LETTER

doi:10.1038/nature19366

Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses

Simon Roux[1], Jennifer R. Brum[1], Bas E. Dutilh[2,3,4], Shinichi Sunagawa[5]†, Melissa B. Duhaime[6], Alexander Loy[7,8], Bonnie T. Poulos[9], Natalie Solonenko[1], Elena Lara[10,11], Julie Poulain[12], Stéphane Pesant[13,14], Stefanie Kandels-Lewis[5,15], Céline Dimier[16,17,18], Marc Picheral[19,20], Sarah Searson[19,20], Corinne Cruaud[12], Adriana Alberti[12], Carlos M. Duarte[21,22], Josep M. Gasol[10], Dolors Vaqué[10], *Tara* Oceans Coordinators*, Peer Bork[5,23], Silvia G. Acinas[10], Patrick Wincker[12,24,25] & Matthew B. Sullivan[1,26]

Ocean microbes drive biogeochemical cycling on a global scale[1]. However, this cycling is constrained by viruses that affect community composition, metabolic activity, and evolutionary trajectories[2,3]. Owing to challenges with the sampling and cultivation of viruses, genome-level viral diversity remains poorly described and grossly understudied, with less than 1% of observed surface-ocean viruses known[4]. Here we assemble complete genomes and large genomic fragments from both surface- and deep-ocean viruses sampled during the *Tara* Oceans and Malaspina research expeditions[5,6], and analyse the resulting 'global ocean virome' dataset to present a global map of abundant, double-stranded DNA viruses complete with genomic and ecological contexts. A total of 15,222 epipelagic and mesopelagic viral populations were identified, comprising 867 viral clusters (defined as approximately genus-level groups[7,8]). This roughly triples the number of known ocean viral populations[4] and doubles the number of candidate bacterial and archaeal virus genera[8], providing a near-complete sampling of epipelagic communities at both the population and viral-cluster level. We found that 38 of the 867 viral clusters were locally or globally abundant, together accounting for nearly half of the viral populations in any global ocean virome sample. While two-thirds of these clusters represent newly described viruses lacking any cultivated representative, most could be computationally linked to dominant, ecologically relevant microbial hosts. Moreover, we identified 243 viral-encoded auxiliary metabolic genes, of which only 95 were previously known. Deeper analyses of four of these auxiliary metabolic genes (*dsrC*, *soxYZ*, *P-II* (also known as *glnB*) and *amoC*) revealed that abundant viruses may directly manipulate sulfur and nitrogen cycling throughout the epipelagic ocean. This viral catalog and functional analyses provide a necessary foundation for the meaningful integration of viruses into ecosystem models where they act as key players in nutrient cycling and trophic networks.

The lack of host-contextualized quantitative surveys of the diversity of microbe-specific viruses in nature is a fundamental block to the incorporation of these viruses into ecosystem models. This is because most naturally occurring microbes and viruses are not currently cultivated and viruses lack a universally conserved marker gene, precluding PCR-based surveys of uncultivated diversity[3]. Although viral metagenomics (viromics) was intended to circumvent these issues, early datasets were fragmented and only suitable for descriptive gene-level analyses—studies that were prohibitively limited by database biases[3]. Subsequent experimental, technological and analytical advances enabled improved viral population ecology, aided by the availability of genomic information[3,9–11]. The 1,148 large viral genome fragments captured in a fosmid library from Mediterranean Sea microbes revealed remarkable viral diversity, with some genomes appearing to be globally distributed based upon the analysis of 6 available viral metagenomes[9]. Similarly, 69 viral reference genomes assembled from single-cell samples helped elucidate the ecological, evolutionary and potential biogeochemical effects of uncultivated viruses infecting an uncultivated anaerobic chemoautotroph[11]. Technological advances mean that metagenomic approaches are now quantitative, at least for double-stranded DNA (dsDNA) templates[3], and can themselves provide genomic information on uncultivated viruses. The 42 surface-ocean viral metagenomes in the *Tara* Oceans Viromes (TOV) dataset is an example of this progress. These data reveal the global underlying structure of these viral communities and identified 5,476 viral populations, of which only 39 were previously known[4].

Here we further identify ocean viral populations, determine and characterize the most abundant and widespread dsDNA ocean viral types, and analyse viral-encoded auxiliary metabolic genes (AMGs) and their distributions to propose new means by which viruses are likely to modulate microbial biogeochemistry. We do so by analysing the Global Oceans Viromes (GOV) dataset, augmenting the TOV dataset with a further 61 samples to represent better the surface and deep oceans. The GOV now totals 104 viromes, with 925 Gb of sequencing data (Supplementary Table 1). Furthermore, our use of upgraded analytical approaches, including cross-assembly[12] and genome-binning[13], improved the genomic representation of sampled viruses (see Supplementary Information for details on the dataset generation

[1]Department of Microbiology, The Ohio State University, Columbus, Ohio 43210, USA. [2]Theoretical Biology and Bioinformatics, Utrecht University, 3584 CH Utrecht, The Netherlands. [3]Centre for Molecular and Biomolecular Informatics, Radboud University Medical Centre, 6525 GA Nijmegen, The Netherlands. [4]Department of Marine Biology, Federal University of Rio de Janeiro, Rio de Janeiro, CEP 21941-902, Brazil. [5]Structural and Computational Biology, European Molecular Biology Laboratory, 69117 Heidelberg, Germany. [6]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan 48109, USA. [7]Division of Microbial Ecology, Department of Microbiology and Ecosystem Science, Research Network Chemistry Meets Microbiology, University of Vienna, A-1090 Vienna, Austria. [8]Austrian Polar Research Institute, A-1090 Vienna, Austria. [9]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA. [10]Department of Marine Biology and Oceanography, Institut de Ciències del Mar (ICM), CSIC Barcelona E0800, Spain. [11]Institute of Marine Sciences (CNR-ISMAR), National Research Council, 30122 Venezia, Italy. [12]CEA - Institut de Génomique, GENOSCOPE, 91057 Evry, France. [13]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, 28359 Bremen, Germany. [14]MARUM, Bremen University, 28359 Bremen, Germany. [15]Directors' Research, European Molecular Biology Laboratory, 69117 Heidelberg, Germany. [16]CNRS, UMR 7144, EPEP, Station Biologique de Roscoff, 29680 Roscoff, France. [17]Sorbonne Universités, UPMC Université Paris 06, UMR 7144, Station Biologique de Roscoff, 29680 Roscoff, France. [18]Institut de Biologie de l'École Normale Supérieure, École Normale Supérieure, Paris Sciences et Lettres Research University, CNRS UMR 8197, INSERM U1024, F-75005 Paris, France. [19]CNRS, UMR 7093, Laboratoire d'océanographie de Villefranche, Observatoire Océanologique, 06230 Villefranche-sur-mer, France. [20]Sorbonne Universités, UPMC Université Paris 06, UMR 7093, Observatoire Océanologique, 06230 Villefranche-sur-mer, France. [21]Mediterranean Institute of Advanced Studies, CSIC-UiB, 21-07190 Esporles, Mallorca, Spain. [22]King Abdullah University of Science and Technology, Red Sea Research Center, Thuwal 23955-6900, Saudi Arabia. [23]Max-Delbrück-Centre for Molecular Medicine, 13092 Berlin, Germany. [24]CNRS, UMR 8030, 91057 Evry, France. [25]Université d'Evry, UMR 8030, 91057 Evry, France. [26]Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, Ohio 43210, USA. †Present address: Department of Biology, Institute of Microbiology, ETH Zurich, 8093 Zurich, Switzerland.
*A list of participants and their affiliations appears in the Supplementary Information.

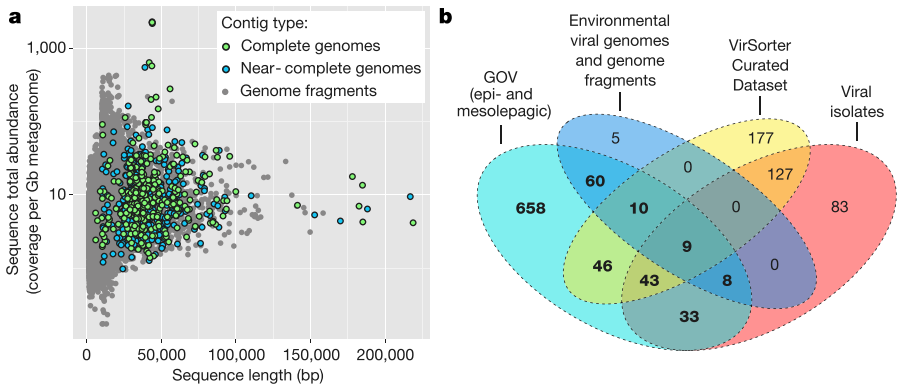


Figure 1 | Composition of the Global Ocean Viromes (GOV) dataset. **a**, The size of viral contigs (*x* axis) and their cumulative coverage across the GOV dataset (*y* axis). Contigs corresponding to complete (345 contigs) or near-complete (425 contigs) genomes are indicated. For clarity, only contigs associated with a viral population (24,412 contigs) are displayed. **b**, Distribution of all viral clusters according to the origin of their members. Viral genomes (or fragments) in a viral cluster can originate from isolated viral genomes, the VirSorter Curated Dataset[8] (viral genomes identified *in silico* from microbial genomes), environmental viral genomes and genome fragments (from fosmid libraries, for example), or the GOV dataset. Viral clusters that include at least one GOV sequence and are further analysed in this study are highlighted in bold.

process). From 1,380,834 contigs, which recruited 67% of the reads, we identified 15,280 viral populations (Fig. 1a; Supplementary Fig. 1 contains an explanation of viral population definition). This expands the number of known ocean viral populations nearly threefold over the prior TOV dataset[4] and improves average contig lengths and genomic context 2.5-fold for populations known to the TOV (Supplementary Table 2). Rarefaction analyses show that, while mesopelagic viral communities remain under-sampled, sampling of epipelagic viral communities now appears to be near-complete (Extended Data Fig. 1a). Because bathypelagic communities were underrepresented owing to cellular contamination, we focused the remaining analyses on 15,222 non-bathypelagic viral populations.

We first categorized viral populations into viral clusters using shared gene-content information and network analytics[7] (see Supplementary Fig. 1 for viral cluster definition schematic). This method starts with genome fragments (those fragments $\geq$10 kb) and results in viral clusters approximately equivalent to known viral genera[7,8]. Clustering of the 15,222 GOV viral populations with 15,929 publicly available bacterial and archaeal viruses revealed 1,259 viral clusters (see Extended Data Fig. 2, Supplementary Table 3 and Supplementary Information for comparisons with alternative classification methods). Of these viral clusters, 658 included sequences that were exclusive to GOV, approximately doubling the number of known bacterial and archaeal virus genera[8], and another 209 contained at least one GOV sequence (Fig. 1b). As with viral populations, rarefaction analyses suggested that viral-cluster diversity was under-sampled in mesopelagic waters, but near-completely sampled in epipelagic waters (Extended Data Fig. 1b).

We next identified the most abundant and widespread viral clusters based on read-recruitment of viral cluster members. In each sample, a fraction of the viral clusters was identified as abundant based on their cumulative contribution to sample diversity (Simpson index estimates state that abundant viral clusters represent 80% of the total sample diversity, Extended Data Fig. 1c). By these criteria, only 38 out of 867 observed viral clusters were abundant in two or more stations, together recruiting an average of 50% and 35% of reads from viral populations for epipelagic and mesopelagic samples, respectively (Supplementary Table 3). Of these 38 abundant viral clusters, 4 were also relatively ubiquitous as they were abundant in more than 25 stations, and 62 of the 91 non-bathypelagic samples were dominated by 1 of these 4 viral clusters (Fig. 2a, b). Among the 38 abundant viral clusters, only 2 corresponded to well-studied viruses from the T4 superfamily[14,15] (viral cluster VC_2, one of the four ubiquitous viral clusters) and the T7 virus genus[16] (VC_9). In total, eight viral clusters represented known but unclassified viral isolates, 10 included viruses known only from environmental sequencing[9,10], and the remaining 18 viral clusters were previously unreported (Fig. 2c and Extended Data Fig. 3).

Once we developed this global map of the dominant dsDNA viral types in the oceans, we next sought to identify the range of hosts that these viruses infect. This is challenging, as culture-based methods insufficiently capture naturally occurring diversity, whereas metagenomic approaches broadly survey viral diversity but often without host information. Fortunately, sequence-based approaches are emerging that examine similarities between (i) viral genomes and host CRISPR spacers[17], (ii) viral and microbial genomes due to integrated prophages or gene transfers[9] and (iii) viral and host genome nucleotide signatures (here, tetranucleotide frequencies[8]; see Supplementary Table 4 and Supplementary Information for discussion of the accuracy and sensitivity of *in silico* host prediction methods). We applied all three methods to the GOV to predict hosts at the phylum level (or class level for Proteobacteria) (Supplementary Table 5) and then summarized these results at the viral cluster level. This led to host-range predictions for 392 out of 867 viral clusters—all with confidence assessed by comparison to a null model (Supplementary Fig. 2 and Supplementary Table 3).

The hosts of the 38 globally abundant viral clusters were largely restricted to abundant and widespread epipelagic-ocean microbes that were previously identified via $_{mi}$TAG-based operational taxonomic unit (OTU) counts in *Tara* Oceans microbial metagenomes[18]. Notably, the four ubiquitous and abundant viral clusters were predicted to infect seven of the eight globally abundant microbial groups (these were Actinobacteria, Alpha-, Delta-, and Gammaproteobacteria, Bacteroidetes, Cyanobacteria, Deferribacteres) (Fig. 2c and Extended Data Fig. 4). The eighth abundant microbial group, Euryarchaeota, was not linked to these 4 viral clusters, but was predicted as a host for 3 of the 34 other abundant viral clusters (VC_3, VC_27, and VC_63; Extended Data Fig. 3). Among the 38 abundant viral clusters, the number of viral clusters that was predicted to infect a given microbial host phylum (or Proteobacterial class) was positively correlated with the global richness of the host rather than its relative abundance (Extended Data Fig. 4b). This suggests that widespread and abundant hosts that are minimally diverse (for example, Cyanobacteria) provide few viral niches, whereas more diverse host groups, even at lower abundance (for example, Betaproteobacteria), provide more opportunity for viral niche differentiation, probably because ocean viruses appear to be globally distributed[4]. Thus, these host associations provide critically needed empirical support for hypotheses derived from global virus–host network models[19].

Having mapped viral diversity and predicted virus–host pairings, we next sought to identify the virus-encoded AMGs that could modify

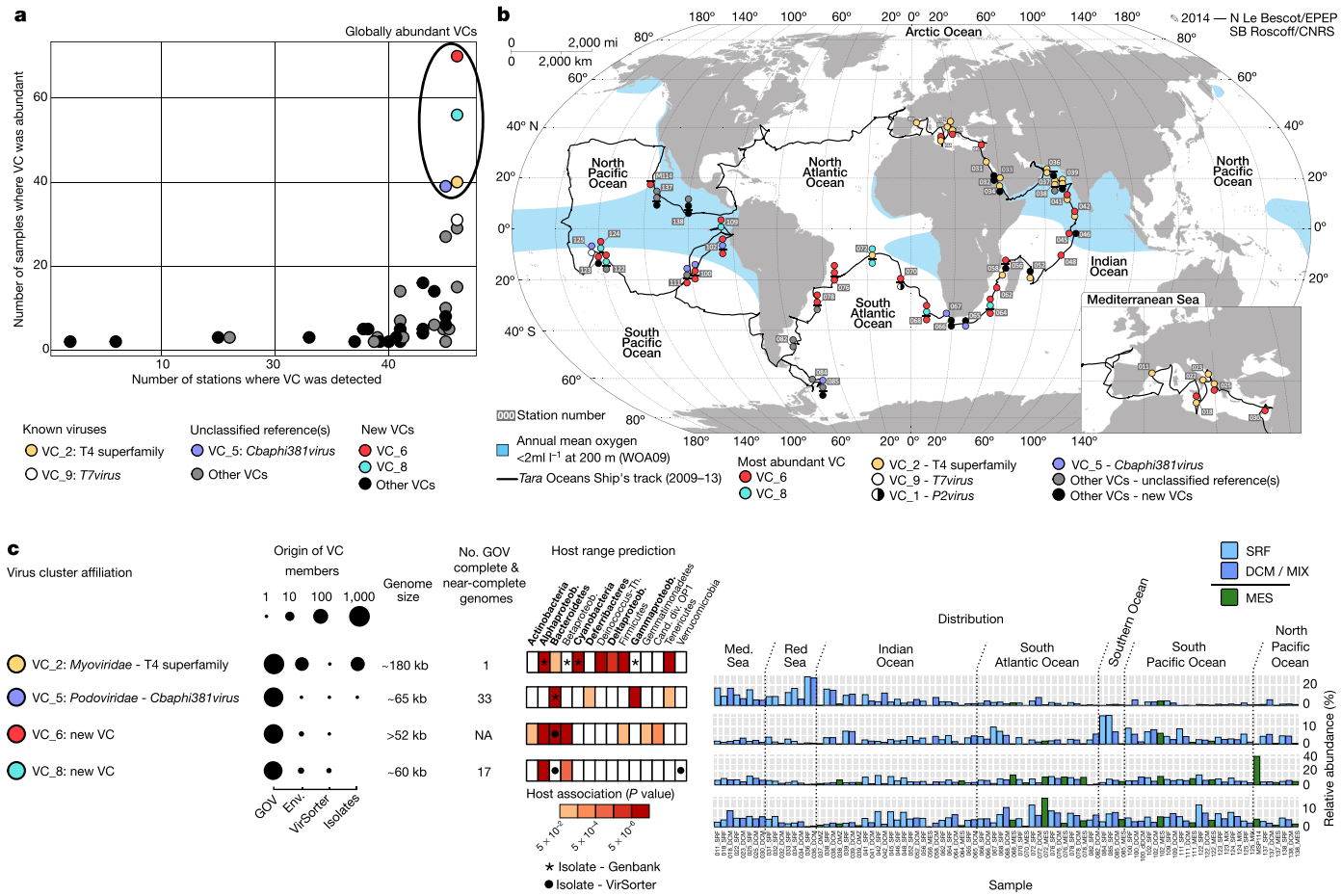


Figure 2 | Characterization of the dominant oceanic viral clusters. **a**, Distribution and abundance of the 38 recurrently abundant viral clusters (VCs) according to the total number of stations in which members of the viral cluster were detected (*x* axis) and the number of samples in which the viral cluster was detected in the abundant fraction (*y* axis). 'Known viruses' are viral clusters with International Committee on Taxonomy of Viruses (ICTV)-classified reference sequences, 'unclassified reference(s)' are viral clusters with isolate genomes lacking ICTV classification, and 'New VCs' are composed solely of environmental sequences. **b**, GOV samples with their most abundant viral cluster mapped to station locations. Samples are stacked vertically when multiple depths are available, with a horizontal line separating epipelagic from mesopelagic layers. Map modified with permission from N. Le Bescot, EPEP, CNRS Station Biologique Roscoff. **c**, Summary of the four globally abundant viral cluster affiliations, origin of viral cluster members (Env: environmental viral sequences), estimated genome sizes, predicted host ranges and distributions (relative abundances are indicated as a percentage of the viral populations identified). The abundant epipelagic microbial groups (representing >1% of the microbial OTU abundance of epipelagic samples) are highlighted in bold. Alphaproteob., Alphaproteobacteria; Betaproteob, Betaproteobacteria; Cand div OP1, Candidate division OP1; Deinococcus-Th, Deinococcus-Thermus; Deltaproteob, Deltaproteobacteria; Gammaproteob, Gammaproteobacteria, Med. Sea, Mediterranean Sea. Oceanic basins are indicated for viral cluster distributions. SRF, Surface; DCM/MIX, Deep chlorophyll maximum/bottom of mixed layer when no deep chlorophyll maximum was observed (stations 123, 124, and 125); MES, Mesopelagic.

host metabolism during infection, probably affecting biogeochemistry. To maximize AMG detection, all 298,383 viral contigs >1.5 kb were examined, including small contigs not associated with a viral population. This revealed 243 putative AMGs (Supplementary Table 6). While 95 of these AMGs were known[20], others offer insights into how viruses may directly manipulate microbial metabolism. Here we focus on four, *dsrC*, *soxYZ*, *P-II* and *amoC*, because of their putative roles in sulfur or nitrogen cycling (see Extended Data Table 1, Supplementary Figs 3–6 and Supplementary Information for functional affiliation of these AMGs). Three of these are not characterized in viruses and one, *dsrC*, has only been observed in viruses from anoxic deep-sea environments[11,21].

Sulfur oxidation in seawater involves two central microbial pathways, dissimilatory sulfur reductase (Dsr) and sulfur oxidation (Sox)[22]. Analyses of GOV AMGs revealed that epipelagic viruses encode key genes for each of these pathways. First, 11 *dsrC*-like genes were identified in viral contigs (Extended Data Fig. 5). The Dsr operon is used by sulfate/sulfite-reducing microbes in anoxic environments, as well as sulfur-oxidizing bacteria in both oxic and anoxic environments[22] (Fig. 3a). DsrC, via its conserved C-terminal motif ($Cys_BX_{10}Cys_A$), provides sulfur to the DsrAB sulfite reductase for processing, thus dictating sulfur metabolism rates[23]. Other DsrC-like proteins (also known as TusE) lack Cys_B and instead participate in tRNA modification[24]. In GOV, four clades of DsrC-like sequences were similar to TusE (DsrC-1 to DsrC-4), whereas the fifth (DsrC-5) was similar to bona fide DsrC (Extended Data Fig. 5, Extended Data Table 1, Supplementary Fig. 3, and Supplementary Information). Second, four *soxYZ* genes were identified on viral contigs (Extended Data Fig. 6). Like DsrC, SoxYZ is an important sulfur carrier that harbours a conserved functional motif identified in all GOV SoxYZ proteins[25] (Fig. 3a, Supplementary Fig. 4, and Supplementary text).

The presence of other AMGs suggests that marine viruses may manipulate nitrogen cycling. We found 10 GOV contigs that encoded *P-II*, a gene widespread across bacteria and archaea and central in nitrogen metabolism regulation[26] (Fig. 3b). There were three AMG clades (*P-II-1*, *P-II-2*, and *P-II-4*) that displayed *P-II*-conserved motifs and had predicted structures similar to bona fide *P-II*. The fourth clade (*P-II-3*) is, however, functionally ambiguous as it lacks a conserved motif (Supplementary Fig. 5, and Supplementary Information). There were two *P-II* AMG clades (*P-II-1* and *P-II-4*) that were proximal to

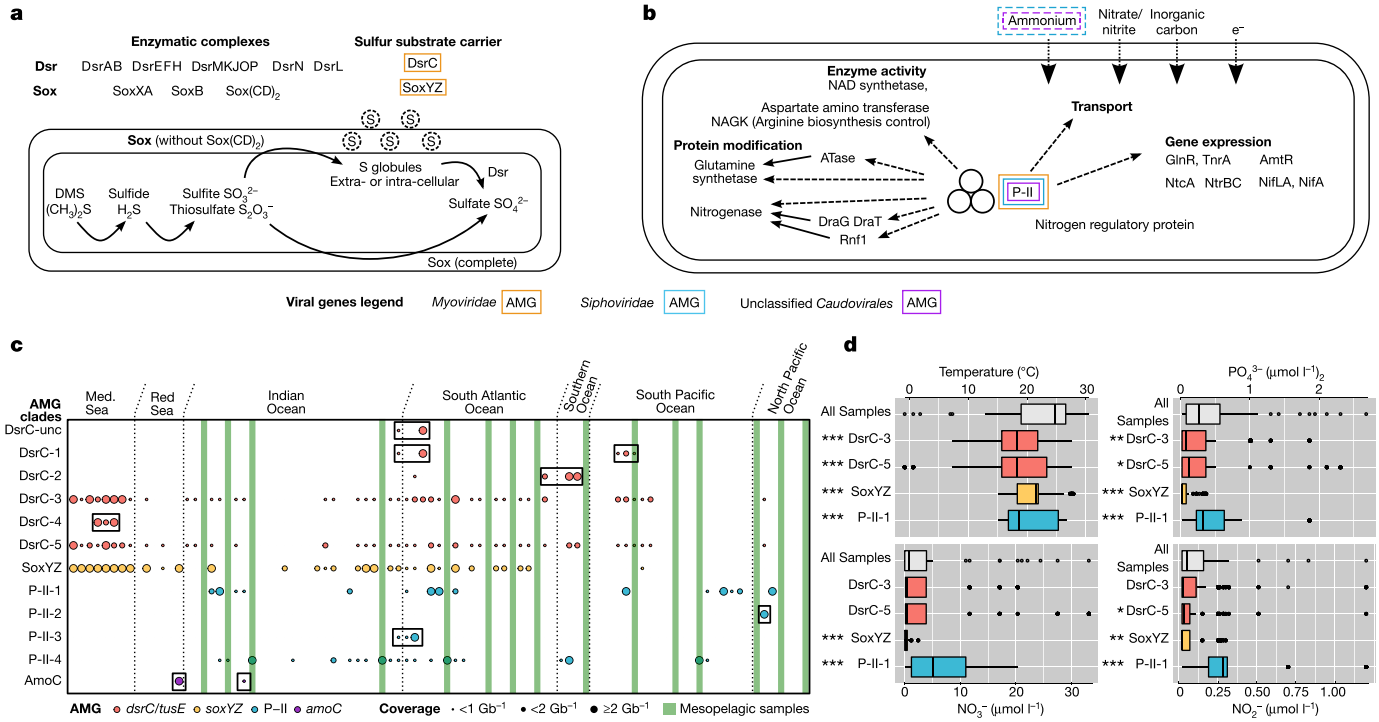


Figure 3 | Characterization and distribution of viral AMGs involved in sulfur and nitrogen cycles. a, **b**, Schematics for microbial sulfur oxidation pathways involving *dsr* and *sox*, the two main gene clusters (**a**), and the central role of the P-II protein in cell regulation (**b**, adapted from images in refs 26, 31). AMG colour outlines indicate viral taxonomic affiliation. Ammonium transporters detected next to viral P-II are highlighted with a dashed outline. **c**, Distribution of viral AMG clades with mesopelagic samples highlighted in green and geographically restricted clades outlined. **d**, Temperature and nutrient conditions in which widespread epipelagic AMGs tend to be most abundant. For each environmental parameter, the range across all epipelagic samples is displayed alongside distributions representing the range of values in which each AMG clade was detected, weighted by the AMG coverage across these samples (see Extended Data Fig. 9 for underlying coverage data). Distributions that differ significantly from the 'all samples' distribution (by two-sided Kolmogorov–Smirnov test) are indicated with asterisks, $*P < 0.05$, $**P < 0.001$, $***P < 0.00001$; boxes represent the first and third quartiles around the median.

the ammonium transporter gene *amt* in GOV contigs (Extended Data Fig. 7). In bacteria, such an arrangement is a signature of *P-II*-like genes that specifically activate alternative nitrogen-production and ammonia-uptake pathways during nitrogen starvation[26]. There was also one GOV contig that included *amoC*, a gene encoding the C subunit of ammonia monooxygenase, suggesting a role in ammonia oxidation[27]. While functional annotation is challenging for these genes[27] and functional motifs are not yet known, the translated AMG was 94% identical to functional AmoC in the phylum Thaumarchaeota—a level of identity only observed among expressed and functional AMGs (Extended Data Fig. 8, Supplementary Fig. 6 and Supplementary Information).

Next, we investigated the origin, evolutionary history, and diversity of these AMGs in epipelagic viruses (Supplementary Information contains additional discussion about taxonomic affiliation and host-prediction for AMG-containing GOV sequences). The 15 GOV contigs that encoded *dsrC* or *soxYZ* genes were, when affiliated, all associated with members of the abundant and ubiquitous T4 superfamily-containing VC_2 (Extended Data Figs 5, 6 and Extended Data Table 1). Phylogenies suggested that these viruses obtained AMGs from sulfur-oxidizing proteobacterial hosts, probably in a single transfer event in the case of *soxYZ* and in two events for *dsrC* (Extended Data Figs 5, 6). Among *dsrC* gene products, the bona fide sulfur-oxidizing DsrC-5 was most closely related to a clade of uncultivated sulfur-oxidizing Gammaproteobacteria (MED13k09, Supplementary Fig. 7). These bacteria are widespread in the epipelagic ocean[28] and are suspected to degrade dimethyl sulfide, a key reduced sulfur species involved in ocean-to-atmosphere sulfur transport and in cloud formation. If confirmed, the infection of these bacteria by DsrC5-encoding viruses would affect critical sulfur-cycling steps throughout surface waters. By contrast, phylogenies suggest that *P-II* AMGs originated from diverse viruses (six viral clusters including the abundant VC_2 and VC_12), and were acquired independently at least four times from Bacteroidetes, Proteobacteria, and possibly Verrucomicrobia (Extended Data Fig. 7 and Supplementary Information). Although a single *amoC* AMG offers only preliminary evaluation of its evolutionary history, this *amoC*-encoding contig appears to represent novel and rare archaeal dsDNA viruses (VC_623) that are predicted to infect ammonia-oxidizing Thaumarchaeota, a phylum known for its major role in global nitrification[29] (Extended Data Fig. 8).

Finally, we investigated the ecology of viruses that encode these AMGs by mapping their distribution across GOVs. We found that seven AMG clades were geographically restricted (*dsrC-unc*, *dsrC-1*, *dsrC-2*, *dsrC-4*, *P-II-2*, *P-II-3*, and *amoC*), whereas five were widespread throughout epipelagic (*dsrC-3*, *dsrC-5*, *soxYZ*, *P-II-1*) or mesopelagic (*P-II-4*) waters (Fig. 3c). All widespread epipelagic AMGs were detected in waters of mid-range temperatures. In contrast, *dsrC-5* and *soxYZ* were predominantly detected in low-nutrient conditions, while *P-II-1* was predominantly detected in high-nutrient conditions (Fig. 3d, Extended Data Fig. 9). Thus, we propose that viruses utilize DsrC-5 or SoxYZ to boost sulfur oxidation rates when infecting sulfur oxidizers in low-nutrient conditions, and P-II under high-nutrient conditions. The latter could be useful to viruses through the activation of high-energy-cost alternative nitrogen-producing pathways typically used only under nitrogen-starvation conditions[26]. Consistent with this, metatranscriptomes from three low-nutrient stations (11_SRF in the Mediterranean Sea, 39_DCM in the Arabian Sea, and 151_SRF in the Atlantic Ocean) revealed expression of viral homologues of *dsrC* and *soxYZ* but not of *P-II* (Extended Data Table 1).

Overall, this systematically collected and processed GOV dataset provides a critical resource for marine microbiology. This map of global

dsDNA ocean viral diversity at the level of population and viral cluster and within viral-encoded AMGs brings global ecological context to abundant surface- and deep-ocean viruses. It will also help to interpret future genomic and metagenomic datasets and help select experimental systems to develop. Together with recent experimental, bioinformatic and theoretical advances[3,12,30], this fundamental resource will accelerate the understanding and prediction of the roles and planetary impacts of viruses in nature.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 8 February; accepted 12 August 2016.
Published online 21 September 2016.

1. Falkowski, P. G., Fenchel, T. & Delong, E. F. The microbial engines that drive Earth's biogeochemical cycles. *Science* **320,** 1034–1039 (2008).
2. Rohwer, F. & Thurber, R. V. Viruses manipulate the marine environment. *Nature* **459,** 207–212 (2009).
3. Brum, J. R. & Sullivan, M. B. Rising to the challenge: accelerated pace of discovery transforms marine virology. *Nat. Rev. Microbiol.* **13,** 147–159 (2015).
4. Brum, J. *et al.* Patterns and ecological drivers of ocean viral communities. *Science* **348,** 1261498 (2015).
5. Karsenti, E. *et al.* A holistic approach to marine eco-systems biology. *PLoS Biol.* **9,** e1001177 (2011).
6. Duarte, C. M. Seafaring in the 21st century: the Malaspina 2010 circumnavigation expedition. *Limnol. Oceanogr.* **24,** 11–14 (2015).
7. Lima-Mendez, G., Van Helden, J., Toussaint, A. & Leplae, R. Reticulate representation of evolutionary and functional relationships between phage genomes. *Mol. Biol. Evol.* **25,** 762–777 (2008).
8. Roux, S., Hallam, S. J., Woyke, T. & Sullivan, M. B. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *eLife* **4,** 1–20 (2015).
9. Mizuno, C. M., Rodriguez-Valera, F., Kimes, N. E. & Ghai, R. Expanding the marine virosphere using metagenomics. *PLoS Genet.* **9,** e1003987 (2013).
10. Chow, C.-E. T., Winget, D. M., White, R. A., III, Hallam, S. J. & Suttle, C. A. Combining genomic sequencing methods to explore viral diversity and reveal potential virus-host interactions. *Front. Microbiol.* **6,** 265 (2015).
11. Roux, S. *et al.* Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-genomics. *eLife* **3,** e03125 (2014).
12. Dutilh, B. E. *et al.* A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* **5,** 4498 (2014).
13. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* **31,** 533–538 (2013).
14. Sullivan, M. B. *et al.* Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. *Environ. Microbiol.* **12,** 3035–3056 (2010).
15. Zhao, Y. *et al.* Abundant SAR11 viruses in the ocean. *Nature* **494,** 357–360 (2013).
16. Labrie, S. J. *et al.* Genomes of marine cyanopodoviruses reveal multiple origins of diversity. *Environ. Microbiol.* **15,** 1356–1376 (2013).
17. Andersson, A. F. & Banfield, J. F. Virus population dynamics and acquired virus resistance in natural microbial communities. *Science* **320,** 1047–1050 (2008).
18. Sunagawa, S. *et al.* Ocean plankton. Structure and function of the global ocean microbiome. *Science* **348,** 1261359 (2015).
19. Flores, C. O., Valverde, S. & Weitz, J. S. Multi-scale structure and geographic drivers of cross-infection within marine bacteria and phages. *ISME J.* **7,** 520–532 (2013).
20. Hurwitz, B. L., Brum, J. R. & Sullivan, M. B. Depth-stratified functional and taxonomic niche specialization in the 'core' and 'flexible' Pacific Ocean Virome. *ISME J.* **9,** 472–484 (2015).
21. Anantharaman, K. *et al.* Sulfur oxidation genes in diverse deep-sea viruses. *Science* **344,** 757–760 (2014).
22. Friedrich, C. G., Bardischewsky, F., Rother, D., Quentmeier, A. & Fischer, J. Prokaryotic sulfur oxidation. *Curr. Opin. Microbiol.* **8,** 253–259 (2005).
23. Santos, A. A. *et al.* A protein trisulfide couples dissimilatory sulfate reduction to energy conservation. *Science* **350,** 1541–1545 (2015).
24. Venceslau, S. S., Stockdreher, Y., Dahl, C. & Pereira, I. A. C. The "bacterial heterodisulfide" DsrC is a key protein in dissimilatory sulfur metabolism. *Biochim. Biophys. Acta* **1837,** 1148–1164 (2014).
25. Dahl, C., Franz, B., Hensen, D., Kesselheim, A. & Zigann, R. Sulfite oxidation in the purple sulfur bacterium Allochromatium vinosum: identification of SoeABC as a major player and relevance of SoxYZ in the process. *Microbiology* **159,** 2626–2638 (2013).
26. Huergo, L. F., Chandra, G. & Merrick & M. P. (II) signal transduction proteins: nitrogen regulation and beyond. *FEMS Microbiol. Rev.* **37,** 251–283 (2013).
27. Stahl, D. A. & de la Torre, J. R. Physiology and diversity of ammonia-oxidizing archaea. *Annu. Rev. Microbiol.* **66,** 83–101 (2012).
28. Loy, A. *et al.* Reverse dissimilatory sulfite reductase as phylogenetic marker for a subgroup of sulfur-oxidizing prokaryotes. *Environ. Microbiol.* **11,** 289–299 (2009).
29. Pester, M., Schleper, C. & Wagner, M. The Thaumarchaeota: an emerging view of their phylogeny and ecophysiology. *Curr. Opin. Microbiol.* **14,** 300–306 (2011).
30. Weitz, J. S. *et al.* A multitrophic model to quantify the effects of marine viruses on microbial food webs and ecosystem processes. *ISME J.* **9,** 1352–1364 (2015).
31. Arcondéguy, T., Jack, R. & Merrick & M. P. (II) signal transduction proteins, pivotal players in microbial nitrogen control. *Microbiol. Mol. Biol. Rev.* **65,** 80–105 (2001).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. Weitz for advice on statistics, C. Pelikan for help with the DsrAB phylogenetic tree, C. Dahl for discussion regarding DsrC function, and members of the Sullivan and the V. Rich laboratories for suggestions and comments on this manuscript. We acknowledge support from UA high-performance computing and the Ohio Supercomputer Center. Sponsors and support for *Tara* Oceans and Malaspina expeditions are listed in the Supplementary Information. This viral research was funded by a National Science Foundation grant (1536989) and Gordon and Betty Moore Foundation grants (3790, 2631) to M.B.S., and the French Ministry of Research and Government through the 'Investissements d'Avenir' program OCEANOMICS (ANR-11-BTBR-0008) and France Genomique (ANR-10-INBS-09-08). Virus researchers were partially supported by the Water, Environmental and Energy Solutions Initiative and the Ecosystem Genomics Institute (S.R.), the Netherlands Organization for Scientific Research Vidi grant 864.14.004 and CAPES/BRASIL (B.E.D.), and the Austrian Science Fund (project P25111-B22, A.L.). Sequencing was provided by Genoscope (*Tara* Oceans) and DOE JGI (Malaspina). All authors approved the final manuscript. This article is contribution number 43 of the *Tara* Oceans expedition.

Author Contributions S.R. and M.B.S. designed the study. C.D., M.P. and S.Se. contributed extensively to sampling collection. S.K.-L. managed the logistics of the *Tara* Oceans project. B.T.P., N.S. and E.L. performed the viral-specific processing of the samples. J.P., C.C., A.A. and P.W. led the sequencing of viral samples. S.R., S.Su. and B.E.D. led the assembly of raw data. S.R., S.Su., M.B.D. and M.B.S. analysed the genomic diversity data. S.R., A.L., J.R.B. and M.B.S. analysed the AMGs data. S.R., J.R.B., B.E.D, S.Su., M.B.D., A.L., S.P., P.B., S.G.A., C.D., J.M.G., D.V. and M.B.S. provided constructive comments, revised and edited the manuscript. *Tara* Oceans Coordinators provided constructive criticism throughout the study. All authors discussed the results and commented on the manuscript.

Author Information All data are fully and freely available from the date of publication, with no restrictions, at EBI, PANGAEA, and iVirus. All of the samples, analyses, publications, and ownership of data are free from legal entanglement or restriction of any sort by the nations in whose waters *Tara* Oceans expedition sampled. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.B.S. (mbsulli@gmail.com).

METHODS

No statistical methods were used to predetermine sample size. These experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

***Tara* Oceans expedition sample collection and processing.** Between 10 October 2009 and 12 December 2011, 90 samples were collected at 45 locations throughout the world's oceans (Supplementary Table 1) through the *Tara* Oceans expedition[32]. These included samples from the following range of depths: surface, deep chlorophyll maximum, bottom of mixed layer when no deep chlorophyll maximum was observed (stations 123, 124, and 125), and mesopelagic samples. The sampling stations were located in 7 oceans and seas, 4 different biomes and 14 Longhurst oceanographic provinces (Supplementary Table 1). For *Tara* station 100, two different peaks of chlorophyll were observed, so two samples were taken at the shallow (100_DCM) and deep (100_dDCM) chlorophyll maximum. For each sample, 20 l of seawater was 0.22 μm-filtered and viruses were concentrated from the filtrate using iron chloride flocculation[33] followed by storage at 4 °C. After resuspension in ascorbic-EDTA buffer (0.1 M EDTA, 0.2 M Mg, 0.2 M ascorbic acid, pH 6.0), viral particles were concentrated using Amicon Ultra 100 kDa centrifugal devices (Millipore), treated with DNase I (100 U ml^{-1}) followed by the addition of 0.1 M EDTA and 0.1 M EGTA to halt enzyme activity and extracted as previously described[34]. In brief, viral particle suspensions were treated with Wizard PCR Preps DNA Purification Resin (Promega) at a ratio of 0.5 ml sample to 1 ml resin, and eluted with Tris-EDTA buffer (10 mM Tris, pH 7.5, 1 mM EDTA) using Wizard Minicolumns. Extracted DNA was Covaris-sheared and size-selected to 160–180 bp sequence lengths, followed by amplification and ligation according to standard Illumina protocol. Sequencing was performed with a HiSeq 2000 system (101 bp, paired end reads).

Temperature, salinity, and oxygen data were collected from each station using an SBE 911plus CTD with Searam recorder and an SBE 43 dissolved oxygen sensor (Sea-Bird Electronics). Nutrient concentrations were determined using segmented flow analysis[35] and included nitrite, phosphate, nitrite-plus-nitrate, and silica. Nutrient concentrations below the detection limit (0.02 μmol kg^{-1}) are reported as 0.02 μmol kg^{-1}. All data from the *Tara* Oceans expedition are available from the European Nucleotide Archive (ENA) (for nucleotide data) and from PANGAEA (for environmental, biogeochemical, taxonomic and morphological data)[36–38].

Malaspina expedition sample collection and processing. Thirteen bathypelagic samples and one mesopelagic sample were collected between 19 April 2011 and 11 July 2011 during the Malaspina 2010 global circumnavigation expedition covering the Pacific and the North Atlantic Oceans. All samples were taken at 4,000 m depth with the exception of two samples from stations 81 and 82, which were collected at 3,500 m and 2,150 m, respectively (Supplementary Table 1). Additionally, station M114 was sampled at the OMZ region at 294 m depth. For each sample, 80 l of seawater was 0.22 μm-filtered and viruses were concentrated from the filtrate using iron chloride flocculation[33], followed by storage at 4 °C. More details about sampling and additional variables used in the Malaspina expedition can be found in ref. 39. Further processing was performed as for the *Tara* Oceans samples other than Illumina sequencing (151 bp, paired end reads).

Contigs assembly. An overview of the contig-generation process is provided in Supplementary Fig. 8. The first step involved the generation of a set of contigs using as many reads as possible from the 104 oceanic viromes. These viromes including 74 epipelagic and 16 mesopelagic samples from the *Tara* Oceans expedition[5] and 1 mesopelagic and 13 bathypelagic samples from the Malaspina expedition[6]. This set of contigs was generated through an iterative cross-assembly[12], using MOCAT[40] and Idba_ud[41], (Supplementary Fig. 8) as follows: (i) high-quality reads were first assembled sample-by-sample with the MOCAT pipeline as described previously[18]; (ii) all reads not mapping (Bowtie 2 (ref. 42), options: -sensitive, -X 2000, -non-deterministic, other parameters at default) to a MOCAT contig (by which we denote 'scaftigs', that is, contigs that were extended and linked using the paired-end information of sequencing read[42]) were assembled sample-by-sample with Idba_ud (iterative *k*-mer assembly, with *k*-mer length increasing from 20 to 100 bp in steps of 20); (iii) all reads that remained unmapped to any contig were then pooled by Longhurst province (that is, unmapped reads from samples corresponding to the same Longhurst province were gathered) and assembled with Idba_ud (with the same parameters as above); and (iv) all remaining reads unmapped from every sample were gathered for a final cross-assembly (using Idba_ud). This resulted in 10,845,515 contigs (Supplementary Fig. 8b).

Genome binning and re-assembly. As the contigs assembled from the marine viral metagenomes could still contain redundant sequences derived from the same (or closely related) populations, we set out to merge contigs derived from the same population into clusters representing population genomes. To this end, contig sequences were first clustered at 95% global average nucleotide identity (ANI) with cd-hit-est[43] (options: -c 0.95 -G 1 -n 10 -mask NX) (Supplementary Fig. 8b), resulting in 10,578,271 non-redundant genome fragments. Next, we used co-abundance (that is, the correlation between abundance profiles estimated by reads mapping) and nucleotide-usage profiles of the non-redundant contigs to further identify contigs derived from the same populations using Metabat[44]. In brief, Metabat uses Pearson correlation between coverage profiles (determined from the mapping of high-quality reads of each sample to the contigs with Bowtie 2 (ref. 42), options: -sensitive, -X 2000, -non-deterministic, other parameters at default) and tetranucleotide frequencies to identify contigs originating from the same genome (Metabat parameters: 98% minimum correlation, mode 'sensitive'; see Supplementary Text for more detail about the selection of these parameters). The 8,744 bins generated, including 3,376,683 contigs, were further analysed, alongside 623,665 contigs that were not included in any genome bin but were ≥1.5 kb.

In an attempt to better assemble these genome bins, two additional sets of contigs were generated for each genome bin (beyond the set of initial contigs binned by Metabat[44]). These were based on the *de novo* assembly of: (i) all reads mapping to the contigs in the genome bin, and (ii) only reads from the sample displaying the highest coverage for the genome bin (both assemblies with Idba_ud[41]; Supplementary Fig. 8c). The latter assembly might be expected to lead to the 'cleanest' genome assembly because it includes the minimum between-sample sequence variation, lowering the probability of generating a chimaeric contig[45]. The former assembly may be necessary if the virus is locally rare, so that sequences from multiple metagenomes are needed to achieve complete genome coverage. Thus, if the assembly from the single 'highest-coverage' sample was improved or equivalent to the initial assembly (that is, the longest contig in the new assembly representing ≥95% of the longest contig in the initial assembly), this set of contigs was selected as the sequence for this bin ($n = 6{,}423$). This optimal single-sample assembly was thus privileged compared to a cross-assembly (either based on the initial contigs or on the re-assembly of all sequences aligned to that bin). Otherwise, the 'all samples' bin re-assembly was selected if it was equivalent to or better than the initial assembly (longest contig representing ≥95% of the longest initial contig, $n = 999$). The assumption that cross-assembly would be needed for locally rare viruses without a high-coverage sample was confirmed by the comparison between the highest coverage of these two types of bins. On average, bins for which the 'optimal' assembly was selected displayed a maximum coverage of 5.47× per Gb of metagenome, while the bins for which the 'cross-assembly' was selected displayed a maximum coverage of 1.37× per Gb of metagenome (Supplementary Table 2). Finally, if both re-assemblies yielded a longest contig smaller (<95%) than the one in the initial assembly, the bin was considered to be a false-positive (that is, binning of contigs from multiple genomes, $n = 1{,}356$), and contigs from the initial assembly were considered as 'unbinned' (263,006 contigs, added to the 623,665 contigs ≥1.5 kb initially retained as 'unbinned').

Identification of viral contigs and delineation of viral populations. Despite efforts to remove cellular DNA completely during sample preparation, the resulting viral metagenomic datasets can only ever be enriched for viruses[46]. Thus, assembled sequences in the GOV dataset were *in silico* filtered *a posteriori* to identify and remove any clearly non-viral signal. In this way, our purification methods should have greatly enriched for viruses, but the *in silico* decontamination step served as a back-up for problematic samples. Together these two filters mean that virtually no known cellular signal should have been considered in our analyses. For the *in silico* cleaning step, VirSorter[47] was used to identify and remove microbial contigs using the 'virome decontamination' mode, with every contig ≥10 kb that was not identified as viral considered to be a microbial contig. Sequences predicted to be from prophages were manually curated to distinguish actual prophages (that is, viral regions within a microbial contig) from contigs that belonged to a viral genome and were wrongly predicted as a prophage. Contigs originating from a eukaryotic virus were identified based on best BLAST hit affiliation of the contig-predicted genes against NCBI RefseqVirus (see Supplementary Text).

The genome bins were affiliated as microbial (if 1 or more contigs were identified as microbial, $n = 1{,}763$), eukaryotic virus (if contigs affiliated as eukaryotic virus comprised more than 10 kb or more than 25% of the genome bin total length, $n = 962$) or viral (that is, archaean and bacterial viruses, $n = 4{,}341$), with the 356 remaining bins that lacked a contig long enough for an accurate affiliation considered as 'unknown' (see Supplementary Text).

Viral bins were then refined to evaluate whether they corresponded to a single viral population or to a mix. To that end, the Pearson correlation and Euclidean distance between abundance profiles (that is, the profile of the average coverage depth of a contig across the 104 samples) of bin members and the bin seed (that is, the largest contig) were computed, and a single-copy viral marker gene (*terL*) was identified in binned contigs (Supplementary Fig. 8e). Thresholds were chosen to maximize the number of bins with exactly one *terL* gene and minimize the number of bins with multiple *terL* genes (Supplementary Fig. 8g). For each bin, contigs with a Pearson correlation coefficient to the bin seed of <0.96 or a Euclidean distance to the seed of >1.05 were removed from the bin, and added to the pool of unbinned contigs. Eventually, every bin still displaying multiple *terL* genes after

this refinement step were split and all corresponding contigs added to the pool of 'unbinned' contigs (Supplementary Fig. 8e).

The final set of contigs was formed by compiling: (i) all contigs belonging to a viral bin, (ii) 'unbinned' viral contigs (that is, contigs affiliated to archaeal and bacterial virus and not part of any genome bin), and (iii) viral contigs identified in microbial or eukaryote virus bins (considered as 'unbinned' contigs, Supplementary Fig. 8f). Within this set of contigs, all viral bins were considered as viral populations, as well as every unbinned viral contig of ≥10 kb, leading to a total of 15,222 epipelagic and mesopelagic populations, and 58 bathypelagic populations (Supplementary Fig. 1, Supplementary Table 2 and Supplementary Information). In this study, we focus only on the 15,222 epipelagic and mesopelagic populations, totaling 24,353 contigs. For the detection of AMGs, we added to these populations all short epipelagic and mesopelagic unbinned viral contigs (<10 kb), totalling 298,383 contigs.

Dataset of publicly available viral genomes and genome fragments. Genomes of viruses associated with a bacterial or archaeal host were downloaded from NCBI RefSeq (1,680 sequences, v70, 05-26-2015; http://www.ncbi.nlm.nih.gov/refseq/). To complete this dataset of reference genomes, viral genomes and genome fragments available in GenBank (http://www.ncbi.nlm.nih.gov/genbank/) but not in RefSeq were downloaded (July 2015) and manually curated to select only bacterial and archaeal viruses (1,017 sequences). These included viral genomes not yet added to RefSeq, as well as genome fragments from fosmid libraries generated from seawater samples[9,10]. Mycophage sequences (available at http://phagesdb.org[48]) were downloaded in July 2015 and included as well if not already in RefSeq (734 sequences). Finally, 12,498 viral genome fragments from the VirSorter Curated Dataset, identified in publicly available microbial genome sequencing projects, were added to the database[8].

Genome (fragments) clustering through gene-content based network analysis. Proteins predicted from 14,650 large GOV contigs (≥10 kb and ≥10 genes), were added to all proteins from the publicly available viral genomes and genomes fragments gathered, and compared through all-vs-all blastp, with a threshold of 10^{-5} for *E*-value and 50 for bit score. Protein clusters were then defined using MCL (Markov Cluster Algorithm, using default parameters for clustering of proteins, similarity scores as log-transformed *E*-value, and 2 for MCL inflation[49]). We then used vContact (https://bitbucket.org/MAVERICLab/ vcontact) to first calculate a similarity score between every pair of genomes and/or contigs based on the number of protein clusters shared between the two sequences (as in refs 7, 8), and then compute an MCL clustering of the genomes/contigs based on these similarity scores (thresholds of 1 for similarity score, MCL inflation of 2). The resulting viral clusters (clusters including ≥2 contigs and/or genomes), consistent with a clustering based on whole-genome BLAST comparison, corresponded approximately to genus-level taxonomy, with rare cases closer to subfamily-level taxonomy (Extended Data Fig. 2 and Supplementary Information). A total of 1,259 viral clusters were obtained, with 867 including at least one GOV sequence. Notably, however, automatically defined viral clusters serve only as a starting point for assigning viral taxonomy. Current ICTV convention for formal taxonomic consideration of these viral clusters would require the manual comparison of genomes and genome fragments to identify signature genes, compare phylogenetic signals and, ideally, observe morphological features of corresponding viruses, although this process is currently being reviewed as advanced computational analytics and genome datasets, such as those presented here, are being developed.

Viral contig annotation. A functional annotation of all GOV-predicted proteins was based on a comparison to the PFAM domain database v.27 (ref. 50) with HmmSearch[51] (threshold of 30 for bit score and 10^{-3} for *E*-value). Additional putative structural proteins were identified through a BLAST comparison to the protein clusters detected in the viral metaproteomics dataset[52]. This metaproteomics dataset led to the annotation of 13,547 hypothetical proteins lacking a PFAM annotation. A taxonomic annotation of the predicted proteins was performed based on a blastp against proteins from archaeal and bacterial viruses from NCBI RefSeq and GenBank (threshold of 50 for bit score and an *E*-value of 10^{-3}).

Viral clusters were affiliated based on isolate genome members, where available. When multiple isolates were included in the viral cluster, the viral cluster was affiliated to the corresponding subfamily or genus of these isolates (excluding all 'unclassified' cases). This was the case for VC_2 (T4 superfamily[14,15]), and VC_9 (T7 virus[16]). When only one, or a handful of, affiliated isolate genomes were included in the viral cluster and lacked genus-level classification, a candidate name was derived from the isolate (if there were several isolates it was derived from the first one isolated). This was the case for VC_5 (*Cbaphi381virus*; ref. 53), VC_12 (*P12024virus*; ref. 54), VC_14 (*MED4-117virus*), VC_19 (*HMO-2011virus*; ref. 55), VC_31 (*RM378virus*; ref. 56), VC_36 (*GBK2virus*; ref. 57), VC_47 (*Cbaphi142virus*; ref. 53) and VC_277 (*vB_RglS_P106Bvirus*; ref. 58). Otherwise, viral clusters were considered as 'new viral clusters'.

Phage proteomic tree computation and visualization. All publicly available complete genomes (see above), all complete (circular) and near-complete (extrachromosomal genome fragment >50 kb with a terminase) from the VirSorter Curated Dataset and all complete and near-complete GOV contigs were compared to generate a phage proteomic tree, as previously described[9,59]. In brief, a proteomic similarity score was calculated for each pair of genome based on an all-versus-all tblastx similarity as the sum of bit scores of significant hits between two genomes ($E \leq 0.001$, bit score ≥ 30, identity percentage ≥ 30). To normalize for different genome sizes, each genome was also compared to itself to generate a self-score, and the distance between two different genomes was calculated as a Dice coefficient as previously[9]. That is, for two genomes A and B with a proteomic similarity score of AB, the corresponding distance *d* would be: $d = 1 - (2 \times AB)/(AA + BB)$; with AA and BB being the self-score of genomes A and B respectively. For clarity, the tree displayed in Extended Data Fig. 2 includes only non-GOV sequences found in a viral cluster with GOV sequence(s) or within a distance $d < 0.5$ to a GOV sequence, totalling 1,522 reference sequences. iTOL[60,61] was used to visualize and display the tree. Detection and estimation of abundance for viral contigs and populations

The presence and relative abundance of a viral contig in a sample was determined based on the mapping of high-quality reads to the contig sequences, computed with Bowtie 2 (options: -sensitive, -X 2000, -non-deterministic, default parameters otherwise[62]), as previously described[4]. A contig was considered to be detected in a metagenome if more than 75% of its length was covered by aligned reads derived from the corresponding sample. A normalized coverage for the contig was then computed as the average contig coverage (that is, the number of nucleotides mapped to the contig divided by the contig length) normalized by the total number of base pairs sequenced in this sample. The detection and relative abundance of a viral population was based on the coverage of its contigs; that is, a population was considered as detected in a sample if more than 75% of its cumulated length was covered, and its normalized coverage was computed as the average normalized coverage of its contigs.

Relative abundance of viral clusters. The relative abundance of viral clusters was calculated based on the coverage of its members within the 15,222 viral populations identified. If a population included contigs that were all linked to the same viral cluster, or that were linked to a single viral cluster (except for unclustered contigs owing to short length), this population coverage was added to the total of the corresponding viral cluster. In the rare cases where the link between population and viral cluster was ambiguous because different contigs within a population pointed towards different viral clusters ($n = 475$, that is, 3.1% of the populations), the population coverage was equally split between these viral clusters. Finally, if no contig in the population belonged to any viral cluster ($n = 2{,}605$, 17% of the populations), the population coverage was added to the 'unclustered' category. Eventually, for each sample, the cumulative coverage of a viral cluster was normalized by the total coverage of all populations to calculate a relative abundance of the viral cluster among viral populations.

The selection of abundant viral clusters within a sample was based on the contribution of the viral cluster to the sample diversity as measured by the Simpson index. For each sample, the overall Simpson index was first calculated with all viral clusters. Following this, viral clusters were sorted by decreasing relative abundance and progressively added to a new calculation of the Simpson index. Viral clusters considered as abundant were the ones which, once cumulated, represented 80% of the sample diversity (that is, a Simpson index ≥80% of the sample total Simpson index; Extended Data Fig. 1c). The 38 viral clusters that were identified as abundant in at least 2 different stations were selected as 'recurrently abundant viral clusters in the GOV dataset' (Fig. 2 and Extended Data Fig. 3).

Host prediction and diversity. Three different approaches were used to link viral contigs and putative host genomes: blastn similarity, CRISPR spacer similarity and tetranucleotide frequency similarities. An overview of the contig-generation process is provided in Supplementary Fig. 8, and an extended discussion about the efficiency and raw results of these host prediction methods is provided in Supplementary Information, Supplementary Table 4, and ref. 63. A list of all host predictions by viral sequence is available in Supplementary Table 5.

Generation of host database. A genome database of putative hosts for the epipelagic and mesopelagic GOV viruses was generated, including all archaeal and bacterial genomes annotated as 'marine' from NCBI RefSeq and WGS (both times only sequences ≥5 kb, 184,663 sequences from 4,452 genomes, downloaded in August 2015), and all contigs ≥5 kb from the 139 *Tara* Oceans microbial metagenomes corresponding to the bacterial and archaeal size fraction (791,373 sequences)[18]. For these microbial metagenomic contigs, a first blastn alignment was computed to compare with all GOV contigs, and exclude from the putative host dataset all metagenomic contigs with a significant similarity to a viral GOV sequence (thresholds of 50 for bit score, 0.001 for *E*-value, and 70% for identity percentage) on ≥90% of their length, as these are likely to be sequences of viral origin sequenced in the bacteria and archaea size fraction (these represented 2.2%

of the contigs in the assembled microbial metagenomes). The taxonomic affiliation of NCBI genomes was taken from the NCBI taxonomy. For *Tara* Oceans contigs, a last common ancestor (LCA) affiliation was generated for each contig based on genes affiliation[18], if three or more genes on the contig were affiliated.

BLAST-based identification of sequence similarity between viral contigs and host genome. All GOV viral contigs were compared to all archaeal and bacterial genomes and genome fragments with a blastn (threshold of 50 for bit score and 0.001 for *E*-value), to identify regions of similarity between a viral contig and a microbial genome, indicative of a prophage integration or horizontal gene transfer[63]. A host prediction was made when: (i) a NCBI genomes displayed a region similar to a GOV viral contig $\geq$5 kb at $\geq$70% identity, or (ii) when a *Tara* Oceans microbial metagenomic contig ($\geq$5 kb) displayed a region similar to a GOV viral contig $\geq$2.5 kb at $\geq$70% identity.

Matches between GOV viral contigs and CRISPR spacers. CRISPR arrays were predicted for all putative host genomes and genome fragments (NCBI microbial genomes and *Tara* Oceans microbial metagenomic contigs) with MetaCRT[64,65]. CRISPR spacers were extracted, and all spacers with ambiguous bases or low complexity (that is, consisting of 4–6 bp repeat motifs) were removed. All remaining spacers were matched to viral contigs with fuzznuc[66], with no mismatches allowed, which, although rarely, observed yields highly accurate host predictions[63] (Supplementary Table 4).

Nucleotide composition similarity: comparison of tetranucleotide frequency. Bacterial and archaeal viruses tend to have a genome composition close to the genome composition of their host, a signal that can be used to predict viral–host pairs[8,63,67]. Here, canonical tetranucleotide frequencies were observed for all viral and host sequences using Jellyfish[68] and mean absolute error (that is, the average of absolute differences) between tetranucleotide-frequency vectors were computed with in-house Perl and Python scripts for each pair of viral and host sequence as previously reported[8]. A GOV viral contig was then assigned to the closest sequence (that is, lowest distance '*d*') from the pool of NCBI genomes if $d < 0.001$ (because both the tetranucleotide-frequency signal and the taxonomic affiliation of these complete genomes are more robust than for metagenomic contigs), and otherwise assigned to the closest (that is, lowest distance) *Tara* Oceans microbial contig if $d < 0.001$.

Summarizing host prediction at the viral-cluster level. Overall, 3,675 GOV contigs could be linked to a putative host group among the 24,353 GOV contigs associated with an epipelagic or mesopelagic viral population. To summarize these affiliations at the viral cluster level, a Poisson distribution was used to estimate the number of expected false-positive associations for each viral cluster–host group combination based on: (i) the global probability of obtaining a host prediction across all pairs of viral and host sequences tested and for all methods (5.8×10^{-8}), (ii) the number of potential predictions generated for the viral cluster, corresponding to 3 times the number of sequences in the viral cluster (to take into account the three methods) and (iii) the number of sequences from the host group in the database (Supplementary Fig. 2). By comparing the number of links observed between a viral cluster and a host group to this expected value, which takes into account the bias in database (that is, some host groups will be over- or under-represented in our set of archaeal and bacterial genomes and genome fragments) and the bias linked to the variable number of sequences in viral clusters, we can determine if the number of associations observed for any combination of viral cluster and host group is likely to be due to chance alone (and calculate the associated *P* value).

Microbial community diversity and richness indexes. Diversity and richness indices for putative host populations were based on the OTU abundance matrix generated from the analysis of $_{mi}$TAGs in *Tara* Oceans microbial metagenomes[18]. These indexes were computed for each host group at the same taxonomic level as the host prediction (that is, the phylum level, except for Proteobacteria where the class level is used). The R package vegan[69] was used to estimate for each group: (i) a global Chao index (that is, including all OTUs from all samples) through the function estaccumR, (ii) a sample-by-sample Chao index with the function estimateR, and (iii) Sorensen indexes between all pairs of samples with the function betadiver. Diversity indices presented in Extended Data Fig. 4 are based solely on epipelagic samples as the 38 viral clusters identified as abundant were mostly retrieved in epipelagic samples. Candidate division OP1 was excluded from this analysis because no OTU affiliated to this phylum was identified.

Detection of AMGs. Predicted proteins from all GOV viral contigs were compared to the PFAM domain database (hmmsearch[51], threshold of 40 for bit score and 0.001 for *E*-value), and all PFAM domains detected were classified into 8 categories: 'structural', 'DNA replication, recombination, repair, nucleotide metabolism', 'transcription, translation, protein synthesis', 'lysis', 'membrane transport, membrane-associated', 'metabolism', 'other', and 'unknown' (as in ref 20). Four AMGs (similar to a domain from the 'metabolism' category) were then selected for further study owing to their central role in sulfur (*dsrC* and *soxYZ*) or nitrogen (*P-II*, *amoC*) cycle, and the fact that these had never been detected in a surface ocean viral genome thus far (*dsrC*/*tusE*-like genes have been detected in deep water viruses[11,21]). To evaluate if an AMG was 'known', a list of PFAM domain detected in NCBI RefSeqVirus and Environmental Phages was computed based on a similar hmmsearch comparison (threshold of 40 for bit score and 0.001 for *E*-value), and augmented by manual annotation of AMGs from refs 20, 70. These corresponded, for the most part, to photosynthesis and carbon metabolism AMGs previously described in cyanophages[71–75]. The complete list of PFAM domains detected in GOV viral contigs is available in Supplementary Table 6.

Phylogenetic tree generation and contigs map comparison. Sequences similar to the four AMGs described in the previous paragraph were recruited from the *Tara* Oceans microbial metagenomes[18], based on a blastp of all predicted proteins from microbial metagenome to the viral AMGs identified (threshold of 100 for bit score, 10^{-5} for *E*-value, except for *P-II* where a threshold of 170 for bit score was used because of the high number of sequences recruited). The viral AMG sequences were also compared to NCBI nr database (blastp, threshold of 50 for bit score and 10^{-3} for *E*-value) to recruit relevant reference sequences (up to 20 for each viral AMG sequence). These sets of viral AMGs and related protein sequences were then aligned with Muscle[76], the alignment manually curated to remove poorly aligned positions with Jalview[77], and two trees were computed from the same curated alignment: a maximum-likelihood tree with FastTree (v2.7.1, model WAG, other parameters set to default[78]) and a bayesian tree with MrBayes (v3.2.5, mixed evolution models, other parameters set to default, 2 MCMC chains were run until the average standard deviation of split frequencies was $<$0.015, relative burn-in of 25% used to generate the consensus tree[79]). In all cases except for AmoC, the mixed model used by MrBayes was 100% WAG, confirming that this model was well suited for archaeal and bacterial virus protein trees. Manual inspection revealed only minor differences between each pair of trees, so a Shimodaira–Hasegawa (SH) test was used to determine which tree best fitted the sequence alignment, using the R library phangorn[80]. Itol[60] was used to visualize and display these trees, in which branches with supports $<$40% were collapsed. Annotated interactive trees are available online at http://itol.embl.de/shared/Siroux. Contigs map comparison were generated with Easyfig[81], following the same method used for the viral clusters (see Supplementary Information).

Functional characterization of putative AMGs. Conserved motifs were identified on the different AMGs based on the literature: *dsrC*-conserved motifs were obtained from ref. 24, *soxYZ* conserved residues were identified from the PFAM domains PF13501 and PF08770, and *P-II* conserved motifs identified from PROSITE documentation PDOC00439. A 3D structure could also be predicted for *P-II* AMGs by I-TASSER[82] (default parameters), the quality of these predictions being confirmed with ProSA web server[83]. To further confirm the functionality of these genes, selective constraint on these AMGs was evaluated through pN/pS calculation, as previously[84]. In brief, synonymous (pS) and non-synonymous (pN) SNPs were observed in each AMG, and compared to expected ratio of synonymous and non-synonymous SNPs under a neutral evolution model for these genes. The interpretation of pN/pS is similar as for dN/dS analyses, with the operation of purifying selection leading to pN/pS values $<$1. Finally, AMG transcripts were searched in metatranscriptomic datasets, generated by the *Tara* Oceans consortium (ENA Id ERS1092158, ERS488920, and ERS494518). To generate these metatranscriptomes, bacterial rRNA depletion was carried out on 240–500 ng total RNA using Ribo-Zero Magnetic Kit for Bacteria (Epicentre) for 0.2–1.6 μm and 0.22–3 μm filters. The Ribo-Zero depletion protocol was modified to be adapted to low RNA input amounts[85]. Depleted RNA was used to synthetize cDNA with SMARTer Stranded RNA-Seq Kit (Clontech)[85]. Metatranscriptomic libraries were quantified by quantitive PCR using the KAPA Library Quantification Kit for Illumina Libraries (KapaBiosystems) and library profiles were assessed using the DNA High Sensitivity LabChip kit on an Agilent Bioanalyzer (Agilent Technologies). Libraries were sequenced on Illumina HiSeq2000 instrument (Illumina) using 100-base-length read chemistry in a paired-end mode. High-quality reads were then mapped to viral contigs containing *dsrC*, *soxYZ*, *P-II*, or *amo*C genes with SOAPdenovo2[42] within MOCAT[40] (options 'screen' and 'filter' with length and identity cutoffs of 45% and 95%, respectively, and paired-end filtering set to 'yes'), and coverage was defined for each gene as the number of base pairs mapped divided by gene length (including only those reads mapped to the predicted coding strand).

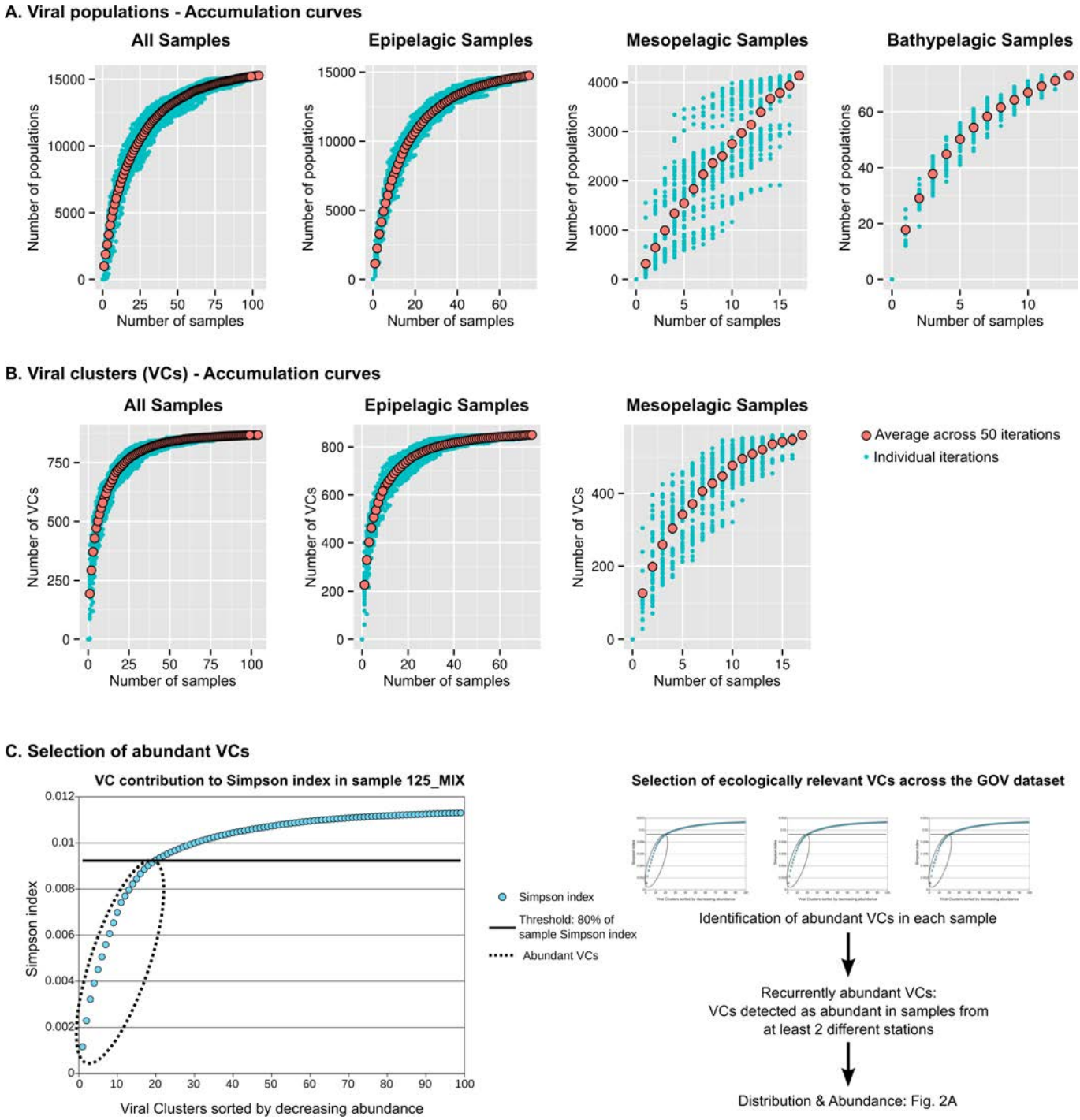
Distribution of AMGs and association with geochemical metadata. The distribution and relative abundance of AMGs was based on the readmapping and normalized coverage of the contig that included the AMG. To get a range of temperature and nutrient concentrations for the widespread AMGs (those detected in $>$5 stations) that takes into account both the samples in which these AMGs were detected and the differences in normalized coverage, a set of samples was selected through a weighted random selection with replacement, with the weight of each

sample corresponding to the normalized coverage of the AMG. This ensured that a range of temperature or nutrient concentration values associated with the distribution and abundance of the AMG could be generated for each AMG and each environmental parameter tested. The number of samples randomly selected for each AMG was the same as the total number of samples for which a value of this parameter was available.

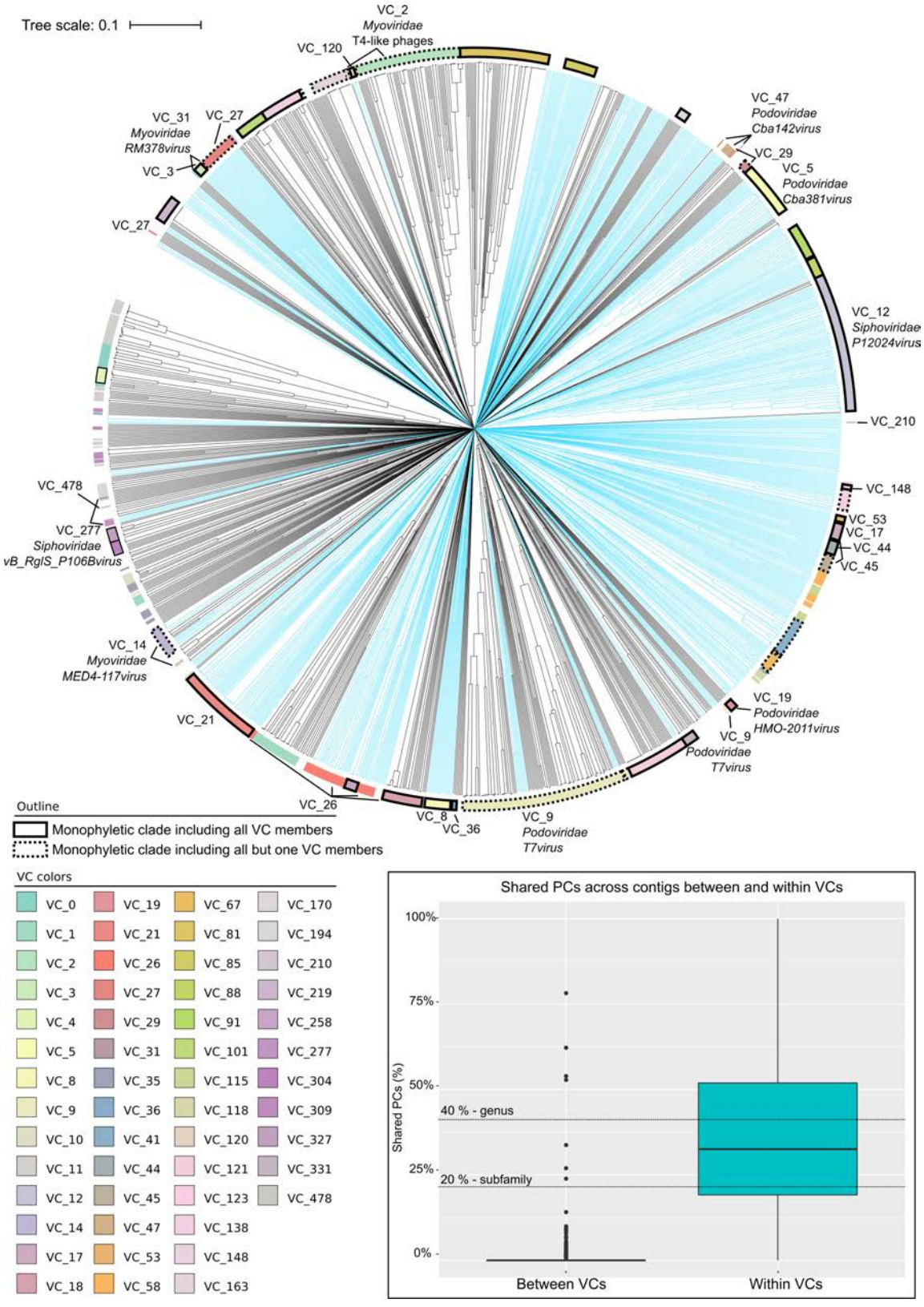
Code availability. Scripts used in this manuscript are available on the Sullivan laboratory bitbucket under project GOV_Ecogenomics (http://bitbucket.org/MAVERICLab/gov_ecogenomics/overview). Scripts used in the assessment of microbial diversity are gathered in the directory Host_diversity, the ones used for host predictions are in Host_prediction, and the scripts used to identify abundant viral clusters are in Virus_clusters_prevalence.

Data availability. All raw reads are available through ENA (*Tara* Oceans) or IMG (Malaspina) using the dataset identifiers listed in Supplementary Table 1. Processed data are available through iVirus (http://mirrors.iplantcollaborative.org/browse/iplant/home/shared/iVirus/GOV/), including all sequences from assembled contigs, lists of viral populations and associated annotated sequences as GenBank files, viral clusters composition and characteristics, map comparisons of genomes and contigs of the 38 abundant viral clusters and host predictions for viral contigs.

32. Pesant, S. *et al.* Open science resources for the discovery and analysis of *Tara* Oceans data. *Sci. Data* **2,** 150023 (2015).
33. John, S. G. *et al.* A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environ. Microbiol. Rep.* **3,** 195–202 (2011).
34. Hurwitz, B. L., Deng, L., Poulos, B. T. & Sullivan, M. B. Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. *Environ. Microbiol.* **15,** 1428–1440 (2013).
35. Aminot, A., Kérouel, R. & Coverly, S. in *Practical Guidelines for the Analysis of Seawater* (ed. O. Wurl) 143–176 (CRC Press, 2009).
36. *Tara* Oceans Consortium & *Tara* Oceans Expedition. Registry of all samples from the *Tara* Oceans Expedition (2009–2013). http://dx.doi.org/10.1594/PANGAEA.842197 (2015).
37. *Tara* Oceans Consortium & *Tara* Oceans Expedition. Environmental context of all samples from the *Tara* Oceans Expedition (2009–2013). http://dx.doi.org/10.1594/PANGAEA.853810 (2015).
38. *Tara* Oceans Consortium & *Tara* Oceans Expedition. Biodiversity context of all samples from the *Tara* Oceans Expedition (2009–2013). http://dx.doi.org/10.1594/PANGAEA.853809 (2015).
39. Salazar, G. *et al.* Global diversity and biogeography of deep-sea pelagic prokaryotes. *ISME J.* **10,** 596–608 (2016). 10.1038/ismej.2015.137.
40. Kultima, J. R. *et al.* MOCAT: a metagenomics assembly and gene prediction toolkit. *PLoS One* **7,** e47656 (2012).
41. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28,** 1420–1428 (2012).
42. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* **1,** 18 (2012).
43. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22,** 1658–1659 (2006).
44. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3,** e1165 (2015).
45. Mavromatis, K. *et al.* Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat. Methods* **4,** 495–500 (2007).
46. Roux, S., Krupovic, M., Debroas, D., Forterre, P. & Enault, F. Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. *Open Biol.* **3,** 130160 (2013).
47. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3,** e985 (2015).
48. Pope, W. H. *et al.* Whole genome comparison of a large collection of mycobacteriophages reveals a continuum of phage genetic diversity. *eLife* **4,** e06416 (2015).
49. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30,** 1575–1584 (2002).
50. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42,** D222–D230 (2014).
51. Eddy, S. R. Accelerated Profile HMM Searches. *PLOS Comput. Biol.* **7,** e1002195 (2011).
52. Brum, J. R. *et al.* Illuminating structural proteins in viral "dark matter" with metaproteomics. *Proc. Natl Acad. Sci. USA* **113,** 2436–2441 (2016).
53. Holmfeldt, K. *et al.* Twelve previously unknown phage genera are ubiquitous in global oceans. *Proc. Natl Acad. Sci. USA* **110,** 12798–12803 (2013).
54. Kang, I., Jang, H. & Cho, J.-C. Complete genome sequences of two *Persicivirga* bacteriophages, P12024S and P12024L. *J. Virol.* **86,** 8907–8908 (2012).
55. Kang, I., Oh, H.-M., Kang, D. & Cho, J.-C. Genome of a SAR116 bacteriophage shows the prevalence of this phage type in the oceans. *Proc. Natl Acad. Sci. USA* **110,** 12343–12348 (2013).
56. Hjorleifsdottir, S., Aevarsson, A., Hreggvidsson, G. O., Fridjonsson, O. H. & Kristjansson, J. K. Isolation, growth and genome of the Rhodothermus RM378 thermophilic bacteriophage. *Extremophiles* **18,** 261–270 (2014).
57. Marks, T. J. & Hamilton, P. T. Characterization of a thermophilic bacteriophage of *Geobacillus kaustophilus*. *Arch. Virol.* **159,** 2771–2775 (2014).
58. Halmillawewa, A. P., Restrepo-Córdoba, M., Yost, C. K. & Hynes, M. F. Genomic and phenotypic characterization of *Rhizobium gallicum* phage vB_RglS_P106B. *Microbiology* **161,** 611–620 (2015).
59. Rohwer, F. & Edwards, R. The Phage Proteomic Tree: a genome-based taxonomy for phage. *J. Bacteriol.* **184,** 4529–4535 (2002).
60. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23,** 127–128 (2007).
61. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* **39,** W475–8 (2011).
62. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
63. Edwards, R. A., McNair, K., Faust, K., Raes, J. & Dutilh, B. E. Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiol. Rev.* **40,** 258–272 (2016).
64. Bland, C. *et al.* CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8,** 209 (2007).
65. Rho, M., Wu, Y.-W., Tang, H., Doak, T. G. & Ye, Y. Diverse CRISPRs evolving in human microbiomes. *PLoS Genet.* **8,** e1002441 (2012).
66. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16,** 276–277 (2000).
67. Ogilvie, L. A. *et al.* Genome signature-based dissection of human gut metagenomes to extract subliminal viral sequences. *Nat. Commun.* **4,** 2420 (2013).
68. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27,** 764–770 (2011).
69. Oksanen, J. *et al.* The vegan package version 2.4-0; https://cran.r-project.org/web/packages/vegan/index.html (2016).
70. Sharon, I. *et al.* Comparative metagenomics of microbial traits within oceanic viral communities. *ISME J.* **5,** 1178–1190 (2011).
71. Thompson, L. R. *et al.* Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proc. Natl Acad. Sci. USA* **108,** E757–E764 (2011).
72. Dammeyer, T., Bagby, S. C., Sullivan, M. B., Chisholm, S. W. & Frankenberg-Dinkel, N. Efficient phage-mediated pigment biosynthesis in oceanic cyanobacteria. *Curr. Biol.* **18,** 442–448 (2008).
73. Lindell, D., Jaffe, J. D., Johnson, Z. I., Church, G. M. & Chisholm, S. W. Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* **438,** 86–89 (2005).
74. Lindell, D. *et al.* Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. *Nature* **449,** 83–86 (2007).
75. Sullivan, M. B. *et al.* Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biol.* **4,** e234 (2006).
76. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5,** 113 (2004).
77. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25,** 1189–1191 (2009).
78. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* **5,** e9490 (2010).
79. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17,** 754–755 (2001).
80. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27,** 592–593 (2011).
81. Sullivan, M. J., Petty, N. K. & Beatson, S. A. Easyfig: a genome comparison visualizer. *Bioinformatics* **27,** 1009–1010 (2011).
82. Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protocols* **5,** 725–738 (2010).
83. Wiederstein, M. & Sippl, M. J. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res.* **35,** W407–10 (2007).
84. Schloissnig, S. *et al.* Genomic variation landscape of the human gut microbiome. *Nature* **493,** 45–50 (2013).
85. Alberti, A. *et al.* Comparison of library preparation methods reveals their impact on interpretation of metatranscriptomic data. *BMC Genomics* **15,** 912 (2014).

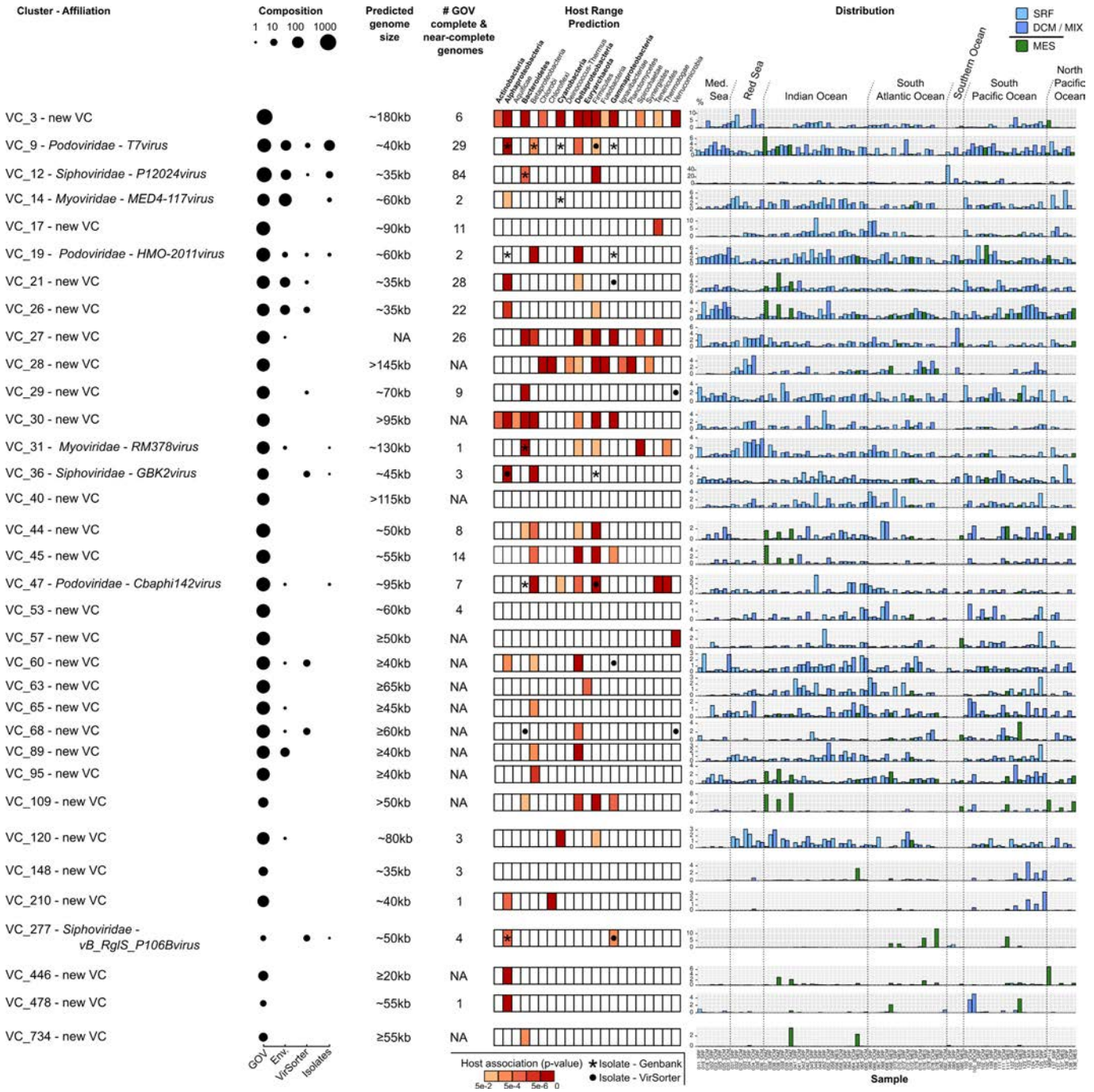


Extended Data Figure 1 | Accumulation curves of populations and viral clusters and identification of abundant viral clusters in GOV samples. **a**, **b**, Accumulation curves for viral populations (**a**) and viral clusters (**b**) were computed from 50 randomly shuffled samples (blue dots) for all samples, epipelagic, mesopelagic, or bathypelagic subsets. For each curve, the average of 50 iterations is displayed with red dots. **c**, Schematic of the selection process of abundant viral clusters. For each sample, viral clusters accounting for (up to) 80% of the sample diversity (as assessed by their Simpson index) was considered as abundant. On the left is an example for sample 125_MIX. Viral clusters detected as abundant in at least two different stations were included in the 38 viral clusters described in Fig. 2 and Extended Data Fig. 3.

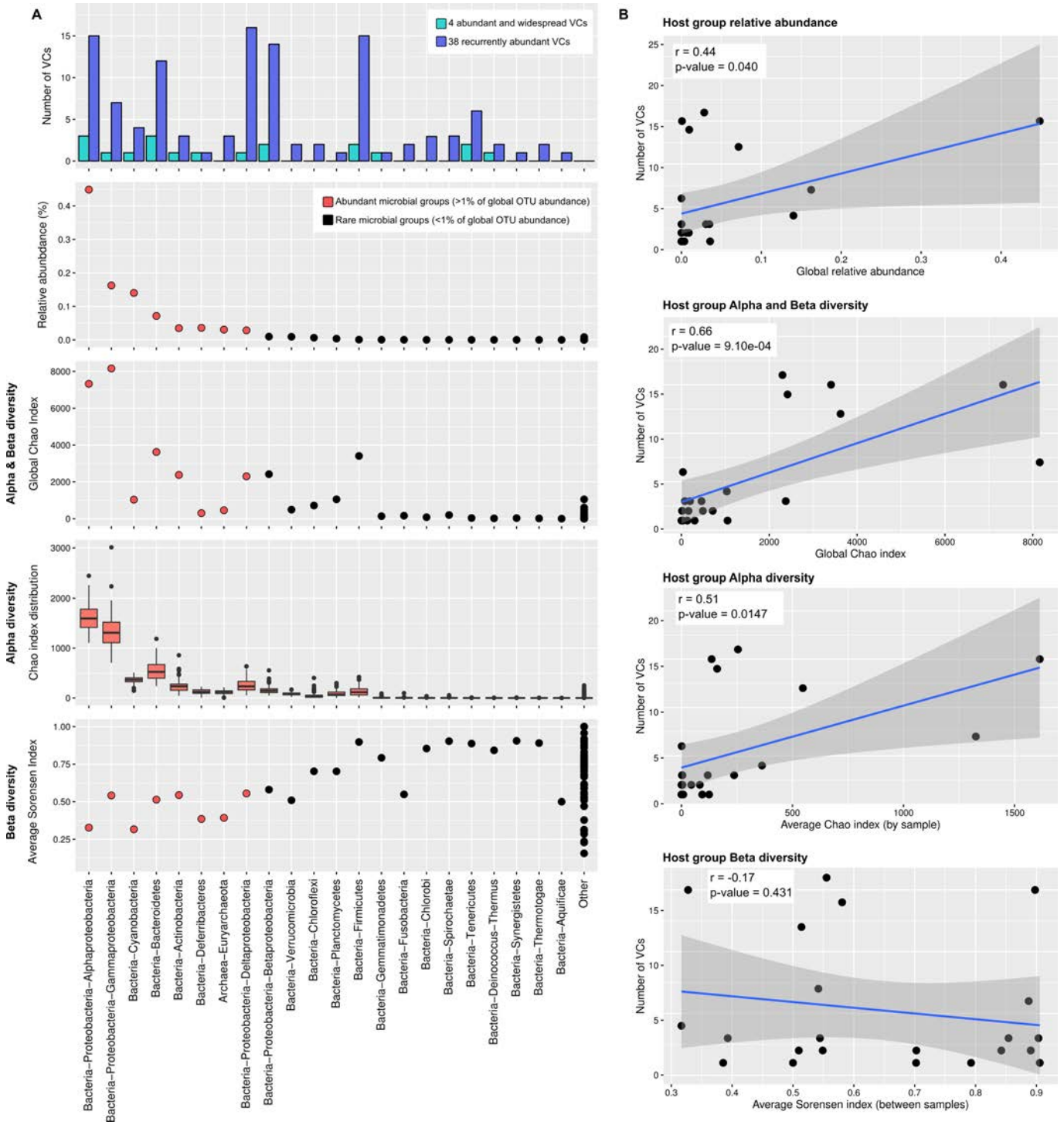


Extended Data Figure 2 | See next page for caption.

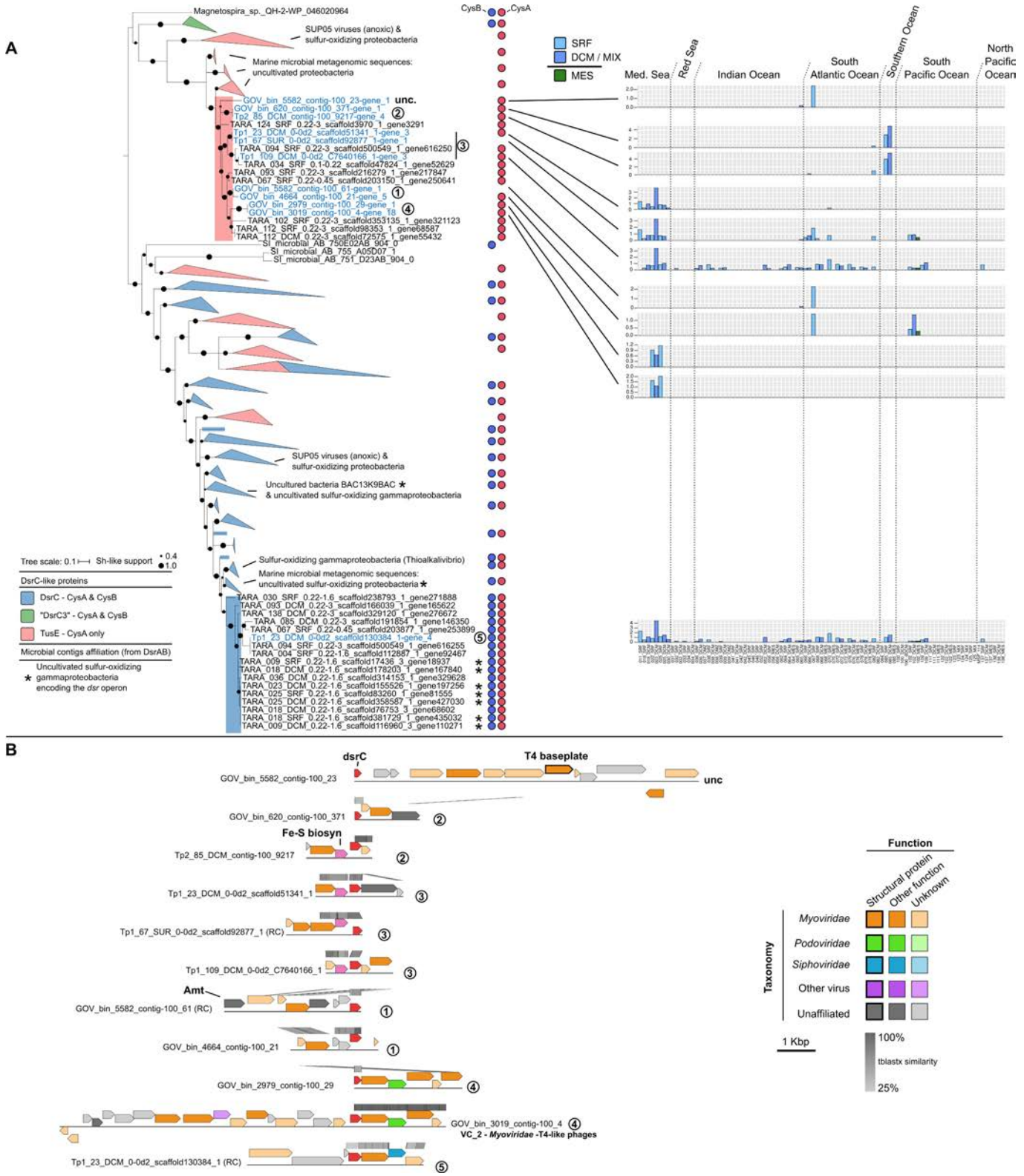
Extended Data Figure 2 | Comparison of viral clusters with other classification methods (phage proteomic tree and percentage of shared genes). The phage proteomic tree includes the 756 GOV complete and near-complete genomes from epipelagic and mesopelagic samples and the closest reference genomes from RefSeq and environmental phages ($d < 0.5$ to a GOV sequence or found in the same viral cluster as a GOV sequence). Branches of monophyletic clades that include more than 3 GOV and/or uncultivated marine sequences with no isolate reference are highlighted in blue. All viral clusters with more than 8 representatives in the tree or part of the 38 abundant viral clusters are indicated by the colours of the outer ring. The name and affiliation (if available) of the 38 abundant viral clusters are indicated next to the viral cluster on the coloured ring. Viral clusters in which members were gathered in single monophyletic clades are indicated with a solid black outline, while viral clusters for which all-but-one member were gathered in a single monophyletic clade are highlighted with a dashed black outline. Distribution of the percentage number of shared genes estimated based on the number of shared protein clusters for viral genome/contigs pairs either between different viral clusters or within viral clusters (bottom right). On average, 73% and 39% of sequences within a viral cluster shared more than 20% and 40% of their genes, respectively, which represent the current thresholds currently accepted for sub-family and genus designations. Similarly, 83% of sequences within a viral cluster were consistently affiliated in the phage proteomic tree as they formed a monophyletic group that included only members of the particular viral cluster. Thus all three classification methods are largely consistent for the GOV dataset (see Supplementary Information).



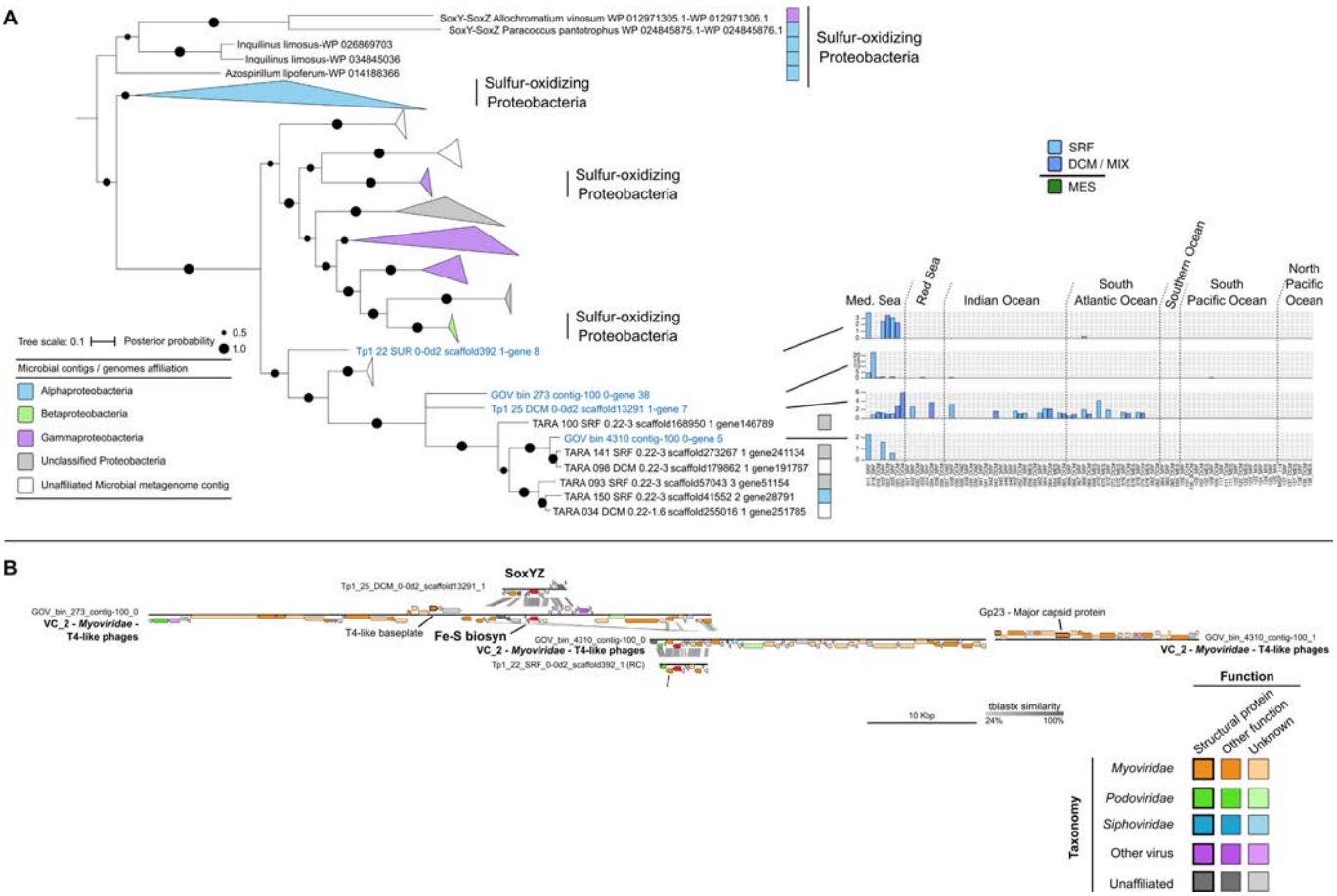
Extended Data Figure 3 | Summary of 34 of the 38 abundant viral clusters. Summaries are given for the 34 abundant viral clusters not summarized in Fig. 2. Predicted genome size is based on the set of isolates and circular contigs in the viral cluster. NA (not applicable) corresponds to viral clusters either without any circular contigs, or for which the relative standard deviation of estimated genome size across the different isolate(s) and/or circular contigs is greater than 15%. Host association values are based on the number of cluster members associated with each host group. Statistical significance of this number of predictions was evaluated by comparison with an expected number of associations calculated using a Poisson distribution. Host associations based on known isolates are indicated with a star (for associations based on cultivated isolates) or a dot (for associations based on the detection of a cluster member in a microbial genome from the VirSorter Curated Dataset). The abundant epipelagic microbial groups (representing >1% of the microbial OTUs in epipelagic samples) are highlighted in bold. Distribution and relative abundance of viral clusters are based on the cumulated coverage of viral cluster members among sample viral populations. The main oceanic basins are indicated for each set of sample.



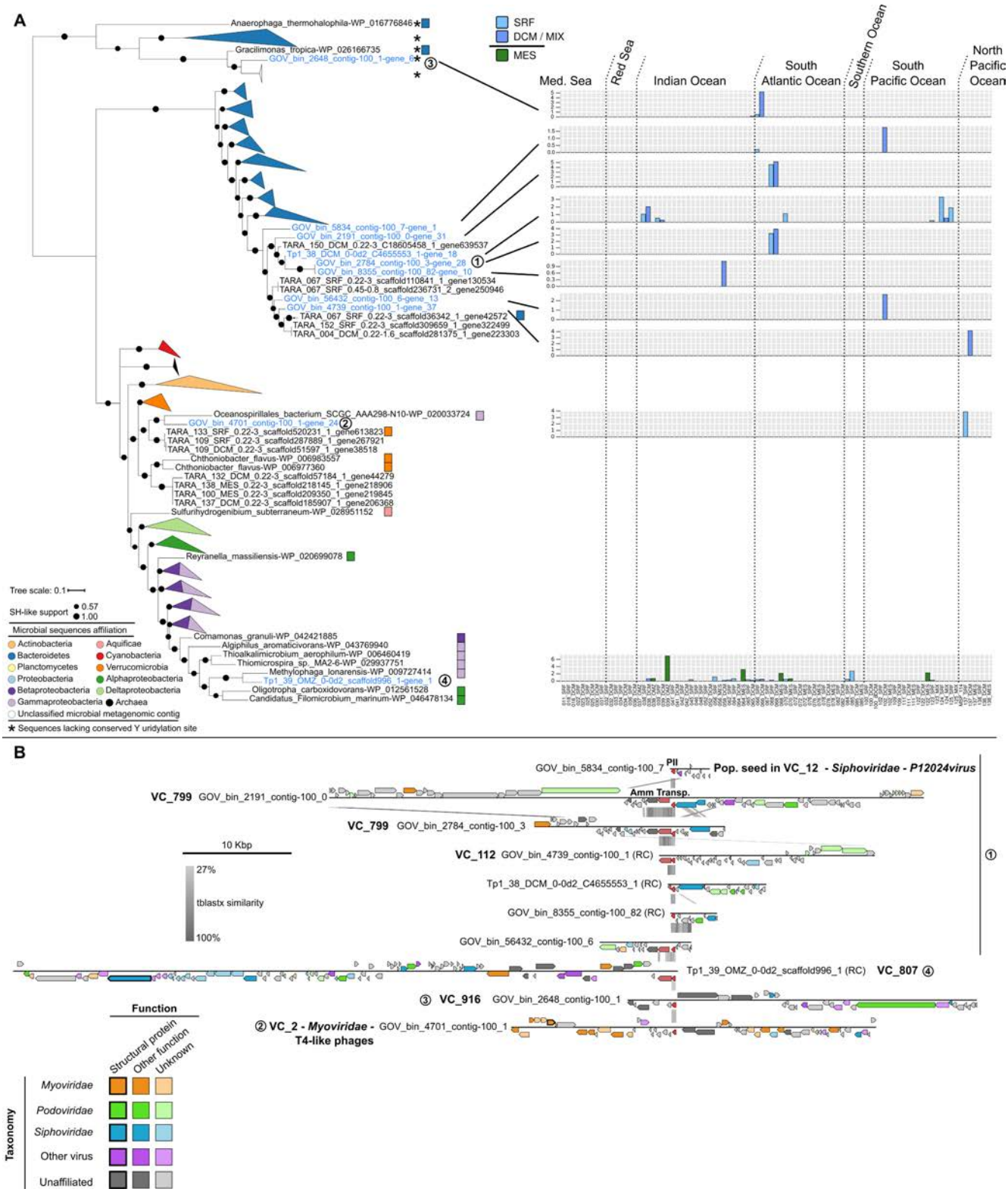
Extended Data Figure 4 | Association between abundant viral clusters and abundance and diversity of host groups. a, Abundance and diversity of bacterial and archaeal host groups associated with the 38 abundant viral clusters (see Fig. 2a). For each host group (at the phylum level, except for Proteobacteria where the class level is used), the different panels display, from top to bottom: (i) the number of viral clusters associated with this host group; (ii) the global relative abundance of this group estimated from the microbial metagenomic OTU counts; (iii) the global diversity of this group based on a Chao index computation including all *Tara* Oceans microbial metagenome samples (that is, including both alpha and beta diversity); (iv) the distribution of Chao indexes by sample for this group (the alpha diversity); and (v) the average Sorensen index between pairs of samples that include at least one OTU of this group (the beta diversity). OTU counts were derived from the 109 epipelagic microbial metagenomes described previously[18]. **b**, Pearson correlations between host-group relative abundance or diversity indices (global Chao index, average Chao index across samples and average Sorensen index across samples) and the number of viral clusters.



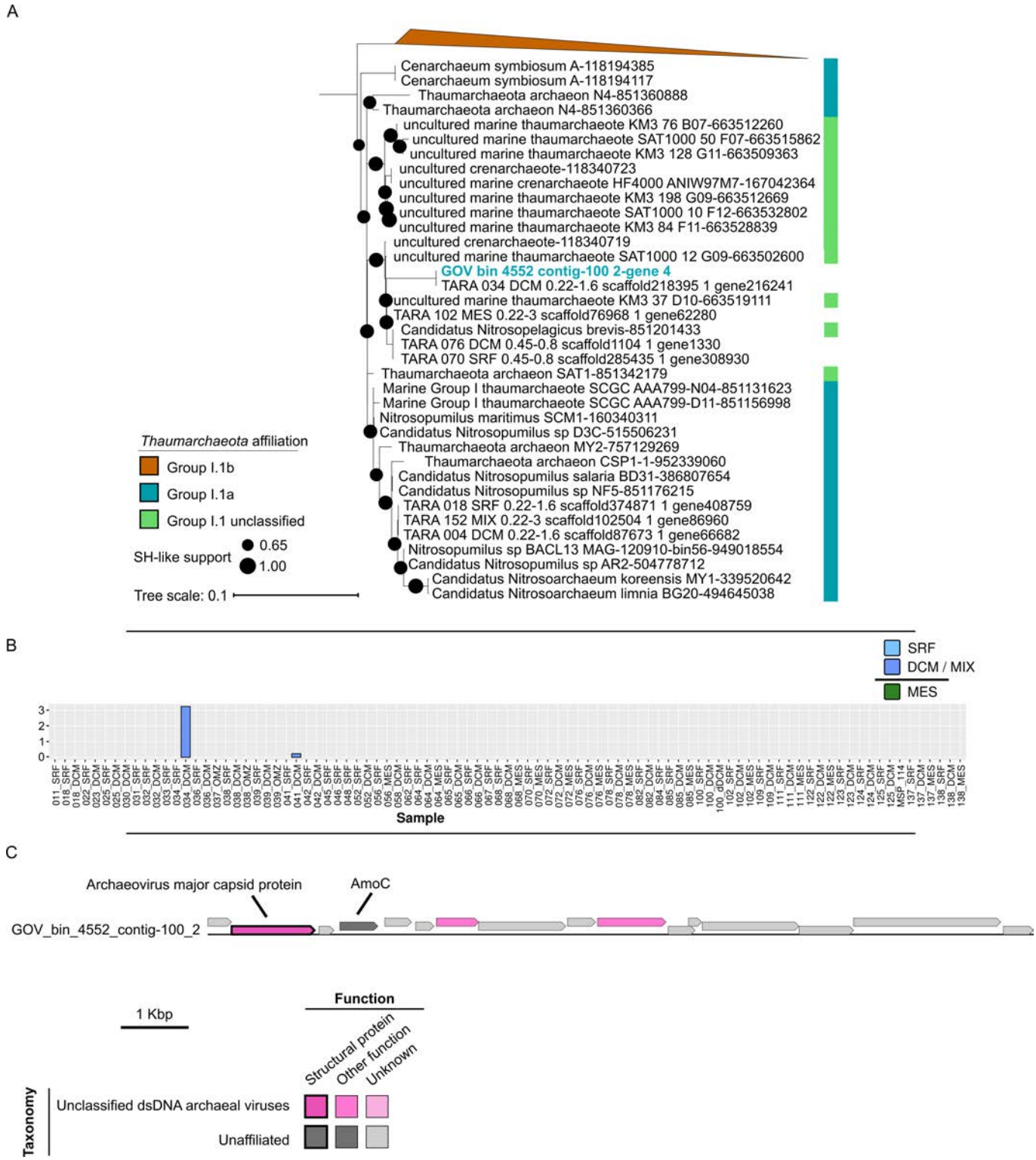
Extended Data Figure 5 | Diversity, distribution, and genome context of *dsrC* genes in GOV contigs. a, Maximum-likelihood tree (from an amino-acid alignment) including the 11 viral DsrC and microbial sequences from microbial metagenomes and NCBI nr database. The presence of conserved cysteine residues (termed CysA and CysB, as in ref. 24) is indicated with coloured circles next to each sequence or clade. The corresponding type of DsrC-like protein is indicated by the colouring of the branch or clade. The microbial metagenomic contigs affiliated to uncultivated, marine sulfur-oxidizing Gammaproteobacteria (as confirmed by complementary phylogenetic analysis of DsrAB; Supplementary Fig. 7) are indicated by stars. Viral AMG sequences are highlighted in blue, internal nodes and SH-like supports are represented by proportional circles (all nodes with support <0.40 were collapsed). Each *dsrC* AMG is associated with an abundance profile (right) that displays the relative abundance of the contig across the 91 epipelagic and mesopelagic samples (based on normalized coverage—that is, contig coverage per Gb of metagenome). **b**, Comparison of *dsrC*-containing contigs maps. A T4-like marker gene (T4 baseplate) is indicated on the maps, alongside putative AMGs (Fe–S biosyn, iron–sulfur cluster biosynthesis; Amt, ammonia transporter).



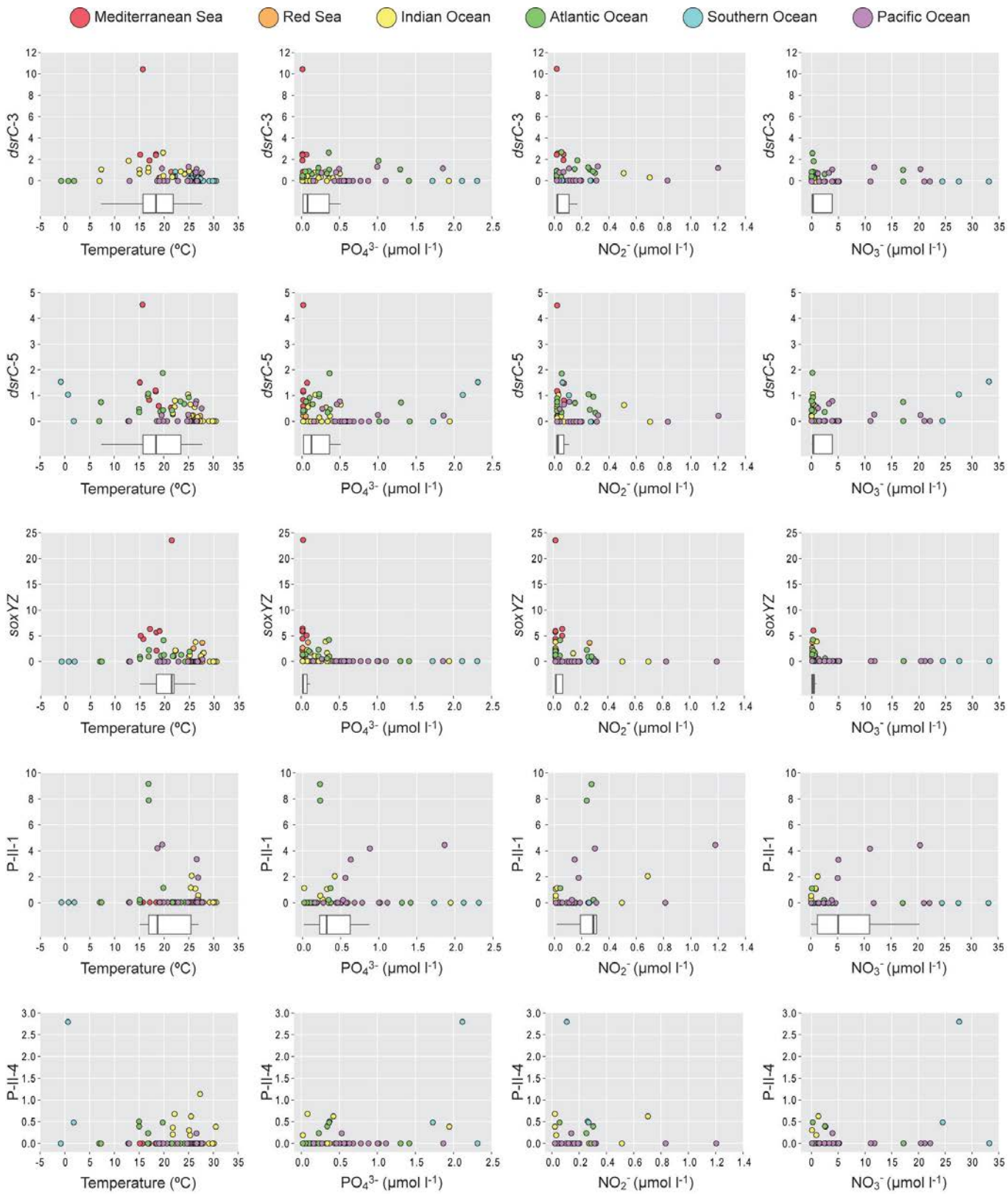
Extended Data Figure 6 | Diversity, distribution, and genome context of *soxYZ* genes in GOV contigs. a, Bayesian tree from an amino-acid alignment, including the four viral *soxYZ* and microbial sequences from microbial metagenomes and the NCBI nr database. The affiliation of microbial clades (either from the NCBI reference or from the LCA affiliation of metagenomic contigs) is indicated by the colouring of the grouped clades or by a coloured square next to the sequence. Viral AMG sequences are highlighted in blue, posterior probabilities are represented by proportional circles (all nodes with posterior probability <0.40 were collapsed). Clades including sulfur-oxidizing proteobacteria are indicated on the tree. Each *soxYZ* AMG is associated with an abundance profile (on the right) displaying the relative abundance of the contig across the 91 epipelagic and mesopelagic samples (based on normalized coverage; that is, contig coverage per Gb of metagenome). **b**, Comparison of *soxYZ*-containing contigs maps. For contig GOV_bin_4310_contig-100_0, the second largest contig from the same bin (GOV_bin_4310_contig-100_1) is displayed. T4-like marker genes (*gp23* and the gene encoding T4 baseplate) are indicated on the maps alongside putative AMGs.



Extended Data Figure 7 | Diversity, distribution, and genome context of *P-II* genes in GOV contigs. a, Maximum-likelihood tree from an amino-acid alignment that includes the 10 viral *P-II* and microbial sequences from microbial metagenomes and the NCBI nr database. The affiliation of microbial clades (either from the NCBI reference or from the LCA affiliation of metagenomic contigs) is indicated by the colouring of the grouped clades or by a coloured square next to the sequence. Sequences lacking the conserved uridylation site of *P-II* (Supplementary Fig. 5) are highlighted with a star next to the sequence name or clade. Viral AMG sequences are highlighted in blue, internal nodes SH-like supports are represented by proportional circles (all nodes with support <0.40 were collapsed). Each *P-II* AMG is associated with an abundance profile (right) displaying the relative abundance of the contig across the 91 epipelagic and mesopelagic samples (based on normalized coverage; that is, contig coverage per Gb of metagenome). **b**, Comparison of *P-II*-containing contig maps. Ammonia transporter genes linked to *P-II* are indicated on the map (dark red). When available, the viral-cluster affiliation of each contig is indicated next to the contig name. Contig GOV_bin_5834_contig-100_7 is too short to be clustered based on a shared protein cluster network, however the seed contig of its population was clustered (in VC_12, *Siphoviridae P12024virus*), hence the indication of this seed contig affiliation.



Extended Data Figure 8 | Diversity, distribution, and genome context of *amoC* gene in GOV contigs. a, Maximum-likelihood tree (from an amino-acid alignment) including the GOV *amoC* AMG and microbial sequences from microbial metagenomes and NCBI nr database. The affiliation of microbial clades (either from the NCBI reference or from the LCA affiliation of metagenomic contigs) is indicated by the colouring of the grouped clades or by a coloured square next to the sequence. Viral AMG sequence is highlighted in blue, internal nodes and SH-like supports are represented by proportional circles (all nodes with support <0.40 were collapsed). **b**, Abundance profile displaying the relative abundance of the contig across the 91 epipelagic and mesopelagic samples (based on normalized coverage; that is, contig coverage per Gb of metagenome). **c**, Map of the *amoC*-containing contig.



Extended Data Figure 9 | Normalized coverage of contigs harbouring AMG as a function of the temperature and nutrient concentrations of the corresponding samples. AMGs are grouped by clade based on their phylogeny (see Extended Data Figs 5–7) and their coverages are cumulated if multiple contigs are included in a clade. Plots display the cumulated normalized coverage of a clade (*y* axis) as function of the temperature or nutrient concentration (*x* axis) across all epipelagic samples for geographically unrestricted clades (that is, clades found in >5 samples, see Fig. 3c). Mesopelagic samples were excluded from the analysis since the AMG signal was detected in epipelagic samples. Samples are colour-coded according to ocean and sea regions (Supplementary Table 1). The calculated preferential range of temperature or nutrient concentration is displayed below each plot for epipelagic AMGs (*P-II-4* distribution could not be linked to specific environmental conditions, but this AMG is the only one consistently retrieved in mesopelagic samples).

Extended Data Table 1 | Summary of genes and contigs characteristics for new viral *dsrC*, *soxYZ*, *P-II*, and *amoC* AMGs

AMG	Contig	# genes in contig	AMG Clade	Conserved motif	Predicted function	Viral Cluster	Contig predicted host (prediction method)	pN/pS (selection pressure)	Metatranscriptomic signal					
									TARA_011_SRF		TARA_039_DCM		TARA_151_SRF	
									AMG gene coverage	# other genes covered	AMG gene coverage	# other genes covered	AMG gene coverage	# other genes covered
dsrC	GOV_bin_5582_contig-100_61	8	DsrC-1	CysA	tRNA modification (TusE-like)	NA	NA	0.17	0	-	0	-	0	-
dsrC	GOV_bin_4664_contig-100_21	6	DsrC-1	CysA	tRNA modification (TusE-like)	NA	NA	0.00	0	-	0	-	0	-
dsrC	GOV_bin_620_contig-100_371	4	DsrC-2	CysA	tRNA modification (TusE-like)	NA	NA	NA	0	-	0	-	0	-
dsrC	Tp2_85_DCM_contig-100_9217	5	DsrC-2	CysA	tRNA modification (TusE-like)	NA	NA	0.05	0	-	0	-	0	-
dsrC	Tp1_67_SUR_0-0d2_scaffold92877_1	5	DsrC-3	CysA	tRNA modification (TusE-like)	NA	NA	NA	0	-	0	-	1.20	1
dsrC	Tp1_23_DCM_0-0d2_scaffold51341_1	5	DsrC-3	CysA	tRNA modification (TusE-like)	NA	NA	0.13	10.52	1	0	-	1.30	3
dsrC	Tp1_109_DCM_0-0d2_C7640166_1	5	DsrC-3	CysA	tRNA modification (TusE-like)	NA	NA	0.07	0	-	0	-	0.59	1
dsrC	GOV_bin_3019_contig-100_4	22	DsrC-4	CysA	tRNA modification (TusE-like)	**VC_2 – T4-like superfamily**	NA	0.19	0	-	0	-	0	-
dsrC	GOV_bin_2979_contig-100_29	6	DsrC-4	CysA	tRNA modification (TusE-like)	NA	NA	NA	0	-	0	-	0	-
dsrC	Tp1_23_DCM_0-0d2_scaffold130384_1	7	DsrC-5	CysA & CysB	sulfur oxidation (bona fide DsrC)	NA	NA	0.21	0.15	0	0	-	0	-
dsrC	GOV_bin_5582_contig-100_23	13	unclustered	CysA	tRNA modification (TusE-like)	NA	NA	NA	0	-	0	-	0	-
soxYZ	Tp1_22_SUR_0-0d2_scaffold392_1	11	SoxYZ	SoxY SoxZ S interaction motif	sulfur oxidation	NA	NA	0.13	1.21	6	0	-	0	-
soxYZ	GOV_bin_273_contig-100_0	64	SoxYZ	SoxY SoxZ S interaction motif	sulfur oxidation	**VC_2 – T4-like superfamily**	Bacteroidetes (tetranucleotide)	0.13	0.63	53	0	-	0	-
soxYZ	GOV_bin_4310_contig-100_0	51	SoxYZ	SoxY SoxZ S interaction motif	sulfur oxidation	**VC_2 – T4-like superfamily**	Alphaprot. (blastn)	0.27	2.37	45	0	-	0	-
soxYZ	Tp1_25_DCM_0-0d2_scaffold13291_1	15	SoxYZ	SoxY SoxZ S interaction motif	sulfur oxidation	NA	NA	0.21	0	-	0.08	2	0.51	0
P-II	GOV_bin_2191_contig-100_0	67	P-II-1	uridylation site & P-II C-terminal	N metabolism regulation	VC_799 – new VC	NA	*0*	0	-	0	-	0	-
P-II	GOV_bin_4739_contig-100_1	38	P-II-1	uridylation site & P-II C-terminal	N metabolism regulation	VC_112 – new VC	Bacteroidetes (tetranucleotide)	NA	0	-	0	-	0	-
P-II	GOV_bin_2784_contig-100_3	36	P-II-1	uridylation site & P-II C-terminal	N metabolism regulation	VC_799 – new VC	NA	NA	0	-	0	-	0	-
P-II	GOV_bin_56432_contig-100_6	18	P-II-1	uridylation site & P-II C-terminal	N metabolism regulation	NA	NA	0.18	0	-	0	-	0	-
P-II	GOV_bin_8355_contig-100_82	10	P-II-1	uridylation site & P-II C-terminal	N metabolism regulation	NA	NA	0.05	0	-	0	-	0	-
P-II	GOV_bin_5834_contig-100_7	11	P-II-1	uridylation site & P-II C-terminal	N metabolism regulation	**VC_12 – *P12024virus***	NA	0.29	0	-	0	-	0	-
P-II	Tp1_38_DCM_0-0d2_C4655553_1	19	P-II-1	uridylation site & P-II C-terminal	N metabolism regulation	NA	NA	0.03	0	-	0	-	0	-
P-II	GOV_bin_4701_contig-100_1	53	P-II-2	uridylation site & P-II C-terminal	N metabolism regulation	**VC_2 – T4-like superfamily**	NA	0.11	0	-	0	-	0	-
P-II	GOV_bin_2648_contig-100_1	33	P-II-3	NA	NA	VC_916 – new VC	Bacteroidetes (tetranucleotide)	0.11	0	-	0	-	0	-
P-II	Tp1_39_OMZ_0-0d2_scaffold996_1	87	P-II-4	uridylation site & P-II C-terminal	N metabolism regulation	VC_807 – new VC	Gammaprot. (tetranucleotide & blastn)	0.67	0	-	0	-	0	-
amoC	GOV_bin_4552_contig-100_2	16	AmoC	NA	Ammonia oxidation	VC_623 – new VC	NA	0.15	0	-	0	-	0	-

Each gene is linked to its contig and, where available, to the corresponding viral cluster and predicted host (from BLAST hit, CRISPR spacer similarity, or nucleotide composition similarity). Widespread and abundant viral clusters are highlighted in bold. In addition, the calculated pN/pS of each gene is indicated (measuring the strength of selection pressure occurring for this gene, the gene with a pN/pS not representing a strong purifying selection is highlighted in red), as well as the coverage of these and other genes found in the contigs in 3 metatranscriptomic samples from 3 open-ocean *Tara* Ocean stations (cases where the AMG coverage is both >0.5 and associated with the coverage of other genes from the same viral contig are highlighted in green). Alphaprot, Alphaproteobacteria; Gammaprot, Gammaproteobacteria.