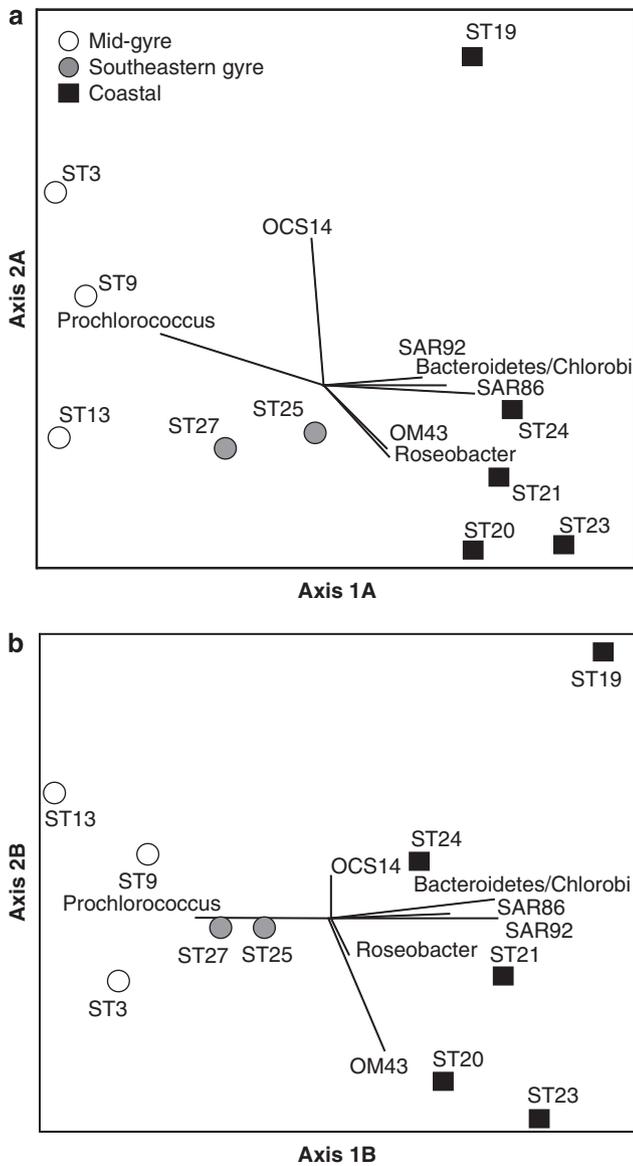
**Figure 3** Sample ordinations in nonmetric multidimensional scaling (NMS) joint plots of MS/MS spectra from proteins identified at open ocean and coastal stations in the South Atlantic. (**a**) Sample ordination of MS/MS spectra from 2273 unannotated proteins identified using the GOS data set. (**b**) Sample ordination of MS/MS spectra from 939 proteins identified and pooled by consensus annotation. Line lengths along ordination axes in the NMS joint plots indicate correlation strengths of 16S rDNA clones from 31 bacterial lineages identified by phylogenetic inference (Supplementary Table 1).

and functional annotations across multiple samples. Correlations with NMS ordination Axis 1B indicated that nutrient transport functions shifted along an open ocean to coastal gradient in nutrient concentrations. TBDT proteins were identified throughout the South Atlantic. The majority of TBDT proteins were positively correlated with ordination Axis 1B, suggesting that TonB transport systems were enriched in microbial membrane fractions obtained from coastal seawater samples (Figure 3b, Supplementary Tables 2 and 3). The

highest positive correlation with ordination Axis 1B was that of a TBDT most closely related to the gamma-proteobacterium *Shewanella baltica* (0.92). TBDT proteins most closely related to gamma-proteobacteria and members of the Bacteroidetes/Chlorobi were correlated with coastal samples along Axis 1B (average = 0.21 ($n = 35$) and 0.3 ($n = 21$), respectively), whereas TBDT proteins most closely related to alpha-proteobacteria and delta-proteobacteria were correlated with samples from the open ocean (average = −0.03 ($n = 17$) and −0.46 ($n = 2$), respectively). Porins most closely related to *Prochlorococcus* and *candidatus Pelagibacter ubique* were among the dominant proteins identified in open ocean membrane fractions, and were negatively correlated with ordination Axis 1B (−0.83 and −0.84, respectively). Of the 16 porins identified, only one was positively correlated with ordination Axis 1B (0.18), indicating that most porins were more highly expressed in the open ocean. Similarly, all proteins annotated as *Prochlorococcus* were negatively correlated with ordination Axis 1B. Thus, shifts in microbial diversity in the South Atlantic were detected by shifts in transport proteins, demonstrating the potential for proteomics to provide insights into environmental conditions that affect lineage and community functions.

Spectral counts were pooled by broad protein annotation and taxonomic categories to evaluate differences between open ocean and coastal communities suggested by NMS ordination analyses (Figures 4a and b, respectively). Spectra from TBDT proteins were relatively more abundant in coastal waters, whereas spectra from porins, permeases and major facilitator superfamily proteins were relatively more abundant in the open ocean (Figure 4a). We identified 218 spectra from cyanobacterial proteins, the majority of which were from open ocean samples (Figure 4b). Spectra from proteins most closely related to members of the alpha-proteobacteria were identified throughout the South Atlantic, whereas a large fraction of the gamma-proteobacteria and Bacteroidetes/Chlorobi were identified at coastal sites. The majority of spectra from 'unknown' proteins, those that could not be assigned a taxonomic and functional annotation because of poor BLAST expect scores or homology to hypothetical proteins, were from the open ocean.

Dominant heterotrophic and autotrophic bacterial lineages exhibited physiological responses to changing nutrient concentrations in the South Atlantic (Supplementary Figure 3). Spatial patterns in the abundance of *Prochlorococcus* and SAR11 proteins support clone abundance estimates (Supplementary Table 1) and provide insights into the physiologies of these dominant lineages in the South Atlantic (Supplementary Figures 3b and 3c, respectively). For example, uncharacterized bacterial porins made up a high percentage of transport proteins expressed by *Prochlorococcus* and SAR11 cells in the open ocean. Although their transport functions are
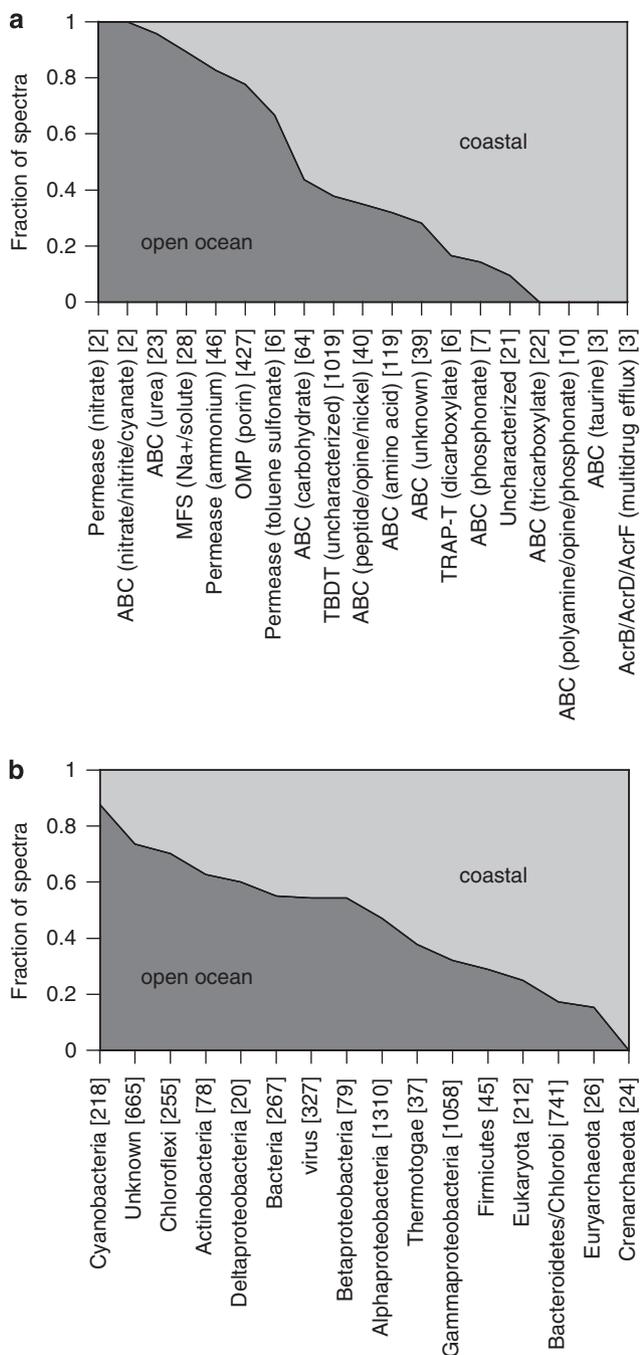
**Figure 4** Fraction of MS/MS spectra from proteins identified at open ocean (stations 3,9,13,25,27) and coastal (stations 19,20,21,23,24) sites. (**a**) Relative abundance and distribution of compound-specific transport proteins. (**b**) Relative abundance and distribution of lineage-specific proteins. Total spectral counts for each category are shown in brackets.



**Figure 5** Archaeal ammonia monooxygenase proteins identified in the Benguela upwelling region. (**a**) Temperature profile with nitrite overlay (contours) plotted in Ocean Data View (Schlitzer, 2002). Nitrite concentrations are in μM (Table 1). Stations as per Figure 1; ordered here by distance from first sampling following the cruise track. (**b**) Tandem mass spectrum of a peptide matching *candidatus Nitrosopumilus maritimus* ammonium monooxygenase A (AmoA) and (**c**) of peptide matching AmoC.

uncertain, they were absent or rare in coastal samples, suggesting that they confer a selective advantage on organisms in oligotrophic systems. Urea ABC transporters and photosystem proteins were among the top five *Prochlorococcus* proteins identified and, with the exception of porins, sodium solute symporters were the most frequently identified SAR11 protein in the open ocean.
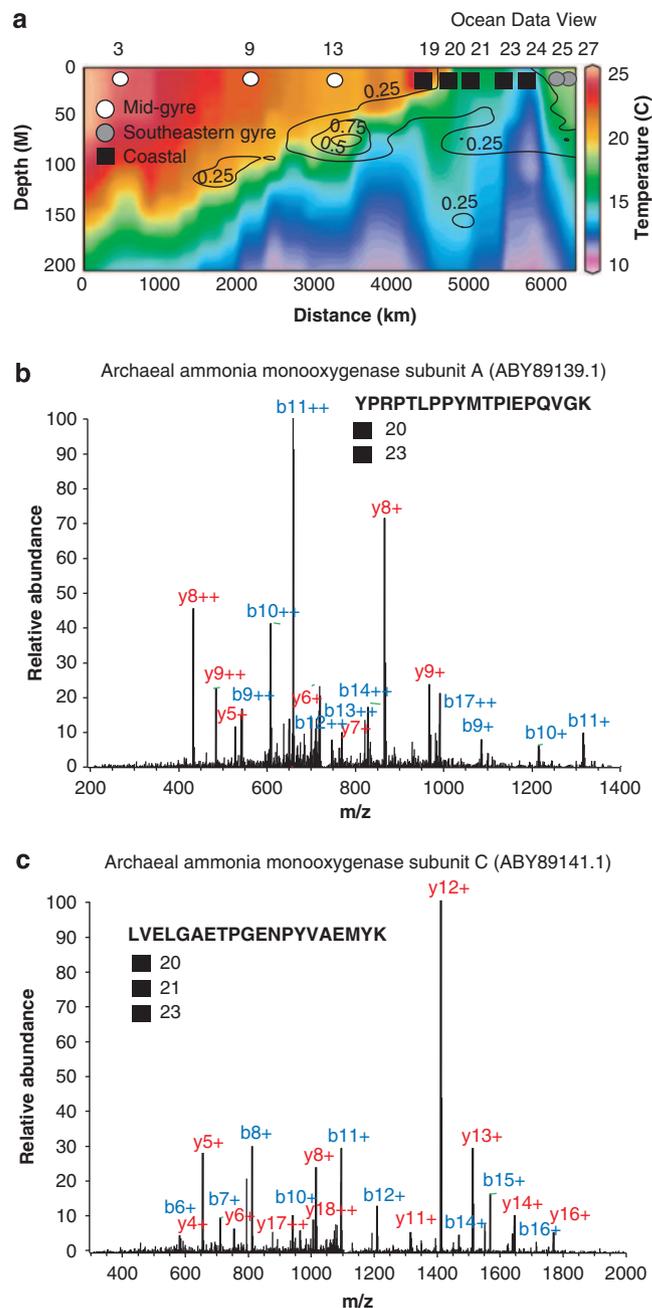
*Nutrient utilization and energy transduction*
We identified proteins from different subunits of archaeal ammonium monooxygenase (Amo), demonstrating the power of metaproteomics to verify archaeal nitrification in the environment. Archaeal Amo proteins from the nitrifying marine
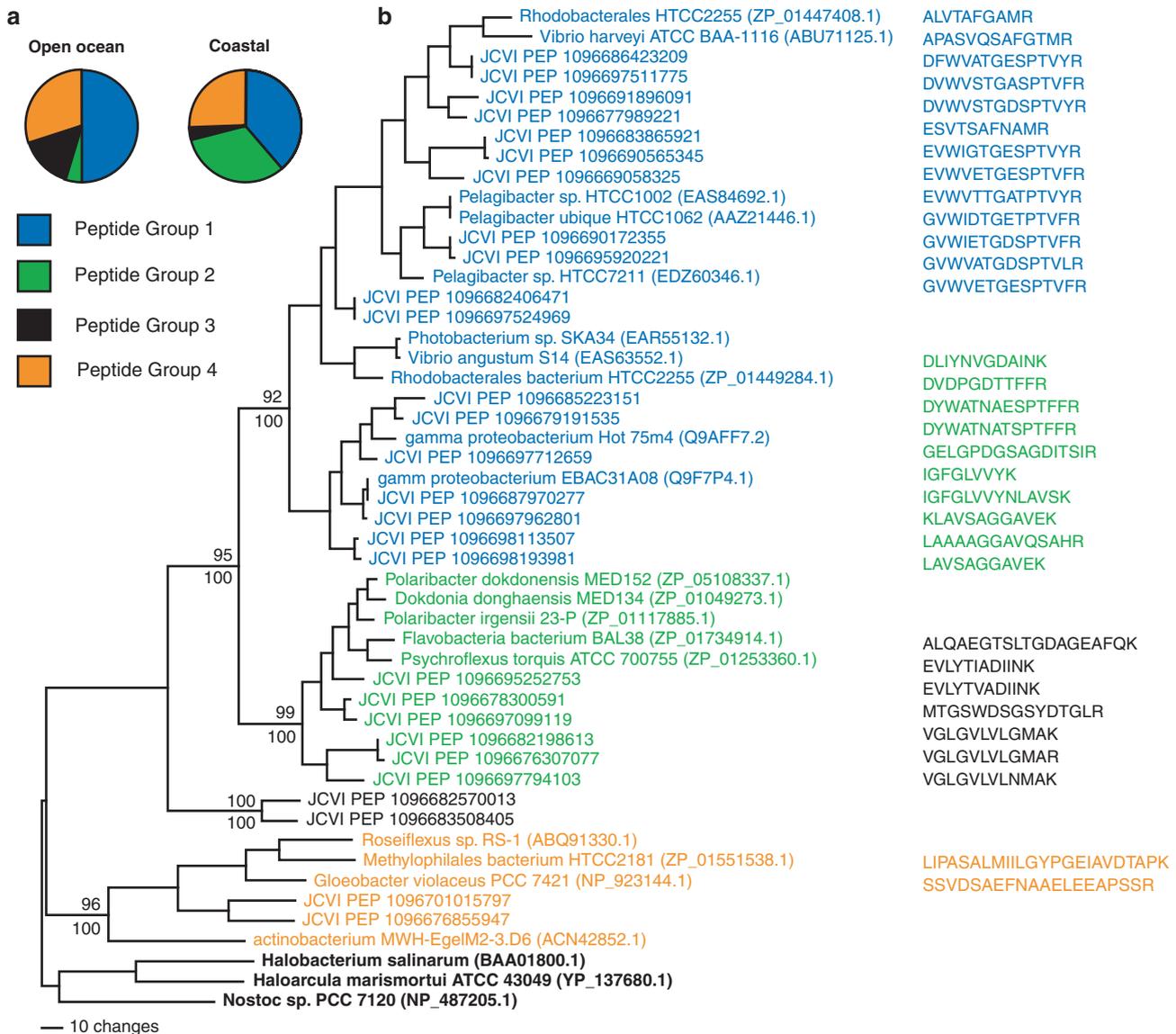
**Figure 6** Rhodopsin protein expression. (**a**) Relative protein abundance estimates for conserved rhodopsin peptide groups identified at open ocean and coastal sites, and (**b**) parsimony tree of rhodopsin sequences identified in the South Atlantic and sequences detected by best Basic Local Alignment Search Tool (BLAST) hits. Parsimony analysis methods were used to identify phylogenetic relationships using 142 parsimony-informative characters using a heuristic search, tree bisection–reconnection, and a starting tree obtained by stepwise addition with random sequence addition with the program PAUP* 4.0 beta (Sunderland, MA, USA). Peptide sequences identified by tandem mass spectra are listed for each protein group (1–4). Partial and redundant GOS proteins were excluded from the analysis and not all GOS proteins in a color group have an exact match to every peptide sequence in that group. Bootstrap values were determined using parsimony (above node) and neighbor-joining (below node) analyses of 100 and 1000 replicates, respectively.

crenarchaeon *candidatus Nitrosopumilus maritimus* were identified at coastal upwelling sites 20, 21 and 23 (Figure 5). Ammonia concentrations at stations where Amo peptides were detected ranged from 0.175–0.4 μM (Table 1).

We also identified four conserved peptide groups from known or suspected light-driven proton pumps (Figure 6). A total of 146 unique GOS protein sequences were detected by proteomic analysis and identified by consensus annotation as rhodopsins. Altogether, 28 representative GOS sequences were aligned with 19 putative rhodopsin sequences from microbial genomes to clarify lineage-specific

patterns of rhodopsin protein expression in the South Atlantic. Sequences most closely related to those found in alpha-proteobacteria and gamma-proteobacteria dominated open ocean profiles (peptide group 1). Their relative contribution decreased from 50% in the open ocean to 39% at coastal stations (Figures 6a). In contrast, sequences most closely related to rhodopsin sequences from *Bacteroidetes* increased from 5% in the open ocean to 32% at coastal sites (peptide group 2). Peptide group 3 consisted of 7 peptides detected in 46 GOS proteins that formed a monophyletic group whose position could not be confidently assigned. Rhodopsins most

closely related to sequences from four different bacterial phyla (*Chloroflexi*, *Proteobacteria*, *Cyanobacteria* and *Actinobacteria*) (peptide group 4) contributed to a significant fraction of rhodopsins at both open ocean and coastal sites (30% and 26%, respectively). This report is the first to verify *in situ* protein expression of light-driven proton pumps outside the alpha-proteobacterial SAR11 clade.

## Discussion

*Marine membrane metaproteomics*
We detected $428 \pm 158$ distinct membrane proteins per sample, fewer than the 1042 proteins from the SAR11 *Prochlorococcus* and *Synechococcus* lineages detected in a previous whole-cell metaproteomic study in the Sargasso Sea (Sowell *et al.*, 2009). This is, in part, because we limited our focus to the membrane fraction, and also because of our higher stringency criteria for peptide identification. In addition, the low biomass and extensive genetic diversity in the marine environment precluded protein coverage that would be typical of samples comprised of relatively few dominant organisms. Even in a much deeper (60 MS/MS runs on a single sample) proteomic survey of bacterioplankton in Puget Sound, only 238 out of 3639 proteins (6.5%) were identified by more than one peptide (R. Morris, unpublished data). These data highlight the need for metaproteomic data to be analyzed in an ecological context. Our approach used multivariate statistics to identify the dominant trends in protein expression from field replicates. Using this approach, we obtained novel insights into the structure and function of microbial communities in the South Atlantic.

In general, distributions of membrane proteins from dominant bacterial lineages were congruent with known patterns of lineage abundance and reflected physical and chemical conditions typical of open ocean and coastal ecosystems (Supplementary Figure 3). However, we identified far fewer marine *Actinobacteria* proteins than predicted from 16 rDNA clone library data. The percentage of marine *Actinobacteria* clones in our libraries (made from cell concentrates rather than whole water) is much higher than previously reported (Jensen and Lauro, 2008). This may be due to enrichment during cell concentration, improved cell lysis or relative differences in membrane protein profiles. Regardless of the reason, our data suggest that members of the marine *Actinobacteria* clade were abundant in our proteomic samples and that protein identifications may have been limited because of sequence divergence from cultured representatives. The majority of the *Actinobacteria* clones detected here belong to the uncultured marine *Actinobacteria* cluster, and share only 81% 16S rRNA identity with *Kribbella flavida*, the closest relative with a genome

sequence. However, phylogenetic analysis of rhodopsins suggests that sequences identified by peptide group 4 may have derived from uncultured marine *Actinobacteria* (Figure 6).

*Light energy transduction*
Genes for light-driven proton pumps (proteorhodopsins) are ubiquitous in the oceans and have been identified in diverse marine Bacteria and Archaea (de la Torre *et al.*, 2003; Frigaard *et al.*, 2006). Proteorhodopsin gene transcripts from diverse microbial taxa were detected in the North Pacific subtropical gyre, indicating their important role in marine ecosystems (Frias-Lopez *et al.*, 2008). However, *in situ* protein expression has only been verified in coastal representatives from the SAR11 clade (Giovannoni *et al.*, 2005). Proteorhodopsins are thought to confer a competitive advantage under low-nutrient conditions by providing cells with a way to generate energy from sunlight. Cultured members of the alpha-proteobacteria and gamma-proteobacteria that have the proteorhodopsin gene do not exhibit increased growth rates or final cell densities under light versus dark growth conditions. However, a *Bacteroidetes* isolate containing proteorhodopsin did show increased cell yields when grown in the light (Gomez-Consarnau *et al.*, 2007). These observations, coupled with the lineage-specific geographic patterns of rhodopsin expression seen here, suggest that the PMF of light-driven proton pumps may serve different functions in different cells.

*Mechanical energy transduction*
TBDTs accounted for 19% of all tandem mass spectra identified in this study and were identified in every sample. In addition, cytoplasmic transmembrane components of the TonB complex (MotA/TolQ/ExbB) were identified in all coastal samples and in two open ocean samples. TonB complexes are known to transport essential compounds across the outer membrane of Gram-negative bacteria using the cytoplasmic membrane PMF (Nikaido, 2003). This is termed mechanical energy transduction because energy from PMF causes a conformational shift in the structure of TonB. TBDT activities were once thought to be restricted to iron complexes (siderophores) and vitamin $B_{12}$ (cobalamin). Recent experimental and bioinformatic studies indicate that nickel, cobalt, copper, maltodextrins, sucrose, thiamin and chito-oligosaccharides are also substrates for TBDTs (Schauer *et al.*, 2008). It can be noted that the *Bacteroides thetaiotaomicron* genome contains over 120 TBDTs, suggesting that the potential range of substrates can be quite broad. At present, we are unable to ascertain the substrate specificity of the distinct TBDT proteins identified here, but their ubiquitous identification in our samples suggests that TBDT activities are an

important mechanism for microbial nutrient acquisition in diverse ecosystems across a broad swath of the South Atlantic.

Basin-scale metaproteomics revealed ubiquitous TBDT and rhodopsin energy transduction processes. Indeed, in some cases, we infer that both proteins were found in the same taxonomic lineage (Supplementary Figure 4). TBDT transport of nickel, maltodextrin and sucrose does not require additional ABC transporter activities to cross the cytoplasmic membrane, suggesting that the PMF is the sole energy source for TonB-activated transport of some compounds (Neugebauer et al., 2005; Blanvillain et al., 2007; Schauer et al., 2007). These observations, coupled with the lineage-specific geographic patterns of rhodopsin expression seen here, suggest that the light-induced PMF may serve functions other than ATP synthesis (Fuhrman et al., 2008).

*Archaeal nitrification*
*Crenarchaea* are known to dominate microbial communities in the twilight zone (~200–1000 m), where they can account for up to 39% of bacterioplankton communities and have important roles in global carbon and nitrogen cycles (Karner et al., 2001; Konneke et al., 2005; Ingalls et al., 2006). Metagenomic studies first revealed a novel gene for ammonia monooxygenase (*amoA*) on an archaeal scaffold, suggesting that some archaea were capable of carrying out the first step in nitrification, the oxidation of ammonia to nitrite (Venter et al., 2004; Treusch et al., 2005). Since then, archaeal *amoA* genes have been detected repeatedly in the environment (Francis et al., 2005; Hallam et al., 2006; Wuchter et al., 2006) and the transcript has been detected in the Peruvian oxygen minimum zone (Lam et al., 2009). However, the importance and extent of archaeal nitrification in the ocean is still under debate (Agogue et al., 2008). The abundance of archaeal nitrification peptides we detected suggests that Archaea can be important nitrifiers in nutrient-rich surface waters.

*Nutrient acquisition*
Transport proteins for urea and ammonia dominated known transport processes in the open ocean (Figure 4a). In contrast, phosphate transport proteins were not detected in the South Atlantic microbial membrane metaproteome. Membrane-associated SAR11, *Prochlorococcus* and *Synechococcus* phosphate transport proteins were identified in a previous metaproteomic study of the subtropical Sargasso Sea (Sowell et al., 2009). Using methods described here, we have detected phosphate transport proteins in the membrane fraction of cultured *Prochlorococcus* (G. Rocap, unpublished data), suggesting that methodological differences in protein extraction and detection approaches

between the two field studies do not account for the discrepancy. Measured phosphate concentrations from the South Atlantic gyre were an order of magnitude higher (0.1–0.4 μM) (Supplementary Table 1) than late summer surface concentrations typical of the subtropical Sargasso Sea (Wu et al., 2000). This suggests that proteomic detection of transport activities can provide insight into the nutrient status of cells in the environment.

*Conclusions*
Applying a function-based comparative environmental proteomics approach on a larger geographical scale than in previous studies and linking results to phylogenetic diversity revealed shifts in nutrient utilization and energy transduction along a natural gradient in nutrient concentrations across the South Atlantic. The dominance of TBDT proteins in the South Atlantic microbial membrane metaproteome and identification of rhodopsins from diverse bacterioplankton lineages present in the open ocean and in coastal waters were notable findings. Lineage-specific contributions of rhodopsins and TBDTs throughout the South Atlantic imply that light and mechanical energy transduction are ubiquitous features benefiting bacterioplankton even as conditions shift. We hypothesize that some cells may have the ability to couple a light-induced PMF with ATP-independent nutrient acquisition. Cells with this combination of capabilities would have a selective advantage in open ocean and many other marine ecosystems where energy and nutrients are limiting. Future studies targeting the marine microbial metaproteome in different oceanic regimes or seasons are likely to provide further insights into the biogeography and functional diversity of these and other microbial processes. In particular, the mechanisms that drive lineage and community functions in the oceans can be correlated with environmental changes.

Our analyses provide several lessons for the future of marine metagenomics and metaproteomics. First, until *de novo* methods for inferring peptide sequence advance significantly, a database that captures the genomic diversity of the community is essential to identify proteins from environmental tandem mass spectra. For example, when tandem mass spectra from this sample suite were searched with the GOS metagenomic library, 6.2 times as many peptides were identified when compared with searches with the GenBank nonredundant database (Supplementary Table 4). The success of similar proteomic studies in other regimes, such as the deep sea, will depend on generation of environment-specific metagenomes. Second, continued efforts to bring additional microbial lineages into culture are critical. Half of the most abundant bacterial lineages identified here were isolated in the last decade using high-throughput cultivation techniques (Supplementary Figure 3a) (Connon and

684

Giovannoni, 2002; Rappe *et al.*, 2002; Cho and Giovannoni, 2004; Lee *et al.*, 2007). Finally, the presence of proteins of unknown function is consistent with a study of microbial community gene expression in the North Pacific gyre, which reported a high number of hypothetical gene transcripts (Frias-Lopez *et al.*, 2008). Additional genome-informed physiological and genetic experiments are required to discover the function of these proteins. Ultimately, a combination of continued sequencing, cultivation and functional genomics will be needed to fully understand the biogeochemical roles of marine microbial communities.

## Acknowledgements

## References

Agogue H, Brink M, Dinasquet J, Herndl GJ. (2008). Major gradients in putatively nitrifying and non-nitrifying Archaea in the deep North Atlantic. *Nature* **456**: 788–791.

Blanvillain S, Meyer D, Boulanger A, Lautier M, Guynet C, Denance N *et al.* (2007). Plant carbohydrate scavenging through tonb-dependent receptors: a feature shared by phytopathogenic and aquatic bacteria. *PLoS ONE* **2**: e224.

Chen JZ, Ryu SY, Gharib SA, Goodlett DR, Schnapp LM. (2008). Exploration of the normal human bronchoalveolar lavage fluid proteome. *Proteomics Clinical Applications* **2**: 585–595.

Cho JC, Giovannoni SJ. (2004). Cultivation and growth characteristics of a diverse group of oligotrophic marine Gammaproteobacteria. *Appl Environ Microbiol* **70**: 432–440.

Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, Delong EF *et al.* (2006). Genomic islands and the ecology and evolution of Prochlorococcus. *Science* **311**: 1768–1770.

Connon SA, Giovannoni SJ. (2002). High-throughput methods for culturing microorganisms in very-low-nutrient media yield diverse new marine isolates. *Appl Environ Microbiol* **68**: 3878–3885.

de la Torre JR, Christianson LM, Beja O, Suzuki MT, Karl DM, Heidelberg J *et al.* (2003). Proteorhodopsin genes are distributed among divergent marine bacterial taxa. *Proc Natl Acad Sci USA* **100**: 12830–12835.

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

Eng JK, Mccormack AL, Yates JR. (1994). An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. *J Am Soc Mass Spectrom* **5**: 976–989.

Francis CA, Roberts KJ, Beman JM, Santoro AE, Oakley BB. (2005). Ubiquity and diversity of ammonia-oxidizing archaea in water columns and sediments of the ocean. *Proc Natl Acad Sci USA* **102**: 14683–14688.

Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW *et al.* (2008). Microbial community gene expression in ocean surface waters. *Proc Natl Acad Sci USA* **105**: 3805–3810.

Frigaard NU, Martinez A, Mincer TJ, DeLong EF. (2006). Proteorhodopsin lateral gene transfer between marine planktonic Bacteria and Archaea. *Nature* **439**: 847–850.

Fuhrman JA, Schwalbach MS, Stingl U. (2008). Proteorhodopsins: an array of physiological roles? *Nat Rev Microbiol* **6**: 488–494.

Fung JM, Morris RM, Adrian L, Zinder SH. (2007). Expression of reductive dehalogenase genes in Dehalococcoides ethenogenes strain 195 growing on tetrachloroethene, trichloroethene, or 2,3-dichlorophenol. *Appl Environ Microbiol* **73**: 4439–4445.

Giovannoni S, Stingl U. (2007). The importance of culturing bacterioplankton in the 'omics' age. *Nat Rev Microbiol* **5**: 820–826.

Giovannoni SJ, Bibbs L, Cho JC, Stapels MD, Desiderio R, Vergin KL *et al.* (2005). Proteorhodopsin in the ubiquitous marine bacterium SAR11. *Nature* **438**: 82–85.

Gomez-Consarnau L, Gonzalez JM, Coll-Llado M, Gourdon P, Pascher T, Neutze R *et al.* (2007). Light stimulates growth of proteorhodopsin-containing marine Flavobacteria. *Nature* **445**: 210–213.

Hallam SJ, Mincer TJ, Schleper C, Preston CM, Roberts K, Richardson PM *et al.* (2006). Pathways of carbon assimilation and ammonia oxidation suggested by environmental genomic analyses of marine Crenarchaeota. *PLoS Biol* **4**: e95.

Ingalls AE, Shah SR, Hansman RL, Aluwihare LI, Santos GM, Druffel ER *et al.* (2006). Quantifying archaeal community autotrophy in the mesopelagic ocean using natural radiocarbon. *Proc Natl Acad Sci USA* **103**: 6442–6447.

Jensen PR, Lauro FM. (2008). An assessment of actino-bacterial diversity in the marine environment. *Antonie Van Leeuwenhoek* **94**: 51–62.

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

Konneke M, Bernhard AE, de la Torre JR, Walker CB, Waterbury JB, Stahl DA. (2005). Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature* **437**: 543–546.

Lam P, Lavik G, Jensen MM, van de Vossenberg J, Schmid M, Woebken D *et al.* (2009). Revising the nitrogen cycle in the Peruvian oxygen minimum zone. *Proc Natl Acad Sci USA* **106**: 4752–4757.

Lee K, Choo YJ, Giovannoni SJ, Cho JC. (2007). Maritimibacter alkaliphilus gen. nov., sp. nov., a genome-sequenced marine bacterium of the Roseobacter clade in the order Rhodobacterales. *Int J Syst Evol Microbiol* **57**: 1653–1658.

Lo I, Denef VJ, Verberkmoes NC, Shah MB, Goltsman D, DiBartolo G *et al.* (2007). Strain-resolved community

proteomics reveals recombining genomes of acidophilic bacteria. *Nature* **446**: 537–541.

Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar *et al.* (2004). ARB: a software environment for sequence data. *Nucleic Acids Res* **32**: 1363–1371.

Morris RM, Fung JM, Rahm BG, Zhang S, Freedman DL, Zinder SH *et al.* (2007). Comparative proteomics of Dehalococcoides spp. reveals strain-specific peptides associated with activity. *Appl Environ Microbiol* **73**: 320–326.

Morris RM, Rappe MS, Connon SA, Vergin KL, Siebold WA, Carlson CA *et al.* (2002). SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **420**: 806–810.

Morris RM, Sowell S, Barofsky D, Zinder S, Richardson R. (2006). Transcription and mass-spectroscopic proteomic studies of electron transport oxidoreductases in Dehalococcoides ethenogenes. *Environ Microbiol* **8**: 1499–1509.

Neugebauer H, Herrmann C, Kammer W, Schwarz G, Nordheim A, Braun V. (2005). ExbBD-dependent transport of maltodextrins through the novel MalA protein across the outer membrane of Caulobacter crescentus. *J Bacteriol* **187**: 8300–8311.

Nikaido H. (2003). Molecular basis of bacterial outer membrane permeability revisited. *Microbiol Mol Biol Rev* **67**: 593–656.

Nunn BL, Aker JR, Shaffer SA, Tsai YH, Strzepek RF, Boyd PW *et al.* (2009). Deciphering diatom biochemical pathways via whole-cell proteomics. *Aquat Microb Ecol* **55**: 241–253.

Nunn BL, Shaffer SA, Scherl A, Gallis B, Wu M, Miller SI *et al.* (2006). Comparison of a Salmonella typhimurium proteome defined by shotgun proteomics directly on an LTQ-FT and by proteome pre-fractionation on an LCQ-DUO. *Brief Funct Genomic Proteomic* **5**: 154–168.

Nunn BL, Timperman AT. (2007). Marine proteomics. *Mar Ecol Prog Ser* **332**: 281–289.

Ram RJ, VerBerkmoes NC, Thelen MP, Tyson GW, Baker BJ, Blake RC *et al.* (2005). Community proteomics of a natural microbial biofilm. *Science* **308**: 1915–1920.

Rappe MS, Connon SA, Vergin KL, Giovannoni SJ. (2002). Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature* **418**: 630–633.

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.

Ryu S, Gallis B, Goo YA, Shaffer SA, Radulovic D, Goodlett DR. (2008). Comparison of a label-free quantitative proteomic method based on Peptide ion current area to the isotope coded affinity tag method. *Cancer Inform* **6**: 243–255.

Schauer K, Gouget B, Carriere M, Labigne A, de Reuse H. (2007). Novel nickel transport mechanism across the bacterial outer membrane energized by the TonB/ExbB/ExbD machinery. *Mol Microbiol* **63**: 1054–1068.

Scherl A, Shaffer SA, Taylor GK, Kulasekara HD, Miller SI, Goodlett DR. (2008). Genome-specific gas-phase fractionation strategy for improved shotgun proteomic profiling of proteotypic peptides. *Anal Chem* **80**: 1182–1191.

Schauer K, Rodionov DA, de Reuse H. (2008). New substrates for TonB-dependent transport: do we only see the 'tip of the iceberg'? *Trends Biochem Sci* **33**: 330–338.

Schlitzer R. (2002). Interactive analysis and visualization of geoscience data with Ocean Data View. *Comput Geosci* **28**: 1211–1218.

Sowell SM, Wilhelm LJ, Norbeck AD, Lipton MS, Nicora CD, Barofsky DF *et al.* (2009). Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea. *ISME J* **3**: 93–105.

Stapels MD, Cho JC, Giovannoni SJ, Barofsky DF. (2004). Proteomic analysis of novel marine bacteria using MALDI and ESI mass spectrometry. *J Biomol Tech* **15**: 191–198.

Treusch AH, Leininger S, Kletzin A, Schuster SC, Klenk HP, Schleper C. (2005). Novel genes for nitrite reductase and Amo-related proteins indicate a role of uncultivated mesophilic crenarchaeota in nitrogen cycling. *Environ Microbiol* **7**: 1985–1995.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM *et al.* (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**: 37–43.

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

VerBerkmoes NC, Denef VJ, Hettich RL, Banfield JF. (2009). Systems biology: functional analysis of natural microbial consortia using community proteomics. *Nat Rev Microbiol* **7**: 196–205.

Wilmes P, Wexler M, Bond PL. (2008). Metaproteomics provides functional insight into activated sludge wastewater treatment. *PLoS One* **3**: e1778.

Wu J, Sunda W, Boyle EA, Karl DM. (2000). Phosphate depletion in the western North Atlantic Ocean. *Science* **289**: 759–762.

Wuchter C, Abbas B, Coolen MJ, Herfort L, van Bleijswijk J, Timmers P *et al.* (2006). Archaeal nitrification in the ocean. *Proc Natl Acad Sci USA* **103**: 12317–12322.

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)