



HAL
open science

Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses

Simon Roux, Jennifer R. Brum, Bas E. Dutilh, Shinichi Sunagawa, Melissa B. Duhaime, Alexander Loy, Bonnie T. Poulos, Natalie Solonenko, Elena Lara, Julie Poulain, et al.

► To cite this version:

Simon Roux, Jennifer R. Brum, Bas E. Dutilh, Shinichi Sunagawa, Melissa B. Duhaime, et al.. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature*, 2016, 537 (7622), pp.689-693. 10.1038/nature19366 . hal-01376822

HAL Id: hal-01376822

<https://hal.sorbonne-universite.fr/hal-01376822>

Submitted on 26 Oct 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **Ecogenomics and potential biogeochemical impacts of globally abundant ocean** 2 **viruses**

3
4 Simon Roux¹, Jennifer R. Brum¹, Bas E. Dutilh^{2,3,4}, Shinichi Sunagawa^{5,‡}, Melissa B. Duhaime⁶,
5 Alexander Loy^{7,8}, Bonnie T. Poulos⁹, Natalie Solonenko¹, Elena Lara^{10,11}, Julie Poulain¹², Stéphane
6 Pesant^{13,14}, Stefanie Kandels-Lewis^{5,15}, Céline Dimier^{16,17}, Marc Picheral^{18,19}, Sarah Searson^{18,19,20},
7 Corinne Cruaud¹², Adriana Alberti¹², Carlos M. Duarte^{21,22}, Josep M. Gasol¹⁰, Dolors Vaqué¹⁰, Tara
8 Oceans Coordinators[†], Peer Bork^{5,23}, Silvia G. Acinas¹⁰, Patrick Wincker^{12,24,25}, Matthew B.
9 Sullivan^{1,26*}

10
11
12 ¹ Department of Microbiology, The Ohio State University, Columbus, OH, USA

13 ² Theoretical Biology and Bioinformatics, Utrecht University, Utrecht, The Netherlands

14 ³ Centre for Molecular and Biomolecular Informatics, Radboud University Medical Centre, Nijmegen, The Netherlands

15 ⁴ Department of Marine Biology, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

16 ⁵ Structural and Computational Biology, European Molecular Biology Laboratory, Heidelberg, Germany

17 ⁶ Department of Ecology and Evolutionary Biology, University of Michigan, MI, USA

18 ⁷ Division of Microbial Ecology, Department of Microbiology and Ecosystem Science, Research Network Chemistry Meets
19 Microbiology, University of Vienna, Vienna, Austria

20 ⁸ Austrian Polar Research Institute, Vienna, Austria

21 ⁹ Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA

22 ¹⁰ Department of Marine Biology and Oceanography, Institut de Ciències del Mar (CSIC), Barcelona, Spain

23 ¹¹ Institute of Marine Sciences (CNR-ISMAR), National Research Council, Venezia, Italy

24 ¹² CEA - Institut de Génomique, GENOSCOPE, Evry, France

25 ¹³ PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany

26 ³³ MARUM, Bremen University, Bremen, Germany

27 ¹⁵ Directors' Research European Molecular Biology Laboratory, Heidelberg, Germany

28 ¹⁶ CNRS, UMR 7144, EPEP, Station Biologique de Roscoff, Roscoff, France

29 ¹⁷ Sorbonne Universités, UPMC Univ Paris 06, UMR 7144, Station Biologique de Roscoff, Roscoff, France

30 ¹⁸ CNRS, UMR 7093, Laboratoire d'océanographie de Villefranche (LOV), Observatoire Océanologique, Villefranche-sur-
31 mer, France

32 ¹⁹ Sorbonne Universités, UPMC Univ Paris 06, UMR 7093, Observatoire Océanologique, Villefranche-sur-mer, France.

33 ²⁰ Department of Oceanography, University of Hawaii, Honolulu, Hawaii, USA

34 ²¹ Mediterranean Institute of Advanced Studies, CSIC-UiB, Esporles, Mallorca, Spain

35 ²² King Abdullah University of Science and Technology (KAUST), Red Sea Research Center (RSRC), Thuwal, Saudi
36 Arabia

37 ²³ Max-Delbrück-Centre for Molecular Medicine, 13092 Berlin, Germany

38 ²⁴ CNRS, UMR 8030, CP5706, Evry, France

39 ²⁵ Université d'Evry, UMR 8030, CP5706, Evry, France

40 ²⁶ Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, OH, USA

41
42 [†] Tara Oceans coordinators and affiliations are listed in the Supplementary Information.

43 [‡] Current address: Institute of Microbiology, ETH Zurich, Zurich, Switzerland

44 * correspondence to mbsulli@gmail.com

45

46 **Abstract**

47 Ocean microbes drive global-scale biogeochemical cycling¹, but do so under constraints imposed by
48 viruses on community composition, metabolic activity, and evolutionary trajectories^{2,3}. Due to
49 sampling and cultivation challenges, genome-level viral diversity remains poorly described and grossly
50 understudied in nature such that <1% of observed surface ocean viruses are ‘known’⁴. Here we
51 assemble complete genomes and large genomic fragments from both surface and deep ocean viruses
52 sampled during the *Tara Oceans* and *Malaspina* research expeditions^{5,6} and analyze the resulting
53 Global Ocean Viromes (GOV) dataset to present a global map of abundant, double stranded DNA
54 (dsDNA) viruses complete with genomic and ecological contexts. A total of 15,222 epi- and
55 mesopelagic viral populations were identified that comprised 867 viral clusters (VCs, approximately
56 genus-level groups^{7,8}). This roughly triples the number of ocean viral populations⁴, doubles candidate
57 bacterial and archaeal virus genera⁸, and near-completely samples epipelagic communities at both the
58 population and VC level. Thirty-eight of the 867 VCs were locally or globally abundant and together
59 accounted for nearly half of the viral populations in any GOV sample. While two thirds of them
60 represent newly described viruses that lacked any cultivated representative, most could be
61 computationally linked to dominant, ecologically relevant microbial hosts. Moreover, we identified 243
62 viral-encoded auxiliary metabolic genes (AMGs), only 95 of which were known. Deeper analyses of
63 four of these AMGs (*dsrC*, *soxYZ*, P-II and *amoC*) revealed that abundant viruses may directly
64 manipulate sulfur and nitrogen cycling throughout the epipelagic ocean. This viral catalog and
65 functional analyses provide a critically-needed foundation to begin meaningfully integrating viruses
66 into ecosystem models as key players in nutrient cycling and trophic networks.

67

68 **Main text**

69 A fundamental bottleneck preventing the incorporation of viruses of microbes into ecosystem
70 models is the lack of host-contextualized quantitative surveys of viral diversity in nature. This is
71 because (i) most naturally-occurring microbes and viruses are not currently cultivated, and (ii) viruses
72 lack a universally conserved marker gene, which precludes PCR-based surveys of uncultivated
73 diversity³. While viral metagenomics (viromics) was introduced to circumvent these issues, early
74 datasets were fragmented and only suitable for descriptive gene-level analyses that were prohibitively
75 database-biased³. Subsequent experimental, technological, and analytical improvements enabled viral
76 population ecology through the availability of genomic information^{3,9-11}. For example, 1,148 large viral
77 genome fragments captured in a fosmid library from Mediterranean Sea microbes revealed remarkable
78 viral diversity, with some genomes appearing globally distributed based upon analysis of six available
79 viral metagenomes⁹. Similarly, 69 viral reference genomes assembled from single-cell samples helped
80 elucidate the ecology, evolution and potential biogeochemical impacts of uncultivated viruses infecting
81 an uncultivated anaerobic chemoautotroph¹¹. Finally, metagenomic approaches are now quantitative, at
82 least for dsDNA templates³, and themselves provide genomic information on uncultivated viruses. For
83 example, 42 surface ocean viral metagenomes in the *Tara Oceans Viromes* (TOV) dataset revealed the
84 global underlying structure of these communities, and identified 5,476 viral populations, only 39 of
85 which were previously known⁴.

86 Here we further identify ocean viral populations, determine and characterize the most abundant and
87 widespread dsDNA ocean viral types, and analyze viral-encoded AMGs and their distributions to
88 propose new means by which viruses likely modulate microbial biogeochemistry. We do so by
89 analyzing the Global Oceans Viromes (GOV) dataset, which augments TOV with 61 samples to better
90 represent the surface and deep oceans, and now totals 104 viromes and 925 Gbp of sequencing data
91 (Supplementary Table 1). Further, upgraded analytical approaches including cross-assembly¹² and
92 genome binning¹³ improved genomic representation of sampled viruses (see Supplementary Text for
93 details on the dataset generation process). From 1,380,834 contigs which recruited 67% of the reads,
94 we identified 15,280 viral populations (Fig. 1A, see Supplementary Fig. 1 for viral population

95 definition explanation). This expands ocean viral populations nearly 3-fold over the prior TOV dataset⁴,
96 while also improving average contig lengths and genomic context 2.5-fold for TOV-known populations
97 (Supplementary Table 2). Rarefaction analyses show that while mesopelagic viral communities remain
98 undersampled, epipelagic viral communities now appear near-completely sampled (Extended Data Fig.
99 1A). Because bathypelagic communities were underrepresented due to cellular contamination, we
100 focused the remaining analyses on 15,222 non-bathypelagic viral populations.

101 We first categorized viral populations into viral clusters, or VCs using shared gene content
102 information and network analytics⁷ (see Supplementary Fig. 1 for VC definition schematic). This
103 method starts from genome fragments ($\geq 10\text{kb}$) and results in VCs approximately equivalent to known
104 viral genera^{7,8}. Clustering of the 15,222 GOV viral populations with 15,929 publicly available bacterial
105 and archaeal viruses revealed 1,259 VCs (see Supplementary Table 3, Supplementary Text & Extended
106 Data Fig. 2 for comparison with alternative classification methods). Of these, 658 included exclusively
107 GOV sequences, which approximately doubles known bacterial and archaeal virus genera⁸, and another
108 209 VCs contained at least one GOV sequence (Fig. 1B). As with viral populations, rarefaction
109 analyses suggested that VC diversity was undersampled in mesopelagic waters, but near-completely
110 sampled in epipelagic waters (Extended Data Fig. 1B).

111 We next identified the most abundant and widespread VCs based on read recruitment of VC
112 members. In each sample, a fraction of the VCs were identified as abundant based on their cumulative
113 contribution to sample diversity (estimated with Simpson Index, abundant VCs represent 80% of the
114 total sample diversity, Extended Data Fig. 1C). By these criteria, only 38 of 867 observed VCs were
115 abundant in two or more stations, and together recruited an average of 50% and 35% of reads from
116 viral populations for epipelagic and mesopelagic samples, respectively (Supplementary Table 3). Four
117 of these 38 abundant VCs were also relatively ubiquitous as they were abundant in more than 25
118 stations, and 62 of the 91 non-bathypelagic samples were dominated by 1 of these 4 VCs (Fig. 2 A &
119 B). Among the 38 abundant VCs, only 2 corresponded to well-studied viruses, from the T4
120 superfamily^{14,15} (VC_2, 1 of the 4 ubiquitous) and the *T7virus* genus¹⁶ (VC_9). Eight represented
121 known but unclassified viral isolates, 10 included viruses known only from environmental
122 sequencing^{9,10}, and the remaining 18 VCs were completely novel (Fig. 2C, Extended Data Fig. 3).

123 Given this global map of the dominant dsDNA viral types in the oceans, we next sought to identify
124 the range of hosts these viruses infect. This is challenging, as culture-based methods insufficiently
125 capture naturally-occurring diversity, whereas metagenomic approaches broadly survey viral diversity
126 but often without host information. Fortunately, sequence-based approaches are emerging that examine
127 similarities between (i) viral genomes and host CRISPR spacers¹⁷, (ii) viral and microbial genomes due
128 to integrated prophages or gene transfers⁹, and (iii) viral and host genome nucleotide signatures (here,
129 tetranucleotide frequencies⁸, see Supplementary Table 4 and Supplementary Text for discussion of the
130 accuracy/sensitivity of *in silico* host prediction methods). We applied all 3 methods to GOV to predict
131 hosts at the phylum level, or class level for Proteobacteria (Supplementary Table 5), then summarized
132 these results at the VC level. This led to host range predictions for 392 of 867 VCs – all with
133 confidence assessed by comparison to a null model (Supplementary Fig. 2 and Supplementary Table 3).

134 The hosts of the 38 globally abundant VCs were largely restricted to abundant and widespread
135 epipelagic-ocean microbes that were previously identified via *mi*Tag-based OTU counts in *Tara* Oceans
136 microbial metagenomes¹⁸. Notably, the 4 ubiquitous and abundant VCs were predicted to infect 7 of the
137 8 globally abundant microbial groups (Actinobacteria, Alpha-, Delta-, and Gammaproteobacteria,
138 Bacteroidetes, Cyanobacteria, Deferribacteres; Fig. 2C, Extended Data Fig. 4). The 8th abundant
139 microbial group, Euryarchaeota, was not linked to these 4 VCs, but was predicted as a host for 3 of the
140 34 other abundant VCs (VC_3, VC_27, and VC_63, Extended Data Fig. 3). Among the 38 abundant
141 VCs, the number of VCs predicted to infect a given microbial host phylum (or class for Proteobacteria)
142 was positively correlated with host global richness rather than relative abundance (Extended Data Fig.
143 4B). This suggests that, likely because ocean viruses appear globally distributed⁴, widespread and

144 abundant hosts that are minimally diverse (e.g. Cyanobacteria) provide few viral niches, whereas more
145 diverse host groups, even at lower abundance (e.g. Betaproteobacteria), provide more opportunity for
146 viral niche differentiation. Hence, these host associations provide critically-needed empirical support
147 for hypotheses derived from global virus-host network models¹⁹.

148 Having mapped viral diversity and predicted virus-host pairings, we next sought to identify virus-
149 encoded AMG that might modify host metabolism during infection and likely impact biogeochemistry.
150 To maximize AMG detection, all 298,383 viral contigs >1.5kb were examined, including small contigs
151 not associated with a viral population. This revealed 243 putative AMGs (Supplementary Table 6).
152 While 95 of these AMGs were known (reviewed in ref. ²⁰), others offer insights into how viruses may
153 directly manipulate microbial metabolisms. Here we focus on 4 (*dsrC*, *soxYZ*, P-II and *amoC*; see
154 Extended Data Table 1, Supplementary Figs. 3-6 and Supplementary Text for functional affiliation of
155 these AMGs) because of their putative roles in sulfur or nitrogen cycling. Three of these are not known
156 in viruses, and one, *dsrC*, has only been observed in viruses from anoxic deep-sea environments^{11,21}.

157 Sulfur oxidation in seawater involves two central microbial pathways – dissimilatory sulfur
158 reductase (Dsr) and sulfur oxidation (Sox)²² – and GOV AMG analyses revealed that epipelagic viruses
159 encode key genes for each. First, 11 *dsrC*-like genes were identified in viral contigs (Extended Data
160 Fig. 5). The Dsr operon is used by sulfate/sulfite-reducing microbes in anoxic environments, as well as
161 sulfur-oxidizing bacteria in oxic and anoxic environments (Fig. 3A)²². DsrC, specifically, provides
162 sulfur to DsrAB-sulfite reductase for processing through a conserved C-terminal motif (Cys_BX₁₀Cys_A),
163 and dictates sulfur metabolism rates²³. Other DsrC-like proteins (also known as TusE) lack Cys_B and
164 instead participate to tRNA modification²⁴. In GOV, four clades of DsrC-like sequences were similar to
165 TusE (DsrC-1 to DsrC-4), whereas the fifth (DsrC-5) was similar to *bona fide* DsrC (Extended Data
166 Fig. 5, Extended Data Table 1, Supplementary Fig. 3, and Supplementary text). Second, 4 *soxYZ* genes
167 were identified on viral contigs (Extended Data Fig. 6). Like DsrC, SoxYZ is an important sulfur
168 carrier harboring a conserved functional motif identified in all GOV SoxYZ proteins (Fig. 3A,
169 Supplementary Fig. 4, and Supplementary text)²⁵.

170 Other AMGs suggest marine viruses may manipulate nitrogen cycling. First, 10 GOV contigs
171 encoded P-II, a gene widespread across bacteria and archaea and central in nitrogen metabolism
172 regulation (Fig. 3B)²⁶. Three AMG clades (P-II-1, P-II-2, and P-II-4) displayed both P-II conserved
173 motifs and had predicted structures similar to bona fide P-II, whereas the fourth clade (P-II-3) is
174 functionally ambiguous as it lacked a conserved motif (Supplementary Fig. 5, and Supplementary text).
175 Second, two P-II AMG clades (P-II-1 and P-II-4) were proximal to an ammonium transporter gene,
176 *amt*, in GOV contigs (Extended Data Fig. 7). In bacteria, such an arrangement is a signature of P-II-like
177 genes that specifically activate alternative nitrogen production and ammonia uptake pathways during
178 nitrogen starvation²⁶. Third, one GOV contig included *amoC*, encoding the subunit C of ammonia
179 monooxygenase, suggesting a role in ammonia oxidation²⁷. While functional annotation is challenging
180 for these genes²⁷, and functional motifs are not yet known, the translated AMG was 94% identical to
181 functional AmoC from Thaumarchaeota – a level of identity only observed among expressed and
182 functional AMGs (Extended Data Fig. 8, Supplementary Fig. 6, and Supplementary text).

183 Next, we investigated the origin, evolutionary history, and diversity of these AMGs in epipelagic
184 viruses (see Supplementary Text for additional discussion about taxonomic affiliation and host
185 prediction for AMG-containing GOV sequences). The 15 GOV contigs encoding *dsrC* or *soxYZ* genes,
186 when affiliated, were all associated with members of the abundant and ubiquitous VC_2 (T4
187 superfamily, Extended Data Fig. 5 and 6, Extended Data Table 1). Phylogenies suggested that these
188 viruses obtained AMGs from S-oxidizing proteobacterial hosts, with likely a single transfer event for
189 *soxYZ* and two for *dsrC* (Extended Data Fig. 5 and 6). Among the latter, the bona fide S-oxidation
190 DsrC-5 was most closely related to a clade of uncultivated S-oxidizing Gammaproteobacteria
191 (MED13k09, Supplementary Fig. 7). These bacteria are widespread in the epipelagic ocean²⁸ and
192 suspected to degrade dimethyl sulfide, a key reduced sulfur species involved in ocean-to-atmosphere

193 sulfur transport and cloud formation. If confirmed, DsrC5-encoding viruses infecting these bacteria
194 would impact critical sulfur cycling steps throughout surface waters. In contrast to sulfur AMGs,
195 phylogenies suggest that P-II AMGs originated from diverse viruses (6 VCs including the abundant
196 VC_2 and VC_12), and were acquired at least 4 times independently from Bacteroidetes,
197 Proteobacteria, and possibly Verrucomicrobia (Extended Data Fig. 7, and Supplementary Text). Finally,
198 while a single *amoC* AMG offers only preliminary evaluation of its evolutionary history, this *amoC*-
199 encoding contig appears to represent novel and rare archaeal dsDNA viruses (VC_623), predicted to
200 infect ammonia-oxidizing Thaumarchaeota, known for their major role in global nitrification²⁹
201 (Extended Data Fig. 8).

202 Finally, we investigated the ecology of viruses encoding these AMGs by mapping their distribution
203 across GOV. Seven AMG clades were geographically restricted (DsrC-unc, DsrC-1, DsrC-2, DsrC-4,
204 P-II-2, P-II-3, and *amoC*), and 5 were widespread throughout epipelagic (DsrC-3, DsrC-5, SoxYZ, P-
205 II-1) or mesopelagic (P-II-4) waters (Fig. 3C). All widespread epipelagic AMGs were detected in
206 waters of mid-range temperatures. In contrast, DsrC-5 and SoxYZ were predominantly detected in low-
207 nutrient conditions, while P-II-1 was predominantly detected in high-nutrient conditions (Fig. 3D,
208 Extended Data Fig. 9). Thus, we hypothesize that viruses utilize DsrC-5 or SoxYZ to boost sulfur
209 oxidation rates when infecting sulfur oxidizers in low-nutrient conditions, and P-II under high-nutrient
210 conditions. The latter could be useful to viruses by activating expensive alternative N-producing
211 pathways typically used only under N-starvation conditions²⁶. Consistent with this, metatranscriptomes
212 from three low-nutrient stations (11_SRF in Mediterranean Sea, 39_DCM in Arabian Sea, and
213 151_SRF in Atlantic Ocean) revealed expression of viral homologs of *dsrC* and *soxYZ* but not of P-II
214 (Extended Data Table 1).

215 Overall, this systematically collected and processed GOV dataset provides a critical resource for
216 marine microbiology. This map of global dsDNA ocean viral diversity, at both the population and VC
217 level, and viral-encoded AMGs brings global ecological context to abundant surface and deep ocean
218 viruses. Both will also help interpret future (meta)genomic datasets and select experimental systems to
219 develop. Together with recent experimental, informatic and theoretical advances^{3,12,30}, this fundamental
220 resource will accelerate the field towards understanding and dynamically predicting the roles and
221 planetary impacts of viruses in nature.

222 **Methods**

223

224 **Sample collection and processing**

225 *Tara Oceans expedition*

226 Ninety samples were collected between October 10, 2009, and December 12, 2011, at 45 locations
227 throughout the world's oceans (Supplementary Table 1) through the *Tara Oceans Expedition*³². These
228 included samples from a range of depths: surface, deep chlorophyll maximum, bottom of mixed layer
229 when no deep chlorophyll maximum was observed (Station 123, 124, and 125), and mesopelagic
230 samples. The sampling stations were located in 7 oceans and seas, 4 different biomes and 14 Longhurst
231 oceanographic provinces (Supplementary Table 1). For TARA station 100, two different peaks of
232 chlorophyll were observed, so two samples were taken at the shallow (100_DCM) and deep
233 (100_dDCM) chlorophyll maximum. For each sample, 20 L of seawater were 0.22 µm-filtered and
234 viruses were concentrated from the filtrate using iron chloride flocculation³³ followed by storage at
235 4°C. After resuspension in ascorbic-EDTA buffer (0.1 M EDTA, 0.2 M Mg, 0.2 M ascorbic acid, pH
236 6.0), viral particles were concentrated using Amicon Ultra 100 kDa centrifugal devices (Millipore),
237 treated with DNase I (100U/mL) followed by the addition of 0.1M EDTA and 0.1M EGTA to halt
238 enzyme activity, and extracted as previously described³⁴. Briefly, viral particle suspensions were treated
239 with Wizard PCR Preps DNA Purification Resin (Promega, WI, USA) at a ratio of 0.5 mL sample to 1
240 mL resin, and eluted with TE buffer (10 mM Tris, pH 7.5, 1 mM EDTA) using Wizard Minicolumns.
241 Extracted DNA was Covaris-sheared and size selected to 160–180 bp, followed by amplification and
242 ligation per the standard Illumina protocol. Sequencing was done on a HiSeq 2000 system (101 bp,
243 paired end reads) at the Genoscope facilities (Paris, France).

244 Temperature, salinity, and oxygen data were collected from each station using a CTD (Sea-Bird
245 Electronics, Bellevue, WA, USA; SBE 911plus with Searam recorder) and dissolved oxygen sensor
246 (Sea-Bird Electronics; SBE 43). Nutrient concentrations were determined using segmented flow
247 analysis³⁵ and included nitrite, phosphate, nitrite plus nitrate, and silica. Nutrient concentrations below
248 the detection limit (0.02 µmol kg⁻¹) are reported as 0.02 µmol kg⁻¹. All data from the Tara Oceans
249 expedition are available from ENA (for nucleotide) and from PANGAEA (for environmental,
250 biogeochemical, taxonomic and morphological data)^{36–38}.

251

252 *Malaspina expedition*

253 Thirteen bathypelagic samples and one mesopelagic sample were collected between April 19, 2011
254 and July 11, 2011 during the Malaspina 2010 global circumnavigation covering the Pacific and the
255 North Atlantic Ocean. All samples were taken at 4,000 m depth except two samples from stations 81
256 and 82 collected at 3,500 and 2,150 m respectively (Supplementary Table 1). Additionally, Station
257 M114 was sampled at the OMZ region at 294 m depth. For each sample, 80 L of seawater were 0.22
258 µm-filtered and viruses were concentrated from the filtrate using iron chloride flocculation³³ followed
259 by storage at 4°C. More details about the sampling and additional variables used in the Malaspina
260 expedition can be found in ref. ³⁹. Further processing was done as for the *Tara Oceans* samples, except
261 that Illumina sequencing was done at DOE JGI Institute (151 bp, paired end reads).

262

263 **Dataset generation**

264 *Contigs assembly*

265 An overview of the contigs generation process is provided in Supplementary Fig. 8. The first step
266 consisted in the generation of a set of contigs using as many reads as possible from the 104 oceanic
267 viromes, including 74 epipelagic and 16 mesopelagic samples from the *Tara Oceans* expedition⁵, and 1
268 mesopelagic and 13 bathypelagic from the Malaspina expedition⁶. This set of contigs was generated
269 through an iterative cross-assembly¹² (using MOCAT⁴⁰ and Idba_ud⁴¹, Supplementary Fig. 8) as
270 follows: (i) high-quality (HQ) reads were first assembled sample by sample with the MOCAT pipeline

271 as described in¹⁸, (ii) all reads not mapping (Bowtie 2⁴², options --sensitive, -X 2000, and --non-
272 deterministic, other parameters at default) to a MOCAT contig (by which we denote ‘scaffigs’, that is,
273 contigs that were extended and linked using the paired-end information of sequencing read⁴³) were
274 assembled sample by sample with Idba_ud (iterative k-mer assembly, with k-mer increasing from 20 to
275 100 by step of 20), (iii) all reads remaining unmapped to any contig were then pooled by Longhurst
276 province (i.e. unmapped reads from samples corresponding to the same Longhurst province were
277 gathered) and assembled with Idba_ud (with the same parameters as above), and (iv) all remaining
278 reads unmapped from every samples were gathered for a final cross-assembly (using Idba_ud). This
279 resulted in 10,845,515 contigs (Supplementary Fig. 8B).

280

281 *Genome binning and re-assembly*

282 The contigs assembled from the marine viral metagenomes might still contain redundant sequences
283 derived from the same, or closely related populations. We set out to merge contigs derived from the
284 same population into clusters representing population genomes. To this end, contig sequences were first
285 clustered at 95% global average nucleotide identity (ANI) with cd-hit-est⁴⁴ (options -c 0.95 -G 1 -n 10 -
286 mask NX, Supplementary Fig. 8B), resulting in 10,578,271 non-redundant genome fragments. Next,
287 we used co-abundance (i.e. correlation between abundance profiles estimated by reads mapping) and
288 nucleotide usage profiles of the non-redundant contigs to further identify contigs derived from the same
289 populations with Metabat⁴⁵. Briefly, Metabat uses Pearson correlation between coverage profiles
290 (determined from the mapping of HQ reads of each sample to the contigs with Bowtie 2⁴², options --
291 sensitive, -X 2000, and --non-deterministic, other parameters at default) and tetranucleotide frequencies
292 to identify contigs originating from the same genome (Metabat parameters: 98% minimum correlation,
293 mode “sensitive”, see Supplementary Text for more detail about the selection of these parameters). The
294 8,744 bins generated, including 3,376,683 contigs, were further analyzed, alongside 623,665 contigs
295 not included in any genome bin but ≥ 1.5 kb.

296 In an attempt to better assemble these genome bins, two additional sets of contigs were generated
297 for each genome bin (beyond the set of initial contigs binned by Metabat⁴⁵), based on the de novo
298 assembly of (i) all reads mapping to the contigs in the genome bin, and (ii) only reads from the sample
299 displaying the highest coverage for the genome bin (both assemblies with Idba_ud⁴¹, Supplementary
300 Fig. 8C). The latter might be expected to lead to the “cleanest” genome assembly because it includes
301 the minimum between-sample sequence variation, lowering the probability of generating chimeric
302 contig⁴⁶. The former may be necessary if the virus is locally rare, so that sequences from multiple
303 metagenomes are needed to achieve complete genome coverage. Thus, if the assembly from the single
304 “highest coverage sample” was improved or equivalent to the initial assembly (longest contig in the
305 new assembly representing $\geq 95\%$ of the longest contig in the initial assembly), this set of contigs was
306 selected as the sequence for this bin (n=6,423). This optimal single-sample assembly was thus
307 privileged compared to a cross-assembly (either based on the initial contigs or on the re-assembly of all
308 sequences aligned to that bin). Otherwise, the “all samples” bin re-assembly was selected if equivalent
309 or better than the initial assembly (longest contig representing $\geq 95\%$ of the longest initial contig,
310 n=999). The assumption that cross-assembly would be needed for locally rare viruses, without a high-
311 coverage sample, was confirmed by the comparison between the highest coverage of these two types of
312 bins: on average, bins for which the “optimal” assembly were selected displayed a maximum coverage
313 of 5.47 per Gb of metagenome, while the bins for which the “cross-assembly” was selected displayed a
314 maximum coverage of 1.37 per Gb of metagenome (Supplementary Table 2). Finally, if both re-
315 assemblies yielded a longest contig smaller ($< 95\%$) than the one in the initial assembly, the bin was
316 considered as a false positive (i.e. binning of contigs from multiple genomes, n=1,356), and contigs
317 from the initial assembly were considered as “unbinned” (263,006 contigs, added to the 623,665
318 contigs ≥ 1.5 kb initially retained as “unbinned”).

319

320 *Identification of viral contigs and delineation of viral populations*

321 Despite efforts to completely remove cellular DNA during sample preparation, the resulting viral
322 metagenomic datasets will only ever be enriched for viruses⁴⁷. Thus, assembled sequences in the GOV
323 dataset were *in silico* filtered *a posteriori* to identify and remove clearly non-viral signal. In this way,
324 our purification methods should have greatly enriched for viruses, but the *in silico* decontamination
325 step served as a back-up for problematic samples. Together these two “filters” mean that virtually no
326 known cellular signal should have been considered in our analyses. For the *in silico* cleaning step,
327 VirSorter⁴⁸ was used to identify and remove microbial contigs using the “virome decontamination”
328 mode, with every contig $\geq 10\text{kb}$ and not identified as a viral contig being considered as a microbial
329 contig. Sequences with a prophage predicted were manually curated to distinguish actual prophages
330 (i.e. viral regions within a microbial contig) from contigs that belonged to a viral genome and were
331 wrongly predicted as a prophage. Contigs originating from an eukaryotic virus were identified based on
332 best BLAST hit affiliation of the contig predicted genes against NCBI RefseqVirus (see Supplementary
333 Text).

334 The genome bins were affiliated as microbial (if 1 or more contigs were identified as microbial,
335 $n=1,763$), eukaryotic virus (if contigs affiliated as eukaryotic virus comprised more than 10kb or more
336 than 25% of the genome bin total length, $n=962$) or viral (i.e. archaeal and bacterial viruses, $n=4,341$),
337 with the 356 remaining bins, lacking a contig long enough for an accurate affiliation, considered as
338 “unknown” (see Supplementary Text).

339 Viral bins were then refined to evaluate if they corresponded to a single or a mix of viral
340 population(s). To that end, the Pearson correlation and Euclidean distance between abundance profiles
341 (i.e. profile of the contig average coverage depth across the 104 samples) of bin members and the bin
342 seed (i.e. the largest contig) were computed, and a single-copy viral marker gene (TerL) was identified
343 in binned contigs (Supplementary Fig. 8E). Thresholds were chosen to maximize the number of bins
344 with exactly one TerL gene and minimize the number of bins with multiple TerL genes (Supplementary
345 Fig. 8G). For each bin, contigs with a Pearson correlation coefficient to the bin seed <0.96 or a
346 Euclidean distance to the seed >1.05 were removed from the bin, and added to the pool of unbinned
347 contigs. Eventually, every bin still displaying multiple TerL genes after this refinement step were split,
348 and all corresponding contigs added to the pool of “unbinned” contigs (Supplementary Fig. 8E).

349 The final set of contigs was formed by compiling (i) all contigs belonging to a viral bin, (ii)
350 “unbinned” viral contigs (i.e. contigs affiliated to archaeal and bacterial virus and not part of any
351 genome bin), and (iii) viral contigs identified in microbial or eukaryote virus bins (considered as
352 “unbinned” contigs, Supplementary Fig. 8F). Within this set of contigs, all viral bins were considered
353 as viral populations, as well as every unbinned viral contig $\geq 10\text{kb}$, leading to a total of 15,222 epi- and
354 mesopelagic populations, and 58 bathypelagic populations (Supplementary Fig. 1, Supplementary
355 Table 2, and Supplementary Text). In this study, we focus only on the 15,222 epi- and mesopelagic
356 populations, totaling 24,353 contigs. For the detection of AMGs, we added to these populations all
357 short epi- and mesopelagic unbinned viral contigs ($<10\text{kb}$), adding up to a total of 298,383 contigs.
358

359 **Sequence clustering and annotations**

360 *Dataset of publicly available viral genomes and genome fragments*

361 Genomes of viruses associated with a bacterial or archaeal host were downloaded from NCBI
362 RefSeq (1,680 sequences, v70, 05-26-2015). To complete this dataset of reference genomes, viral
363 genomes and genome fragments available in Genbank but not in RefSeq were downloaded (July 2015)
364 and manually curated to select only bacterial and archaeal viruses (1,017 sequences). These included
365 viral genomes not yet added to RefSeq, as well as genome fragments from fosmid libraries generated
366 from seawater samples^{9,10}. Mycophage sequences (available from <http://phagesdb.org>⁴⁹) were
367 downloaded (July 2015) and included as well if not already in RefSeq (734 sequences). Finally, 12,498
368 viral genome fragments from the VirSorter Curated Dataset, identified in publicly available microbial

369 genome sequencing projects, were added to the database⁸.

370
371 *Genome (fragments) clustering through gene-content based network analysis*

372 Proteins predicted from 14,650 large GOV contigs (≥ 10 kb and ≥ 10 genes), were added to all
373 proteins from the publicly available viral genomes and genomes fragments gathered, and compared
374 through all-vs-all blastp, with a threshold of 10^{-5} on e-value and 50 on bit score. Protein clusters were
375 then defined using MCL (using default parameters for clustering of proteins, similarity scores as
376 log-transformed e-value, and 2 for MCL inflation⁵⁰). vContact (<https://bitbucket.org/MAVERICLab/vcontact>)
377 vcontact) was then used to calculate a similarity score between every pair of genome and/or contigs
378 based on the number shared of PCs between the two sequences (as in^{7,8}), and then compute a MCL
379 clustering of the genomes/contigs based on these similarity scores (thresholds of 1 on similarity score,
380 MCL inflation of 2). The resulting viral clusters (or VCs, clusters including ≥ 2 contigs and/or
381 genomes), consistent with a clustering based on whole-genome BLAST comparison, corresponded to
382 approximately genus-level taxonomy, with rare cases closer to subfamily-level taxonomy (Extended
383 Data Fig. 2 and Supplementary Text). A total of 1,259 viral clusters were obtained, with 867 including
384 at least one GOV sequence. Notably, however, automatically defined VCs merely serve as a starting
385 place for assigning viral taxonomy. Current ICTV convention for formal taxonomic consideration of
386 these VCs would require manual comparison of genomes and genome fragments to identify signature
387 genes, comparison of phylogenetic signals, and ideally observation of morphological features of
388 corresponding viruses, although this process is currently being reviewed as advanced computational
389 analytics and genome datasets, such as those presented here, are being developed.

390
391 *Viral contigs annotation*

392 A functional annotation of all GOV predicted proteins was based on a comparison to the PFAM
393 domain database (v27⁵¹) with HmmsSearch⁵² (threshold of 30 on bit score and $1e-3$ on e-value), and
394 additional putative structural proteins were identified through a BLAST comparison to protein clusters
395 detected in viral metaproteomics dataset⁵³. This metaproteomics dataset led to the annotation of 13,547
396 hypothetical proteins lacking a PFAM annotation. A taxonomic annotation was performed based on a
397 blastp of the predicted proteins against proteins from archaeal and bacterial viruses from NCBI RefSeq
398 and Genbank (threshold of 50 on bit score and 10^{-3} on e-value).

399 VCs were affiliated based on isolate genome members, when available. When multiple isolates
400 were included in the VC, the VC was affiliated to the corresponding subfamily or genus of these
401 isolates (excluding all “unclassified” cases). This was the case for VC_2 (T4 subfamily^{14,15}), and VC_9
402 (*T7virus*¹⁶). When only one or a handful of affiliated isolate genomes were included in the VC and
403 lacked genus-level classification, a candidate name was derived from the isolate (if several isolates,
404 from the first one isolated). This was the case for VC_5 (*Cbaphi381virus*⁵⁴), VC_12 (*P12024virus*⁵⁵),
405 VC_14 (*MED4-117virus*), VC_19 (*HMO-2011virus*⁵⁶), VC_31 (*RM378virus*⁵⁷), VC_36 (*GBK2virus*⁵⁸),
406 VC_47 (*Cbaphi142virus*⁵⁴), and VC_277 (*vB_RglS_P106Bvirus*⁵⁹). Otherwise, VCs were considered
407 as “new VCs”.

408
409 *“Phage proteomic tree” (i.e. “whole-genome comparison tree”) computation and visualization*

410 All publicly available complete genomes (see above), all complete (circular) and near-complete
411 (extrachromosomal genome fragment >50 kb with a terminase) from the VirSorter Curated Dataset, and
412 all complete and near-complete GOV contigs were compared to generate a phage proteomic tree, as
413 previously described^{9,60}. Briefly, a proteomic similarity score was calculated for each pair of genome
414 based on a all-vs-all tblastx similarity as the sum of bit scores of significant hits between two genomes
415 (e-value ≤ 0.001 , bit score ≥ 30 , identity percentage ≥ 30). To normalize for different genome sizes,
416 each genome was also compared to itself to generate a self-score, and the distance between two
417 different genomes was calculated as a Dice coefficient (as in⁹), i.e. for two genomes A and B with a

418 proteomic similarity score of AB, the corresponding distance d would be $1-(2*AB)/(AA+BB)$, with AA
419 and BB being the self-score of genomes A and B respectively. For clarity, the tree displayed in
420 Extended Data Fig. 2 only include non-GOV sequences found in a VC with GOV sequence(s) or within
421 a distance <0.5 to a GOV sequence, adding for a total of 1,522 reference sequences. iTOL^{61,62} was used
422 to visualize and display the tree.

423

424 **Distribution and relative abundance of viral populations and VCs**

425 *Detection and estimation of abundance for viral contigs and populations*

426 The presence and relative abundance of a viral contig in a sample were determined based on the
427 mapping of HQ reads to the contig sequences, computed with Bowtie 2 (options --sensitive, -X 2000,
428 and --non-deterministic, default parameters otherwise⁴²), as previously described⁴. A contig was
429 considered as detected in a metagenome if more than 75% of its length was covered by aligned reads
430 derived from the corresponding sample. A normalized coverage for the contig was then computed as
431 the average contig coverage (i.e. number of nucleotides mapped to the contig divided by the contig
432 length) normalized by the total number of bp sequenced in this sample. The detection and relative
433 abundance of a viral population was based on the coverage of its contigs: a population was considered
434 as detected in a sample if more than 75% of its cumulated length was covered, and its normalized
435 coverage was computed as the average normalized coverage of its contigs.

436

437 *Relative abundance of VCs*

438 The relative abundance of VCs was calculated based on the coverage of its members within the
439 15,222 viral populations identified. If a population included contigs all linked to the same VC, or
440 linked to a single VC except for unclustered (because too short) contigs, this population coverage was
441 added to the total of the corresponding VC. In the rare cases where the link between population and VC
442 was ambiguous because different contigs within a population pointed toward different VCs ($n=475$, i.e.
443 3.1% of the populations), the population coverage was equally split between these VCs. Finally, if no
444 contig in the population belonged to any VC ($n=2,605$, 17% of the populations), the population
445 coverage was added to the “unclustered” category. Eventually, for each sample, the cumulated coverage
446 of a VC was normalized by the total coverage of all populations to calculate a relative abundance of the
447 VC among viral populations.

448 The selection of abundant VCs within a sample was based on the contribution of the VC to the
449 sample diversity as measured by the Simpson index. For each sample, the overall Simpson index was
450 first calculated with all VCs. Then, VCs were sorted by decreasing relative abundance and
451 progressively added to a new calculation of the Simpson index. VCs considered as abundant were the
452 ones which, once cumulated, represented 80% of the sample diversity (i.e. a Simpson index greater or
453 equal to 80% of the sample total Simpson index, Extended Data Fig. 1C). The 38 VCs identified as
454 abundant in at least 2 different stations were selected as “recurrently abundant VCs in the GOV
455 dataset” (Fig. 2 and Extended Data Fig. 3).

456

457 **Host prediction and diversity**

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

464

465 *Generation of host database*

466 A genome database of putative hosts for the epi- and mesopelagic GOV viruses was generated,

467 including all archaea and bacteria genomes annotated as “marine” from NCBI RefSeq and WGS (both
468 times only sequences $\geq 5\text{kb}$, 184,663 sequences from 4,452 genomes, downloaded in August 2015), and
469 all contigs $\geq 5\text{kb}$ from the 139 *Tara Oceans* microbial metagenomes corresponding to the bacteria and
470 archaea size fraction (791,373 sequences)¹⁸. For these microbial metagenomic contigs, a first blastn
471 was computed to compare them to all GOV contigs, and exclude from the putative host dataset all
472 metagenomic contigs with a significant similarity to a viral GOV sequence (thresholds of 50 on bit
473 score, 0.001 on e-value, and 70% on identity percentage) on $\geq 90\%$ of their length, as these are likely
474 sequences of viral origin sequenced in the bacteria and archaea size fraction (these represented 2.2% of
475 the contigs in the assembled microbial metagenomes). The taxonomic affiliation of NCBI genomes was
476 taken from the NCBI taxonomy. For *Tara Oceans* contigs, a last common ancestor (LCA) affiliation
477 was generated for each contig based on genes affiliation¹⁸, if 3 genes or more on the contig were
478 affiliated.

479 480 *BLAST-based identification of sequence similarity between viral contigs and host genome*

481 All GOV viral contigs were compared to all archaeal and bacterial genomes and genome fragments
482 with a blastn (threshold of 50 on bit score and 0.001 on e-value), to identify regions of similarity
483 between a viral contig and a microbial genome, indicative of a prophage integration or horizontal gene
484 transfer⁶³. A host prediction was made when (i) a NCBI genomes displayed a region similar to a GOV
485 viral contig $\geq 5\text{kb}$ at $\geq 70\%$ id, or (ii) when a *Tara Oceans* microbial metagenomic contig ($\geq 5\text{kb}$)
486 displayed a region similar to a GOV viral contig $\geq 2.5\text{kb}$ at $\geq 70\%$ id.

487 488 *Matches between GOV viral contigs and CRISPR spacers.*

489 CRISPR arrays were predicted for all putative host genomes and genome fragments (NCBI
490 microbial genomes and *Tara Oceans* microbial metagenomic contigs) with MetaCRT^{64,65}. CRISPR
491 spacers were extracted, and all spacers with ambiguous bases or low complexity (i.e. consisting of 4 to
492 6 bp repeat motifs) were removed. All remaining spacers were matched to viral contigs with fuzznuc⁶⁶,
493 with no mismatches allowed, which although rarely observed yields highly accurate host predictions⁶³ (
494 Supplementary Table 4).

495 496 *Nucleotide composition similarity: comparison of tetranucleotide frequency*

497 Bacterial and archaeal viruses tend to have a genome composition close to the genome composition
498 of their host, a signal that can be used to predict viral-host pairs^{8,63,67}. Here, canonical tetranucleotide
499 frequencies were observed for all viral and host sequences using Jellyfish⁶⁸, and mean absolute error
500 (i.e. average of absolute differences) between tetranucleotide frequency vectors were computed with
501 in-house Perl and Python scripts for each pair of viral and host sequence as in ref. ⁸. A GOV viral
502 contig was then assigned to the closest sequence (i.e. lowest distance d) from the pool of NCBI
503 genomes if $d < 0.001$ (because both the tetranucleotide frequency signal and the taxonomic affiliation of
504 these complete genomes are more robust than for metagenomic contigs), and otherwise assigned to the
505 closest (i.e. lowest distance) *Tara Oceans* microbial contig if $d < 0.001$.

506 507 *Summarizing host prediction at the VC level*

508 Overall, 3,675 GOV contigs could be linked to a putative host group among the 24,353 GOV
509 contigs associated with an epi- or mesopelagic viral population. To summarize these affiliations at the
510 VC level, a Poisson distribution was used to estimate the number of expected false positive associations
511 for each VC – host group combination based on (i) the global probability of obtaining a host prediction
512 across all pairs of viral and host sequences tested and for all methods (5.8×10^{-08}), (ii) the number of
513 potential predictions generated for the VC, corresponding to 3 times the number of sequences in the VC
514 (to take into account the three methods), and (iii) the number of sequences from the host group in the
515 database (Supplementary Figure 2). By comparing the number of links observed between a VC and a

516 host group to this expected value, which takes into account the bias in database (i.e. some host groups
517 will be over- or under-represented in our set of archaeal and bacterial genomes and genome fragments)
518 and the bias linked to the variable number of sequences in VCs, we can determine if the number of
519 associations observed for any VC – host group combination is likely to be due to chance alone (and
520 calculate the associated p-value).

521

522 *Microbial community diversity and richness indexes*

523 Diversity and richness indexes for putative host populations were based on the OTU abundance
524 matrix generated from the analysis of *mi*TAGs in *Tara* Oceans microbial metagenomes¹⁸. These indexes
525 were computed for each host group at the same taxonomic level as the host prediction, i.e. the phylum
526 level except for Proteobacteria where the class level is used. The R package *vegan*⁶⁹ was used to
527 estimate for each group (i) a global Chao index (i.e. including all OTUs from all samples) through the
528 function *estaccumR*, (ii) a sample-by-sample Chao index with the function *estimateR*, and (iii)
529 Sorensen indexes between all pairs of samples with the function *betadiver*. Diversity indexes presented
530 in Extended Data Fig 4 are based on epipelagic samples only, as the 38 VCs identified as abundant
531 were mostly retrieved in epipelagic samples. Candidate division OP1 was excluded from this analysis
532 because no OTU affiliated to this phylum was identified.

533

534 **Identification and annotation of putative AMGs**

535 *Detection of AMGs*

536 Predicted proteins from all GOV viral contigs were compared to the PFAM domain database
537 (*hmmsearch*⁵², threshold of 40 on bit score and 0.001 on e-value), and all PFAM domains detected
538 were classified into 8 categories: “structural”, “DNA replication, recombination, repair, nucleotide
539 metabolism”, “transcription, translation, protein synthesis”, “lysis”, “membrane transport, membrane-
540 associated”, “metabolism”, “other”, and “unknown” (as in ref.²⁰). Four AMGs (i.e. similar to a domain
541 from the “metabolism” category) were then selected for further study because of their central role in
542 sulfur (*dsrC* and *soxYZ*) or nitrogen (P-II, *amoC*) cycle, and the fact that these had never been detected
543 in a surface ocean viral genome so far (*dsrC/tusE*-like genes have been detected in deep water
544 viruses^{11,21}). To evaluate if an AMG was “known”, a list of PFAM domain detected in NCBI
545 RefSeqVirus and Environmental Phages was computed based on a similar *hmmsearch* comparison
546 (threshold of 40 on bit score and 0.001 on e-value), and augmented by manual annotation of AMGs
547 from^{20,70}. These corresponded for the most part to photosynthesis and carbon metabolism AMGs
548 previously described in cyanophages^{71–75}. The complete list of PFAM domains detected in GOV viral
549 contigs is available as Supplementary Table 6.

550

551 *Phylogenetic tree generation and contigs map comparison*

552 Sequences similar to these AMGs were recruited from the *Tara* Oceans microbial metagenomes¹⁸
553 based on a *blastp* of all predicted proteins from microbial metagenome to the viral AMGs identified
554 (threshold of 100 on bit score, 10^{-5} on e-value, except for P-II where a threshold of 170 on bit score was
555 used because of the high number of sequences recruited). The viral AMG sequences were also
556 compared to NCBI nr database (*blastp*, threshold of 50 on bit score and 10^{-3} on e-value) to recruit
557 relevant reference sequences (up to 20 for each viral AMG sequence). These sets of viral AMGs and
558 related protein sequences were then aligned with *Muscle*⁷⁶, the alignment manually curated to remove
559 poorly aligned positions with *Jalview*⁷⁷, and two trees were computed from the same curated
560 alignment: a maximum-likelihood tree with *FastTree* (v2.7.1, model WAG, other parameters set to
561 default⁷⁸) and a bayesian tree with *MrBayes* (v3.2.5, mixed evolution models, other parameters set to
562 default, 2 MCMC chains were run until the average standard deviation of split frequencies was <0.015 ,
563 relative burn-in of 25% used to generate the consensus tree⁷⁹). In all cases except *AmoC*, the mixed
564 model used by *MrBayes* was 100% WAG, confirming that this model was well suited for archaeal and

565 bacterial virus protein trees. Manual inspection revealed only minor differences between each pair of
566 trees, so an SH test was used to determine which tree best fitted the sequence alignment, using the R
567 library phangorn⁸⁰. ItoI⁶¹ was used to visualize and display these trees, in which branches with supports
568 <40% were collapsed. Annotated interactive trees are available online at
569 <http://itol.embl.de/shared/Siroux>. Contigs map comparison were generated with Easyfig⁸¹, following
570 the same method as for the VCs (see Supplementary Information).

571

572 *Functional characterization of putative AMGs*

573 Conserved motifs were identified on the different AMGs based on the literature: *dsrc* conserved
574 motifs were obtained from ref. ²⁴, *soxYZ* conserved residues were identified from the PFAM domains
575 PF13501 and PF08770, and P-II conserved motifs from PROSITE documentation PDOC00439. A 3D
576 structure could also be predicted for P-II AMGs by I-TASSER⁸² (default parameters), the quality of
577 these predictions being confirmed with ProSA web server⁸³. To further confirm the functionality of
578 these genes, selective constraint on these AMGs was evaluated through pN/pS calculation, as in ref. ⁸⁴.
579 Briefly, synonymous and non-synonymous SNPs were observed in each AMG, and compared to
580 expected ratio of synonymous and non-synonymous SNPs under a neutral evolution model for this
581 genes. The interpretation of pN/pS is similar as for dN/dS analyses, with the operation of purifying
582 selection leading to pN/pS values < 1. Finally, AMG transcripts were searched in metatranscriptomic
583 datasets generated through the *Tara* Oceans consortium (ENA Id ERS1092158, ERS488920, and
584 ERS494518). For generating these metatranscriptomes, bacterial rRNA depletion was carried out on
585 240–500 ng total RNA using Ribo-Zero Magnetic Kit for Bacteria (Epicentre, Madison, WI) for 0.2–
586 1.6 and 0.22–3µm filters. The Ribo-Zero depletion protocol was modified to be adapted to low RNA
587 input amounts⁸⁵. Depleted RNA was used to synthesize cDNA with SMARTer Stranded RNA-Seq Kit
588 (Clontech, Mountain View, CA)⁸⁵. Metatranscriptomic libraries were quantified by qPCR using the
589 KAPA Library Quantification Kit for Illumina Libraries (KapaBiosystems, Wilmington, MA) and
590 library profiles were assessed using the DNA High Sensitivity LabChip kit on an Agilent Bioanalyzer
591 (Agilent Technologies, Santa Clara, CA). Libraries were sequenced on Illumina HiSeq2000 instrument
592 (Illumina, San Diego, CA) using 100 base-length read chemistry in a paired-end mode. High quality
593 reads were then mapped to viral contigs containing *dsrc*, *soxYZ*, P-II, or *amoC* genes with
594 SOAPdenovo2⁴³ within MOCAT⁴⁰ (options *screen* and *filter* with length and identity cutoffs of 45 and
595 95%, respectively, and paired-end filtering set to *yes*), and coverage was defined for each gene as the
596 number of bp mapped divided by gene length (including only reads mapped to the predicted coding
597 strand).

598

599 *Distribution of AMGs and association with geochemical metadata*

600 The distribution and relative abundance of AMGs was based on the read mapping and normalized
601 coverage of the contig including the AMG. To get a range of temperature and nutrient concentrations
602 for the widespread AMGs (detected in >5 stations) that takes into account both the samples in which
603 these AMGs were detected and the differences in normalized coverage, a set of samples was selected
604 through a weighted random drawing replacement, with the weight of each sample corresponding to the
605 AMG's normalized coverage. That way, a range of temperature or nutrient concentration values
606 associated with the AMG's distribution and abundance could be generated for each AMG and each
607 environmental parameter tested. The number of samples randomly selected for each AMG was the
608 same as the total number of samples for which a value of this parameter was available.

609

610 **Code and data availability**

611 Scripts used in this manuscript are available on the Sullivan lab bitbucket under project
612 “GOV_Ecogenomics” (http://bitbucket.org/MAVERICLab/gov_ecogenomics/overview). Scripts used
613 in the assessment of microbial diversity are gathered in the directory “Host_diversity”, the ones used

614 for host predictions are in “Host_prediction”, and the scripts used to identify abundant VCs are in
615 “Virus_clusters_prevalence”. All raw reads are available through ENA (*Tara Oceans*) or JGI
616 (Malaspina) using the dataset identifiers listed in Supplementary Table 1. Processed data are available
617 through iVirus (<http://mirrors.iplantcollaborative.org/browse/iplant/home/shared/ivirus/GOV/>),
618 including all sequences from assembled contigs, list of viral populations and associated annotated
619 sequences as genbank files, viral clusters composition and characteristics, map comparisons of
620 genomes and contigs of the 38 abundant VCs, and host predictions for viral contigs.

621

622 **References**

- 623 1. Falkowski, P. G., Fenchel, T. & Delong, E. F. The Microbial Engines That Drive Earth’s
624 Biogeochemical Cycles. *Science* **320**, 1034–9 (2008).
- 625 2. Rohwer, F. & Thurber, R. V. Viruses manipulate the marine environment. *Nature* **459**, 207–212
626 (2009).
- 627 3. Brum, J. R. & Sullivan, M. B. Rising to the challenge: accelerated pace of discovery transforms
628 marine virology. *Nat. Rev. Microbiol.* **13**, 147–59 (2015).
- 629 4. Brum, J. *et al.* Patterns and ecological drivers of ocean viral communities. *Science* **348**, 1261498–1–
630 10 (2015).
- 631 5. Karsenti, E. *et al.* A holistic approach to marine eco-systems biology. *PLoS Biol.* **9**, e1001177 (2011).
- 632 6. Duarte, C. M. Seafaring in the 21st Century : The Malaspina 2010 Circumnavigation Expedition.
633 *Limnology and Oceanography Bulletin* **24**, 11–14 (2015).
- 634 7. Lima-Mendez, G., Van Helden, J., Toussaint, A. & Leplae, R. Reticulate representation of
635 evolutionary and functional relationships between phage genomes. *Mol. Biol. Evol.* **25**, 762–77
636 (2008).
- 637 8. Roux, S., Hallam, S. J., Woyke, T. & Sullivan, M. B. Viral dark matter and virus-host interactions
638 resolved from publicly available microbial genomes. *Elife* **4**, 1–20 (2015).
- 639 9. Mizuno, C. M., Rodriguez-Valera, F., Kimes, N. E. & Ghai, R. Expanding the marine virosphere
640 using metagenomics. *PLoS Genet.* **9**, e1003987 (2013).
- 641 10. Chow, C.-E. T., Winget, D. M., White, R. a., Hallam, S. J. & Suttle, C. a. Combining genomic
642 sequencing methods to explore viral diversity and reveal potential virus-host interactions. *Front.*
643 *Microbiol.* **6**, 1–15 (2015).
- 644 11. Roux, S. *et al.* Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed
645 by single-cell- and meta- genomics. *Elife* **3**, 1–20 (2014).
- 646 12. Dutilh, B. E. *et al.* A highly abundant bacteriophage discovered in the unknown sequences of
647 human faecal metagenomes. *Nat. Commun.* **5**, 1–11 (2014).
- 648 13. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential
649 coverage binning of multiple metagenomes. *Nat. Biotechnol.* **31**, 533–8 (2013).
- 650 14. Sullivan, M. B. *et al.* Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-

- 651 like myoviruses from diverse hosts and environments. *Environ. Microbiol.* **12**, 3035–56 (2010).
- 652 15. Zhao, Y. *et al.* Abundant SAR11 viruses in the ocean. *Nature* **494**, 357–360 (2013).
- 653 16. Labrie, S. J. *et al.* Genomes of marine cyanopodoviruses reveal multiple origins of diversity.
654 *Environ. Microbiol.* **15**, 1356–76 (2013).
- 655 17. Andersson, A. F. & Banfield, J. F. Virus population dynamics and acquired virus resistance in
656 natural microbial communities. *Science* **320**, 1047–50 (2008).
- 657 18. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348**, 1–10
658 (2015).
- 659 19. Flores, C. O., Valverde, S. & Weitz, J. S. Multi-scale structure and geographic drivers of cross-
660 infection within marine bacteria and phages. *ISME J.* **7**, 520–32 (2013).
- 661 20. Hurwitz, B. L., Brum, J. R. & Sullivan, M. B. Depth-stratified functional and taxonomic niche
662 specialization in the “core” and “flexible” Pacific Ocean Virome. *ISME J.* **9**, 472–84 (2015).
- 663 21. Anantharaman, K. *et al.* Sulfur Oxidation Genes in Diverse Deep-Sea Viruses. *Science* **344**, 757–
664 760 (2014).
- 665 22. Friedrich, C. G., Bardischewsky, F., Rother, D., Quentmeier, A. & Fischer, J. Prokaryotic sulfur
666 oxidation. *Curr. Opin. Microbiol.* **8**, 253–9 (2005).
- 667 23. Santos, A. A. *et al.* A protein trisulfide couples dissimilatory sulfate reduction to energy
668 conservation. *Science* **350**, 1541–1546 (2015).
- 669 24. Venceslau, S. S., Stockdreher, Y., Dahl, C. & Pereira, I. A. C. The “bacterial heterodisulfide” DsrC
670 is a key protein in dissimilatory sulfur metabolism. *Biochim. Biophys. Acta* **1837**, 1148–64
671 (2014).
- 672 25. Dahl, C., Franz, B., Hensen, D., Kesselheim, A. & Zigann, R. Sulfite oxidation in the purple sulfur
673 bacterium *Allochromatium vinosum*: identification of SoeABC as a major player and relevance
674 of SoxYZ in the process. *Microbiology* **159**, 2626–38 (2013).
- 675 26. Huergo, L. F., Chandra, G. & Merrick, M. P(II) signal transduction proteins: nitrogen regulation and
676 beyond. *FEMS Microbiol. Rev.* **37**, 251–83 (2013).
- 677 27. Stahl, D. A. & de la Torre, J. R. Physiology and diversity of ammonia-oxidizing archaea. *Annu. Rev.*
678 *Microbiol.* **66**, 83–101 (2012).
- 679 28. Loy, A. *et al.* Reverse dissimilatory sulfite reductase as phylogenetic marker for a subgroup of
680 sulfur-oxidizing prokaryotes. *Environ. Microbiol.* **11**, 289–99 (2009).
- 681 29. Pester, M., Schleper, C. & Wagner, M. The Thaumarchaeota: an emerging view of their phylogeny
682 and ecophysiology. *Curr. Opin. Microbiol.* **14**, 300–6 (2011).
- 683 30. Weitz, J. S. *et al.* A multitrophic model to quantify the effects of marine viruses on microbial food
684 webs and ecosystem processes. *ISME J.* **9**, 1352–1364 (2015).

685 31. Arcondéguy, T., Jack, R. & Merrick, M. P II Signal Transduction Proteins, Pivotal Players in
686 Microbial Nitrogen Control. *Microbiol. Mol. Biol. Rev.* **65**, 80–105 (2001).

687

688 **References: Methods section**

689 32. Pesant, S. *et al.* Open science resources for the discovery and analysis of Tara Oceans data. *Sci.*
690 *Data* **2**, 150023 (2015).

691 33. John, S. G. *et al.* A simple and efficient method for concentration of ocean viruses by chemical
692 flocculation. *Environ. Microbiol. Rep.* **3**, 195–202 (2011).

693 34. Hurwitz, B. L., Deng, L., Poulos, B. T. & Sullivan, M. B. Evaluation of methods to concentrate and
694 purify ocean virus communities through comparative, replicated metagenomics. *Environ.*
695 *Microbiol.* **15**, 1428 – 1440 (2012).

696 35. Aminot, A., K erouel, R. & Coverly, S. in *Pract. Guidel. Anal. Seawater* 143–176 (2009).

697 36. Tara Oceans Consortium & Tara Oceans Expedition. Registry of all samples from the Tara Oceans
698 Expedition (2009-2013). (2015). doi:10.1594/PANGAEA.842197

699 37. Tara Oceans Consortium & Tara Oceans Expedition. Environmental context of all samples from the
700 Tara Oceans Expedition (2009-2013). (2015). doi:10.1594/PANGAEA.853810

701 38. Tara Oceans Consortium & Tara Oceans Expedition. Biodiversity context of all samples from the
702 Tara Oceans Expedition (2009-2013). (2015). doi:10.1594/PANGAEA.853809

703 39. Salazar, G. *et al.* Global diversity and biogeography of deep-sea pelagic prokaryotes. *ISME J.* **10**,
704 596–608 (2016). doi:10.1038/ismej.2015.137

705 40. Kultima, J. R. *et al.* MOCAT: a metagenomics assembly and gene prediction toolkit. *PLoS One* **7**,
706 e47656 (2012).

707 41. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-
708 cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428
709 (2012).

710 42. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–9
711 (2012).

712 43. Luo, R. *et al.* SOAPdenovo2 : an empirically improved memory-efficient short-read de novo
713 assembler. *Gigascience* **1**, 1–6 (2012).

714 44. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or
715 nucleotide sequences. *Bioinformatics* **22**, 1658–9 (2006).

716 45. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately
717 reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).

- 718 46. Mavromatis, K., Ivanova, N., Barry, K. & Shapiro, H. Use of simulated data sets to evaluate the
719 fidelity of metagenomic processing methods. *Nat. Methods* **4**, 495–500 (2007).
- 720 47. Roux, S., Krupovic, M., Debroas, D., Forterre, P. & Enault, F. Assessment of viral community
721 functional potential from viral metagenomes may be hampered by contamination with cellular
722 sequences. *Open Biol.* **3**, 130160 (2013).
- 723 48. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial
724 genomic data. *PeerJ* **3**, e985 (2015).
- 725 49. Pope, W. H. *et al.* Whole genome comparison of a large collection of mycobacteriophages reveals a
726 continuum of phage genetic diversity. *Elife* **4**, (2015).
- 727 50. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of
728 protein families. *Nucleic Acids Res.* **30**, 1575–84 (2002).
- 729 51. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–30 (2014).
- 730 52. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
- 731 53. Brum, J. R. *et al.* Illuminating structural proteins in viral “dark matter” with metaproteomics. *Proc.*
732 *Natl. Acad. Sci. U. S. A.* **113**, 2436–2441 (2016).
- 733 54. Holmfeldt, K. *et al.* Twelve previously unknown phage genera are ubiquitous in the global oceans.
734 *Proc. Natl. Acad. Sci. U. S. A.* **110**, 12798–12803 (2013).
- 735 55. Kang, I., Jang, H. & Cho, J.-C. Complete genome sequences of two *Persicivirga* bacteriophages,
736 P12024S and P12024L. *J. Virol.* **86**, 8907–8 (2012).
- 737 56. Kang, I., Oh, H.-M., Kang, D. & Cho, J.-C. Genome of a SAR116 bacteriophage shows the
738 prevalence of this phage type in the oceans. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 12343–8 (2013).
- 739 57. Hjorleifsdottir, S., Aevarsson, A., Hreggvidsson, G. O., Fridjonsson, O. H. & Kristjansson, J. K.
740 Isolation, growth and genome of the *Rhodothermus* RM378 thermophilic bacteriophage.
741 *Extremophiles* **18**, 261–70 (2014).
- 742 58. Marks, T. J. & Hamilton, P. T. Characterization of a thermophilic bacteriophage of *Geobacillus*
743 *kaustophilus*. *Arch. Virol.* **159**, 2771–5 (2014).
- 744 59. Halmillawewa, A. P., Restrepo-Córdoba, M., Yost, C. K. & Hynes, M. F. Genomic and phenotypic
745 characterization of *Rhizobium gallicum* phage vB_RglS_P106B. *Microbiology* **161**, 611–20
746 (2015).
- 747 60. Rohwer, F. & Edwards, R. The Phage Proteomic Tree : a Genome-Based Taxonomy for Phage. *J.*
748 *Bacteriol.* **184**, 4529–4535 (2002).
- 749 61. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display
750 and annotation. *Bioinformatics* **23**, 127–128 (2007).
- 751 62. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic

- 752 trees made easy. *Nucleic Acids Res.* **39**, W475–8 (2011).
- 753 63. Edwards, R. A., McNair, K., Faust, K., Raes, J. & Dutilh, B. E. Computational approaches to
754 predict bacteriophage-host relationships. *FEMS Microbiol. Rev.* **40**, 258-272 (2015).
- 755 64. Bland, C. *et al.* CRISPR recognition tool (CRT): a tool for automatic detection of clustered
756 regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**, 209 (2007).
- 757 65. Rho, M., Wu, Y.-W., Tang, H., Doak, T. G. & Ye, Y. Diverse CRISPRs Evolving in Human
758 Microbiomes. *PLoS Genet.* **8**, e1002441 (2012).
- 759 66. Rice, P., Longden, I. & Bleasby, a. EMBOSS: the European Molecular Biology Open Software
760 Suite. *Trends Genet.* **16**, 276–7 (2000).
- 761 67. Ogilvie, L. a *et al.* Genome signature-based dissection of human gut metagenomes to extract
762 subliminal viral sequences. *Nat. Commun.* **4**, 2420 (2013).
- 763 68. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences
764 of k-mers. *Bioinformatics* **27**, 764–70 (2011).
- 765 69. Oksanen, J. *et al.* *The vegan Package.* (2016).
- 766 70. Sharon, I. *et al.* Comparative metagenomics of microbial traits within oceanic viral communities.
767 *ISME J.* **5**, 1178–90 (2011).
- 768 71. Thompson, L. R. *et al.* Phage auxiliary metabolic genes and the redirection of cyanobacterial host
769 carbon metabolism. *Proc. Natl. Acad. Sci. U. S. A.* **108**, E757–64 (2011).
- 770 72. Dammeyer, T., Bagby, S. C., Sullivan, M. B., Chisholm, S. W. & Frankenberg-Dinkel, N. Efficient
771 phage-mediated pigment biosynthesis in oceanic cyanobacteria. *Curr. Biol.* **18**, 442–8 (2008).
- 772 73. Lindell, D., Jaffe, J. D., Johnson, Z. I., Church, G. M. & Chisholm, S. W. Photosynthesis genes in
773 marine viruses yield proteins during host infection. *Nature* **438**, 86–9 (2005).
- 774 74. Lindell, D. *et al.* Genome-wide expression dynamics of a marine virus and host reveal features of
775 co-evolution. *Nature* **449**, 83–6 (2007).
- 776 75. Sullivan, M. B. *et al.* Prevalence and Evolution of Core Photosystem II Genes in Marine
777 Cyanobacterial Viruses and Their Hosts. *PLoS Biol.* **4**, e234 (2006).
- 778 76. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space
779 complexity. *BMC Bioinformatics* **5**, 113 (2004).
- 780 77. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2--a
781 multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191
782 (2009).
- 783 78. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for
784 large alignments. *PLoS One* **5**, e9490 (2010).

- 785 79. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees.
786 *Bioinformatics* **17**, 754–755 (2001).
- 787 80. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–3 (2011).
- 788 81. Sullivan, M. J., Petty, N. K. & Beatson, S. A. Easyfig: a genome comparison visualizer.
789 *Bioinformatics* **27**, 1009–10 (2011).
- 790 82. Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for automated protein structure
791 and function prediction. *Nat. Protoc.* **5**, 725–38 (2010).
- 792 83. Wiederstein, M. & Sippl, M. J. ProSA-web: interactive web service for the recognition of errors in
793 three-dimensional structures of proteins. *Nucleic Acids Res.* **35**, W407–10 (2007).
- 794 84. Schloissnig, S. *et al.* Genomic variation landscape of the human gut microbiome. *Nature* **493**, 45–
795 50 (2013).
- 796 85. Alberti, A. *et al.* Comparison of library preparation methods reveals their impact on interpretation
797 of metatranscriptomic data. *BMC Genomics* **15**, 912 (2014).

798

799 **Acknowledgements**

800 We thank Joshua Weitz for advice on statistics, Claus Pelikan for help with the DsrAB phylogenetic
801 tree, Christiane Dahl for valuable discussion regarding DsrC function, and members of the Sullivan and
802 Rich labs for suggestions and comments on this manuscript. We acknowledge support from UA high-
803 performance computing and the Ohio Supercomputer Center. Sponsors and support for *Tara* Oceans
804 and Malaspina expeditions are listed in the Supplementary Information. This viral research was funded
805 by a National Science Foundation grant (#1536989) and Gordon and Betty Moore Foundation grants
806 (#3790, #2631) to MBS, and the French Ministry of Research and Government through the
807 “Investissements d’Avenir” program OCEANOMICS (ANR-11-BTBR-0008). Virus researchers were
808 partially supported by the Water, Environmental and Energy Solutions Initiative and the Ecosystem
809 Genomics Institute (SR), the Netherlands Organization for Scientific Research Vidi grant 864.14.004
810 and CAPES/BRASIL (BED), and the Austrian Science Fund (project P25111-B22, AL). The authors
811 declare that all data reported herein are fully and freely available from the date of publication, with no
812 restrictions, at EBI, PANGAEA, and iVirus (see Methods and Supplementary Table 1), and that all of
813 the samples, analyses, publications, and ownership of data are free from legal entanglement or
814 restriction of any sort by the nations in whose waters *Tara* Oceans expedition sampled in. All authors
815 approved the final manuscript. This article is contribution number 43 of the *Tara* Oceans Expedition.

816

817 **Consortia**

818 *Tara* Oceans Consortium Coordinators

819 A list of authors and affiliations appears in the Supplementary Information.

820

821 **Author Contributions**

822 S.R., and M.B.S. designed the study. C.D., M.P., and Sa.S., contributed extensively to sampling
823 collection. S.K-L. managed the logistic of the *Tara* Oceans project. B.T.P., N.S. and E.L. performed the
824 viral-specific processing of the samples. J.P., C.C., A.A., and P.W. led the sequencing of viral samples.
825 S.R., S.S. and B.E.D. led the assembly of raw data. S.R., S.S., M.B.D. and M.B.S. analyzed the
826 genomic diversity data. S.R., A.L., J.R.B. and M.B.S. analyzed the AMGs data. S.R., J.R.B., B.E.D,
827 S.S., 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,

828 revised and edited the manuscript. *Tara* Oceans coordinators provided creative environment and
829 constructive criticism throughout the study. All authors discussed the results and commented on the
830 manuscript.

831

832 **Competing financial interests**

833 The authors declare no competing financial interests.

834

835 **Author Information**

836 Reprints and permissions information is available at www.nature.com/reprints. Correspondence and
837 requests for materials should be addressed to mbsulli@gmail.com.

838

839 **Figure legends**

840

841 **Figure 1: Composition of the Global Ocean Viromes (GOV) dataset.** **A.** Size of viral contigs (x-
842 axis) and cumulative coverage across the GOV dataset (y-axis). Contigs corresponding to complete
843 (345 contigs) or near-complete genomes (425 contigs) are indicated. For clarity, only contigs associated
844 with a viral population (24,412 contigs) are displayed. **B.** Distribution of all viral clusters (VCs)
845 according to the origin of their members. Viral genomes (or fragments) in a VC can originate from
846 isolate viral genomes, the VirSorter Curated Dataset⁸ (viral genomes identified *in silico* from microbial
847 genomes), environmental viral genomes and genome fragments (e.g. from fosmid libraries), or the
848 GOV dataset. VCs including at least one GOV sequence and further analyzed in this study are
849 highlighted in bold.

850

851 **Figure 2: Characterization of the dominant oceanic viral clusters (VCs).** **A.** Distribution and
852 abundance of the 38 recurrently abundant VCs according to the total number of stations in which
853 members of the VC were detected (x-axis), and the number of samples in which the VC was detected in
854 the abundant fraction (y-axis). “Known viruses” are VCs with ICTV-classified reference sequences,
855 “Unclassified reference(s)” are VCs with isolate genomes lacking ICTV classification, and “New VCs”
856 are composed solely of environmental sequences. **B.** GOV samples with their most abundant VC
857 mapped to station locations. Samples are stacked vertically when multiple depths are available, with a
858 horizontal line separating epipelagic from mesopelagic layers. Map modified with permission from N.
859 Le Bescot, EPEP, CNRS Roscoff. **C.** Summary of the 4 globally abundant VCs affiliation, origin of VC
860 members (Env: environmental viral sequences), estimated genome size, predicted host range, and
861 distribution (relative abundance are indicated as % of the viral populations identified). The abundant
862 epipelagic microbial groups (representing >1% of the microbial OTUs abundance of epipelagic
863 samples) are highlighted in bold; Alphaproteob.-Alphaproteobacteria, Betaproteob.-Betaproteobacteria,
864 Deinococcus-Th.-Deinococcus-Thermus, Deltaproteob.-Deltaproteobacteria, Gammaproteob.-
865 Gammaproteobacteria, Cand div OP1-Candidate division OP1. Oceanic basins are indicated for VCs
866 distributions; Med. Sea-Mediterranean Sea.

867

868 **Figure 3: Characterization and distribution of viral Auxiliary Metabolic Genes (AMGs) involved**
869 **in sulfur and nitrogen cycles.** Schematics for **(A)** microbial sulfur oxidation pathways involving the
870 two main gene clusters (*dsr* and *sox*) and **(B)** the central role of the P-II protein in cell regulation
871 (adapted from^{26,31}). AMG color outlines indicate their viral taxonomic affiliation. Ammonium
872 transporters detected next to viral P-II are highlighted with a dashed outline. **C.** Distribution of viral
873 AMG clades, with mesopelagic samples highlighted in green, and geographically restricted clades
874 outlined. **D.** Temperature and nutrient conditions for which widespread epipelagic AMGs tend to be
875 most abundant. For each environmental parameter, the range across all epipelagic samples is displayed
876 alongside distributions representing the range of values where each AMG clade was detected, weighted

877 by the AMG coverage across these samples (see Extended Data Fig. 9 for underlying coverage data).
878 Distributions significantly different from the “All Samples” distribution (two-sided KS-test) are
879 indicated with stars. Boxes represent the first and third quartiles around the median.

880

881 **Extended Data Figure 1: Accumulation curves of populations (A) and viral clusters (VCs, B) and**
882 **identification of abundant VCs in GOV samples (C). A & B.** Accumulation curves were computed
883 from 50 randomly shuffled samples (blue dots), with all, epipelagic, mesopelagic, or bathypelagic
884 subsets of the data. For each curve, the average of 50 iterations is displayed with red dots. C. Schematic
885 of the selection process of abundant VCs. For each sample, VCs accounting for (up to) 80% of the
886 sample diversity (as assessed by Simpson index) were considered as abundant (example for sample
887 125_MIX on the left). VCs detected as abundant in at least two different stations were included in the
888 38 VCs described in Fig. 2 and Extended Data Fig. 3.

889

890 **Extended Data Figure 2: Comparison of VCs with other classification methods: phage proteomic**
891 **tree and percentage of shared genes.** The phage proteomic tree includes the 756 GOV complete and
892 near-complete genomes from epi- and mesopelagic samples, and closest references from RefSeq and
893 Environmental phages ($d < 0.5$ to a GOV sequence or found in the same VC as a GOV sequence).
894 Branches of monophyletic clades including more than 3 GOV and/or uncultivated marine sequences
895 with no isolate reference are highlighted in blue. All VCs with more than 8 representatives in the tree or
896 part of the 38 abundant VCs are indicated with coloring of the outer ring. The name and affiliation (if
897 available) of the 38 abundant VCs are indicated next to the VC on the colored ring. VCs whose
898 members were gathered in a single monophyletic clades are indicated with a solid black outline, while
899 VCs for which all but one members were gathered in a single monophyletic clades are highlighted with
900 a dashed black outline. Inset: distribution of number of shared genes estimated based on the number of
901 shared PCs (protein clusters) for viral genome/contigs pairs either between different VCs or within
902 VCs. On average, 73% and 39% of sequences within a VC shared more than 20% and 40% of their
903 genes, respectively, which represent the current thresholds currently accepted for sub-family and genus
904 designations. Similarly, 83% of sequences within a VC were consistently affiliated in the phage
905 proteomic tree as they formed a monophyletic group including only members of the particular VC.
906 Thus all three classification methods are largely consistent for the GOV dataset (see Supplementary
907 Text).

908

909 **Extended Data Figure 3: Summary of 34 of the 38 abundant viral clusters (VCs, the 4 other**
910 **abundant VCs being the ubiquitous ones presented in Fig. 2).** Predicted genome size is based on the
911 set of isolates and circular contigs in the VC (NA corresponds to VCs without any circular contigs, or
912 for which the relative standard deviation of estimated genome size across the different isolate(s) and/or
913 circular contigs is greater than 15%). Host association values are based on the number of cluster
914 members associated with each host group, the statistical significance of this number of predictions
915 being evaluated by comparison with an expected number of associations calculated from a Poisson
916 distribution. Host associations based on known isolates are indicated with a star (for associations based
917 on cultivated isolates) or a dot (for associations based on the detection of a cluster member in a
918 microbial genome from the VirSorter Curated Dataset). The abundant epipelagic microbial groups
919 (representing $>1\%$ of the microbial OTUs abundance of epipelagic samples) are highlighted in bold.
920 Distribution and relative abundance of VCs are based on the cumulated coverage of VC members
921 among sample viral populations. The main oceanic basins are indicated for each set of sample, Med.
922 Sea-Mediterranean Sea.

923

924 **Extended Data Figure 4: Association between abundant viral clusters (VCs) and host group**
925 **abundance and diversity A.** Abundance and diversity of bacterial and archaeal host groups associated

926 with the 38 abundant VCs (see Fig. 2A). For each host group (phylum level, except for Proteobacteria
927 where the class level is used), the different panels display from top to bottom (i) the number of VCs
928 associated with this host group, (ii) the global relative abundance of this group estimated from the
929 microbial metagenomic OTU counts, (iii) the global diversity of this group based on a Chao index
930 computation including all *Tara* Oceans microbial metagenome samples (i.e. including both Alpha and
931 Beta diversity), (iv) the distribution of Chao indexes by sample for this group (Alpha diversity), and (v)
932 the average Sorensen index between pairs of samples including at least one OTU of this group (Beta
933 diversity). OTU counts were derived from the 109 epipelagic microbial metagenomes described in¹⁸. **B.**
934 Pearson correlations between host group relative abundance or diversity indexes (Global Chao,
935 Average Chao across samples, and Average Sorensen across samples) and the number of VCs.

936

937 **Extended Data Figure 5: Diversity, distribution, and genome context of *dsrC* genes in GOV**
938 **contigs. A.** Maximum-likelihood tree (from an amino-acid alignment) including the 11 viral DsrC and
939 microbial sequences from microbial metagenomes and NCBI nr database. The presence of conserved C
940 residues (named Cys-A & Cys-B, as in ref. ²⁴) is indicated with color circles next to each sequence or
941 clade, and the corresponding type of DsrC-like protein is indicated by coloring the branch or clade. The
942 microbial metagenomic contigs affiliated to uncultivated, marine sulfur-oxidizing
943 Gammaproteobacteria (as confirmed by complementary phylogenetic analysis of DsrAB,
944 Supplementary Fig. 7) are indicated with a star next to the sequence or clade. Viral AMG sequences are
945 highlighted in blue, internal nodes SH-like supports are represented by proportional circles (all nodes
946 with support < 0.40 were collapsed). Each *dsrC* AMG is associated with an abundance profile (on the
947 right) displaying the relative abundance of the contig across the 91 epi- and mesopelagic samples
948 (based on normalized coverage, i.e. contig coverage / Gb of metagenome). **B.** Comparison of *dsrC*-
949 containing contigs maps. T4-like marker gene (T4 baseplate) is indicated on the maps, alongside
950 putative AMGs (Fe-S biosyn for Iron-sulfur cluster biosynthesis, and Amt for Ammonia transporter).

951

952 **Extended Data Figure 6: Diversity, distribution, and genome context of *soxYZ* genes in GOV**
953 **contigs. A.** Bayesian tree (from an amino-acid alignment) including the 4 viral SoxYZ and microbial
954 sequences from microbial metagenomes and NCBI nr database. The affiliation of microbial clades
955 (either from the NCBI reference or from the LCA affiliation of metagenomic contigs) is indicated by
956 coloring of the grouped clades or with a colored square next to the sequence. Viral AMG sequences are
957 highlighted in blue, posterior probabilities are represented by proportional circles (all nodes with
958 posterior probability < 0.40 were collapsed). Clades including sulfur-oxidation proteobacteria are
959 indicated on the tree. Each *soxYZ* AMG is associated with an abundance profile (on the right)
960 displaying the relative abundance of the contig across the 91 epi- and mesopelagic samples (based on
961 normalized coverage, i.e. contig coverage / Gb of metagenome). **B.** Comparison of *soxYZ*-containing
962 contigs maps. For contig GOV_bin_4310_contig-100_0, the second largest contig from the same bin
963 (GOV_bin_4310_contig-100_1) is displayed. T4-like marker genes (Gp23 and T4 baseplate) are
964 indicated on the maps, alongside putative AMGs (Fe-S biosyn: Iron-sulfur cluster biosynthesis).

965

966 **Extended Data Figure 7: Diversity, distribution, and genome context of P-II genes in GOV**
967 **contigs. A.** Maximum-likelihood tree (from an amino-acid alignment) including the 10 viral P-II and
968 microbial sequences from microbial metagenomes and NCBI nr database. The affiliation of microbial
969 clades (either from the NCBI reference or from the LCA affiliation of metagenomic contigs) is
970 indicated by coloring of the grouped clades or with a colored square next to the sequence. The
971 sequences lacking the conserved uridylation site of P-II (Supplementary Fig. 5) are highlighted with a
972 star next to the sequence name or clade. Viral AMG sequences are highlighted in blue, internal nodes
973 SH-like supports are represented by proportional circles (all nodes with support < 0.40 were collapsed).
974 Each P-II AMG is associated with an abundance profile (on the right) displaying the relative abundance

975 of the contig across the 91 epi- and mesopelagic samples (based on normalized coverage, i.e. contig
976 coverage / Gb of metagenome). **B.** Comparison of P-II-containing contigs maps. Ammonia transporter
977 genes linked to P-II are indicated on the map (Amm Transp, dark red). When available, the VC
978 affiliation of each contig is indicated next to the contig name. Contig GOV_bin_5834_contig-100_7 is
979 too short to be clustered based on a shared PC network, however the seed contig of its population was
980 clustered (in VC_12, *Siphoviridae - P12024virus*), hence this seed contig affiliation is indicated.

981

982 **Extended Data Figure 8: Diversity, distribution, and genome context of *amoC* gene in GOV**
983 **contigs.** **A.** Maximum-likelihood tree (from an amino-acid alignment) including the GOV *amoC* AMG
984 and microbial sequences from microbial metagenomes and NCBI nr database. The affiliation of
985 microbial clades (either from the NCBI reference or from the LCA affiliation of metagenomic contigs)
986 is indicated by coloring of the grouped clades or with a colored square next to the sequence. Viral AMG
987 sequence is highlighted in blue, internal nodes SH-like supports are represented by proportional circles
988 (all nodes with support < 0.40 were collapsed). **B.** Abundance profile displaying the relative abundance
989 of the contig across the 91 epi- and mesopelagic samples (based on normalized coverage, i.e. contig
990 coverage / Gb of metagenome). **C.** Map of the *amoC*-containing contig.

991

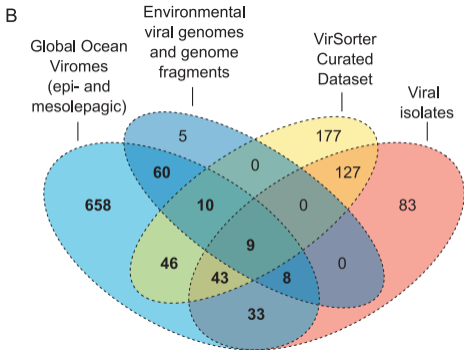
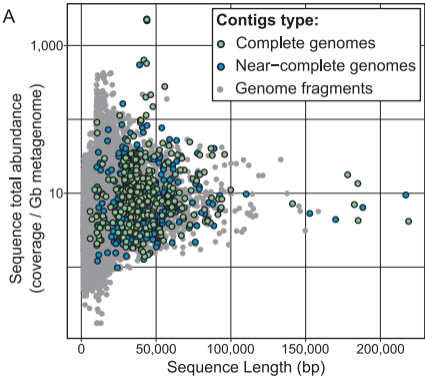
992 **Extended Data Figure 9: Normalized coverage of contigs harboring AMG as function of the**
993 **temperature and nutrient concentrations (NO₂, NO₃, PO₄) of the corresponding samples.** AMGs
994 are grouped by clade based on the phylogeny (see Extended Data Fig. 5-6-7), and coverages are
995 cumulated when a clade included multiple contigs. Plots display the cumulated normalized coverage of
996 a clade (y-axis) as function of the temperature or nutrient concentration (x-axis) across all epipelagic
997 samples (mesopelagic samples were excluded from the analysis since the AMG signal was detected in
998 epipelagic samples), only for clades not geographically restricted (i.e. found in >5 samples, see Fig.
999 3C). Samples are color-coded according to their ocean and sea region (Supplementary Table 1). The
1000 calculated preferential range of temperature or nutrient concentration is displayed below each plot for
1001 the epipelagic AMGs (P-II-4 distribution could not be linked to specific environmental conditions, but
1002 this AMG is the only one consistently retrieved in mesopelagic samples).

1003

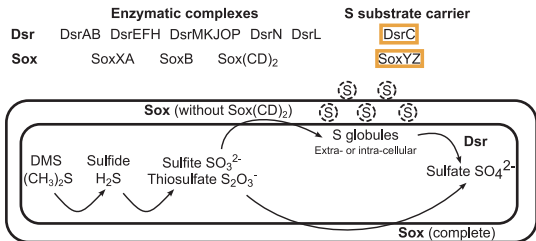
1004 **Extended Data Table 1: Summary of genes and contigs characteristics for new viral DsrC,**
1005 **SoxYZ, and P-II AMGs.** Each gene is linked to its contig, and when available, to the corresponding
1006 viral cluster and predicted host (from BLAST hit, CRISPR spacer similarity, or nucleotide composition
1007 similarity, Alphaprot.-Alphaproteobacteria, Gammaprot.-Gammaproteobacteria). Widespread and
1008 abundant VCs are highlighted in bold. In addition, the calculated pN/pS of each gene is indicated
1009 (measuring the strength of selection pressure occurring for this gene, the gene with a pN/pS not
1010 representing a strong purifying selection is highlighted in red), as well as the coverage of these genes
1011 and other genes in the contigs in 3 metatranscriptomic samples from 3 open ocean Tara stations (cases
1012 where the AMG coverage is >0.5 and associated with the coverage of other genes from the same viral
1013 contig are highlighted in green).

1014

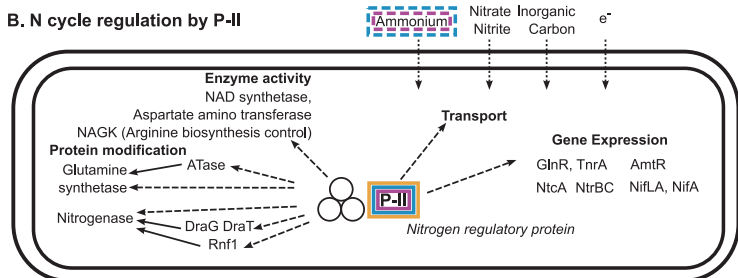
1015



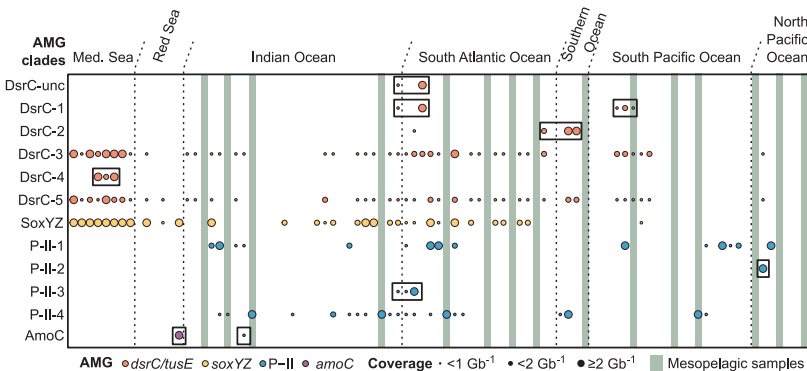
A. Sulfur oxidation pathways



B. N cycle regulation by P-II



C. Sulfur and Nitrogen AMG clades coverage



D. Ranges of ecological conditions for widespread epipelagic AMGs

