environmental microbiology

Environmental Microbiology (2017) **00**(00), 00–00 doi:10.1111/1462-2920.13937

Metaproteomics of marine viral concentrates reveals key viral populations and abundant periplasmic proteins in the oligotrophic deep chlorophyll maximum of the South China Sea

Zhang-Xian Xie,[1†] Feng Chen,[2†] Shu-Feng Zhang,[1] Ming-Hua Wang,[1] Hao Zhang,[1] Ling-Fen Kong,[1] Min-Han Dai,[1] Hua-Sheng Hong,[1] Lin Lin[1] and Da-Zhi Wang[1*]

[1]*State Key Laboratory of Marine Environmental Science/College of the Environment and Ecology, Xiamen University, Xiamen, 361005, China.*

[2]*Institute of Marine and Environmental Technology, University of Maryland Center for Environmental Science, Baltimore, MD, USA.*

Summary

Viral concentrates (VCs), containing bioinformative DNA and proteins, have been used to study viral diversity, viral metagenomics and virus–host interactions in natural ecosystems. Besides viruses, VCs also contain many noncellular biological components including diverse functional proteins. Here, we used a shotgun proteomic approach to characterize the proteins of VCs collected from the oligotrophic deep chlorophyll maximum (DCM) of the South China Sea. Proteins of viruses infecting picophytoplankton, that is, cyanobacteria and prasinophytes, and heterotrophic bacterioplankton, such as SAR11 and SAR116, dominated the viral proteome. Almost no proteins from RNA viruses or known gene transfer agents were detected, suggesting that they were not abundant at the sampling site. Remarkably, nonviral proteins made up about two thirds of VC proteins, including overwhelmingly abundant periplasmic transporters for nutrient acquisition and proteins for diverse cellular processes, that is, translation, energy metabolism and one carbon metabolism. Interestingly, three 56 kDa selenium-binding proteins putatively involved in peroxide reduction from gammaproteobacteria were abundant in the VCs, suggesting active removal of peroxide compounds at DCM. Our study demonstrated that metaproteomics provides a valuable avenue to explore the diversity and structure of the viral community and also the pivotal biological functions affiliated with microbes in the natural environment.

Received 10 May, 2017; revised 15 September, 2017; accepted 16 September, 2017. *For correspondence. E-mail dzwang@xmu.edu.cn; Tel. 86-592-2186016; Fax 86-592-2180655. †These authors contributed equally to this work.

Introduction

The ecological significance of viruses in the marine environment has been addressed for more than two decades (Bergh *et al.*, 1989; Fuhrman, 1999; Wommack and Colwell, 2000; Suttle, 2005). Although viral particles are abundant ($\sim 10^7$ ml^{-1}) in surface seawater, concentrates are often needed to obtain sufficient viral material (i.e., viral genomic DNA) for studying viral ecology, and a variety of concentrating methods have been developed (Suttle *et al.*, 1991; Colombet *et al.*, 2007; Wommack *et al.*, 2009; John *et al.*, 2011; Hurwitz *et al.*, 2013). The tangential flow ultrafiltration method is commonly used to concentrate viruses to facilitate viral isolation (Suttle *et al.*, 1991; Thurber *et al.*, 2009), viral detection (Chen *et al.*, 1996; Short and Suttle, 2002) and the study of virus–microbe interactions (Weinbauer and Peduzzi, 1995; Hewson *et al.*, 2001; Winter *et al.*, 2004). Viral concentrates (VCs) also provide a high yield of nucleic acids for community genome sequencing, and several studies unveil an unexpected genetic diversity of viruses in marine environments (Breitbart *et al.*, 2002; Angly *et al.*, 2006; Williamson *et al.*, 2012; Hurwitz and Sullivan, 2013; Mizuno *et al.*, 2013).

In addition to viral particles, marine VCs contain other noncellular bioactive components that are ecologically important. Viral metagenomics has demonstrated the presence of abundant cellular sequences in VCs, even after treatment with DNase to remove free DNA (Williamson *et al.*, 2012; Hurwitz and Sullivan, 2013; Mizuno *et al.*, 2013). One of the possibilities is that some DNA may be protected by virus-like structures or bounded to other dissolved materials (Jiang and Paul, 1995; Brum, 2005),

which prevent them from being digested by DNase. Recent studies indicate that some virus-size structures, such as gene transfer agents (GTAs) and membrane vesicles (MVs) can be generated by marine bacteria (Biers *et al.*, 2008; Biller *et al.*, 2014). These nonviral particles may be abundant in the ocean (Biller *et al.*, 2014) and play important roles in biogeochemical cycling, microbial evolution and phage resistance (McDaniel *et al.*, 2010; Lang *et al.*, 2012; Biller *et al.*, 2014; Scanlan, 2014; Soler *et al.*, 2015). However, our knowledge of these nonviral 'dark' components in the VCs is still very poor owing to the inherent limitations of current methods (i.e., epifluorescence microscopy or metagenomics) to characterize natural nonviral particles (Forterre *et al.*, 2013; Soler *et al.*, 2015).

Proteins are one of the bio-informative high molecular weight materials present in the VCs. They can be from viral and virus-like particles, or from the cell detritus as a result of viral infection or grazer sloppy feeding, or from extracellular excretion such as MVs and individual enzymes. Metaproteomics is a powerful culture-independent approach that characterizes the expressed proteins in natural microbial communities. Recently, the availability of extensive metagenomic sequences from diverse marine environments (Rusch *et al.*, 2007; Yooseph *et al.*, 2007; Brum *et al.*, 2015) facilitates metaproteomic studies in the ocean (Morris *et al.*, 2010; Sowell *et al.*, 2011; Dong *et al.*, 2013; Georges *et al.*, 2014; Brum *et al.*, 2016). However, up to now, only one metaproteomic report of VCs has been published, in which five abundant protein clusters (PCs) predicted to be viral capsid proteins are the most interesting finding (Brum *et al.*, 2016). In addition, metaproteomics can be complementary to DNA-based viral metagenomics for exploring the diversity of natural VCs, where some types of viruses such as RNA viruses are missing. There have been few proteomic studies to investigate the origin, composition and function of noncellular entities in the VCs.

Here, we characterized the VC proteins (the fraction between 10 kDa and 0.2 μm) collected from the deep chlorophyll maximum (DCM) in the oligotrophic South China Sea (SCS) using a shotgun proteomic approach. Oligotrophic regions cover a large part of the ocean surface. In most oligotrophic oceans, the DCM is common either seasonally or permanently. Frequent viral infection events and close microbial interactions are expected because of the high productivity and microbial biomass in this zone of the water column. We found that among the 39% of the spectra identified in DCM VC samples, many originated from pelagiphages (SAR11 viruses), cyanophages (cyanobacterial viruses) and viruses infecting picophytoeukaryotes. RNA viruses or GTA related proteins were seldom detected. Periplasmic proteins, including substrate binding proteins associated with nutrient transport and the predicted selenium containing proteins, were abundant. Our results provide the first view of the diverse and complex biological origins of proteins in VCs.

Results and discussion

A combination of the local DCM bacterial metagenomic (SEATS-DCM-Bac, size above 0.2 μm) dataset, protein sequences of RNA viruses from the National Center for Biotechnology Information (NCBI) and environmental sequences of dsDNA viruses from the Pacific Ocean (Hurwitz and Sullivan, 2013) and Mediterranean Sea (Mizuno *et al.*, 2013) was used to interpret the complex peptide mixture. Both 3161 and 2836 unique peptides were identified from the two independent VC samples from the DCM of the SCS. After further analysis using a 1% cutoff false discovery rate and the at least two unique peptides matching criterion, a total of 636 nonredundant proteins matching 7220 spectra were identified and 75% of the proteins were shared between the two VC samples (Supporting Information Table S1). The percentage of nonredundant proteins shared between the two VC samples ranged from 64% to 90% in each superkingdom (Supporting Information Table S1). Spectral counts were normalized within samples for protein semi-quantitative analysis. The protein profile patterns based on this quantification between the two VC samples varied a little (Figs 1 and 2). Therefore, all nonredundant proteins were used for further analysis.

Viral proteins

Taxonomic annotation was based mainly on the top BLASTp hit against the NCBI nonredundant database followed by a manual examination. An average of 39% of spectra from the two VCs (Fig. 1A) was assigned to viral proteins, representing 234 out of the total 636 proteins identified (Supporting Information Table S2). More than 50% of the viral spectra were of phage origin (Fig. 1B). As the taxonomic distribution observed in a previous study (Brum *et al.*, 2016), myoviruses, podoviruses and siphoviruses were the major viral families, accounting for 21%, 15% and 3% of the viral spectra (Fig. 1B). Viruses infecting *Synechococcus*, *Prochlorococcus, Pelagibacter* and picophytoeukaryotes contributed approximately 12%, 12%, 6% and 9% of total viral spectra, respectively (Fig. 1C), suggesting that these viruses were abundant in the SCS. This result was consistent with the presence of abundant cyanophages, pelagiphages and Phycodnaviridae in the photic zone of the ocean (Mizuno *et al.*, 2013; Zhao *et al.*, 2013), where host populations are abundant (Supporting Information Table S3) (Zhang *et al.*, 2014). In contrast, only a small proportion of proteins originated from the viruses infecting the Euryarchaeota and Actinobacteria. Around 16% of viral spectra matched homologous proteins

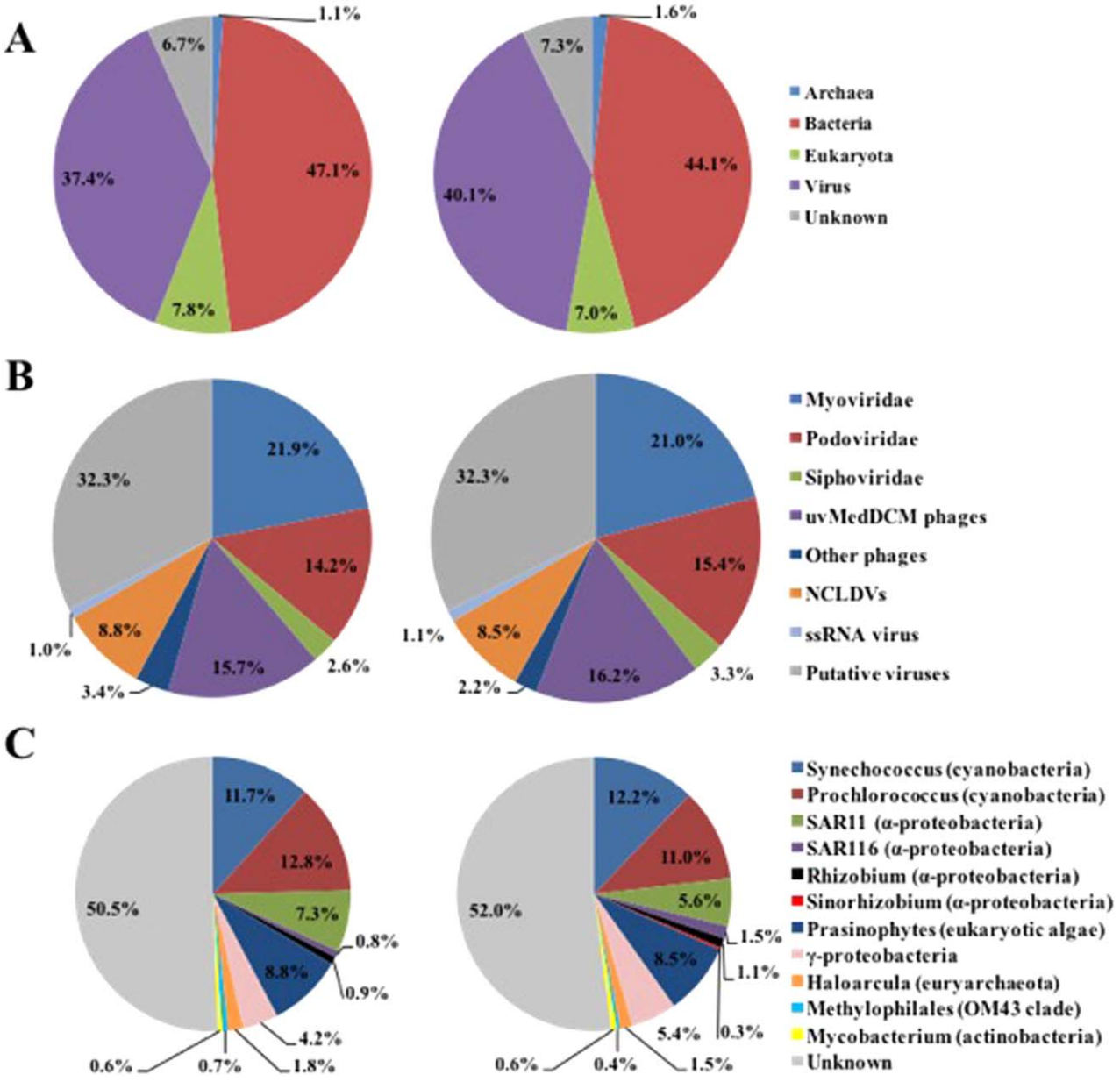

Fig. 1. The relative abundance distribution of viral proteins based on normalized spectral counts in the two VCs collected from the DCM of the SCS. Classified by superkingdoms (A), and viral proteins by major groups (B) and by their host (C). uvMedDCM phages: viral proteins assigned to uncultured phages collected from the Mediterranean DCM layer (Mizuno *et al*., 2013). NCLDV: nucleo-cytoplasmic large DNA viruses most of which were viruses infecting prasinophytes.

from 40 viral contigs from the DCM of the Mediterranean Sea (Fig. 1B), suggesting that these unknown viruses might be related to DCM habitat. Moreover, many of the viral proteins (represented by 32% of the viral spectra) were recruited from bacterial genome segments (Fig. 1B), suggesting that they were from phages integrated into host cells. In addition, two proteins related to cellular metabolism were detected based on putative viral contigs annotated using VirSorter (Supporting Information Table S2), which might be auxiliary metabolic genes expressed in the infected cells but released to the VCs after lysis.

Only one sequence related to an ssRNA virus (Fig. 1B) was observed, although RNA viral reference sequences accounted for 1/3 of the reference database. Furthermore, none of the sequences related to RNA viruses or known GTAs were detected, even with searching against a second database combined SEATS-DCM-Bac dataset with a marine RNA viral metagenomic dataset (Steward *et al*., 2013) and known GTA sequences from NCBI. At present,

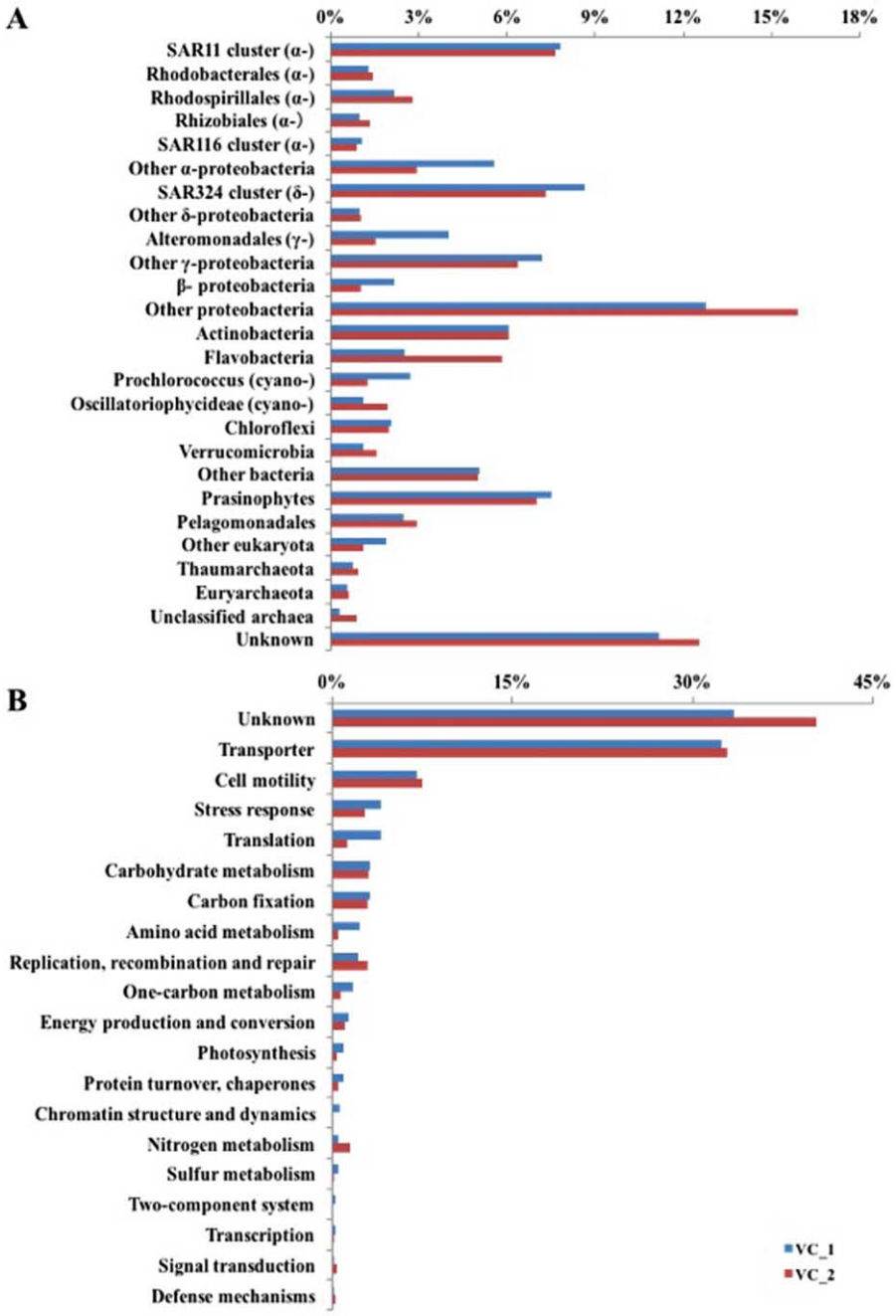

Fig. 2. The relative abundance distribution of nonviral proteins based on normalized spectral counts in the two VCs (VC_1 and VC_2) from the DCM of the SCS, by taxonomic (A) and functional (B) classification, respectively. α-, β-, γ-, δ- and cyano- represent alpha-, beta-, gamma-, delta-proteobacteria and cyanobacteria.

no method of directly enumerating natural RNA viruses or GTAs is available (Lang *et al*., 2009; Lang *et al*., 2012; Forterre *et al*., 2013), hence, their distribution in and contribution to the marine environment are largely unknown. We believe that the protein-based method used in our study is not biased for any specific virions with protein coats, and should provide a reliable estimation of the relative abundance of each major type of virion. The lower cutoff (10 kDa) used in our study compared with the 30 kDa applied in the RNA viral metagenomics (Steward *et al*., 2013) suggested that it is unlikely that we missed small RNA viral particles during protein preparation. Moreover, sequence-library bias could be excluded by a large number of sequences of RNA viruses from Refseq or environments (Steward *et al*., 2013) included in the protein-searching database. Therefore, the outcome that only one protein related to a RNA virus was detected is most likely due to low abundance of RNA viruses in the sampling site. This finding supports the hypothesis that most ocean viruses are DNA viruses (Lang *et al*., 2009; Brum *et al*., 2015). In addition, no detection of sequences related to GTAs could be caused by the lack of

environmental GTA sequences or by the low abundance of known GTAs in the collected VC samples.

We observed that viral structural proteins comprised the majority of the viral proteins (86%) in the VCs (Supporting Information Table S2), which was consistent with findings from a previous viral metaproteomic study (Brum *et al*., 2016). We used the same method as them to map the Pacific Ocean virome (POV) dataset to obtain the eight most abundant PCs which accounted for half of the spectra counts (Supporting Information Fig. S1). All PCs, except DCM_Cluster_1, comprised proteins with pfam domains related to viral structure (Supporting Information Table S2). Among the eight PCs, CAM_CRCL_773 and CAM_CRCL_625 were also found to be abundant in the virion-associated metaproteome reported by Brum *et al*. (2016) but at different abundances. These comparisons showed that some of the same components were found in ocean viral proteomes, but there was variation from site to site.

A major difference between this study and previous work (Brum *et al*., 2016) was that we did not further purify the viral particles from the VCs, which enabled us to explore the contributions and functions of other noncellular entities in the nano-microbial world based on the information found in nonviral proteins (see below).

Nonviral proteins

More than 60% of the spectra identified from both samples were matched to a total of 402 nonredundant proteins (Supporting Information Tables S1 and S4) that could not be assigned to viruses (Fig. 1A). The vast majority of these spectra were bacterial in origin, with a small fraction of eukaryotes, archaea and taxonomic unknown. Bacterial groups of SAR324, SAR11, Actinobacteria, Alteromonadales and Flavobacteria and picoeukaryotic prasinophytes were the major contributors (Fig. 2A), which was consistent with their high abundances at the DCM of the sampling site (Supporting Information Fig. S2) (Zhang *et al*., 2014). Eight percent of nonviral spectra corresponding to 34 proteins were detected from SAR11, an abundant group of small bacteria in the ocean (Morris *et al*., 2002; Rappé *et al*., 2002). The presence of bacterial proteins in the VCs could be associated with the passage of small bacterial cells, for example, SAR11, through the 0.2 μm pore size membrane filter. However, counting of bacterial cells using flow cytometry suggested that bacterial cells were rare in the filtrates ($<10^3$ bacterial cells ml^{-1}, which is below the limit of detection of flow cytometry), thus, proteins from intact bacterial cells should contribute little to the proteome of the VCs. Other microbial processes, such as cell lysis and extracellular excretion, might produce nonviral proteins which are part of the VCs proteomes. For example, 37% of the nonviral proteins (Supporting Information Table S5), such as ABC transporters, ribosomal proteins and molecular chaperone GroEL present in the purified bacterial MVs (Aguilera *et al*., 2014; Altindis *et al*., 2014; Biller *et al*., 2014), were also detected in the VCs, suggesting the potential contribution of MVs to the marine VC proteomes.

Proteins with unknown function and transporter proteins, each accounting for 1/3 of the nonviral spectra, made up the majority of the nonviral proteins (Fig. 2B). These transporters comprised mainly substrate-binding proteins that are located in the periplasm of Gram-negative bacteria or that are membrane-anchored lipoproteins in Gram-positive bacteria and archaea. Diverse substrates were predicted to be affiliated with transporters detected in the VCs, that is, carbohydrates, amino acids, oligopeptides, carboxylic acids, glycine betaine, polyamines, urea, phosphate/phosphonate and iron (Supporting Information Fig. S3), indicating diverse nutrient utilization strategies in the DCM of the SCS.

Approximately 75% of the VC transporter spectra matched transporters for carbohydrates, amino acids, oligopeptides and carboxylic acids and these abundant transporters were commonly found in diverse microbial groups, especially in the SAR11, SAR324 and Actinobacteria (Supporting Information Fig. S3) as previously reported (Sowell *et al*., 2009; Morris *et al*., 2010; Sowell *et al*., 2011). SAR11 transporters for glycine betaine (OpuAC/ProX) were frequently detected in VCs (Supporting Information Table S6) and bacterial metaproteomes (Sowell *et al*., 2009; 2011; Williams *et al*., 2012), indicating that utilization of glycine betaine by natural SAR11 populations might be ubiquitous in the ocean. In SAR11, OpuAC/ProX also functions as a putative 3-dimethylsulphoniopropionate (DMSP) transporter and O-acetylhomoserine (thiol)-lyase is involved in the putative DMSP demethylation pathway (Tripp *et al*., 2008; Sun *et al*., 2016). The detection of these proteins (Supporting Information Table S6) suggested that SAR11 might utilize DMSP as a sulfur and carbon source at the DCM of the SCS. Taurine transporter and adenylylsulfate reductase subunit A (AprA) were identified from SAR11 in the VCs (Supporting Information Table S6), indicating that taurine might be an important source of carbon and nitrogen for SAR11, which was consistent with a previous metaproteomic study (Williams *et al*., 2012). In SAR11, AprA has been proposed to function in the detoxification of sulfite generated from taurine degradation (Williams *et al*., 2012) since SAR11 does not contain a complete assimilatory sulphur reduction pathway (Tripp *et al*., 2008). Moreover, OpuAC/ProX from other alphaproteobacteria and taurine transporters from other bacteria including SAR324 were detected (Supporting Information Table S6), suggesting the utilization of glycine betaine, DMSP and taurine by these bacteria. In contrast to the highly abundant SAR11 phosphonate and phosphate transporters present in the Sargasso Sea with a low

phosphate concentration (Sowell *et al.*, 2009), spectra assigned to phosphonate transporters in the VCs were from Actinobacteria, SAR324 and thaumarchaea, but none from SAR11 (Supporting Information Table S6). This indicated that the SAR11 population at the DCM of the SCS was, as a result of a relatively high phosphate concentration (0.45 μM, Supporting Information Table S3), not subjected to phosphorus.

One *Prochlorococcus* urea transporter was detected (Supporting Information Table S6), suggesting that urea is an important nitrogen source for cyanobacteria in the ocean. A recent study shows that urea transporters are present in the MVs of marine *Prochlorococcus* (Biller *et al.*, 2014). Urea transporters are differentially expressed in the exoproteomes of *Synechococcus* under different conditions, including long- or short-term axenic culture, coculture with a heterotrophic bacterium, dark incubation and short-term infection by cyanophage (Christie-Oleza *et al.*, 2015). These studies suggest that the urea transporter detected in our VCs might be associated with MVs produced from cyanobacteria and involved in nitrogen cycling.

Spectra involved in transport, translation, energy metabolism and nitrogen metabolism were found in the archaea, including marine group II/III euryarchaeota and thaumarchaea (Supporting Information Table S4). Although thaumarchaeal genomes also contain numerous transporters for amino acids, oligopeptides, sugars and taurine (Walker *et al.*, 2010; Swan *et al.*, 2014), only one thaumarchaeal phosphonate ABC transporter was detected while others were euryarchaeal ABC transporters for amino acid or oligopeptide, suggesting different nutrient requirements among different archaeal groups at the DCM. Unlike SAR11 (Sowell *et al.*, 2009), thaumarchaeal phosphonate transporter might not be relevant to environmental phosphorus limitation since it lacked the genes for C-P lyases and hydrolases in the thaumarchaeal genomes (Walker *et al.*, 2010; Swan *et al.*, 2014). The thaumarchaeal phosphonate transporter expressed at the DCM of the SCS with a high phosphate concentration was consistent with its transcript found in a phosphorus-unlimited system (Hollibaugh *et al.*, 2014). Moreover, one of the interesting findings was archaeal copper-containing nitrite reductase (Supporting Information Table S6). This enzyme catalyzes the reduction of nitrite to nitric oxide and is implicated in the production of the greenhouse gas nitrous oxide (N_2O) (Lund *et al.*, 2012). Isotopic signatures of N_2O suggest ammonia-oxidizing archaea are the major source of marine N_2O (Santoro *et al.*, 2011). Our result indicated that the archaea might be involved in the production of N_2O at the DCM of the SCS.

Proteins involved in other biological processes were also detected, such as stress response, cell motility, translation, carbohydrate metabolism, carbon fixation, amino acid metabolism, replication, recombination and repair, C1 metabolism, energy metabolism and photosynthesis (Fig. 2B), coincident with the exported proteins linked to diverse biological processes observed in laboratory cultivated cyanobacteria (Christie-Oleza *et al.*, 2015). Proteins related to the oxidation of C1 and methylated compounds such as carbon-monoxide dehydrogenase and formate dehydrogenase were detected in diverse bacterial groups, including the Chloroflexia, *Roseovarius*, the SAR324 cluster and an uncultured bacterium (Supporting Information Table S6). Dissolved organic carbon in the ocean contains various C1 and methylated compounds, such as methanol, formaldehyde, formate, DMSP and methylamine (Sun *et al.*, 2011). The presence of these proteins in the marine VCs indicated that the oxidation activities of C1 and methylated compounds, typically such as formate, glycine betaine and DMSP, were important microbial processes at the DCM, which was consistent with the frequent detection of proteins involved in the metabolism of C1 and methylated compounds in metaproteomic studies of marine microbial communities (Sowell *et al.*, 2011; Williams *et al.*, 2012; Georges *et al.*, 2014). The majority of proteins from picophytoeukaryotes were related to the functions of carbon fixation, cell motility and photosynthesis, which was likely to be attributed to the release of cell structure from the chloroplasts and flagella after cell lysis or mechanical damage.

Interestingly, many abundant nonviral proteins belonged to the periplasmic proteins (Supporting Information Table S7), suggesting enrichment of the periplasmic proteins in the VCs. We have no clear explanation about this. It is postulated that periplasmic proteins might be incorporated into bacterial vesicles produced through either vesicularization of shattered membrane fragments during natural cell lysis events or shedding from the cell surface during growth (Kesty and Kuehn, 2003; Forterre *et al.*, 2013; Turnbull *et al.*, 2016), since MVs are estimated to have an abundance similar to viral particles in the ocean (Biller *et al.*, 2014). In addition, the substrate binding proteins of ABC transporters are typically present in large amounts in marine bacterial metaproteomes (Sowell *et al.*, 2009; Morris *et al.*, 2010; Sowell *et al.*, 2011), which could partially explain the frequent detection of these transporters (over 80% of the transporter-associated spectra) in the VCs. However, we cannot rule out the possibility that some soluble proteins such as cytoplasmic proteins might occasionally be packed into vesicles or bound to other materials, which protected them from degradation and so accumulated in the seawater.

Abundant bacterial 56-kDa selenium-binding protein (SBP56)

In our study, three proteins containing the SBP56 Pfam domain were detected (Supporting Information Table S6),

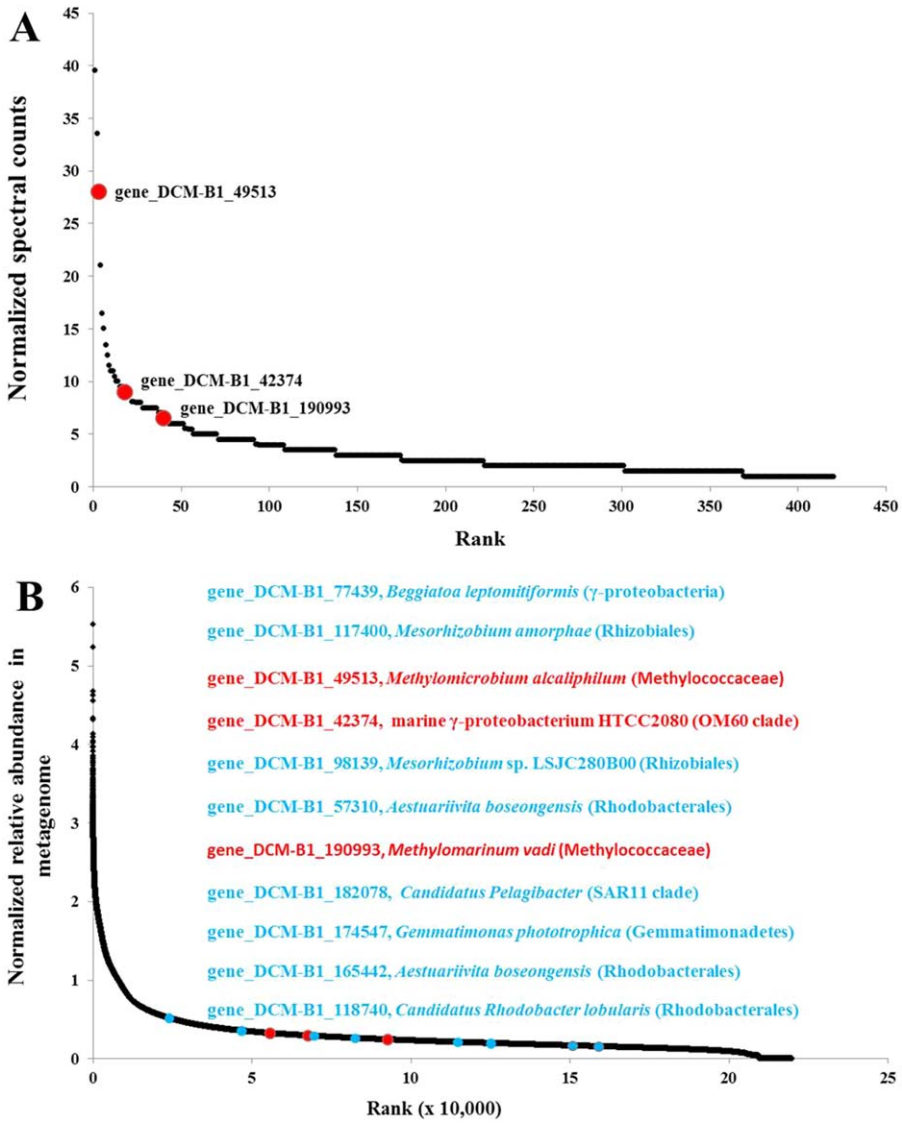

Fig. 3. Rank of each SBP56-like protein among nonviral proteins based on normalized spectral counts (A) and among predicted genes in SEATS-DCM-Bac metagenome based on normalized read counts (B). The SBP56-like proteins that are present in both the metagenome and the metaproteome are shown in red, while those are not detected in the metaproteome are blue. Taxonomic annotations by BLASTp against NCBInr database are shown in the figure. The annotation order from top to bottom corresponds to the coloured dots from left to right.

indicating the existence of SBP56 homologs in the VCs. In contrast to very few SBP56 spectra (from 0 to 3) detected in the previous metaproteomic studies of marine cellular fractions collected above the DCM (Sowell *et al.*, 2009; 2011; Morris *et al.*, 2010; Williams *et al.*, 2012), the three proteins were ranked among the most abundant nonviral proteins in the VCs (Fig. 3A). This suggested that bacterial SBP56 might be an important enzyme at the DCM. However, the bacterial SBP56 family has previously been seldom documented.

Taxonomic annotation showed that these peptides were most closely related to proteins from three gammaproteobacteria, *Methylomarinum vadi*, *Methylomicrobium alcaliphilum* and a marine isolate HTCC2080 in the OM60 group (Fig. 3B), none of which are abundant in the sampling site (Supporting Information Fig. S2) (Zhang *et al.*, 2014). Eight other SBP56 homologs related to other abundant bacteria such as SAR11, Rhodobacterales and Rhizobiales were also present in the protein-searching database, but they did not match any of these peptides (Fig. 4 and Supporting Information Table S8). Searching against all the contigs from the Tara oceans DCM (TARA-DCM, Brum *et al.*, 2015; Sunagawa *et al.*, 2015) indicated that none of the mapped contigs were from the viromic dataset (< 0.22 μm) or contained any virus-associated gene (Fig. 5), suggesting their nonviral origins. Phylogenetic analysis (Fig 4 and Supporting Information Fig. S4) was further used to infer the origin of metaproteome-detected SBP56 homologs based on 84 environmental SBP56 sequences from either the global ocean sampling (GOS) or the TARA-DCM metagenomics datasets (Yooseph *et al.*, 2007; Sunagawa *et al.* 2015), as well as more than 700 bacterial SBP56 genes from diverse bacterial groups. The results indicated that the SBP56-like

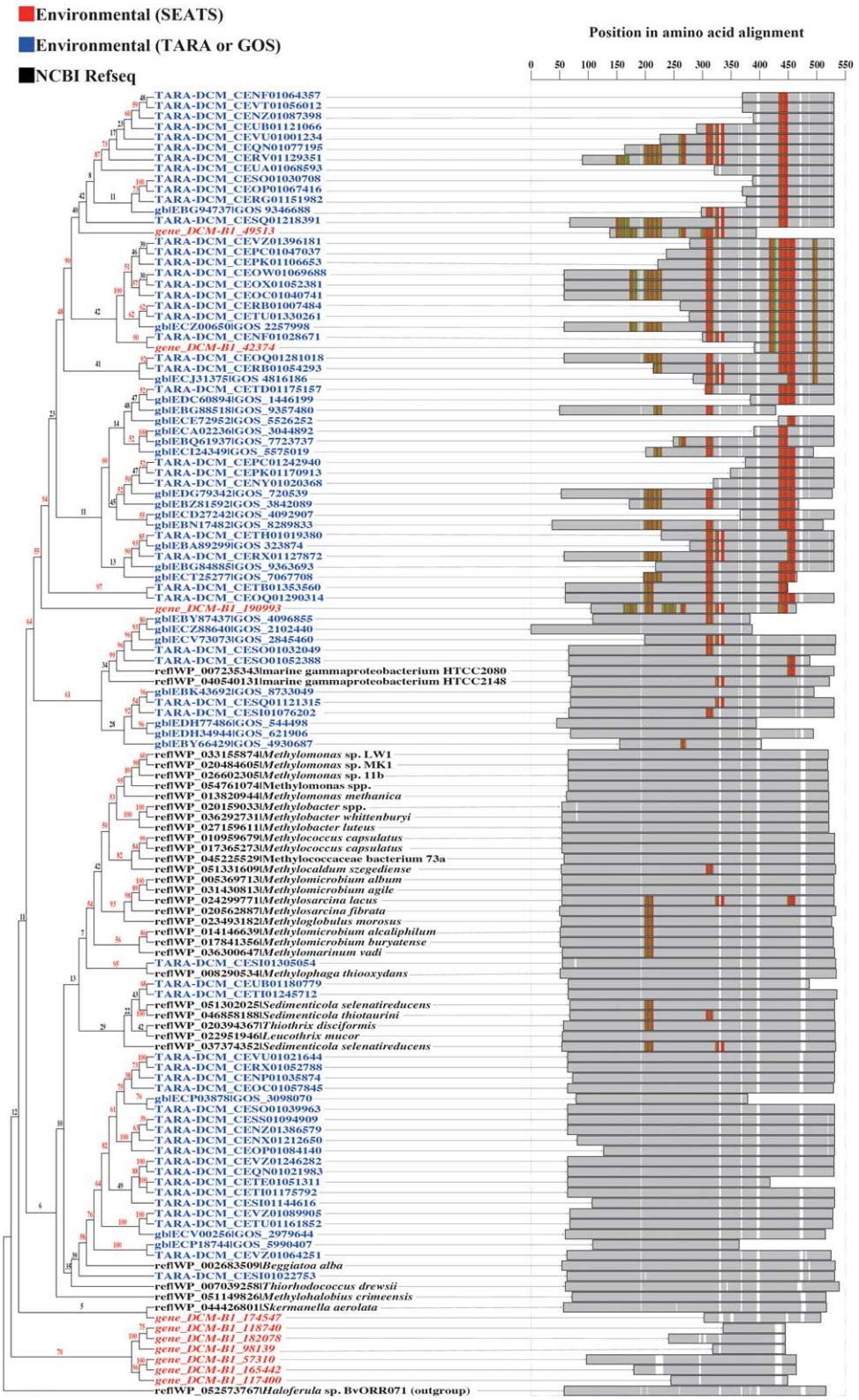

Fig. 4. Phylogenetic analysis and peptide coverage of environmental and reference SBP56-like proteins. The prevalence of a given amino acid residue at that position in the consensus sequence is heat mapped, with red as the most frequent and green as the least frequent residue. Grey means the undetected residues.

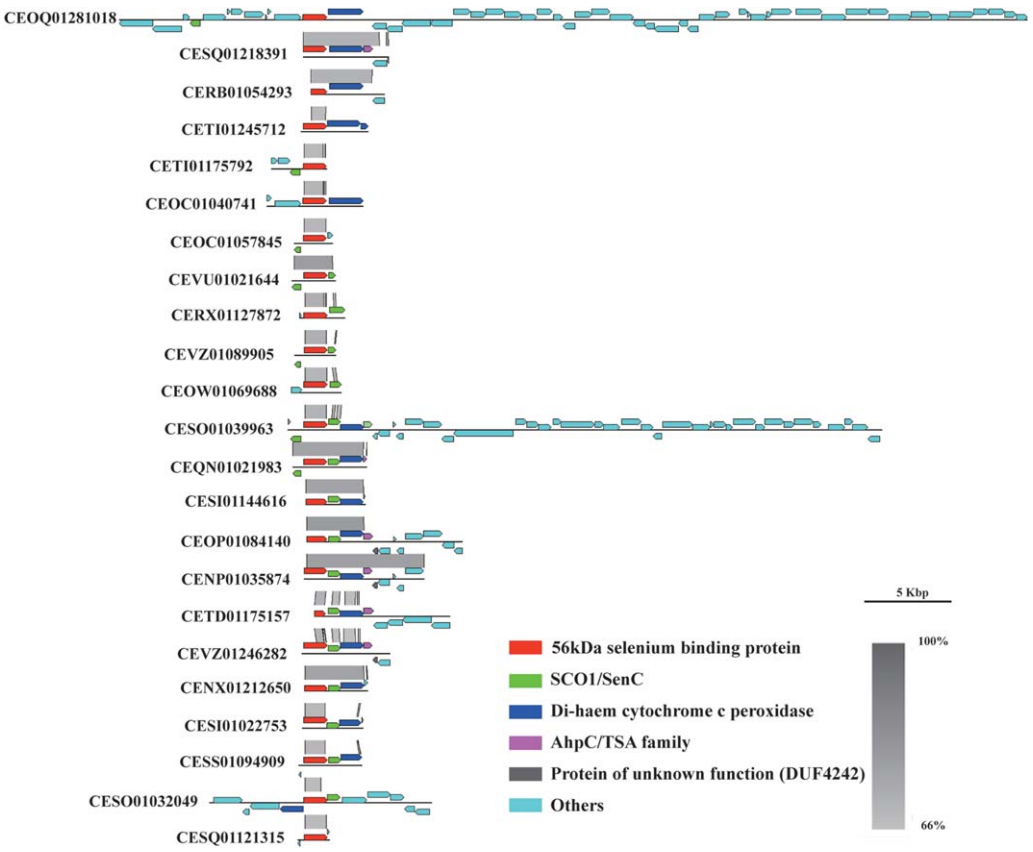

Fig. 5. Gene context for TARA-DCM metagenomic contigs containing SBP56 homologs detected in metaproteomes in this study. Vertical blocks between sequences with gray colour indicate regions of shared similarity shaded based on BLASTn.

proteins detected in the VCs were derived from unknown relatives of the OM60 group.

The function of the SBP56 protein family in bacteria is not yet clear although SBP56 has been demonstrated to have thioredoxin activity or intra-Golgi transport activity in eukaryotes (Tamura and Stadtman, 1996; Porat *et al.*, 2000). Except that selenium binding was annotated for the metaproteome-detected SBP56 structural analog (2ECE, Supporting Information Fig. S5), functional information is limited based on model-based function prediction using the I-TASSER prediction server (Zhang, 2008). Both TARA-DCM contigs (Sunagawa *et al.*, 2015) (Fig. 5) and reference genomes (Supporting Information Fig. S6) that contain metaproteome-detected SBP56 homologs were retrieved for genomic context analysis leading to further functional prediction (Mavromatis *et al.*, 2009). Almost all SBP56 was present next to ORFs encoding for cytochrome c peroxidase (CCP) that could reduce peroxides. Next to CCP, several contigs also contained peroxiredoxin ORF belonging to the AhpC/TSA family that was related to alkyl hydroperoxide reductase and thiol specific antioxidant. Therefore, genomic context analysis suggested that SBP56 was functionally related to the removal of peroxide compounds. Moreover, subcellular location predictions using SignalP and CELLO for SBP56 homologs (Supporting Information Table S9) implied that the SBP56 detected in our study were likely periplasmic proteins. Hence, the abundance variation between our study and previous studies (Sowell *et al.*, 2009; 2011; Morris *et al.*, 2010; Williams *et al.*, 2012) indicated that specific microbial groups occupying the DCM might produce SBP56 and secrete this protein to the periplasm in response to peroxide stress. Overall, information on the origin, expression profile, cellular location and function of bacterial SBP56 in the ocean are lacking and more efforts should be devoted to this protein and its ecological significance.

Concluding remarks

Our metaproteomic analysis provided insights into the origin and function of proteins in marine VCs. Viral proteins in the VCs indicate the presence of viruses of abundant bacterial groups, consistent with observations found in metagenomic studies (Mizuno *et al.*, 2013; Brum *et al.*, 2015). RNA viruses and known GTA particles might not be abundant in the DCM of the SCS. An important but not

previously reported observation is that periplasmic proteins, including substrate binding proteins and SBP56 proteins, were abundant in the VC protein pool. The presence of abundant periplasmic proteins, as well as other diverse proteins that have cytoplasmic functions, such as oxidation of C1 and methylated compound, indicates the possible presence of these enzymes in the 'nonbacterial' world. However, it should be pointed out that several factors, that is, different preparation methods of VC (Hurwitz *et al.*, 2013), coverage of reference database (Morris *et al.*, 2010) and inherent differences in microbial genetics (Christie-Oleza *et al.*, 2015), could affect VC composition and protein identification and lead to findings different from the VC proteomes described in a previous study (Brum *et al.*, 2016). Therefore, comprehensive metaproteomic studies of VCs from various marine environments will not only deepen our understanding of viral diversity and community structure, but also provide new insights into the potential ecological functions affiliated with extracellular biomaterials in the ocean.

Experimental procedures

Sampling in the SCS

Two seawater samples, 110 L each, were collected from the DCM layer (75 m) at the SEATS station (18.0°N, 116.0°E) in the SCS during a summer 2012 cruise (ancillary data provided in Supporting Information Table S3). To avoid the loss of unassembled phage structures and other potential virus-like structures, the size fraction between 10 kDa and 0.2 μm was collected as the VC sample in this study. The procedures for protein sample preparation from the VC were modified from the methods described elsewhere (Chen *et al.*, 1996; Dong *et al.*, 2013). Briefly, seawater samples were filtered on board using a GF/F glass fiber (Whatman, 0.7 μm pore size) and subsequently a 0.2 μm pore size filter (Millipore, Durapore membrane filter). A final concentration of 0.01% (w:v) sodium dodecyl sulfate (SDS) and 5 mM of sodium azide were added to the filtrate to solubilize the proteins and inhibit bacterial growth. VCs were prepared from these filtrates with tangential flow filtration using a Millipore Pellicon 2 tangential flow filtration system equipped with a 10 kDa regenerated cellulose filter. Approximately 450 ml of concentrated filtrate was obtained and stored at −80°C until protein extraction. To extract proteins from the VCs, the filtrates were further concentrated to around 25 ml in a 50 ml stirred ultrafiltration cell (Millipore) with a regenerated cellulose filter (Millipore, 5 kDa) and desalted with 25 ml of desalting buffer (35 mM NH_4HCO_3, 0.01% SDS [w:v]) on ice. After repeating the desalting step three times, approximately 10 ml of desalted sample was obtained for protein extraction.

Protein extraction from VCs

The desalted sample was precipitated with ice-cold 20% trichloroacetic acid in acetone (threefold the filtrate volume) at −20°C overnight. After centrifugation at 12 000 *g* for 30 min at 4°C, the pellets were rinsed three times in ice-cold acetone and then air dried and solubilized in rehydration buffer containing 7 M urea, 2 M thiourea and 3-((3-cholamidopropyl) dimethylammonium)-1-propane-sulfonate (4% w:v). Finally, the supernatant containing the extracted proteins was obtained after centrifugation at 17 000 *g* for 30 min at room temperature and approximately 40 μg each was obtained based on the Bradford protein assay.

Protein digestion and fractionation

After reduction with dithiothreitol and alkylation with iodoacetamide, the protein extracts were processed using the FASP procedure (Wiśniewski *et al.*, 2009) with 10 kDa Microcon filtration devices (Millipore). Briefly, the buffer was removed after loading the protein solution on the device. Proteins retained on the device were rinsed three times with 100 μl of 0.5 M triethylammonium bicarbonate (TEAB) solution. After centrifugation, 1 μg of sequencing grade trypsin (Promega) in 100 μl 0.5 M TEAB solution was added at two intervals 12 h apart to digest protein to peptides at 37°C (the ratio of trypsin over protein was around 1:20 [w/w], in total 2 μg trypsin). The peptides were collected after centrifugation and dried for reconstitution with 2 ml buffer A [25 mM KH_2PO_4 in 25% acetonitrile (ACN), pH 3.0] and loaded onto a 4.6 × 250 mm Ultremex SCX column containing 5 μm particles (Phenomenex) for SCX chromatographic fractionation using the Shimadzu LC-20AB HPLC system. The peptides were eluted at a 1 ml/min flow rate with a gradient of buffer A for 10 min, 5%–35% buffer B (25 mM KH_2PO_4, 1 M KCl in 25% ACN, pH 3.0) for 30 min, then 35%–80% buffer B for 1 min. The system was then maintained in 80% buffer B for 3 min before equilibrating with buffer A for 10 min prior to the next injection. Elution was monitored by measuring absorbance at 214 nm. Fractions were collected every 1 min and the eluted peptides were finally pooled into 10 fractions, desalted with a Strata X C18 column (Phenomenex) and vacuum-dried.

LC-ESI-MS/MS analysis

Each fraction was suspended in buffer C [5% ACN, 0.1% formic acid (FA)] and centrifuged at 20 000 *g* for 10 min. The final concentration of peptide in each fraction was around 0.5 μg/μl. The supernatant (8 μl) was loaded on a Shimadzu LC-20AD nano-HPLC using the autosampler onto a C18 trap column and the peptides were eluted onto an analytical C18 column (inner diameter 75 μm, 18 cm) packed in-house. The samples were loaded at 8 μl/min for 4 min, then the 40 min gradient was run at 300 nl/min starting from 2% to 35% buffer D (95% ACN, 0.1% FA), followed by 5 min linear gradient to 80%, continued at 80% for 4 min and finally returned to 5% for 1 min.

The peptides were subjected to nanoelectrospray ionization followed by tandem mass spectrometry (MS/MS) in a Q EXACTIVE mass spectrometer (ThermoFisher Scientific, San Jose, CA) coupled online to the HPLC. Intact peptides were detected in the Orbitrap at a resolution of 70 000. Peptides were selected for MS/MS using a high-energy collision dissociation operating mode with a normalized collision energy setting of 27.0; and ion fragments were detected in the

Orbitrap at a resolution of 17 500. A data-dependent procedure that alternated between one MS scan followed by 15 MS/MS scans was applied for the 15 most abundant precursor ions above a threshold ion count of 20 000 in the MS survey scan with a following Dynamic Exclusion duration of 15 s. The electrospray voltage applied was 1.6 kV and the heated capillary was at 280°C. Automatic gain control (AGC) was used to optimize the spectra generated by the Orbitrap. The AGC target for full MS was $3e^6$, and $1e^5$ for MS/MS. For MS scans, the m/z scan range was from 350 to 2000 Da.

Protein identification

The instrument data file for each fraction was merged and transformed to MGF files using Proteome Discoverer (ver. 1.3.0.339; ThermoFisher Scientific, San Jose, CA). Peptide and protein identifications were performed using the Mascot search engine (ver. 2.3.0; Matrix Science, London, United Kingdom) against a combined database, DB1, which contained 5 440 190 nonredundant sequences integrated from the following datasets: the environmental protein-coding sequences (eCDSs) predicted from the local microbial metagenomics (SEATS-DCM-Bac, 219 878 sequences, see the section of metagenomic sequencing, assembly and database construction) of the DCM layer at the same time, the POV (~4.1 million) dataset (Hurwitz and Sullivan, 2013), the Mediterranean DCM metaviromic dataset (uvMedDCM-Vir, 38 758 sequences) (Mizuno *et al.*, 2013) and protein sequences of ds-, ss-RNA viruses and retro-transcribing viruses (1 847 660 sequences downloaded from the NCBI on July 27, 2014). Database searching was restricted to tryptic peptides. For protein identification, a mass tolerance of 20 ppm was permitted for intact peptide masses and 0.6 Da for fragmented ions, with allowance for one missed cleavage in the trypsin digests, carbamidomethyl as the fixed modification and oxidation as the variable modification. The charge states of peptides were set to +1, +2 and +3. Specifically, an automatic decoy database search was performed in Mascot by choosing the decoy checkbox in which a reverse database is generated and tested for raw spectra as well as the real database. To reduce the probability of false peptide identification, only peptides of ion scores at the 95% confidence interval using a Mascot probability analysis greater than the identity score were counted as identified and the false discovery rate was calculated to be around 1% for protein identification. Proteins matched with at least two unique peptides were finally accepted as confident identifications for further bioinformatic analysis.

To evaluate the contributions of marine RNA viruses and GTAs to the proteome of the VCs, an independent search was run against another database using the same searching parameters as DB1. In general, searching against a large database is computationally intensive. Therefore, 438 132 eCDSs from the coastal RNA viral metagenomics (Steward *et al.*, 2013) and 9597 GTA protein sequences from the NCBI (downloaded on July 29, 2014) as well as our SEATS-DCM-Bac dataset were merged into 665 929 nonredundant sequences as a small reference database, DB2, to quickly test the existence of proteins related to RNA viruses or GTAs. Since all sequences confidently identified were from the common dataset (SEATS-DCM-Bac) between DB1 and DB2, the sequences of RNA viruses and GTAs in the DB2 was not combined with the DB1 for further searching against, and the analysis in this study was based mainly on the identifications generated from DB1.

Protein annotation

The protein annotations and taxonomic assignments were performed as previously described (Morris *et al.*, 2010). Briefly, all eCDSs identified were searched against the NCBInr database downloaded in October 2014 using the BLASTp. For each eCDS identified, the BLAST expected score of the best BLAST hit $<10^{-5}$ was selected for protein description and taxonomic assignment, otherwise, the eCDS was annotated as 'unknown'. Once sequences were grouped together, consensus annotations for protein description and taxonomic assignment were performed. Usually, all the best BLAST hits in one protein group were the same. If not, taxonomic identity and protein description assignments were made at the most specific level. After the taxonomic consensus annotations, sequences not assigned to viruses were further manually examined. If a sequence contained only viral structural protein domain or were present in a viral contig scanned using VirSorter (Roux *et al.*, 2015), it was reclassified as a viral protein, otherwise annotation was not changed. Note that sequences from the uvMedDCM-Vir dataset (Mizuno *et al.*, 2013) had previously been annotated as genes in the viral contigs, so these proteins once detected were classified as viral proteins in default.

Functional annotations were based on Clusters of Orthologous Groups of proteins (Tatusov *et al.*, 2003), the Kyoto Encyclopedia of Genes and Genomes (http://www.genome.jp/kegg/) (Kanehisa and Goto, 2000) and Pfam (version 27.0, http://pfam.xfam.org/) (Finn *et al.*, 2014). Protein subcellular location was predicted using CELLO v. 2.5 (http://cello.life.nctu.edu.tw/) (Yu *et al.*, 2006). For subcellular location prediction, the most likely location was selected as its annotation. Thus, each sequence had three predictions based on the organism (Gram-negative or positive bacteria and eukaryotes) from which it was derived. The consensus taxonomic annotation was taken into account. For example, if the consensus taxonomic annotation was a Gram-negative bacterium, the subcellular location prediction for the Gram-negative bacterium was regarded as its annotation. If the taxonomic annotation for an identified sequence was not clear, such as 'bacteria' or 'unknown', only a uniform prediction was considered as the precise annotation, otherwise, it was annotated as 'unknown'. If the taxonomic annotation was an 'archaea', its prediction was considered as the correct annotation when it had at least two identical predictions, otherwise it was annotated as 'unknown'.

Using cd-hit-2d ('-g 1 –n 4 –d 0 –T 24 –M 45000'; 80% coverage and 60% percent identity), all viral proteins were further mapped to PCs of POV. If a viral protein was clustered with POV proteins with the same cluster id, the cluster id was used as the PC id of that viral protein. Then the remaining viral proteins were self-clustered and new cluster ids were created.

Quantitative analysis based on spectral counts

To decrease the bias due to common spectra shared between different proteins, spectral counts for each protein were

normalized following the equation used in a previous study (Brum *et al.*, 2016). To provide relative quantification within or between replicates, percentages of normalized spectral counts for each protein category were used in Figs 1 and 2 Supporting Information Figs S1 and S3.

Metagenomic sequencing, assembly and database construction

A local metagenomic library is of great help in the protein identification of metaproteomics (Morris *et al.*, 2010). We assumed that all cellular proteins in the VCs should be encoded by genes present in the cellular fraction. Because of the limited volume of seawater available, we created only a local metagenomic library (SEATS-DCM-Bac) of cellular size, and several viromic libraries were introduced to solve the viral database. Two DNA samples with 10 L each of the same seawater were collected on polycarbonate filters (Millipore, 0.2 μm pore size) and extracted using a CTAB (cetyl/hexadecyl trimethyl ammonium bromide) protocol as previously described (Lin *et al.*, 2010). DNA sequencing was performed on an Illumina HiSeq 2000 platform following the paired-end sample preparation protocol. Approximately, 55 million high quality reads with 10 900 Mbp were generated. Assembly was performed using SOAPdenovo (version 1.06, http://soap.genomics.org.cn/soapdenovo.html). ORFs were predicted using MetaGeneMark (Version 2.10, http://exon.gatech.edu/GeneMark/metagenome/Prediction) and redundant ORFs were removed using cd-hit (version 4.6.1, http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi). Finally, 219 878 predicted ORFs were obtained, comprised mainly of bacterial sequences with a small fraction of archaeal, eukaryotic, viral and taxonomic unclassified sequences (Supporting Information Fig. S2). These nucleotide sequences were further translated into amino acid sequences for protein identification. To calculate the relative abundance of each predicted gene, the number of reads mapping to each gene (H) were normalized in the following steps: (1) Each H was divided by the length of the gene (L) to give a length normalized abundance for each gene (H/L); (2) each H/L was divided by the sum of H/L for all genes from the sample (N) to generate a sample-size normalized abundance for each gene (H/L/N); (3) each H/L/N was rescaled to the mean of H/L/N across the sample as the final relative abundance of each gene.

Data analysis of SBP56-like proteins

Homologous sequences of metaproteome-detected SBP56-like proteins were obtained using BLASTp against the metagenomic protein database (env_nr) or Refseq protein database on the NCBI website (Supporting Information Table S9). Environmental contigs containing homologs of the three SBP56-like proteins (e-value $<$1e-20, identity $>$ 70% and alignment length longer than query sequence length) were retrieved using tBLASTn against all TARA-DCM metagenomic assembly datasets and annotated on the interface of the US Department of Energy Systems Biology Knowledgebase (KBase) (Arkin *et al.*, 2016). SBP56-like sequences in the contigs were extracted.

To perform phylogenetic analysis, the three SBP56-like sequences and their homologs from SEATS-DCM-Bac, GOS and TARA-DCM datasets as well as an out-group sequence were multiple aligned using ClustalW (Chenna *et al.*, 2003), and trimmed using trimAl (Capella-Gutiérrez *et al.*, 2009). Finally, a maximum likelihood tree was generated using the RAxML program (Stamatakis, 2014) based on the best-fit models of evolution determined with the ProtTest program (Abascal *et al.*, 2005). Identical aligned sequences were excluded from the analysis and a bootstrap convergence test was conducted with the option of autoMRE in the RAxML. Aligned sequences without trimming were used for peptide coverage analysis and shown together with the tree using the script of a previous study (Sowell *et al.*, 2009). To further support the phylogenetic lineage of the three SBP56-like protein in the small tree, a total of 789 reference sequences were retrieved from the Refseq protein database on the NCBI website with the keywords 'selenium-binding protein' or 'selenium binding protein' within the bacterial domain and were included to create the larger phylogenetic tree of Supporting Information Fig. S4 using the same method.

To explore the function of the metaproteome-detected SBP56-like proteins, several analyses were conducted: first, the three amino acid sequences were subjected to structural prediction using the I-TASSER online server (Zhang, 2008) with default settings; then, gene context analysis of TARA-DCM contigs was made, based on the KBase annotations and visualized using the Easyfig program (Sullivan *et al.*, 2011); and finally, gene cluster analysis was performed on the reference genomes that contained SBP56-like peptides detected in the metaproteome based on the IMG chromosomal cassette by Pfam (Mavromatis *et al.*, 2009).

Acknowledgements

We thank the captain and crew of the R/V DongFangHong 2 for their assistance, and Bingzhang Chen in Xiamen University for providing bacterial data of the DCM in the SEATS station during the same cruise. John Hodgkiss is thanked for his help with English. Grieg F. Steward, Francisco Rodriguez-Valera and Matthew B. Sullivan are thanked for kindly providing the protein datasets of coastal RNA viruses, uvMedDCM-Vir and POV, respectively. Special thanks are given to Steve J. Giovannoni, Luis Bolanos, Zach Landry and Jimmy Saw from Oregon State University for improving the manuscript. This study was supported by the National Natural Science Foundation of China through grants 41425021 and 40821063, and the Ministry of Science and Technology of the People's Republic of China through grant 2015CB954003. D.-Z. Wang was also supported by the 'Ten Thousand Talents Program' for leading talents in science and technological innovation. Z.-X. Xie was awarded a visiting scholarship at Oregon State University from 2015 to 2017 by the China Scholarship Council, 201406310092. The authors declare no conflict of interest.

References

Abascal, F., Zardoya, R., and Posada, D. (2005) ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21:** 2104–2105.

Aguilera, L., Toloza, L., Gimenez, R., Odena, A., Oliveira, E., Aguilar, J., *et al*. (2014) Proteomic analysis of outer membrane vesicles from the probiotic strain *Escherichia coli* Nissle 1917. *Proteomics* **14:** 222–229.

Altindis, E., Fu, Y., and Mekalanos, J.J. (2014) Proteomic analysis of *Vibrio cholerae* outer membrane vesicles. *Proc Natl Acad Sci USA* **111:** E1548–E1556.

Angly, F.E., Felts, B., Breitbart, M., Salamon, P., Edwards, R.A., Carlson, C., *et al*. (2006) The marine viromes of four oceanic regions. *PLoS Biol* **4:** e368.

Arkin, A.P., Stevens, R.L., Cottingham, R.W., Maslov, S., Henry, C.S., Dehal, P., *et al*. (2016) The DOE systems biology knowledgebase (KBase). *bioRxiv.* https://doi.org/10.1101/096354

Bergh, Ø., Børsheim, K.Y., Bratbak, G., and Heldal, M. (1989) High abundance of viruses found in aquatic environments. *Nature* **340:** 467–468.

Biers, E.J., Wang, K., Pennington, C., Belas, R., Chen, F., and Moran, M.A. (2008) Occurrence and expression of gene transfer agent genes in marine bacterioplankton. *Appl Environ Microbiol* **74:** 2933–2939.

Biller, S.J., Schubotz, F., Roggensack, S.E., Thompson, A.W., Summons, R.E., and Chisholm, S.W. (2014) Bacterial vesicles in marine ecosystems. *Science* **343:** 183–186.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segall, A.M., Mead, D., *et al*. (2002) Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci USA* **99:** 14250–14255.

Brum, J.R. (2005) Concentration, production and turnover of viruses and dissolved DNA pools at Stn ALOHA, North Pacific Subtropical Gyre. *Aquat Microb Ecol* **41:** 103–113.

Brum, J.R., Ignacio-Espinoza, J.C., Roux, S., Doulcier, G., Acinas, S.G., Alberti, A., *et al*. (2015) Patterns and ecological drivers of ocean viral communities. *Science* **348:** 1261498.

Brum, J.R., Ignacio-Espinoza, J.C., Kim, E.-H., Trubl, G., Jones, R.M., Roux, S., *et al*. (2016) Illuminating structural proteins in viral "dark matter" with metaproteomics. *Proc Natl Acad Sci USA* **113:** 2436–2441.

Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25:** 1972–1973.

Chen, F., Suttle, C.A., and Short, S.M. (1996) Genetic diversity in marine algal virus communities as revealed by sequence analysis of DNA polymerase genes. *Appl Environ Microbiol* **62:** 2869–2874.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., and Thompson, J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res* **31:** 3497–3500.

Christie-Oleza, J.A., Armengaud, J., Guerin, P., and Scanlan, D.J. (2015) Functional distinctness in the exoproteomes of marine Synechococcus. *Environ Microbiol* **17:** 3781–3794.

Colombet, J., Robin, A., Lavie, L., Bettarel, Y., Cauchie, H., and Sime-Ngando, T. (2007) Virioplankton 'pegylation': use of PEG (polyethylene glycol) to concentrate and purify viruses in pelagic ecosystems. *J Microbiol Methods* **71:** 212–219.

Dong, H.-P., Wang, D.-Z., Xie, Z.-X., Dai, M.-H., and Hong, H.-S. (2013) Metaproteomic characterization of high molecular weight dissolved organic matter in surface seawaters in the South China Sea. *Geochim Cosmochim Acta* **109:** 51–61.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., *et al*. (2014) Pfam: the protein families database. *Nucleic Acids Res* **42:** D222–D230.

Forterre, P., Soler, N., Krupovic, M., Marguet, E., and Ackermann, H.W. (2013) Fake virus particles generated by fluorescence microscopy. *Trends Microbiol* **21:** 1–5.

Fuhrman, J.A. (1999) Marine viruses and their biogeochemical and ecological effects. *Nature* **399:** 541–548.

Georges, A.A., El-Swais, H., Craig, S.E., Li, W.K., and Walsh, D.A. (2014) Metaproteomic analysis of a winter to spring succession in coastal northwest Atlantic Ocean microbial plankton. *ISME J* **8:** 1301–1313.

Hewson, I., O'Neil, J.M., Heil, C.A., Bratbak, G., and Dennison, W.C. (2001) Effects of concentrated viral communities on photosynthesis and community composition of co-occurring benthic microalgae and phytoplankton. *Aquat Microb Ecol* **25:** 1–10.

Hollibaugh, J.T., Gifford, S.M., Moran, M.A., Ross, M.J., Sharma, S., and Tolar, B.B. (2014) Seasonal variation in the metatranscriptomes of a Thaumarchaeota population from SE USA coastal waters. *ISME J* **8:** 685–698.

Hurwitz, B.L., and Sullivan, M.B. (2013) The Pacific Ocean Virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLoS One* **8:** e57355.

Hurwitz, B.L., Deng, L., Poulos, B.T., and Sullivan, M.B. (2013) Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. *Environ Microbiol* **15:** 1428–1440.

Jiang, S.C., and Paul, J.H. (1995) Viral contribution to dissolved DNA in the marine environment as determined by differential centrifugation and kingdom probing. *Appl Environ Microbiol* **61:** 317–325.

John, S.G., Mendez, C.B., Deng, L., Poulos, B., Kauffman, A.K.M., Kern, S., *et al*. (2011) A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environ Microbiol Rep* **3:** 195–202.

Kanehisa, M., and Goto, S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28:** 27–30.

Kuehn, M.J., and Kesty, N.C. (2005) Bacterial outer membrane vesicles and the host-pathogen interaction. *Gene Dev* **19:** 2645–2655.

Lang, A.S., Rise, M.L., Culley, A.I., and Steward, G.F. (2009) RNA viruses in the sea. *FEMS Microbiol Rev* **33:** 295–323.

Lang, A.S., Zhaxybayeva, O., and Beatty, J.T. (2012) Gene transfer agents: phage-like elements of genetic exchange. *Nat Rev Microbiol* **10:** 472–482.

Lin, S., Zhang, H., Zhuang, Y., Tran, B., and Gill, J. (2010) Spliced leader-based metatranscriptomic analyses lead to recognition of hidden genomic features in dinoflagellates. *Proc Natl Acad Sci USA* **107:** 20033–20038.

Lund, M.B., Smith, J.M., and Francis, C.A. (2012) Diversity, abundance and expression of nitrite reductase (nirK)-like genes in marine thaumarchaea. *ISME J* **6:** 1966–1977.

Mavromatis, K., Chu, K., Ivanova, N., Hooper, S.D., Markowitz, V.M., and Kyrpides, N.C. (2009) Gene context analysis in the Integrated Microbial Genomes (IMG) data management system. *PLoS One* **4:** e7979.

McDaniel, L.D., Young, E., Delaney, J., Ruhnau, F., Ritchie, K.B., and Paul, J.H. (2010) High frequency of horizontal gene transfer in the oceans. *Science* **330:** 50–50.

Mizuno, C.M., Rodriguez-Valera, F., Kimes, N.E., and Ghai, R. (2013) Expanding the marine virosphere using metagenomics. *PLoS Genet* **9:** e1003987.

Morris, R.M., Rappe, M.S., Connon, S.A., Vergin, K.L., Siebold, W.A., Carlson, C.A., and Giovannoni, S.J. (2002) SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **420:** 806–810.

Morris, R.M., Nunn, B.L., Frazar, C., Goodlett, D.R., Ting, Y.S., and Rocap, G. (2010) Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction. *ISME J* **4:** 673–685.

Porat, A., Sagiv, Y., and Elazar, Z. (2000) A 56-kDa selenium-binding protein participates in intra-Golgi protein transport. *J Biol Chem* **275:** 14457–14465.

Rappé, M.S., Connon, S.A., Vergin, K.L., and Giovannoni, S.J. (2002) Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature* **418:** 630–633.

Roux, S., Enault, F., Hurwitz, B.L., and Sullivan, M.B. (2015) VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3:** e985.

Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., *et al*. (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5:** e77.

Santoro, A.E., Buchwald, C., McIlvin, M.R., and Casciotti, K.L. (2011) Isotopic signature of N2O produced by marine ammonia-oxidizing archaea. *Science* **333:** 1282–1285.

Scanlan, D. (2014) Bacterial vesicles in the ocean. *Science* **343:** 143–144.

Short, S.M., and Suttle, C.A. (2002) Sequence analysis of marine virus communities reveals that groups of related algal viruses are widely distributed in nature. *Appl Environ Microbiol* **68:** 1290–1296.

Soler, N., Krupovic, M., Marguet, E., and Forterre, P. (2015) Membrane vesicles in natural environments: a major challenge in viral ecology. *ISME J* **9:** 793–796.

Sowell, S.M., Wilhelm, L.J., Norbeck, A.D., Lipton, M.S., Nicora, C.D., Barofsky, D.F., *et al*. (2009) Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea. *ISME J* **3:** 93–105.

Sowell, S.M., Abraham, P.E., Shah, M., Verberkmoes, N.C., Smith, D.P., Barofsky, D.F., and Giovannoni, S.J. (2011) Environmental proteomics of microbial plankton in a highly productive coastal upwelling system. *ISME J* **5:** 856–865.

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

Steward, G.F., Culley, A.I., Mueller, J.A., Wood-Charlson, E.M., Belcaid, M., and Poisson, G. (2013) Are we missing half of the viruses in the ocean? *ISME J* **7:** 672–679.

Sullivan, M.J., Petty, N.K., and Beatson, S.A. (2011) Easyfig: a genome comparison visualizer. *Bioinformatics* **27:** 1009–1010.

Sun, J., Steindler, L., Thrash, J.C., Halsey, K.H., Smith, D.P., Carter, A.E., *et al*. (2011) One carbon metabolism in SAR11 pelagic marine bacteria. *PLoS One* **6:** e23973.

Sun, J., Todd, J.D., Thrash, J.C., Qian, Y., Qian, M.C., Temperton, B., *et al*. (2016) The abundant marine bacterium Pelagibacter simultaneously catabolizes dimethylsulfoniopropionate to the gases dimethyl sulfide and methanethiol. *Nat Microbiol* **1:** 16065.

Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., *et al*. (2015) Structure and function of the global ocean microbiome. *Science* **348:** 1261359.

Suttle, C.A. (2005) Viruses in the sea. *Nature* **437:** 356–361.

Suttle, C.A., Chan, A.M., and Cottrell, M.T. (1991) Use of ultrafiltration to isolate viruses from seawater which are pathogens of marine phytoplankton. *Appl Environ Microbiol* **57:** 721–726.

Swan, B.K., Chaffin, M.D., Martinez-Garcia, M., Morrison, H.G., Field, E.K., Poulton, N.J., *et al*. (2014) Genomic and metabolic diversity of Marine Group I Thaumarchaeota in the mesopelagic of two subtropical gyres. *PLoS One* **9:** e95380.

Tamura, T., and Stadtman, T.C. (1996) A new selenoprotein from human lung adenocarcinoma cells: purification, properties, and thioredoxin reductase activity. *Proc Natl Acad Sci USA* **93:** 1006–1011.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., *et al*. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4:** 41.

Thurber, R.V., Haynes, M., Breitbart, M., Wegley, L., and Rohwer, F. (2009) Laboratory procedures to generate viral metagenomes. *Nat Protoc* **4:** 470–483.

Tripp, H.J., Kitner, J.B., Schwalbach, M.S., Dacey, J.W., Wilhelm, L.J., and Giovannoni, S.J. (2008) SAR11 marine bacteria require exogenous reduced sulphur for growth. *Nature* **452:** 741–744.

Turnbull, L., Toyofuku, M., Hynen, A.L., Kurosawa, M., Pessi, G., and Petty, N.K., *et al*. (2016) Explosive cell lysis as a mechanism for the biogenesis of bacterial membrane vesicles and biofilms. *Nat Commun* **7:** 11220.

Walker, C.B., La Torre, J.R., Klotz, M.G., Urakawa, H., Pinel, H., Arp, D.J., *et al*. (2010) *Nitrosopumilus maritimus* genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. *Proc Natl Acad Sci USA* **107:** 8818–8823.

Weinbauer, M.G., and Peduzzi, P. (1995) Effect of virus-rich high molecular weight concentrates of seawater on the dynamics of dissolved amino acids and carbohydrates. *Mar Ecol Prog Ser* **127:** 245–253.

Williams, T.J., Long, E., Evans, F., DeMaere, M.Z., Lauro, F.M., Raftery, M.J., *et al*. (2012) A metaproteomic assessment of winter and summer bacterioplankton from Antarctic Peninsula coastal surface waters. *ISME J* **6:** 1883–1900.

Williamson, S.J., Allen, L.Z., Lorenzi, H.A., Fadrosh, D.W., Brami, D., Thiagarajan, M., *et al*. (2012) Metagenomic exploration of viruses throughout the Indian Ocean. *PLoS One* **7:** e42047.

Winter, C., Smit, A., Herndl, G.J., and Weinbauer, M.G. (2004) Impact of virioplankton on archaeal and bacterial community richness as assessed in seawater batch cultures. *Appl Environ Microbiol* **70:** 804–813.

Wiśniewski, J.R., Zougman, A., Nagaraj, N., and Mann, M. (2009) Universal sample preparation method for proteome analysis. *Nat Methods* **6:** 359–362.

Wommack, K.E., and Colwell, R.R. (2000) Virioplankton: viruses in aquatic ecosystems. *Microbiol Mol Biol Rev* **64:** 69–114.

Wommack, K.E., Williamson, K.E., Helton, R.R., Bench, S.R., and Winget, D.M. (2009) Methods for the isolation of viruses from environmental samples. In *Bacteriophages: Methods and Protocols, Volume 1: Isolation, Characterization, and Interactions*. Clokie, M.R.J., and Kropinski, A.M. (eds). Totowa, NJ: Humana Press, pp. 3–14.

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

Yu, C.S., Chen, Y.C., Lu, C.H., and Hwang, J.K. (2006) Prediction of protein subcellular localization. *Proteins* **64:** 643–651.

Zhang, Y. (2008) I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **9:** 40.

Zhang, Y., Zhao, Z., Dai, M., Jiao, N., and Herndl, G.J. (2014) Drivers shaping the diversity and biogeography of total and active bacterial communities in the South China Sea. *Mol Ecol* **23:** 2260–2274.

Zhao, Y., Temperton, B., Thrash, J.C., Schwalbach, M.S., Vergin, K.L., Landry, Z.C., *et al*. (2013) Abundant SAR11 viruses in the ocean. *Nature* **494:** 357–360.

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

The metagenomic data is deposited to the Sequence Read Archive at the NCBI (http://www.ncbi.nlm.nih.gov/) under the study accession SRP066135, and the mass spectrometry proteomic data have been deposited to the ProteomeXchange Conssortium via the PRIDE parther repository (www.ebi.ac.uk/pride/archive/) with identifier PXD007174.

Fig. S1. Percentages of virus-associated protein clusters among the viral proteins indicate 8 viral protein clusters abundantly present in the two VCs from the DCM of SCS.

Fig. S2. Taxonomic distribution based on local DCM metagenomic analysis (SEATS-DCM-Bac, size above 0.2 μm). α-, β-, γ- and δ- represent alpha-, beta-, gamma-, delta-proteobacteria.

Fig. S3. Spectral counts based distributions of predicted substrates (A) and taxonomic annotations (B) for all transporters identified in the two VCs (left pies for VC_1, right pies for VC_2) collected from the DCM of the SCS.

Fig. S4. Phylogenetic tree of bacterial SBP56 proteins. Nodes with bootstrap values equal or larger than 50 are highlighted in red and size of dot is proportional to the bootstrap value.

Fig. S5. Representative structural model for the three SBP56-like proteins (A. gene_DCM-B1_49513; B. gene_DCM-B1_42374; C. gene_DCM-B1_190993) detected in the VCs. The coloured ribbon indicates the predicted structures while the purple backbone shows the template protein, a hypothetical selenium-binding protein from *Sulfolobus tokodaii* (2ECE).

Fig. S6. Gene context of chromosomal cassettes (red dashed boxes) from the reference genomes (Figure 4) that contain SBP56 proteins matched metaproteome-detected peptides. Genes in each cassette are listed following the order from left to right in the Supporting Information Table S10 and SBP56 is indicated by small red box under the centerline. Small colour boxes next to the bottom red dashed line indicated the functional categories.

Table S1. Number and percentage of proteins identified in each superkingdom in the two independent proteomes of VCs (VC_1 and VC_2) collected from the DCM of the SCS.

Table S2 (see Excel file). Viral proteins with two or more unique peptides matching identified in the VCs collected from the DCM of the SCS after searching against DB1.

Table S3. Ancillary data of the DCM in the SouthEast Asian Time-Series (SEATS) station in the SCS during the cruise of summer 2012.

Table S4 (see Excel file). Nonviral proteins, with two or more unique peptides matching, identified in the VCs collected from the DCM of the SCS after searching against DB1.

Table S5 (see Excel file). List of nonviral proteins potentially linked to membrane vesicles.

Table S6. List of nonviral proteins discussed in the main text.

Table S7 (see Excel file). List of nonviral proteins potentially located at periplasmic space. Noted that proteins from gram-positive bacteria and archaea are excluded in the list, since these microbes do not have periplasmic space.

Table S8 (see Excel file). Peptides and spectra assigned to SBP56-like proteins detected in the two VCs.

Table S9 (see Excel file). Homologous sequences of identified SBP56-like proteins from NCBI refseq and environmental nonredundant databases.

Table S10 (see Excel file). Gene list of SBP56 protein containing chromosomal cassettes from the reference genomes.