

# Ecology of inorganic sulfur auxiliary metabolism in widespread bacteriophages

**Kristopher Kieft**

University of Wisconsin–Madison

**Zhichao Zhou**

University of Wisconsin-Madison

**Rika Anderson**

Carleton College

**Alison Buchan**

University of Tennessee

**Barbara Campbell**

Clemson University <https://orcid.org/0000-0002-3329-6413>

**Steven Hallam**

University of British Columbia <https://orcid.org/0000-0002-4889-6876>

**Matthias Hess**

University of California Davis <https://orcid.org/0000-0003-0321-0380>

**Matthew Sullivan**

The Ohio State University <https://orcid.org/0000-0001-8398-8234>

**David Walsh**

Concordia University

**Simon Roux**

DOE Joint Genome Institute

**Karthik Anantharaman** (✉ [karthik@bact.wisc.edu](mailto:karthik@bact.wisc.edu))

University of Wisconsin-Madison <https://orcid.org/0000-0002-9584-2491>

---

## Article

**Keywords:** Microbial sulfur metabolism, auxiliary metabolic genes (AMGs), bacteriophages, biogeochemistry

**Posted Date:** September 8th, 2020

**DOI:** <https://doi.org/10.21203/rs.3.rs-68034/v1>

**License:**  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

---

**Version of Record:** A version of this preprint was published at Nature Communications on June 9th, 2021.  
See the published version at <https://doi.org/10.1038/s41467-021-23698-5>.

1 **Ecology of inorganic sulfur auxiliary metabolism in widespread bacteriophages**

2  
3  
4 Kristopher Kieft<sup>1#</sup>, Zhichao Zhou<sup>1#</sup>, Rika E. Anderson<sup>2</sup>, Alison Buchan<sup>3</sup>, Barbara J. Campbell<sup>4</sup>, Steven J.  
5 Hallam<sup>5,6,7,8,9</sup>, Matthias Hess<sup>10</sup>, Matthew B. Sullivan<sup>11</sup>, David A. Walsh<sup>12</sup>, Simon Roux<sup>13</sup>, Karthik  
6 Anantharaman<sup>1\*</sup>

7  
8 **Affiliations:**

9 <sup>1</sup> Department of Bacteriology, University of Wisconsin–Madison, Madison, WI, 53706, USA

10 <sup>2</sup> Biology Department, Carleton College, Northfield, Minnesota, USA

11 <sup>3</sup> Department of Microbiology, University of Tennessee, Knoxville, TN, 37996, USA

12 <sup>4</sup> Department of Biological Sciences, Life Science Facility, Clemson University, Clemson, SC, 29634, USA

13 <sup>5</sup> Department of Microbiology & Immunology, University of British Columbia, Vancouver, British  
14 Columbia V6T 1Z3, Canada

15 <sup>6</sup> Graduate Program in Bioinformatics, University of British Columbia, Genome Sciences Centre,  
16 Vancouver, British Columbia V5Z 4S6, Canada

17 <sup>7</sup> Genome Science and Technology Program, University of British Columbia, Vancouver, BC V6T  
18 1Z4, Canada

19 <sup>8</sup> Life Sciences Institute, University of British Columbia, Vancouver, British Columbia, Canada

20 <sup>9</sup> ECOSCOPE Training Program, University of British Columbia, Vancouver, British Columbia,  
21 Canada V6T 1Z3

22 <sup>10</sup> Department of Animal Science, University of California Davis, Davis, CA, 95616, USA

23 <sup>11</sup> Department of Microbiology, The Ohio State University, Columbus, OH, 43210, USA

24 <sup>12</sup> Groupe de recherche interuniversitaire en limnologie, Department of Biology, Concordia University,  
25 Montréal, QC, H4B 1R6, Canada

26 <sup>13</sup> DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA

27  
28 <sup>#</sup>These authors contributed equally

29 <sup>\*</sup>Corresponding author

30 Email: [karthik@bact.wisc.edu](mailto:karthik@bact.wisc.edu)

31 Address: 4550 Microbial Sciences Building, 1550 Linden Dr., Madison, WI, 53706

32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44

45 **ABSTRACT**

46

47 Microbial sulfur metabolism contributes to biogeochemical cycling on global scales. Sulfur metabolizing  
48 microbes are infected by phages that can encode auxiliary metabolic genes (AMGs) to alter sulfur  
49 metabolism within host cells but remain poorly characterized. Here we identified 191 phages derived from  
50 twelve environments that encoded 227 AMGs for oxidation of sulfur and thiosulfate (*dsrA*, *dsrC/tusE*, *soxC*,  
51 *soxD* and *soxYZ*). Evidence for retention of AMGs during niche-differentiation of diverse phage  
52 populations provided evidence that auxiliary metabolism imparts measurable fitness benefits to phages with  
53 ramifications for ecosystem biogeochemistry. Gene abundance and expression profiles of AMGs suggested  
54 significant contributions by phages to sulfur and thiosulfate oxidation in freshwater lakes and oceans, and  
55 a sensitive response to changing sulfur concentrations in hydrothermal environments. Overall, our study  
56 provides novel insights on the distribution, diversity and ecology of phage auxiliary metabolism associated  
57 with sulfur and reinforces the necessity of incorporating viral contributions into biogeochemical  
58 configurations.

59

60 **INTRODUCTION**

61

62 Viruses that infect bacteria (bacteriophages, or phages) are estimated to encode a larger repertoire  
63 of genetic capabilities than their bacterial hosts and are prolific at transferring genes throughout microbial  
64 communities<sup>1-4</sup>. The majority of known phages have evolved compact genomes by minimizing non-coding  
65 regions, reducing the average length of encoded proteins, fusing proteins and retaining few non-essential  
66 genes<sup>5,6</sup>. Despite their reduced genome size and limited coding capacity, phages are known for their ability  
67 to modulate host cells during infection, take over cellular metabolic processes and proliferate through a  
68 bacterial population, typically through lysis of host cells<sup>7,8</sup>. Phage-infected hosts, termed virocells, take on  
69 a distinct physiology compared to an uninfected state<sup>9</sup>. As many as 30-40% of all bacteria are assumed to  
70 be in a virocell state, undergoing phage-directed metabolism<sup>10,11</sup>. This has led to substantial interest in  
71 understanding the mechanisms that provide phages with the ability to redirect nutrients within a host and  
72 ultimately how this manipulation may affect microbiomes and ecosystems.

73

74 One such mechanism by which phages can alter the metabolic state of their host is through the  
75 activity of phage-encoded auxiliary metabolic genes (AMGs)<sup>12,13</sup>. AMGs are typically acquired from the  
76 host cell and can be utilized during infection to augment or redirect specific metabolic processes within the  
77 host cell<sup>14-16</sup>. These augmentations likely function to maintain or drive specific steps of a metabolic pathway  
78 and can provide the phage with sufficient fitness advantages to retain these genes over time<sup>12,17</sup>. Two notable  
79 examples of AMGs are core photosystem II proteins *psbA* and *psbD*, which are commonly encoded by  
80 phages infecting Cyanobacteria in both freshwater and marine environments, and responsible for  
81 supplementing photosystem function in virocells during infection<sup>18-21</sup>. *PsbA* and *PsbD* play important roles  
82 in maintenance of photosynthetic energy production over time within the host; this energy is subsequently  
83 utilized for the production of resources (e.g., nucleotides) for phage propagation<sup>12,14</sup>. Other descriptions of  
84 AMGs include those for sulfur oxidation in the pelagic oceans<sup>16,22</sup>, methane oxidation in freshwater lakes<sup>23</sup>,  
85 ammonia oxidation in surface oceans<sup>24</sup>, carbon utilization (e.g., carbohydrate hydrolysis) in soils<sup>25,26</sup>, and  
86 marine ammonification<sup>27</sup>. Beyond these examples, the combined effect of phage auxiliary metabolism on  
87 ecosystems scales has yet to be fully explored or implemented into conceptualizations of microbial  
community functions and interactions.

88 Dissimilatory sulfur metabolism (DSM) encompasses both reduction (e.g., sulfate to sulfide) and  
89 oxidation (e.g., sulfide or thiosulfate to sulfate) and accounts for the majority of sulfur metabolism on  
90 Earth<sup>28</sup>. Bacteria capable of DSM (termed as sulfur microbes) are phylogenetically diverse, spanning 13  
91 separate phyla, and can be identified throughout a range of natural and human systems, aquatic and  
92 terrestrial biomes, aerobic or anaerobic environments, and in the light or dark<sup>29</sup>. Since DSM is often coupled  
93 with primary production and the turnover of buried organic carbon, understanding these processes is  
94 essential for interpreting the biogeochemical significance of both microbial- and phage-mediated nutrient  
95 and energy transformations<sup>29</sup>. Phages of DSM-mediating microorganisms are not well characterized beyond  
96 the descriptions of phages encoding *dsrA* and *dsrC* genes infecting known sulfur oxidizers from the SUP05  
97 group of Gammaproteobacteria<sup>16,22</sup>, and viruses encoding *dsrC* and *soxYZ* genes associated with  
98 proteobacterial hosts in the epipelagic ocean<sup>30</sup>. Despite the identification of DSM AMGs across multiple  
99 host groups and environments, there remains little context for their global diversity and roles in the  
100 biogeochemical cycling of sulfur. Characterizing the ecology, function and roles of phages associated with  
101 DSM is crucial to an integral understanding of the mechanisms by which sulfur species are transformed  
102 and metabolized.

103 Here we leveraged publicly available metagenomic and metatranscriptomic data to identify phages  
104 capable of manipulating DSM within host cells. We identified 191 phages encoding AMGs for oxidation  
105 and disproportionation of reduced sulfur species, such as elemental sulfur and thiosulfate, in coastal ocean,  
106 pelagic ocean, hydrothermal vent, human, and terrestrial environments. We refer to these phages encoding  
107 AMGs for DSM as *sulfur phages*. These sulfur phages represent different taxonomic clades of  
108 *Caudovirales*, namely from the families *Siphoviridae*, *Myoviridae* and *Podoviridae*, with diverse gene  
109 contents, and evolutionary history. Using paired viral-host gene coverage measurements from  
110 metagenomes recovered from hydrothermal environments, freshwater lakes, and *Tara* Ocean samples, we  
111 provide evidence for the significant contribution of viral AMGs to sulfur and thiosulfate oxidation.  
112 Investigation of metatranscriptomic data suggested that phage-directed sulfur oxidation activities showed  
113 significant increases with the increased substrate supplies in hydrothermal ecosystems, which indicates  
114 rapid and sensitive responses of virocells to altered environmental conditions. Overall, our study provides  
115 novel insights on the distribution, diversity, and ecology of phage-directed dissimilatory sulfur and  
116 thiosulfate metabolisms and reinforces the need to incorporate viral contributions into assessments of  
117 biogeochemical cycling.

## 118 119 **RESULTS**

### 120 121 **Unique sulfur phages encode AMGs for oxidation of elemental sulfur and thiosulfate**

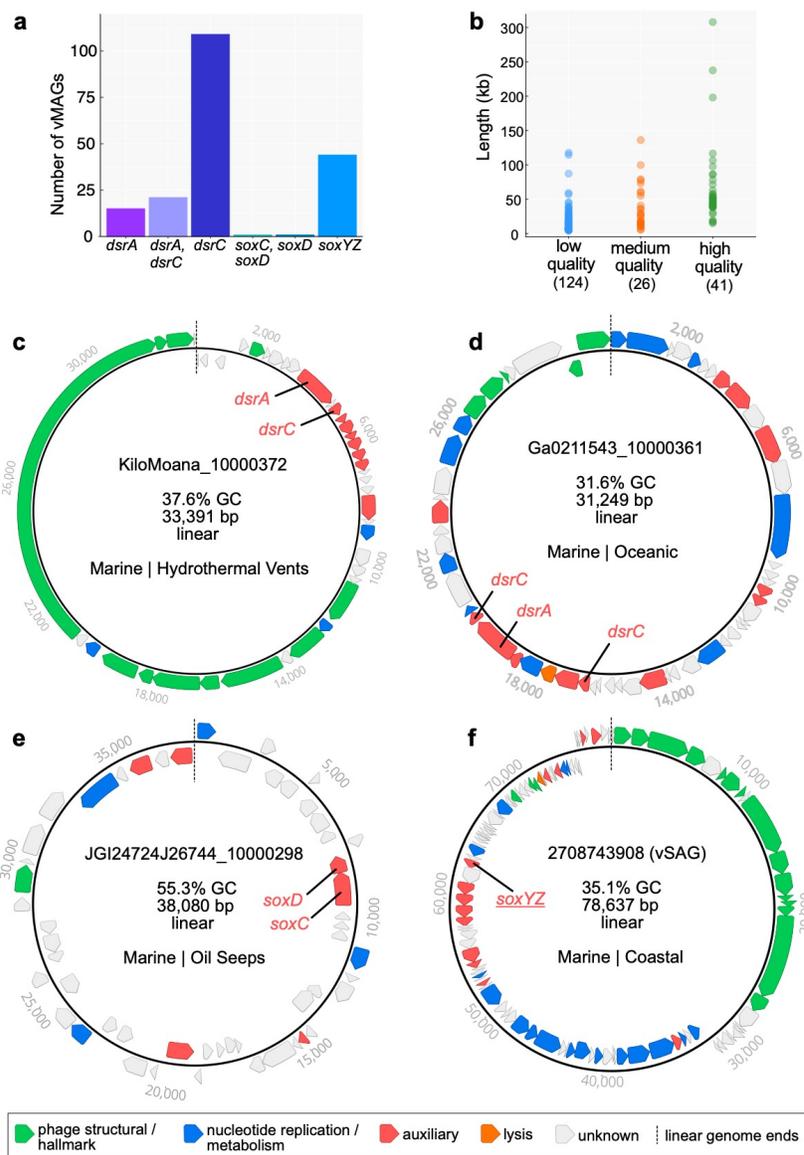
122 We queried the Integrated Microbial Genomes/Viruses (IMG/VR v2.1) database for phages  
123 encoding genes associated with pathways for dissimilatory sulfur oxidation and reduction processes. We  
124 identified 190 viral metagenome-assembled genomes (vMAGs) and one viral single-amplified genome<sup>31</sup>  
125 carrying genes encoding for reverse dissimilatory sulfite reductase subunits A and C (*dsrA* and *dsrC*),  
126 thiouridine synthase subunit E (*tusE*, a homolog of *dsrC*), sulfane dehydrogenase subunits C and D (*soxC*,  
127 *soxD*), and fused sulfur carrier proteins Y and Z for thiosulfate oxidation (*soxYZ*). While phages carrying  
128 *dsrA*, *dsrC/tusE* and *soxYZ* have been previously described in specific marine environments, this is the first  
129 report of *soxC* and *soxD* encoded on viral genomes. Each identified vMAG encoded between one to four  
130 total DSM AMGs for a total of 227 AMGs (Fig. 1a, Supplementary Table 1). The vMAGs ranged in length  
131 from 5 kb to 308 kb, with an average length of approximately 31 kb and a total of 83 sequences greater than

132 20 kb. The vMAGs consisted of 124 low-, 26 medium- and 41 high-quality draft scaffolds according to  
 133 quality estimations based on gene content (Fig. 1b). Only one vMAG was a complete circular genome and  
 134 was identified as previously described<sup>22</sup>. The majority of viruses in this study, with the exception of several  
 135 vMAGs encoding *tusE*-like AMGs were predicted to have an obligate lytic lifestyle on the basis of encoded  
 136 proteins functions.

137 The vMAGs displayed  
 138 unique and diverse genomic  
 139 arrangements, regardless of the  
 140 encoded AMG(s). However, in  
 141 most cases the encoded AMGs  
 142 were found within auxiliary gene  
 143 cassettes, separate from structural  
 144 and nucleotide metabolism  
 145 cassettes (Fig. 1c, d, e, f).  
 146 Auxiliary cassettes in phages  
 147 typically encode genes that are  
 148 not essential for productive  
 149 propagation but can provide  
 150 selective advantages during  
 151 infection, such as in specific  
 152 nutrient limiting conditions or to  
 153 overcome metabolic  
 154 bottlenecks<sup>32</sup>. This genomic  
 155 arrangement suggests that the role  
 156 of DSM AMGs is related to host  
 157 modulation rather than essential  
 158 tasks such as  
 159 transcription/translation, genome  
 160 replication or structural assembly.

### 162 Validation of conserved amino 163 acid residues and domains in 164 AMG proteins

165 Validating AMG protein  
 166 sequences ensures that their  
 167 identification on vMAG genomes  
 168 represents accurate annotations  
 169 (i.e., predicted biological  
 170 function). We used *in silico*  
 171 approaches for protein validation  
 172 by aligning AMG protein  
 173 sequences with biochemically  
 174 validated reference sequences  
 175 from isolate bacteria or phages



**Fig. 1** Dataset summary statistics and representative genome organization diagrams of vMAGs. **a** The number of vMAGs, 191 total, encoding single or multiple DSM AMGs. **b** Estimated vMAG genome qualities as a function of scaffold lengths. vMAGs encoding **c** *dsrA* and *dsrC*, **d** *dsrA* and two *dsrC*, **e** *soxC* and *soxD*, and **f** *soxYZ*. For **c**, **d**, **e** and **f** linear vMAG scaffolds are visualized as circular with the endpoints indicated by dashed lines, and predicted open reading frames are colored according to VIBRANT annotation functions.

176 and assessed the presence or absence of functional domains and conserved amino acid residues. We  
177 highlighted cofactor coordination/active sites, cytochrome c motifs, substrate binding motifs, siroheme  
178 binding sites, cysteine motifs, and other strictly conserved residues (collectively termed *residues*). Finally,  
179 we assessed if phage AMG sequences are under selection pressures to be retained.

180 Conserved residues identified on AMG protein sequences include: DsrA: substrate binding (R,  
181 KxKxK, R, HeR) and siroheme binding (CxgxxxC, CxxdC) (Supplementary Fig. 1); DsrC: strictly  
182 conserved cysteine motifs (Cxxxgxp xxxC) (Supplementary Fig. 2); SoxYZ: substrate binding cysteine  
183 (ggCs) and variable cysteine motif (CC) (Supplementary Fig. 3); SoxC: cofactor coordination/active sites  
184 (XxH, D, R, XxK) (Supplementary Fig. 4); SoxD: cytochrome c motifs (CxxCHG, CMxxC)  
185 (Supplementary Fig. 5). The identification of these residues on the majority of AMG protein sequences  
186 suggests they are as a whole functional. However, there are several instances of AMGs potentially encoding  
187 non-functional or distinctively different genes. For example, only 23 DsrC AMG protein sequences  
188 contained both of the strictly conserved cysteine motifs, 112 contained only the second cysteine motif, 1  
189 contained only the first cysteine motif, and another 5 contained neither. The lack of strictly conserved  
190 cysteine motifs in phage DsrC has been hypothesized to represent AMGs with alternate functions during  
191 infection<sup>16</sup>, but this hypothesis has yet to be validated. Likely, most DsrC AMG protein sequences lacking  
192 one or more cysteine residues functionally serve as TusE, a related sulfur transfer protein for tRNA thiol  
193 modifications<sup>33</sup>. Indeed, several vMAGs originating from the human oral microbiome encode *tusE*-like  
194 AMGs that flank additional *tus* genes (Supplementary Fig. 2 and Supplementary Table 2). Further examples  
195 of missing residues include two vMAGs encoding *soxD* in which one is missing the first cytochrome c  
196 motif, and both are missing the second cytochrome c motif (Supplementary Fig. 5). This initially suggests  
197 the presence of non-functional SoxD, but this notion is contested by the presence of conserved residues in  
198 SoxC. Functional SoxC, encoded adjacent to *soxD* in one of the vMAGs, suggests that both likely retain  
199 function. It has been shown that phage proteins divergent from respective bacterial homologs can retain  
200 their original anticipated activity or provide additional functions<sup>34</sup>. Overall, with the notable exception of  
201 118 *tusE*-like AMGs, *in silico* analyses of AMG protein sequences suggests vMAGs encode functional  
202 metabolic proteins.

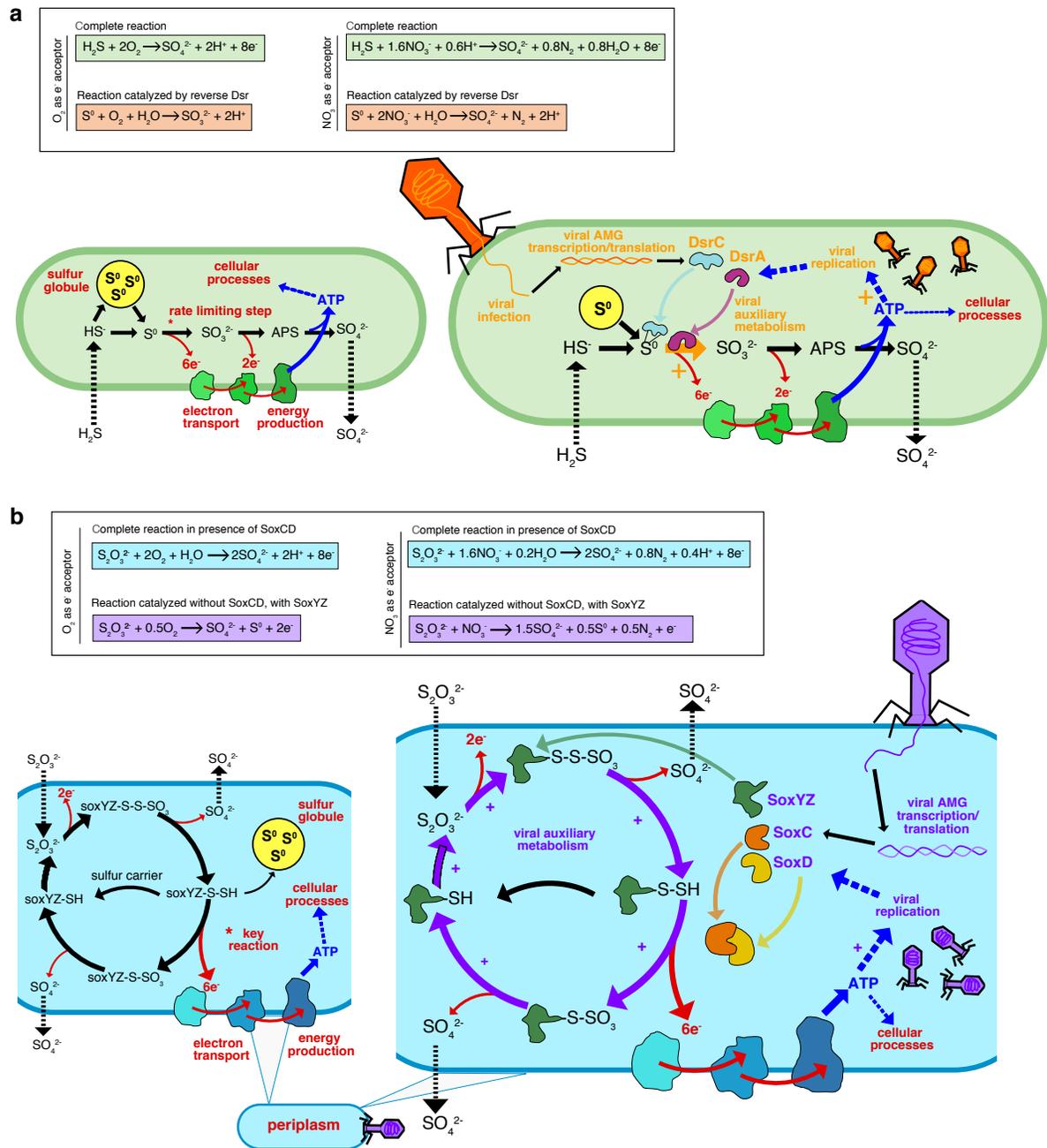
203 To understand selective pressures on AMGs, we calculated the ratio of non-synonymous to  
204 synonymous nucleotide differences ( $dN/dS$ ) in phage AMGs and their bacterial homologs to assess if phage  
205 genes are under purifying selection. A calculated  $dN/dS$  ratio below 1 indicates a gene, or genome as a  
206 whole, is under selective pressures to remove deleterious mutations. Therefore,  $dN/dS$  calculation of vMAG  
207 AMGs resulting in values below 1 would indicate that the viruses selectively retain the AMG. Calculation  
208 of  $dN/dS$  for vMAG *dsrA*, *dsrC* and *soxYZ* AMGs resulted in values below 1, suggesting AMGs are under  
209 purifying selection (Supplementary Fig. 6).

210

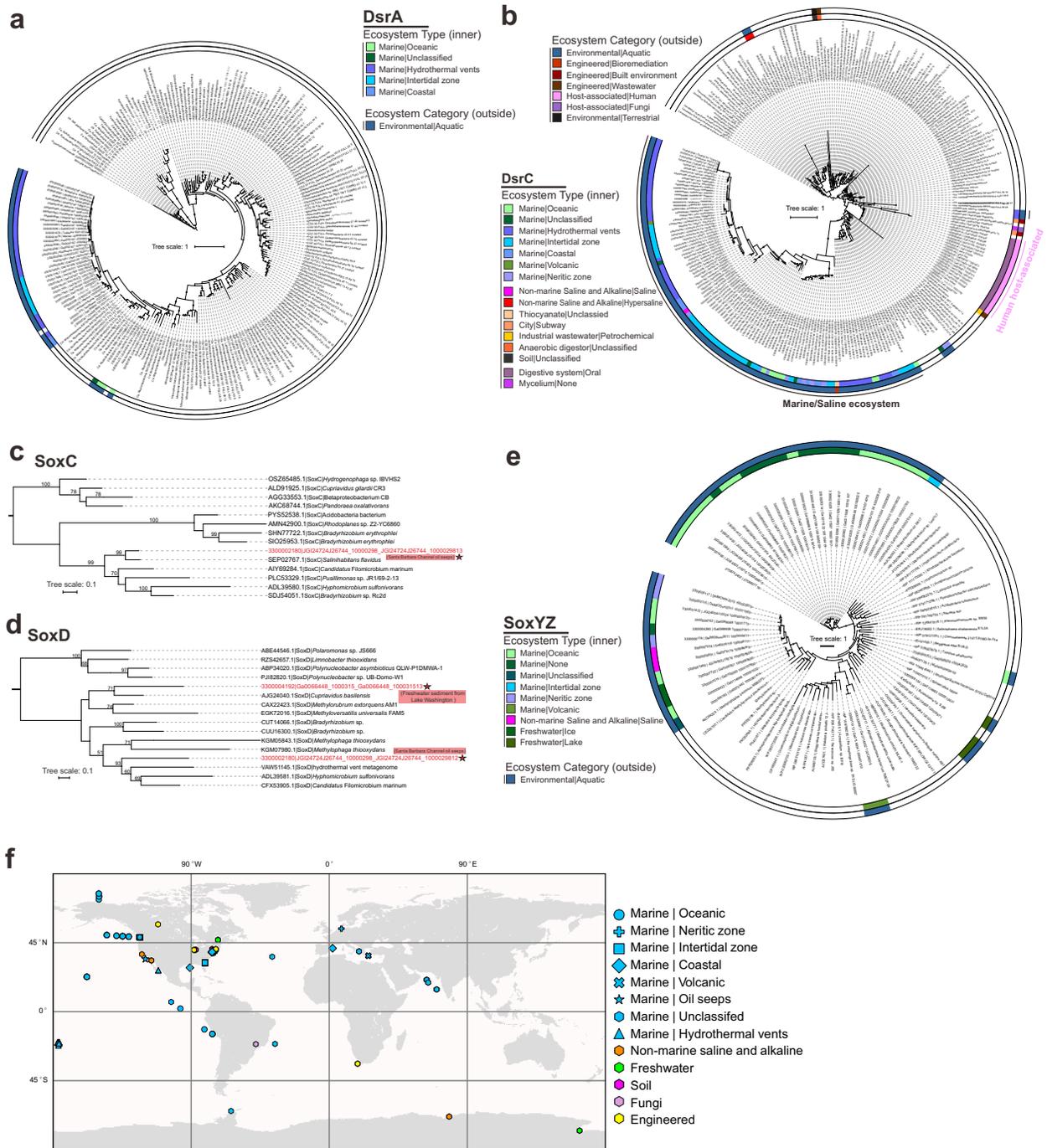
### 211 **DSM AMGs likely manipulate key steps in sulfur oxidation pathways to redistribute energy**

212 As previously stated, DSM AMGs encoded by the vMAGs likely function specifically for the  
213 manipulation of sulfur transformations in the host cell during infection. To better understand the  
214 implications of this manipulation, we constructed conceptual diagrams of both sulfur (i.e., *dsr* AMGs)  
215 oxidation and thiosulfate (i.e., *sox* AMGs) oxidation/disproportionation, with oxygen or nitrate as the  
216 electron acceptor, in both uninfected and infected hosts (Fig. 2).

217 To understand the potential advantages of carrying *dsrC* and *dsrA* AMGs specifically, each step in  
218 the sulfide oxidation pathway needs consideration. During host-only sulfide oxidation<sup>35</sup>, sulfide diffusing  
219 into the cell is converted into elemental sulfur by a sulfide:quinone oxidoreductase (e.g., *sqr*) and in some



**Fig. 2** Conceptual diagrams of viral DsrA, DsrC, SoxC, SoxD and SoxYZ auxiliary metabolism. **a** Microbial dissimilatory oxidation of hydrogen sulfide and stored inorganic sulfur. The resulting production of ATP utilized for cellular processes and growth and the pathway's rate limiting step is indicated with an asterisk (top). Viral infection and manipulation of sulfur oxidation by encoded DsrA or DsrC to augment the pathway's rate limiting step and increase energy yield towards viral replication (bottom). **b** Microbial dissimilatory oxidation of thiosulfate or storage of inorganic sulfur in the periplasm. The resulting production of ATP is utilized for cellular processes and the pathway's key energy yielding reaction indicated with an asterisk (top). Viral infection and manipulation of thiosulfate oxidation by encoded SoxC, SoxD or SoxYZ to augment the entire pathway and the key energy yielding step to increase energy yield towards viral replication (bottom). For **a** and **b** cellular processes are shown in red, sulfur oxidation pathway is shown in black, energy flow is shown in blue, and viral processes are shown in orange (a) or purple (b).



**Fig. 3** Phylogenetic tree of AMG proteins and distribution of phage genomes (on a world map). **a, b** Phylogenetic trees of phage DsrA and DsrC **c, d, e** SoxC, SoxD, SoxYZ. Ultrafast bootstrap (UFBoot) support values (> 50%) are labelled on the nodes. **c, d** Phage gene encoded protein sequences are labeled with stars and their environmental origin information is labeled accordingly. **f** World map showing distribution of phage genomes that contain the sulfur-related AMGs. Studies on human systems are excluded from the map.

221  
222  
223  
224

cases the pathway can begin directly with the import of elemental sulfur. The elemental sulfur can be stored in localized sulfur globules until it is metabolized through the sulfide oxidation pathway<sup>36</sup>. During sulfide oxidation, elemental sulfur carried by the sulfur carrier protein DsrC is oxidized into sulfite by the enzyme

225 complex DsrAB. This step is estimated to be the rate limiting step in the complete pathway and yields the  
226 most electrons (six electrons) for ATP generation. Rate limitation is caused by either the saturation of the  
227 DsrAB enzyme complex or the DsrC carrier<sup>37,38</sup>. The final steps in sulfide/sulfur oxidation involve further  
228 oxidation of sulfite into adenosine 5-phosphosulfate (APS) and then sulfate by an APS reductase (e.g.,  
229 *aprAB*) and sulfate adenylyltransferase, respectively (e.g., *sat*) which yields two electrons<sup>35</sup>. The obtained  
230 ATP can then be utilized for cellular processes. In contrast, during phage infection involving the modulation  
231 of sulfide oxidation, the rate limiting step (i.e., co-activity of DsrC and DsrA) can be supplemented by  
232 phage DsrC and/or DsrA to potentially increase the rate and ATP yield of the reaction as well as utilize any  
233 stored elemental sulfur<sup>22</sup>. This influx of ATP could then be effectively utilized for phage propagation (e.g.,  
234 phage protein production, genome replication or genome encapsidation) (Fig. 2a).

235 Likewise, the normal state of thiosulfate oxidation/disproportionation may be augmented by phages  
236 encoding *soxYZ*, *soxC* and *soxD*. During host-only thiosulfate oxidation<sup>39</sup>, thiosulfate is transported into the  
237 cell where the two thiol groups, transported by SoxYZ, undergo a series of oxidation reactions. A portion  
238 of the carried sulfur, after yielding two electrons, will be transported out of the cell as sulfate. The remaining  
239 carried sulfur may either be stored in elemental sulfur globules or proceed to the key energy yielding step.  
240 The key energy yielding step bypasses the storage of elemental sulfur and utilizes the SoxCD enzyme  
241 complex to produce six electrons for ATP yield<sup>35,40</sup>. During phage infection involving the modulation of  
242 thiosulfate oxidation/disproportionation, the entire pathway can be supported by phage SoxYZ sulfur  
243 carriers in order to continuously drive elemental sulfur storage, which could then be oxidized by the Dsr  
244 complex. However, there is no evidence that phages benefit from coupling the *sox* and *dsr* pathways since  
245 no vMAGs were found to encode both a *sox* and *dsr* AMG simultaneously. Finally, phage SoxCD may be  
246 utilized to drive the pathway to the key energy yielding step. As with the *dsr* pathway, the resulting ATP  
247 would be utilized for phage propagation (Fig. 2b).

248

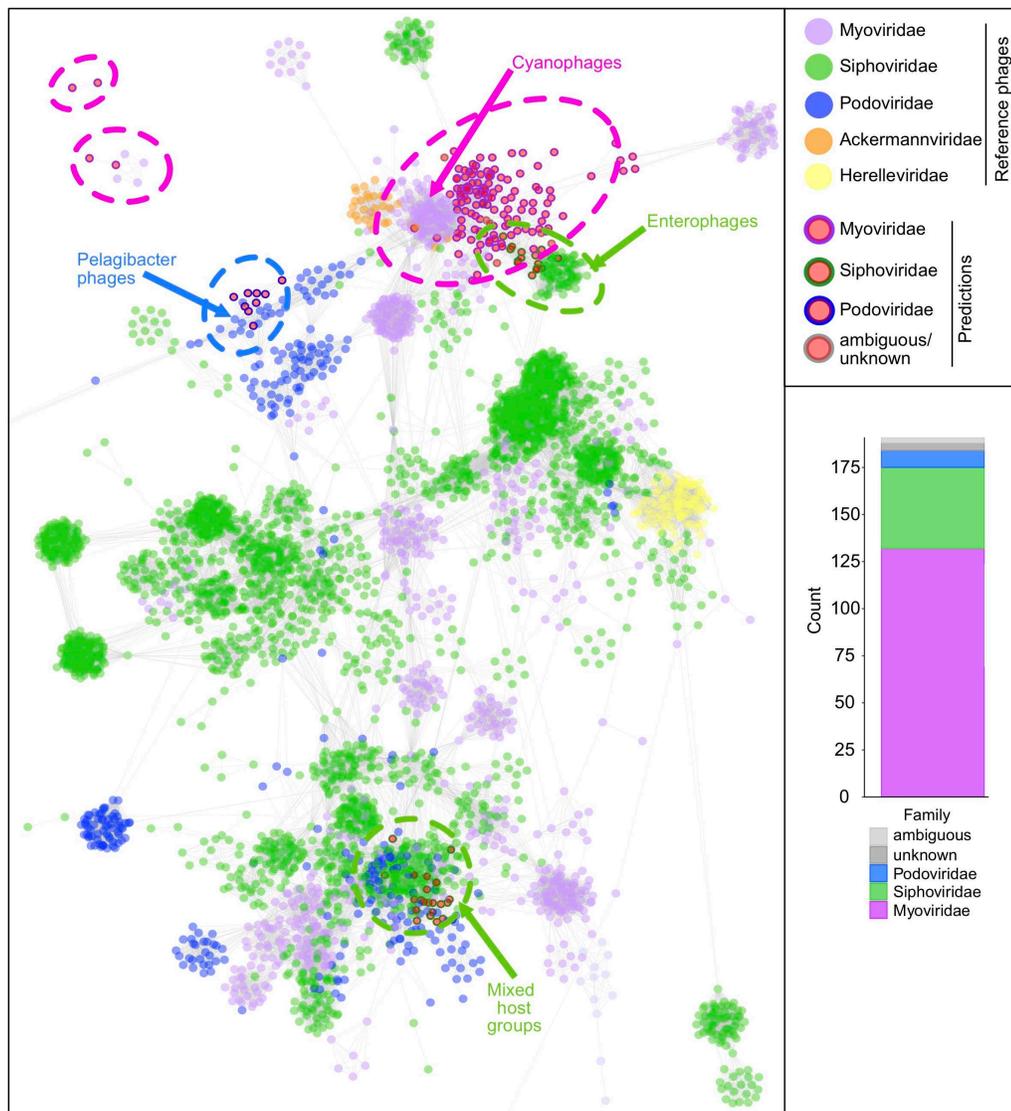
### 249 **Sulfur phages are widely distributed in the environment**

250 Next, we studied the ecological and distribution patterns of vMAGs encoding DSM AMGs. We  
251 characterized their diverse ecology and distribution patterns in various environments by building  
252 phylogenetic trees using the identified AMG and reference microbial proteins, and parsing environmental  
253 information of vMAG metadata from the IMG/VR database. We identified vMAGs encoding *dsrA* mainly  
254 in a few ocean environments, while more widely distributed vMAGs encoding *dsrC* were found in in ocean,  
255 saline, oil seep-associated, terrestrial, engineered, and symbiotic environments (Fig. 3a, b). For *soxC* and  
256 *soxD*, we only identified vMAGs encoding these AMGs in two metagenome datasets, one from Santa  
257 Barbara Channel oil seeps (vMAG encoding both *soxC* and *soxD*) and another from freshwater sediment  
258 from Lake Washington (Fig. 3c, d). The vMAGs encoding *soxYZ* were discovered in aquatic environments,  
259 consisting of different ocean, saline and freshwater ecosystem types (Fig. 3e). In addition to vMAG  
260 distribution amongst diverse ecosystem types we identified wide biogeographic distribution across the  
261 globe (Fig. 3f). Collectively, these DSM AMGs are ecologically and biogeographically ubiquitous, and  
262 potentially assist host functions in many different environment types and nutrient conditions (including  
263 both natural and engineered environments).

264

### 265 **Sulfur phages are taxonomically diverse within the order *Caudovirales***

266 We applied two approaches to taxonomically classify and cluster the identified vMAGs. First, we  
 267 used a reference database similarity search to assign each vMAG to one of 25 different prokaryote-infecting  
 268 viral families (see Methods). The majority of vMAGs were assigned to *Myoviridae* (132 vMAGs; 69%),  
 269 *Siphoviridae* (43 vMAGs; 22%) and *Podoviridae* (9 vMAGs; 5%). These three families represent dsDNA  
 270 phages belonging to the order *Caudovirales*. The remaining seven vMAGs were identified as ambiguous  
 271 *Caudovirales* (3 vMAGs; 1.5%) and unknown at both the order and family levels (4 vMAGs; 2%).  
 272 However, based on the data presented here and previous classifications<sup>16,22,30</sup>, the seven unclassified  
 273 vMAGs likely belong to one of the three major *Caudovirales* families (Fig. 4).



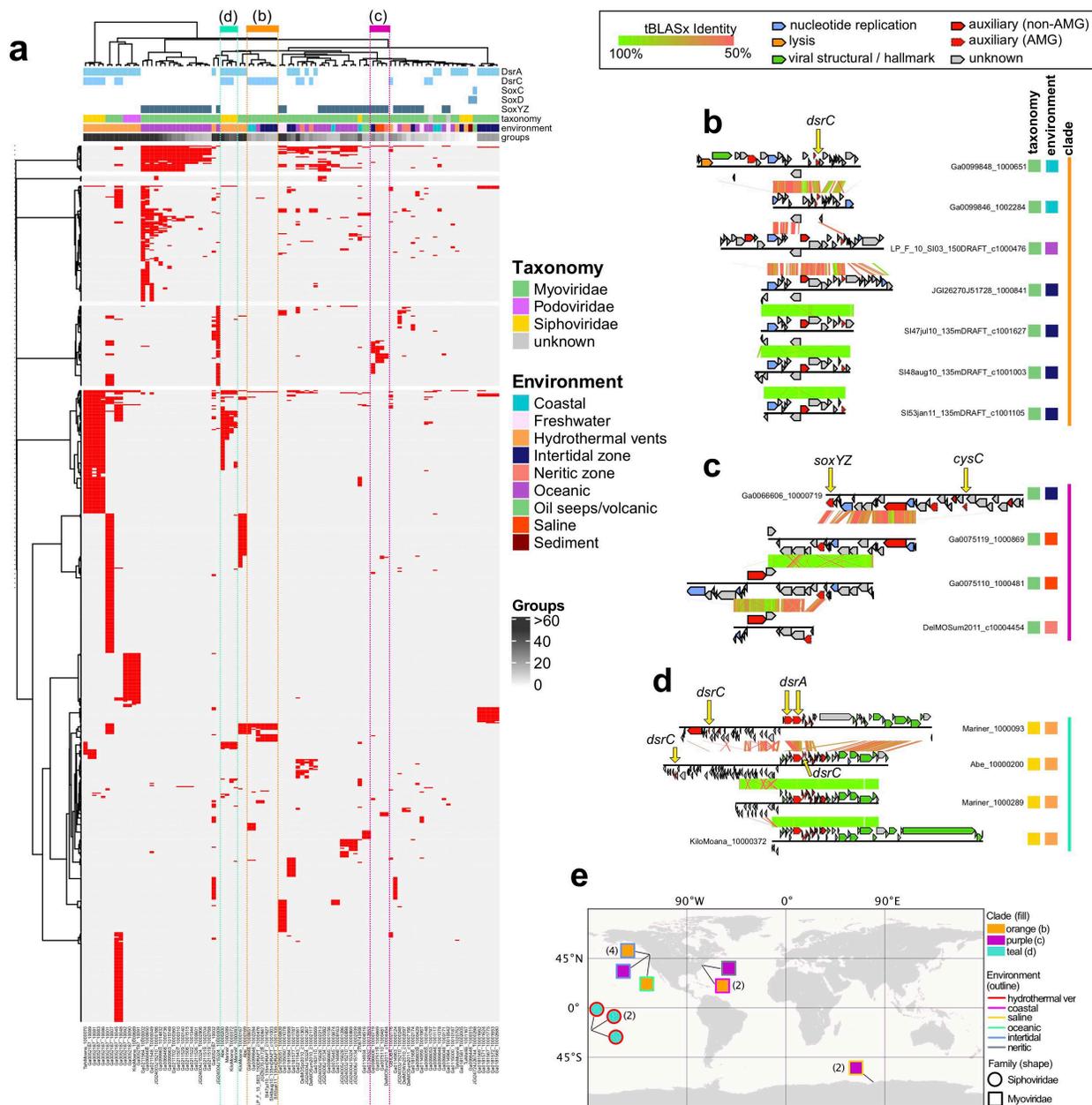
**Fig. 4** Taxonomic assignment of vMAGs and protein network clustering with reference phages. In the protein network each dot represents a single vMAG (circles with outlines) or reference phage (circles without outlines), and dots are connected by lines respective to shared protein content. Genomes (i.e., dots) having more similarities will be visualized by closer proximity and more connections. Cluster annotations depicted by dotted lines were approximated manually. vMAG taxonomy was colored according to predictions by a custom reference database and script, shown by bar chart insert.

274 In accordance with these results we constructed a protein sharing network of the vMAGs with  
275 reference viruses from the NCBI GenBank database (Fig. 4). The vMAGs arranged into four main clusters  
276 with reference *Myoviridae*, *Siphoviridae* and *Podoviridae*, and four individual vMAGs were arranged  
277 outside of main clusters. Of the seven vMAGs with ambiguous/unknown predictions, six clustered with  
278 *Myoviridae* and *Siphoviridae* vMAGs and reference phages, further suggesting their affiliation with major  
279 *Caudovirales* families. On the basis of these findings, we hypothesize that the function(s) of DSM AMG  
280 during infection is most likely constrained by specific host sulfur metabolisms rather than viral taxonomy.  
281 The broad distribution of DSM AMG across *Caudovirales* further suggests that this modulatory  
282 mechanism is established across multiple taxonomic clades of phages, either arising independently or  
283 acquired via gene transfer. Most vMAGs clustered with reference phages that infect *Pelagibacter*,  
284 Cyanobacteria and Enterobacteria, with one cluster represented by a mixed group of host ranges. However,  
285 it is likely that host range stems beyond these indicated taxa, suggested by the inclusion of a SUP05-  
286 infecting vMAG<sup>22</sup> within the *Pelagibacter* cluster. In the present state of the reference databases, this type  
287 of protein sharing network cannot be used to reliably predict the host range of these uncultivated vMAGs.  
288

### 289 **Sulfur phages display diversification across environments and genetic mosaicism**

290 To further assess the diversity of the identified vMAGs and their evolutionary history, we analyzed  
291 shared protein groups as well as gene arrangements between individual vMAGs. All predicted proteins  
292 from 94 of the vMAGs, excluding vMAGs encoding only *tusE*-like AMGs, were clustered into protein  
293 groups (see Methods). A total of 887 protein groups representing 3677 proteins were generated, roughly  
294 corresponding to individual protein families. Only a few protein groups were globally shared amongst the  
295 vMAGs, including common phage proteins (e.g., *phoH*, *nifU*, *iscA*, nucleases, helicases, lysins, RNA/DNA  
296 polymerase subunits, ssDNA binding proteins and morphology-specific structural proteins) (Fig. 5a). This  
297 result is consistent with that of taxonomic clustering, further highlighting the diversity of phage genomes  
298 that encode DSM AMGs. A lack of universally shared protein groups likewise suggests the DSM AMGs  
299 function independently of other host metabolic pathways and likely strictly serve to supplement host DSM  
300 pathways.

301 Most vMAGs that formed clades according to shared protein groups could be explained by shared  
302 taxonomy and/or source environment. For example, 16 *Myoviridae* vMAGs encoding *soxYZ* from oceanic  
303 environments clustered together, only differing according to their total number of representative protein  
304 groups (Fig. 5a). There were exceptions, such as seven *dsrC*-encoding vMAGs which displayed variable  
305 pairwise protein similarity (at a 50% identity cutoff) and variation in the location of their *dsrC* gene within  
306 their genome, despite a clearly shared and distinctive synteny of other genes (Fig. 5b). The seven vMAGs  
307 originated from three different marine environment types (coastal, oceanic and intertidal) and were all  
308 predicted to be myoviruses (Fig. 5b). This diversity is likely explained by the retention of the *dsrC* gene  
309 over time despite components of the genome undergoing genetic exchange, recombination events or  
310 mutation accumulation. Phages are well known to display genetic mosaicism, or the exchange and  
311 diversification of genes and gene regions<sup>32,41</sup>. The same conclusion can be made with myoviruses encoding  
312 *soxYZ* from different marine environments (intertidal, saline and neritic) (Fig. 5c) as well as siphoviruses  
313 encoding both *dsrC* and *dsrA* from hydrothermal environments (Fig. 5d). In addition to distribution amongst  
314 diverse environmental categories these genetically mosaic vMAGs, per protein sharing clade, are  
315 geographically dispersed (Fig. 5e). Additionally, one vMAG (Ga0066606\_10000719) encoding *soxYZ* also  
316 encodes the assimilatory sulfur metabolism AMG *cysC* (Fig. 5b). This presents an interesting discontinuity  
317 suggesting that this particular vMAG, as well as three others encoding *cysC* (Ga0052187\_10001,



**Fig. 5** vMAG protein grouping and genome alignments. **a** vMAG hierarchical protein grouping where each row represents a single protein group (887 total) and each column represents a single vMAG (94 total). Metadata for encoded AMGs, estimated taxonomy, source environment and number of protein groups per vMAG is shown. Clades respective of **b**, **c** and **d** are depicted by colored dotted lines. Genome alignments of **b** seven divergent Myoviridae vMAGs encoding *dsrC* from diverse environments, **c** four divergent Myoviridae vMAGs encoding *soxYZ* from diverse environments, and **d** four divergent Siphoviridae vMAGs encoding *dsrA* and *dsrC* from hydrothermal environments. For the genome alignments, each black line represents a single genome and arrows represent predicted proteins which are colored according to VIBRANT annotations; genomes are connected by lines representing tBLASTx similarity. **e** Map of geographic distribution of 15 vMAGs depicted in **b**, **c** and **d**, annotated with respective clade, source environment and taxonomic family.

319 Ga0052187\_10007 and JGI24004J15324\_10000009), target both dissimilatory and assimilatory sulfur  
320 metabolism simultaneously to more generally affect sulfur metabolism in the host.

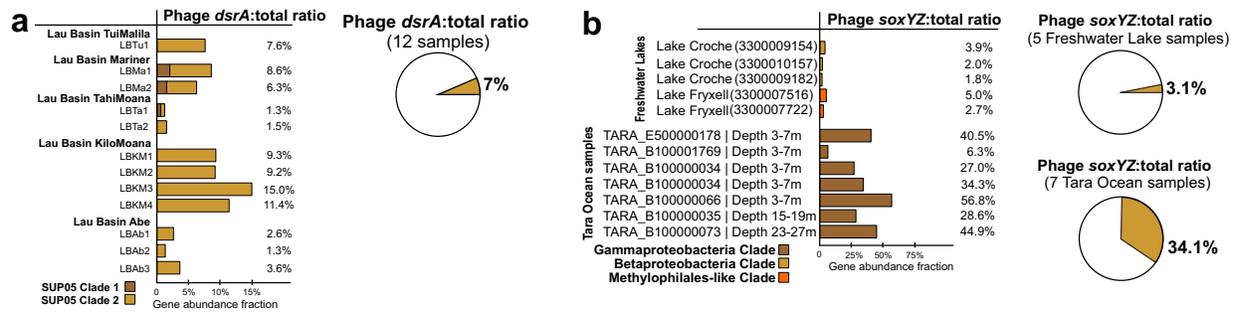
321

### 322 **Estimates of sulfur phage contributions to sulfur oxidation**

323 We utilized metagenomic datasets containing the vMAGs to calculate the ratio of phage:total genes  
324 for each AMG. The phage:total gene ratios within a community and for each predicted phage-host pair can  
325 be used to estimate phage contributions to sulfur and thiosulfate oxidation/disproportionation. By mapping  
326 metagenomic reads to AMGs and putative bacterial hosts within the metagenome, we obtained the vMAG  
327 AMG to total gene ratios, which represents the relative contribution of AMG functions to the representative  
328 metabolism such as sulfur oxidation (Supplementary Tables 3, 4, Supplementary Fig. 7). We calculated  
329 vMAG *dsrA* (Fig. 6a) and *soxYZ* (Fig. 6b) gene coverage ratios in hydrothermal, freshwater lake, and *Tara*  
330 Ocean metagenomic datasets. We identified phage-host gene pairs which contained vMAG AMGs and their  
331 corresponding host genes from the phylogenetic tree of DsrA and SoxYZ (Supplementary Figs. 8, 9). Our  
332 results show that phage *dsrA* contributions in hydrothermal environments arise primarily from the SUP05  
333 Clade 2; and those of phage *soxYZ* are niche-specific, with Lake Croche, Lake Fryxell, and *Tara* Ocean  
334 samples mainly represented by the Betaproteobacteria Clade, Methylophilales-like Clade, and  
335 Gammaproteobacteria Clade, respectively. This indicates the specificity of specific groups of AMGs being  
336 distributed and potentially functioning in each environment. The average phage:total gene coverage ratios  
337 also differ in individual groups, with phage *soxYZ*:total ratio in *Tara* Ocean samples being the highest  
338 (34%), followed by phage *dsrA*:total ratio in hydrothermal samples (7%) and phage *soxYZ*:total ratio in  
339 freshwater lakes (3%). Phage *soxYZ*, the sulfur carrier gene, in the oceans have higher phage:total gene  
340 coverage ratio compared to *dsrA*, a component of the catalytic core of *dsr* complex, in the other two  
341 environments. Along with observations associated with phage *dsrC*, our results suggest that AMGs  
342 encoding sulfur carriers rather than catalytic subunits appear to be more favored by phages. While the  
343 limited environment types and sulfur AMGs studied here do not provide sufficient statistical confidence to  
344 generalize these results, nevertheless, higher abundance of sulfur carrier genes could still be a common  
345 phenomenon in virocells. Additionally, notably although gene abundance ratios do not necessarily represent  
346 function contributions, this scenario still provides a reasonable estimation, suggesting considerable sulfur-  
347 oxidizing contributions of phage sulfur AMGs in corresponding virocells.

348 Subsequently, the phage:host AMG coverage ratios for individual phage-host pairs were calculated  
349 to estimate the potential functional contribution within each environmental sample (Figs. 7a, b,  
350 Supplementary Tables 3, 4, Supplementary Figs. 10, 11). By taking average ratios of groups of *dsrA* phage-  
351 host pairs in SUP05 Clade 1 and SUP05 Clade 2, and *soxYZ* phage-host pair in freshwater lake and *Tara*  
352 Ocean samples, we found that within each pair the phage:total gene coverage ratios were generally higher  
353 than ~50%. These within-pair phage:total gene coverage ratios are much higher than the above phage:total  
354 ratios in the whole community. *Tara* Ocean samples also have the highest average phage:total gene  
355 coverage ratios of individual phage-host pairs among these three environments, as with the pattern of ratios  
356 in the whole community.

357 The above analyses suggest that DSM AMGs likely contribute significantly to function of host-  
358 driven metabolisms on the scale of both community level and individual phage-host pairs, while the ratio  
359 of contribution varies greatly for each environment and each niche-specific AMG. Importantly, phage-  
360 encoded *soxYZ* have a high gene coverage contribution to pelagic ocean microbial communities, which  
361 highlights the functional significance of phage-driven sulfur cycling metabolisms, and that of thiosulfate  
362 oxidation/disproportionation as a whole in this environment, which remains critically under-studied<sup>16,42</sup>.



**Fig. 6** Phage to total *dsrA* and *soxYZ* gene coverage ratios. **a** Viral *dsrA* to total (viral and bacterial *dsrA* gene together) gene coverage ratios. The contribution of viral *dsrA* genes from different SUP05 Gammaproteobacteria clades is shown in different colors. The average viral *dsrA*:total ratio was calculated from 12 samples. **b** Viral *soxYZ* to total gene coverage ratios. The contribution of viral *soxYZ* genes from three different clades is shown in different colors. Genes from Freshwater Lake and *Tara* Ocean samples were compared separately, and the average viral *soxYZ*:total ratios were calculated and compared separately as for Freshwater Lake and *Tara* Ocean samples.

363

### 364 Rapid alteration of sulfur phage *dsrA* activity across geochemical gradients

365

366

367

368

369

370

371

372

373

374

375

376

377

378

379

380

381

382

383

384

385

386

387

388

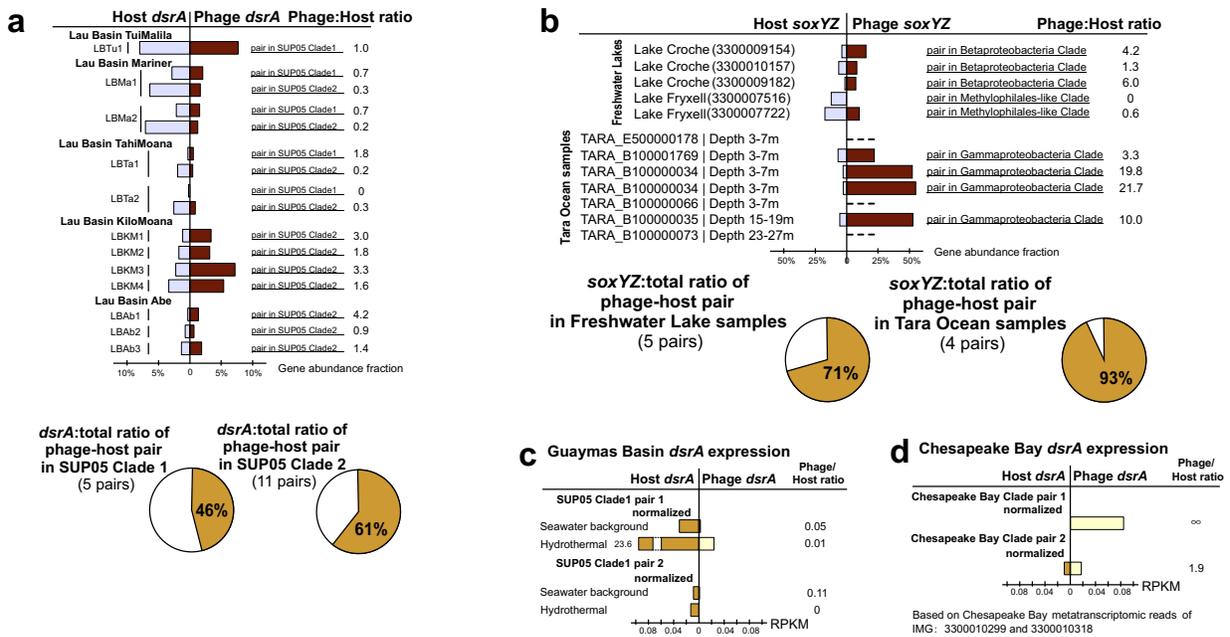
389

390

Since DSM AMGs are associated with critical energy generating metabolism in microorganisms, we wanted to study the ability of sulfur phages to respond to changing geochemistry, involving virocell-driven biogeochemical cycling. In hydrothermal ecosystems, reduced chemical substrates such as  $H_2S$ ,  $S^0$ ,  $CH_4$ , and  $H_2$  display sharp chemical gradients as they are released from high-temperature vents and dilute rapidly upon mixing with cold seawater. Microorganisms in deep-sea environments respond to such elevated concentrations of reduced sulfur compounds by upregulating their metabolic activity in hydrothermal environments<sup>43,44</sup>. These characteristics make hydrothermal and background deep-sea environments a contrasting pair of ecological niches to investigate alteration of AMG expression. We used transcriptomic profiling to study gene expression in phage:host pairs recovered from hydrothermal vents in Guaymas Basin and background deep-sea samples in the Gulf of California (Supplementary Table 3, Supplementary Figs. 10, 11). Sulfur phage *dsrA* expression measured in reads per kilobase of transcript (RPKM) varied from 0.03-3 in the background deep-sea to 0.40-39 in hydrothermal environments (Supplementary Table 3d). Average phage *dsrA* expression ratio of hydrothermal to background was 15 (Supplementary Table 3d). Limited by coding gene repertoire and their biology, phages themselves do not have the ability to independently sense and react to sulfur compounds. However, our results suggest that sulfur phage activities, occurring within a virocell, are closely coupled to changing geochemistry with higher observed activity in environments with greater concentration of reduced sulfur compounds.

Although phage *dsrA* occupies considerable portions of total *dsrA* gene abundance in hydrothermal environments, freshwater lake, and *Tara* Ocean environments (46-71%), their expression levels vary across different environments. In Guaymas Basin hydrothermal environments, as reflected by two pairs of SUP05 Clade 1 phage and host *dsrA* genes, phage to host *dsrA* gene ratios varied from 0 to 0.11 (Fig. 7c). In contrast, in Chesapeake Bay, as reflected by two pairs of phage and host *dsrA* genes (Chesapeake Bay *dsrA* clade), phage to host *dsrA* gene ratios varied from 1.9 to infinity. The low abundance of phage *dsrA* in hydrothermal metatranscriptomes is in sharp contrast to the high abundance of phage *dsrA* in hydrothermal metagenomes (observed at Guaymas Basin and Lau Basin) (Fig. 7a, c). One explanation for this observation is that this scenario could be an accident but not representing real phage gene expression patterns in

391 hydrothermal systems, possibly occurring in a situation when phage activity was very high just prior to  
 392 sampling. In this scenario, the majority of hosts/virocells might have lysed post viral infection.  
 393



**Fig. 7** Phage to host *dsrA* and *SoxYZ* gene coverage ratios and *dsrA* gene expression comparison between phage and host pairs. **a** Phage *dsrA* to total gene coverage ratios of each phage-host pair. Average phage *dsrA*:total ratios of phage-host pairs in SUP05 Clade 1 and Clade 2 were calculated by 5 and 11 pairs of genes, respectively. **b** Phage *soxYZ* to total gene coverage ratios of each phage-host pair. The contribution of phage *soxYZ* genes from three different clades is shown in different colors. Average phage *dsrA*:total ratios of phage-host pairs in Freshwater Lakes and Tara Ocean were calculated separately. **c** Phage to host *dsrA* gene expression comparison in Guaymas Basin metatranscriptomes. The same database was used for mapping both hydrothermal and background metatranscriptomic datasets **d** Phage to host *dsrA* gene expression comparison in Chesapeake Bay metatranscriptomes. The same database was used for mapping all Chesapeake Bay metatranscriptomic datasets. Gene expression levels are shown in RPKM normalized by gene sequence depth and gene length.

394  
 395

## DISCUSSION

396 Since the first descriptions of viral metabolic reprogramming using AMGs<sup>13</sup> there has been interest  
 397 in the extent and overall impact of viral auxiliary metabolism on global energy flows and ecosystem nutrient  
 398 availability<sup>45</sup>. Through metagenomic surveys and investigation, we have expanded the current  
 399 understanding of viral auxiliary metabolism impacting dissimilatory sulfur oxidation processes.  
 400 Specifically, we have shown that diverse lineages of phages are involved in these processes, we have  
 401 investigated their biogeography, ecology, and evolutionary history, and we estimated their potential effects  
 402 on microbiomes. From this, several hypotheses and new questions regarding viral auxiliary metabolism and  
 403 sulfur cycling can be addressed.

404 First, our findings support previous hypotheses that viral metabolism targets key or bottleneck steps  
 405 in host metabolic pathways. *DsrA*, *DsrC*, *SoxYZ*, *SoxC*, and *SoxD* all alleviate bottlenecks in sulfur and  
 406 thiosulfate oxidation/disproportionation<sup>22,46</sup>. We did not identify other genes in sulfur oxidation pathways  
 407 such as sulfide:quinone oxidoreductase, flavocytochrome *c* cytochrome/flavoprotein subunits, APS  
 408 reductase subunits, sulfate adenylyltransferase, *dsrB*, or *soxAB* for other necessary steps of sulfur oxidation.  
 409

410 However, this poses the additional question of why DsrB, the dimer pair to DsrA, has yet to be identified  
411 as an AMG. Likely, encoding *dsrA* provides a significantly greater fitness advantage to phages in  
412 comparison to *dsrB*. Furthermore, sulfur carriers, rather than enzymes, appear to be more favored by  
413 phages. In total, 174 vMAGs in this study encoded at least one sulfur carrier (*dsrC*, *tusE*-like, *soxYZ*) with  
414 only the remaining 17 encoding catalytic subunits of enzymes (*dsrA*, *soxC*, *soxD*). Phage sulfur carriers like  
415 *soxYZ* were observed to be more abundant in whole community and that catalytic subunits such as *dsrA*.  
416 This may be due to the greater need for sulfur carriers (e.g., *dsrC*) to drive dissimilatory sulfur  
417 transformations. Evidence for this hypothesis is provide by observations that sulfur carriers are often  
418 constitutively expressed in host cells in comparison to respective catalytic components (e.g., *dsrA*)<sup>38,47</sup>. By  
419 providing transcripts and proteins of these important pathway components during infection, phages  
420 encoding DSM AMGs may benefit more from obtaining greater energy and self-catalyzing substrates  
421 within a virocell.

422 The data presented by vMAG protein clustering and genome alignments (Fig. 5) supports the  
423 hypothesis that the DSM AMGs are retained on fast evolving phage genomes, pointing specifically to a  
424 role of the AMG in increasing phage replication abilities and fitness. Although the mechanism of dispersion  
425 is unknown for most of the vMAGs it is likely that a single AMG transfer event occurred within each clade  
426 based on retention of similar gene arrangements at AMG locations in the respective genomes. This suggests  
427 that the AMG were retained despite niche (i.e., geographic and environmental) differentiation of individual  
428 vMAG populations. It has been postulated that AMGs, like other phage genes, must provide a significant  
429 fitness advantage in order to be retained over time on an evolving phage genome<sup>12</sup>.

430 Taken together, these observations support the conclusion that viral auxiliary metabolism targets  
431 key steps in host metabolic pathways for finely tuned manipulation of energy production or nutrient  
432 acquisition. Although the fitness effects of DSM AMGs have not been quantified, the geographical  
433 distribution of identified vMAGs and retention of AMGs by phages despite constrained coding capacity  
434 strongly suggests a significant fitness benefit of encoding DSM AMGs. The exact fitness benefit achieved  
435 from encoding DSM AMGs remains elusive without cultured representatives of phage-host pairs. Since  
436 DSM AMGs have been identified on phages from all three major *Caudovirales* families it is likely that the  
437 fitness benefits deal specifically with sulfur oxidation and electron yield from bolstering the speed or  
438 efficiency of the pathway. It is most likely that the phages benefit primarily in the short term and during  
439 active lytic infection due to the abundance of DSM AMGs on lytic phage genomes. Yet, the presence of  
440 assimilatory sulfate reduction genes (i.e., *cysC*) in conjunction with DSM genes provides an example of a  
441 possible exception with a more general sulfur manipulation, highlighting the necessity of further  
442 investigations into viral auxiliary metabolism.

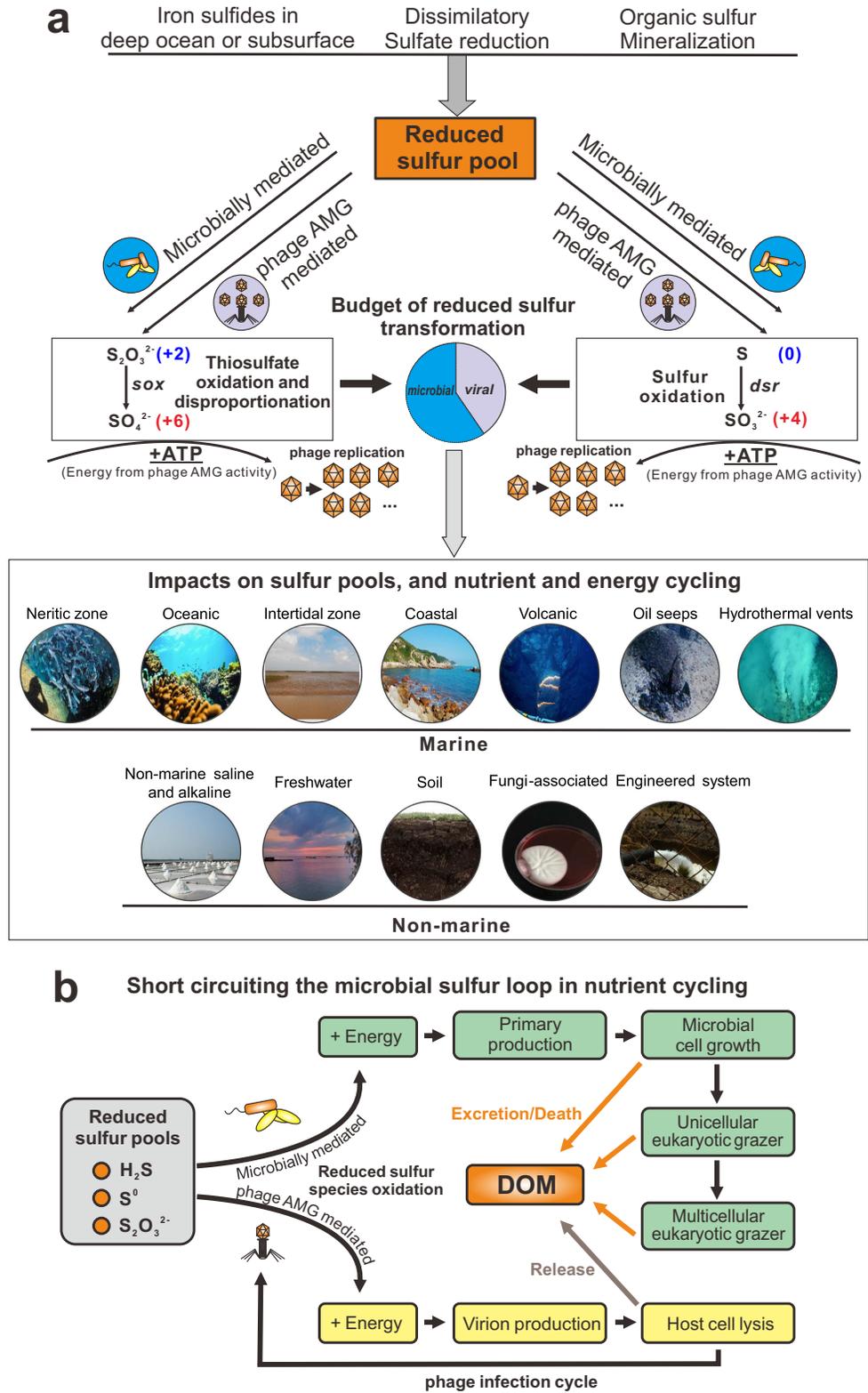
443 The abundance of phage DSM AMGs in metagenomes and metatranscriptomes as measured by  
444 phage:total gene coverage ratios suggest that phage-mediated reduced sulfur transformations can contribute  
445 significantly to fluxes and budgets of sulfur within the community (Fig. 8, Supplementary Figs. 7, 12).  
446 Within each phage-host pair, phage genes contribute to over half of gene coverage associated with the sulfur  
447 and thiosulfate oxidation pathways, which highlights the underappreciated role of phages encoding DSM  
448 AMGs in remodeling sulfur cycling, especially for the oxidation of reduced sulfur. Reduced sulfur  
449 compounds such as H<sub>2</sub>S, S<sup>0</sup>, and S<sub>2</sub>O<sub>3</sub><sup>2-</sup> are abundant in hydrothermal systems with hydrothermal fluids at  
450 Guaymas Basin containing aqueous H<sub>2</sub>S concentrations of up to ~6 mmol/kg (endmember measurement),  
451 while that of background seawater is negligible<sup>43,48</sup>. Previously reported estimates of energy budgets for  
452 sulfur oxidizing bacteria in the Guaymas Basin hydrothermal system suggest that up to 3400 J/kg is  
453 available for microbial metabolism, of which up to 83% may derive from sulfur oxidation<sup>43</sup>. Sulfur phage

454 *dsrA* expression levels (arising from virocells) were elevated in hydrothermal systems in comparison to the  
455 background deep-sea, hinting at significant contributions of virocells mediating phage-driven sulfur  
456 oxidation to overall energy budgets by. Conservatively assuming that 40% of all sulfur-oxidizing SUP05  
457 Gammaproteobacteria are infected by sulfur phages (in line with observations of phage infections in the  
458 pelagic oceans), it may be estimated that 1129 J/Kg of energy for microbial metabolism representing 1/3 of  
459 all energy available from hydrothermal vent fluids may in fact be transformed by virocells containing sulfur  
460 phages. Phages are thus an integral component of the sulfur biogeochemical cycle with the ability to  
461 manipulate microbial metabolism associated with multiple reduced sulfur compounds which can impact  
462 sulfur budgets at ecosystem scales. It is therefore essential that future assessments of biogeochemical  
463 cycling incorporate the role of phages and their impacts on sulfur pools. Limited by the resolution of omics-  
464 based approach in this study, finer scale phage-host interactions and activities could not be achieved, which  
465 justifies the necessity to reinforce fine-scale phage AMG activity research within host cells in future.

466 Across diverse environments on the Earth, the reduced sulfur pool includes sources of deep ocean  
467 or subsurface deposited iron sulfides, and reduced sulfur species from dissimilatory sulfate reduction and  
468 organic sulfur mineralization (Fig. 8a). Sulfur phage AMG-assisted metabolism contributes to the  
469 redistribution of sulfur-generated energy and can alter its budgets, which have so far only been attributed  
470 to microbial processes (Fig. 8a). Within virocells, phage mediated sulfur oxidation will take advantage of  
471 gene components of sulfur-metabolizing pathways, express transcripts, and produce enzymes to re-direct  
472 energy for the use of phage replication (Fig. 8a). Globally distributed sulfur phages are widely distributed  
473 across various environments and impose significant impacts on the sulfur pools, and nutrient and energy  
474 cycling (Fig. 8a). At the same time, phage AMG mediated sulfur oxidation can short-circuit the microbial  
475 sulfur loop from reduced sulfur pools to dissolved organic matter (DOM) (Fig. 8b). Without viral infection,  
476 energy generated by reduced sulfur pools would typically be used for primary production to fuel microbial  
477 cell growth, and then transferred higher up the food chain to grazers. Through cell excretion effects, cell  
478 death and nutrient release, DOM produced from sulfur-based primary production would be released to the  
479 environment. However, during infection by sulfur phages, energy generated in virocells by reduced sulfur  
480 pools could be used towards phage reproduction and propagation. After virion production and packaging,  
481 lytic phages would lyse the host cell, and release DOMs into the environment. This DSM AMG mediated  
482 approach thereby short-circuits the microbial sulfur loop.

483 In conclusion, we have described the distribution, diversity and ecology of phage auxiliary  
484 metabolism associated with sulfur and demonstrated the abundance and activity of sulfur phages in the  
485 environment, yet many questions remain unanswered. Future research will involve unraveling mechanisms  
486 of sulfur phage and host interaction, remodeling of sulfur metabolism at the scale of individual virocells,  
487 microbial communities and ecosystems, and constraining sulfur budgets impacted by sulfur phages.

488



**Fig. 8** Conceptual figure indicating the ecology and function of AMGs in sulfur metabolisms. **a** DSM AMG effect on the budget of reduced sulfur transformation. **b** Diagram of virus-mediated metabolism short circuiting the microbial sulfur loop in nutrient cycling.

490

## 491 MATERIALS AND METHODS

492

### 493 vMAG acquisition and validation

494 The Integrated Microbial Genomes and Virome (IMG/VR) database<sup>49,50</sup> was queried for *sox* and  
495 *dsr* gene annotations (v2.1, October 2018). A total of 192 unique vMAGs greater than 5kb in length were  
496 identified. For consistency between these vMAGs, open reading frames were predicted using Prodigal (-p  
497 meta, v2.6.3)<sup>51</sup>. Each of the 192 vMAGs were validated as phage using VIBRANT<sup>52</sup> (v1.2.1, virome mode),  
498 VirSorter<sup>53</sup> (v1.0.3, virome decontamination mode, virome database) and manual validation of viral  
499 hallmark annotations (Supplementary Table 5). To identify lysogenic vMAGs, annotations were queried  
500 for the key terms “integrase”, “recombination”, “repressor” and “prophage”. Annotations of validated  
501 vMAGs are provided in Supplementary Table 2. Five vMAGs not identified by either program were  
502 manually verified as phage according to VIBRANT annotations (i.e., KEGG, Pfam and VOG databases)  
503 by searching for viral hallmark genes, greater ratio of VOG to KEGG annotations and a high proportion of  
504 unannotated proteins. Note, not all vMAGs were predicted as phage by VIBRANT, but all vMAGs were  
505 given full annotation profiles. One scaffold was determined to be non-viral and remove based on the  
506 presence of many bacterial-like annotations and few viral-like annotations. Validation (including software-  
507 guided and manually inspected procedures) produced a total of 191 vMAGs encoding 227 DSM AMG. It  
508 is of note that the DSM AMGs carried by three vMAGs (Ga0121608\_100029, Draft\_10000217 and  
509 Ga0070741\_10000875) could not be definitely ruled out as encoded within microbial contamination. This  
510 was determined based on the high density of non-phage annotations surrounding the AMGs in conjunction  
511 with the presence of an integrase annotation, suggesting the possibility of phage integration near the AMG.

512

### 513 Taxonomy of vMAGs

514 Taxonomic assignment of vMAGs was conducted using a custom reference database and script. To  
515 construct the reference database, NCBI GenBank<sup>54</sup> and RefSeq<sup>55</sup> (release July 2019) were queried for  
516 “prokaryotic virus”. A total of 15,238 sequences greater than 3kb were acquired. Sequences were  
517 dereplicated using mash and nucmer<sup>56</sup> at 95% sequence identity and 90% coverage. Dereplication resulted  
518 in 7,575 sequences. Open reading frames were predicted using Prodigal (-p meta, v2.6.3) for a total of  
519 458,172 proteins. Taxonomy of each protein was labeled according to NCBI taxonomic assignment of the  
520 respective sequence. DIAMOND<sup>57</sup> (v0.9.14.115) was used to construct a protein database. Taxonomy is  
521 assigned by DIAMOND BLASTp matches of proteins from an unknown phage sequence to the constructed  
522 database at the classifications of Order, Family and Sub-family. Assignment consists of reference protein  
523 taxonomy matching to each classification at the individual and all protein levels to hierarchically select the  
524 most likely taxonomic match rather than the most common (i.e., not recruitment of most common match).  
525 Taxonomic assignments are available for 25 Families and 29 Sub-families for both bacterial and archaeal  
526 viruses. The database, script and associated files used to assign taxonomy are provided. To construct the  
527 protein network diagram vConTACT2<sup>58</sup> (v0.9.5, default parameters) was used to cluster vMAGs with  
528 reference viruses from NCBI from the families *Ackermannviridae*, *Herelleviridae*, *Inoviridae*,  
529 *Microviridae*, *Myoviridae*, *Podoviridae* and *Siphoviridae* as well as several archaea-infecting families. The  
530 network was visualized using Cytoscape<sup>59</sup> (v3.7.2) and colored according to family affiliation.

531

### 532 World map distribution of vMAGs

533 IMG/VR Taxon Object ID numbers respective of each vMAGs were used to identify global  
534 coordinates of studies according to IMG documentation. Coordinates were mapped using Matplotlib  
535 (v3.0.0) Basemap<sup>60</sup> (v1.2.0). Human studies were excluded from coordinate maps.

536

### 537 **Sequence alignments and conserved residues**

538 Protein alignments were performed using MAFFT<sup>61</sup> (v7.388, default parameters). Visualization of  
539 alignments was done using Geneious Prime 2019.0.3. N- and C-terminal ends of protein alignments were  
540 manually removed, and gaps were stripped by 90% (SoxD and SoxYZ) or 98% (DsrA and DsrC/TusE) for  
541 clarity. Amino acid residues were highlighted by pairwise identity of 90% (SoxC and SoxYZ) or 95%  
542 (DsrA, DsrC/TusE and SoxD). An identity graph, generated by Geneious, was fitted to the alignment to  
543 visualize pairwise identity of 100% (green), 99-30% (yellow) and 29-0% (red). Conservation of domains  
544 and amino acid residues was assessed according to annotations by The Protein Data Bank.

545 To calculate  $dN/dS$  ratios between vMAG AMG pairs, dRep<sup>63</sup> (v2.6.2) was used to compare AMG  
546 sequences of *dsrA* (n = 39), *dsrC* (n = 141) and *soxYZ* (n = 44) separately (dRep compare --SkipMash --  
547 S\_algorithm goANI). A custom auxiliary script (dnds\_from\_drep.py<sup>64</sup>) was used to calculate  $dN/dS$  ratios  
548 from the dRep output between various AMG pairs. Resulting  $dN/dS$  values were plotted using Seaborn<sup>65</sup>  
549 (v0.8.1) and Matplotlib. Phage AMG pairs and respective  $dN/dS$  values can be found in Supplementary  
550 Table 6.

551

### 552 **vMAG protein grouping**

553 All protein sequences of 94 vMAGs, excluding those with non-validated DsrC (i.e., potentially  
554 TusE-like) AMGs according to the conserved CxxxxxxxxxC motif, were grouped using mmseqs2<sup>66</sup> (--  
555 min-seq-id 0.3 -c 0.6 -s 7.5 -e 0.001). Groups containing at least two different representative vMAGs were  
556 retained (887 groups total). A presence/absence heatmap was made using the R package  
557 “ComplexHeatmap”<sup>67</sup> and hierarchically grouped according to the ward.D method. Metadata for AMG,  
558 taxonomy and source environment were laid over the grouped columns. Two vMAGs,  
559 Ga0066448\_1000315 and JGI24724J26744\_10000298, were not represented by any of the 887 retained  
560 clusters. vMAG alignments were done using EasyFig<sup>68</sup> (v2.2.2).

561

### 562 **vMAG genome structure and organization**

563 vMAGs representative of each AMG family were selected. Annotations were performed using  
564 VIBRANT and the best scoring annotation was used. Genomes were visualized using Geneious Prime and  
565 manually colored according to function.

566

### 567 **AMG protein phylogenetic tree reconstruction**

568 The DSM protein reference sequences were downloaded from NCBI nr database (accessed May  
569 2019) by searching names and results were manually filtered. The curated results were clustered by 70%  
570 sequence similarity using CD-HIT<sup>69</sup> (v4.7). These representative sequences from individual clusters were  
571 aligned with the corresponding vMAG AMG protein sequences using MAFFT (default settings).  
572 Alignments were subjected to phylogenetic tree reconstruction using IQ-TREE<sup>70</sup> (v1.6.9) with the following  
573 settings: -m MFP -bb 100 -s -redo -mset WAG,LG,JTT,Dayhoff -mrate E,I,G,I+G -mfreq FU -wbtl  
574 (“LG+G4” was chosen as the best-fit tree reconstruction model). The environmental origin information of  
575 each vMAG AMG was used to generate the stripe ring within the phylogenetic tree in the operation frame  
576 of iTOL<sup>71</sup> online server.

577  
578  
579  
580  
581  
582  
583  
584  
585  
586  
587  
588  
589  
590  
591  
592  
593  
594  
595  
596  
597  
598  
599  
600  
601  
602  
603  
604  
605  
606  
607  
608  
609  
610  
611  
612  
613  
614  
615  
616  
617  
618  
619  
620

### **Metagenomic mapping and gene coverage ratio calculation**

The metagenomic reads were first dereplicated by a custom Perl script and trimmed by Sickle<sup>72</sup> (v1.33, default settings). The QC-passed metagenomic reads were used to map against the collection of genes of investigated metagenomic assemblies by Bowtie2<sup>73</sup> (v2.3.4.1). The gene coverage for each gene was calculated by “jgi\_summarize\_bam\_contig\_depths” command within metaWRAP<sup>74</sup> (v1.0.2). The phage:total gene coverage ratio was calculated by adding up all the phage and bacterial gene coverage values and using it to divide the summed phage gene coverage values.

We identified the phage-host gene pairs in the phylogenetic tree containing AMG and their bacterial counterpart gene encoding proteins. We assigned the phage-host gene pairs according to the following two criteria: 1) The phage and host gene encoding proteins are phylogenetically close in the tree; the branches containing them should be neighboring branches. 2) They should be from the same metagenomic dataset, which means that AMGs and bacterial host genes are from the same environment sample. The identified phage-host gene pairs were labelled accordingly in the phylogenetic tree.

For the gene coverage ratio calculation of phage genes and bacterial genes within a phage-host pair, we first calculated the phage:total gene coverage ratio and bacterial:total gene coverage ratio using the same method as described above; and then, in order to avoid the influence of numbers of phage or bacterial genes, we normalized the above two ratio values by the number of phage and bacterial genes, respectively. Finally, the normalized phage:host gene coverage ratio of this phage-host pair was calculated by comparing these two ratio values, accordingly.

Additionally, reads mapping performance was re-checked by comparing original mapping results (using Bowtie 2 “-very-sensitive” option) to the mapping results that only include reads with one mismatch (Supplementary Fig. 7). Checking results have justified the reliability of our original mapping performance and our gene coverage ratio calculation.

### **Metatranscriptomic mapping**

The metatranscriptomic reads were first dereplicated by a custom Perl script and trimmed by Sickle (default settings), and then subjected to rRNA-filtering using SortMeRNA<sup>75</sup> (v2.0) with the 8 default rRNA databases (including prokaryotic 16S rRNA, 23S rRNA; eukaryotic 18S rRNA, 28S rRNA; and Rfam 5S rRNA and 5.8S rRNA). QC-passed metagenomic reads were mapped against the collection of AMGs using Bowtie2 (-very-sensitive). The gene expression level in Reads Per Kilobase per Million mapped reads (RPKM) was calculated by normalizing the sequence depth (per million reads) and the length of the gene (in kilobases).

### **Data availability**

All IMG/VR sequences are available at <https://img.jgi.doe.gov/cgi-bin/vr/main.cgi> and [https://genome.jgi.doe.gov/portal/pages/dynamicOrganismDownload.jsf?organism=IMG\\_VR](https://genome.jgi.doe.gov/portal/pages/dynamicOrganismDownload.jsf?organism=IMG_VR). Sequences from identified vMAGs are available publicly and described in Supplementary Tables 1 and 2. All sequences and custom analysis scripts used in this study are also available at [https://github.com/AnantharamanLab/Kieft\\_and\\_Zhou\\_et\\_al.\\_2020](https://github.com/AnantharamanLab/Kieft_and_Zhou_et_al._2020).

## 621 REFERENCES

- 622
- 623 1. Clokie, M. R., Millard, A. D., Letarov, A. V. & Heaphy, S. Phages in nature. *Bacteriophage* **1**, 31–45  
624 (2011).
- 625 2. Edwards, R. A., McNair, K., Faust, K., Raes, J. & Dutilh, B. E. Computational approaches to predict  
626 bacteriophage–host relationships. *FEMS Microbiol Rev* **40**, 258–272 (2016).
- 627 3. Louca, S., Mazel, F., Doebeli, M. & Parfrey, L. W. A census-based estimate of Earth’s bacterial and  
628 archaeal diversity. *PLOS Biology* **17**, e3000106 (2019).
- 629 4. Jiang, S. C. & Paul, J. H. Gene Transfer by Transduction in the Marine Environment. *APPL.*  
630 *ENVIRON. MICROBIOL.* **64**, 8 (1998).
- 631 5. Russell, P. W. & Müller, U. R. Construction of bacteriophage luminal diameterX174 mutants with  
632 maximum genome sizes. *J Virol* **52**, 822–827 (1984).
- 633 6. Hatfull, G. F. *et al.* Comparative genomic analysis of 60 Mycobacteriophage genomes: genome  
634 clustering, gene acquisition, and gene size. *J. Mol. Biol.* **397**, 119–143 (2010).
- 635 7. Hurwitz, B. L. & U’Ren, J. M. Viral metabolic reprogramming in marine ecosystems. *Current*  
636 *Opinion in Microbiology* **31**, 161–168 (2016).
- 637 8. Hurwitz, B. L., Hallam, S. J. & Sullivan, M. B. Metabolic reprogramming by viruses in the sunlit and  
638 dark ocean. *Genome Biology* **14**, R123 (2013).
- 639 9. Howard-Varona, C. *et al.* Phage-specific metabolic reprogramming of virocells. *ISME J* 1–15 (2020)  
640 doi:10.1038/s41396-019-0580-z.
- 641 10. Suttle, C. A. Marine viruses — major players in the global ecosystem. *Nature Reviews Microbiology*  
642 **5**, 801–812 (2007).
- 643 11. Heldal, M. & Bratbak, G. Production and decay of viruses in aquatic environments. *Mar. Ecol. Prog.*  
644 *Ser.* **72**, 205–212 (1991).
- 645 12. Bragg, J. G. & Chisholm, S. W. Modeling the Fitness Consequences of a Cyanophage-Encoded  
646 Photosynthesis Gene. *PLOS ONE* **3**, e3550 (2008).
- 647 13. Mann, N. H., Cook, A., Millard, A., Bailey, S. & Clokie, M. Bacterial photosynthesis genes in a  
648 virus. *Nature* **424**, 741 (2003).
- 649 14. Thompson, L. R. *et al.* Phage auxiliary metabolic genes and the redirection of cyanobacterial host  
650 carbon metabolism. *PNAS* **108**, E757–E764 (2011).
- 651 15. Breitbart, M., Thompson, L., Suttle, C. & Sullivan, M. Exploring the Vast Diversity of Marine  
652 Viruses. *Oceanography* **20**, 135–139 (2007).
- 653 16. Roux, S. *et al.* Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by  
654 single-cell- and meta-genomics. *eLife Sciences* **3**, e03125 (2014).
- 655 17. Lindell, D., Jaffe, J. D., Johnson, Z. I., Church, G. M. & Chisholm, S. W. Photosynthesis genes in  
656 marine viruses yield proteins during host infection. *Nature* **438**, 86–89 (2005).
- 657 18. Lindell, D. *et al.* Transfer of photosynthesis genes to and from Prochlorococcus viruses. *Proc. Natl.*  
658 *Acad. Sci. U.S.A.* **101**, 11013–11018 (2004).
- 659 19. Lindell, D. *et al.* Genome-wide expression dynamics of a marine virus and host reveal features of co-  
660 evolution. *Nature* **449**, 83–86 (2007).
- 661 20. Ruiz-Perez, C. A., Tsementzi, D., Hatt, J. K., Sullivan, M. B. & Konstantinidis, K. T. Prevalence of  
662 viral photosynthesis genes along a freshwater to saltwater transect in Southeast USA. *Environmental*  
663 *Microbiology Reports* **11**, 672–689 (2019).
- 664 21. Sullivan, M. B. *et al.* Prevalence and Evolution of Core Photosystem II Genes in Marine  
665 Cyanobacterial Viruses and Their Hosts. *PLoS Biology* **4**, e234 (2006).
- 666 22. Anantharaman, K. *et al.* Sulfur Oxidation Genes in Diverse Deep-Sea Viruses. *Science* **344**, 757–760  
667 (2014).
- 668 23. Chen, L.-X. *et al.* Large Freshwater Phages with the Potential to Augment Aerobic Methane  
669 Oxidation. <http://biorxiv.org/lookup/doi/10.1101/2020.02.13.942896> (2020)  
670 doi:10.1101/2020.02.13.942896.

- 671 24. Ahlgren, N. A., Fuchsman, C. A., Rocap, G. & Fuhrman, J. A. Discovery of several novel,  
672 widespread, and ecologically distinct marine Thaumarchaeota viruses that encode amoC nitrification  
673 genes. *The ISME Journal* **13**, 618–631 (2019).
- 674 25. Emerson, J. B. *et al.* Host-linked soil viral ecology along a permafrost thaw gradient. *Nature*  
675 *Microbiology* **3**, 870 (2018).
- 676 26. Trubl, G. *et al.* Soil Viruses Are Underexplored Players in Ecosystem Carbon Processing. *mSystems*  
677 **3**, e00076-18 (2018).
- 678 27. Cassman, N. *et al.* Oxygen minimum zones harbour novel viral communities with low diversity:  
679 Viral community characteristics of an oxygen minimum zone. *Environmental Microbiology* **14**,  
680 3043–3065 (2012).
- 681 28. Andreae, M. O. Ocean-atmosphere interactions in the global biogeochemical sulfur cycle. *Marine*  
682 *Chemistry* **30**, 1–29 (1990).
- 683 29. Anantharaman, K. *et al.* Expanded diversity of microbial groups that shape the dissimilatory sulfur  
684 cycle. *The ISME Journal* **12**, 1715 (2018).
- 685 30. Roux, S. *et al.* Ecogenomics and potential biogeochemical impacts of globally abundant ocean  
686 viruses. *Nature* **537**, 689–693 (2016).
- 687 31. Martinez-Hernandez, F. *et al.* Single-virus genomics reveals hidden cosmopolitan and abundant  
688 viruses. *Nat Commun* **8**, 15892 (2017).
- 689 32. Hatfull, G. F. & Hendrix, R. W. Bacteriophages and their Genomes. *Curr Opin Virol* **1**, 298–303  
690 (2011).
- 691 33. Ikeuchi, Y., Shigi, N., Kato, J., Nishimura, A. & Suzuki, T. Mechanistic Insights into Sulfur Relay by  
692 Multiple Sulfur Mediators Involved in Thiouridine Biosynthesis at tRNA Wobble Positions.  
693 *Molecular Cell* **21**, 97–108 (2006).
- 694 34. Dammeyer, T., Bagby, S., Sullivan, M., Chisholm, S. & Frankenberg-Dinkel, N. Efficient phage-  
695 mediated pigment biosynthesis in oceanic cyanobacteria. *Curr Biol* **18**, (2008).
- 696 35. Ghosh, W. & Dam, B. Biochemistry and molecular biology of lithotrophic sulfur oxidation by  
697 taxonomically and ecologically diverse bacteria and archaea. *FEMS Microbiol Rev* **33**, 999–1043  
698 (2009).
- 699 36. Marshall, K. T. & Morris, R. M. Isolation of an aerobic sulfur oxidizer from the SUP05/Arctic96BD-  
700 19 clade. *The ISME Journal* **7**, 452–455 (2013).
- 701 37. Grimm, F., Dobler, N. & Dahl, C. Regulation of *dsr* genes encoding proteins responsible for the  
702 oxidation of stored sulfur in *Allochromatium vinosum*. *Microbiology* **156**, 764–773 (2010).
- 703 38. Bradley, A. S., Leavitt, W. D. & Johnston, D. T. Revisiting the dissimilatory sulfate reduction  
704 pathway. *Geobiology* **9**, 446–457 (2011).
- 705 39. Hensen, D., Sperling, D., Trüper, H. G., Brune, D. C. & Dahl, C. Thiosulphate oxidation in the  
706 phototrophic sulphur bacterium *Allochromatium vinosum*. *Mol. Microbiol.* **62**, 794–810 (2006).
- 707 40. Friedrich, C. G. *et al.* Novel genes coding for lithotrophic sulfur oxidation of *Paracoccus*  
708 *pantotrophus* GB17. *J. Bacteriol.* **182**, 4677–4687 (2000).
- 709 41. Hatfull, G. F. Bacteriophage Genomics. *Curr Opin Microbiol* **11**, 447–453 (2008).
- 710 42. Warwick-Dugdale, J., Buchholz, H. H., Allen, M. J. & Temperton, B. Host-hijacking and planktonic  
711 piracy: how phages command the microbial high seas. *Virol J* **16**, (2019).
- 712 43. Anantharaman, K., Breier, J. A., Sheik, C. S. & Dick, G. J. Evidence for hydrogen oxidation and  
713 metabolic plasticity in widespread deep-sea sulfur-oxidizing bacteria. *PNAS* (2012)  
714 doi:10.1073/pnas.1215340110.
- 715 44. Anantharaman, K., Breier, J. A. & Dick, G. J. Metagenomic resolution of microbial functions in  
716 deep-sea hydrothermal plumes across the Eastern Lau Spreading Center. *The ISME Journal* **10**, 225–  
717 239 (2016).
- 718 45. Zimmerman, A. E. *et al.* Metabolic and biogeochemical consequences of viral infection in aquatic  
719 ecosystems. *Nature Reviews Microbiology* **18**, 21–34 (2020).
- 720 46. Breitbart, M. Marine Viruses: Truth or Dare. *Annual Review of Marine Science* **4**, 425–448 (2012).

- 721 47. Haveman, S. A. *et al.* Gene Expression Analysis of Energy Metabolism Mutants of *Desulfovibrio*  
722 *vulgaris* Hildenborough Indicates an Important Role for Alcohol Dehydrogenase. *Journal of*  
723 *Bacteriology* **185**, 4345–4353 (2003).
- 724 48. Von Damm, K. L., Edmond, J. M., Measures, C. I. & Grant, B. Chemistry of submarine hydrothermal  
725 solutions at Guaymas Basin, Gulf of California. *Geochimica et Cosmochimica Acta* **49**, 2221–2237  
726 (1985).
- 727 49. Paez-Espino, D. *et al.* IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses.  
728 *Nucleic Acids Res.* **45**, D457–D465 (2017).
- 729 50. Paez-Espino, D. *et al.* IMG/VR v.2.0: an integrated data management and analysis system for  
730 cultivated and environmental viral genomes. *Nucleic Acids Res.* **47**, D678–D686 (2019).
- 731 51. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification.  
732 *BMC Bioinformatics* **11**, 119 (2010).
- 733 52. Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: automated recovery, annotation and curation of  
734 microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome*  
735 **8**, 90 (2020).
- 736 53. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial  
737 genomic data. *PeerJ* **3**, (2015).
- 738 54. Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res*  
739 **44**, D67–D72 (2016).
- 740 55. O’Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic  
741 expansion, and functional annotation. *Nucleic Acids Res* **44**, D733–D745 (2016).
- 742 56. Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLOS Computational*  
743 *Biology* **14**, e1005944 (2018).
- 744 57. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature*  
745 *Methods* **12**, 59–60 (2015).
- 746 58. Jang, H. B. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by  
747 gene-sharing networks. *Nature Biotechnology* **1** (2019) doi:10.1038/s41587-019-0100-8.
- 748 59. Shannon, P. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction  
749 Networks. *Genome Research* **13**, 2498–2504 (2003).
- 750 60. Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing In Science & Engineering* **9**, 90–95  
751 (2007).
- 752 61. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7:  
753 Improvements in Performance and Usability. *Mol Biol Evol* **30**, 772–780 (2013).
- 754 62. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).
- 755 63. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic  
756 comparisons that enables improved genome recovery from metagenomes through de-replication. *The*  
757 *ISME Journal* **11**, 2864–2868 (2017).
- 758 64. Olm, M. R. *et al.* Consistent Metagenome-Derived Metrics Verify and Delineate Bacterial Species  
759 Boundaries. *mSystems* **5**, (2020).
- 760 65. Michael Waskom *et al.* *mwaskom/seaborn: v0.8.1 (September 2017)*. (Zenodo, 2017).  
761 doi:10.5281/zenodo.883859.
- 762 66. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis  
763 of massive data sets. *Nature Biotechnology* **35**, 1026–1028 (2017).
- 764 67. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in  
765 multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
- 766 68. Sullivan, M. J., Petty, N. K. & Beatson, S. A. Easyfig: a genome comparison visualizer.  
767 *Bioinformatics* **27**, 1009–1010 (2011).
- 768 69. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation  
769 sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

- 770 70. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective  
771 Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol* **32**, 268–274  
772 (2015).
- 773 71. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display  
774 and annotation. *Bioinformatics* **23**, 127–128 (2007).
- 775 72. Joshi, N. & Fass, J. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files.  
776 <https://github.com/najoshi/sickle> (2011).
- 777 73. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359  
778 (2012).
- 779 74. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP—a flexible pipeline for genome-resolved  
780 metagenomic data analysis. *Microbiome* **6**, 158 (2018).
- 781 75. Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in  
782 metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012).
- 783  
784

### 785 **Acknowledgements**

786 We thank the University of Wisconsin—Office of the Vice Chancellor for Research and Graduate  
787 Education, University of Wisconsin—Department of Bacteriology, and University of Wisconsin—  
788 College of Agriculture and Life Sciences for their support. K.K. is supported by a Wisconsin  
789 Distinguished Graduate Fellowship Award from the University of Wisconsin-Madison. This work was  
790 partly supported by the National Science Foundation grant OCE-0961947 to MBS. The work conducted  
791 by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the  
792 U.S. Department of Energy under contract no. DE-AC02-05CH11231.

793  
794

### 795 **Author Contributions**

796 K.K., Z.Z., S.R. and K.A. designed the study. K.K. and S.R. identified the genomes. K.K., Z.Z., and K.A.  
797 conducted the analyses. K.K., Z.Z., and K.A. drafted the manuscript. All authors reviewed the results,  
798 revised and approved the manuscript.

799  
800

### 801 **Competing Interests**

802 The authors declare no competing interests.

803  
804

### 805 **Figure and Table Captions**

806

807 **Fig. 1** Dataset summary statistics and representative genome organization diagrams of vMAGs. **a** The  
808 number of vMAGs, 191 total, encoding single or multiple DSM AMGs. **b** Estimated vMAG genome  
809 qualities as a function of scaffold lengths. vMAGs encoding **c** *dsrA* and *dsrC*, **d** *dsrA* and two *dsrC*, **e** *soxC*  
810 and *soxD*, and **f** *soxYZ*. For **c**, **d**, **e** and **f** linear vMAG scaffolds are visualized as circular with the endpoints  
811 indicated by dashed lines, and predicted open reading frames are colored according to VIBRANT  
812 annotation functions.

813

814 **Fig. 2** Conceptual diagrams of viral DsrA, DsrC, SoxC, SoxD and SoxYZ auxiliary metabolism. **a**  
815 Microbial dissimilatory oxidation of hydrogen sulfide and stored inorganic sulfur. The resulting production

816 of ATP utilized for cellular processes and growth and the pathway's rate limiting step is indicated with an  
817 asterisk (top). Viral infection and manipulation of sulfur oxidation by encoded DsrA or DsrC to augment  
818 the pathway's rate limiting step and increase energy yield towards viral replication (bottom). **b** Microbial  
819 dissimilatory oxidation of thiosulfate or storage of inorganic sulfur in the periplasm. The resulting  
820 production of ATP is utilized for cellular processes and the pathway's key energy yielding reaction  
821 indicated with an asterisk (top). Viral infection and manipulation of thiosulfate oxidation by encoded SoxC,  
822 SoxD or SoxYZ to augment the entire pathway and the key energy yielding step to increase energy yield  
823 towards viral replication (bottom). For **a** and **b** cellular processes are shown in red, sulfur oxidation pathway  
824 is shown in black, energy flow is shown in blue, and viral processes are shown in orange (**a**) or purple (**b**).  
825

826 **Fig. 3** Phylogenetic tree of AMG proteins and distribution of phage genomes (on a world map). **a, b**  
827 Phylogenetic trees of phage DsrA and DsrC **c, d, e** SoxC, SoxD, SoxYZ. Ultrafast bootstrap (UFBoot)  
828 support values (> 50%) are labelled on the nodes. **c, d** Phage gene encoded protein sequences are labeled  
829 with stars and their environmental origin information is labeled accordingly. **f** World map showing  
830 distribution of phage genomes that contain the sulfur-related AMGs. Studies on human systems are  
831 excluded from the map.  
832

833 **Fig. 4** Taxonomic assignment of vMAGs and protein network clustering with reference phages. In the  
834 protein network each dot represents a single vMAG (circles with outlines) or reference phage (circles  
835 without outlines), and dots are connected by lines respective to shared protein content. Genomes (i.e., dots)  
836 having more similarities will be visualized by closer proximity and more connections. Cluster annotations  
837 depicted by dotted lines were approximated manually. vMAG taxonomy was colored according to  
838 predictions by a custom reference database and script, shown by bar chart insert.  
839

840 **Fig. 5** vMAG protein grouping and genome alignments. **a** vMAG hierarchical protein grouping where each  
841 row represents a single protein group (887 total) and each column represents a single vMAG (94 total).  
842 Metadata for encoded AMGs, estimated taxonomy, source environment and number of protein groups per  
843 vMAG is shown. Clades respective of **b, c** and **d** are depicted by colored dotted lines. Genome alignments  
844 of **b** seven divergent Myoviridae vMAGs encoding *dsrC* from diverse environments, **c** four divergent  
845 Myoviridae vMAGs encoding *soxYZ* from diverse environments, and **d** four divergent Siphoviridae  
846 vMAGs encoding *dsrA* and *dsrC* from hydrothermal environments. For the genome alignments, each black  
847 line represents a single genome and arrows represent predicted proteins which are colored according to  
848 VIBRANT annotations; genomes are connected by lines representing tBLASTx similarity. **e** Map of  
849 geographic distribution of 15 vMAGs depicted in **b, c** and **d**, annotated with respective clade, source  
850 environment and taxonomic family.  
851

852 **Fig. 6** Phage to total *dsrA* and *soxYZ* gene coverage ratios. **a** Viral *dsrA* to total (viral and bacterial *dsrA*  
853 gene together) gene coverage ratios. The contribution of viral *dsrA* genes from different SUP05  
854 Gammaproteobacteria clades is shown in different colors. The average viral *dsrA*:total ratio was calculated  
855 from 12 samples. **b** Viral *soxYZ* to total gene coverage ratios. The contribution of viral *soxYZ* genes from  
856 three different clades is shown in different colors. Genes from Freshwater Lake and *Tara* Ocean samples  
857 were compared separately, and the average viral *soxYZ*:total ratios were calculated and compared separately  
858 as for Freshwater Lake and *Tara* Ocean samples.  
859

860 **Fig. 7** Phage to host *dsrA* and *SoxYZ* gene coverage ratios and *dsrA* gene expression comparison between  
861 phage and host pairs. **a** Phage *dsrA* to total gene coverage ratios of each phage-host pair. Average phage  
862 *dsrA*:total ratios of phage-host pairs in SUP05 Clade 1 and Clade 2 were calculated by 5 and 11 pairs of  
863 genes, respectively. **b** Phage *soxYZ* to total gene coverage ratios of each phage-host pair. The contribution  
864 of phage *soxYZ* genes from three different clades is shown in different colors. Average phage *dsrA*:total  
865 ratios of phage-host pairs in Freshwater Lakes and *Tara* Ocean were calculated separately. **c** Phage to host  
866 *dsrA* gene expression comparison in Guaymas Basin metatranscriptomes. The same database was used for  
867 mapping both hydrothermal and background metatranscriptomic datasets **d** Phage to host *dsrA* gene  
868 expression comparison in Chesapeake Bay metatranscriptomes. The same database was used for mapping  
869 all Chesapeake Bay metatranscriptomic datasets. Gene expression levels are shown in RPKM normalized  
870 by gene sequence depth and gene length.

871  
872 **Fig. 8** Conceptual figure indicating the ecology and function of AMGs in sulfur metabolisms. **a** DSM AMG  
873 effect on the budget of reduced sulfur transformation. **b** Diagram of virus-mediated metabolism short  
874 circuiting the microbial sulfur loop in nutrient cycling.

875  
876 **Supplementary Figure 1. DsrA Protein alignment and identified conserved residues in microbial and**  
877 **phage sequences.** Highlighted amino acids indicate pairwise identity of  $\geq 95\%$  and colored boxes indicate  
878 substrate binding motifs (pink) and strictly conserved siroheme binding motifs (blue). An identity graph  
879 (top) was fitted to the alignments to visualize pairwise identity at the following thresholds: 100% (green),  
880 99-30% (yellow, scaled) and 29-0% (red, scaled).

881  
882 **Supplementary Figure 2. DsrC protein alignment and conserved residues in microbial and phage**  
883 **sequences.** Highlighted amino acids indicate pairwise identity of  $\geq 95\%$  and colored boxes indicate strictly  
884 conserved residues (blue) or lack of conserved residues (gray). An identity graph (top) was fitted to the  
885 alignments to visualize pairwise identity at the following thresholds: 100% (green), 99-30% (yellow,  
886 scaled) and 29-0% (red, scaled).

887  
888 **Supplementary Figure 3. SoxYZ protein alignment and conserved residues in microbial and phage**  
889 **sequences.** Highlighted amino acids indicate pairwise identity of  $\geq 90\%$  and colored boxes indicate  
890 substrate binding cysteine (blue) and cysteine motif (pink). An identity graph (top) was fitted to the  
891 alignments to visualize pairwise identity at the following thresholds: 100% (green), 99-30% (yellow,  
892 scaled) and 29-0% (red, scaled).

893  
894 **Supplementary Figure 4. SoxC protein alignment and conserved residues in microbial and phage**  
895 **sequences.** Highlighted amino acids indicate pairwise identity of  $\geq 90\%$  and colored boxes indicate cofactor  
896 coordination / active site (blue). An identity graph (top) was fitted to the alignments to visualize pairwise  
897 identity at the following thresholds: 100% (green), 99-30% (yellow, scaled) and 29-0% (red, scaled).

898  
899 **Supplementary Figure 5. SoxD protein alignment and conserved residues in microbial and phage**  
900 **sequences.** Highlighted amino acids indicate pairwise identity of  $\geq 95\%$  and colored boxes indicate  
901 cytochrome c motif (blue). An identity graph (top) was fitted to the alignments to visualize pairwise identity  
902 at the following thresholds: 100% (green), 99-30% (yellow, scaled) and 29-0% (red, scaled).

903

904 **Supplementary Figure 6. Calculation of the ratio of non-synonymous to synonymous ( $dN/dS$ )**  
905 **nucleotide differences of AMGs.** Comparison of  $dN/dS$  ratios between vMAG AMG pairs for *dsrA*, *dsrC*  
906 and *soxYZ*. Each point represents a single comparison pair. Values below 1 suggest purifying selection  
907 pressures.

908  
909 **Supplementary Figure 7. Mapping quality checks for phage and bacterial sulfur AMGs.** **a** Result for  
910 phage and bacterial *dsrA* genes in the metagenome IMG: 3300001676. The phage-host pair contains one  
911 phage *dsrA* (TuiMalila\_10011672) and two bacterial *dsrA* (TuiMalila\_10106401, TuiMalila\_10061351).  
912 Both the original mapping result and the mapping results including reads with one mismatch were  
913 compared. The normalized phage / bacteria gene coverage ratios were calculated for both of the above  
914 settings. The normalized phage/bacteria gene coverage ratio based on the original mapping result are shown  
915 in Fig. 7a. **b** Result for phage and bacterial *soxYZ* gene in the metagenome of IMG: 3300009154. The  
916 phage-host pair contains one phage *soxYZ* (Ga0114963\_1000012431) and one bacterial *soxYZ*  
917 (Ga0114963\_108352751). Both the original mapping result and the mapping results including reads with  
918 one mismatch were compared. The normalized phage/bacteria gene coverage ratios were calculated for both  
919 of the above settings. The normalized phage/bacteria gene coverage ratios based on the original mapping  
920 results are shown in Fig. 7b. Filtering steps to only retain reads with only one mismatch was conducted by  
921 mapped.py (<https://github.com/christophertbrown/bioscripts/blob/master/ctbBio>) with the settings of "-m 1  
922 -p both". Mapping results were visualized by Geneious Prime v2020.1.2.

923  
924 **Supplementary Figure 8. Phylogenetic tree of phage and bacterial DsrA and phage-host pairs from**  
925 **hydrothermal environments.** The phage and bacterial *dsrA* encoding proteins from the metagenomes  
926 studied in this project were aligned with reference sequences. The phylogenetic tree was reconstructed by  
927 IQ-TREE v1.6.9 with settings as described in the methods. Branches with over 90% UFBoot bootstrap  
928 values were labeled with closed circles. Phage *dsrA* genes are labeled in red. The phage-host gene pairs  
929 (linked with dash lines) were labeled accordingly in the tree. The hydrothermal metagenomes (12  
930 metagenomes in total) are from five locations in Lau Basin, southwest Pacific Ocean. IMG metagenome  
931 ID samples are available in Supplementary Table 3 ("Phage and bacterial *dsrA* gene abundance  
932 percentage").

933  
934 **Supplementary Figure 9. Phylogenetic tree of phage and bacterial SoxYZ and phage-host gene pairs**  
935 **from Freshwater Lake and Tara Ocean samples.** The phage and bacterial *soxYZ* encoding proteins from  
936 the metagenomes studied in this project were aligned with reference sequences. The phylogenetic tree was  
937 reconstructed by IQ-TREE v1.6.9 with settings as described in the methods. Branches with over 90%  
938 UFBoot bootstrap values were labeled with closed circles. Phage *soxYZ* genes are labeled in red. The phage-  
939 host gene pairs (linked with dash lines) were labeled accordingly in the tree. The IMG metagenome IDs of  
940 Freshwater Lake and Tara Ocean samples are available in Supplementary Table 4.

941  
942 **Supplementary Figure 10. Phylogenetic tree of phage and bacterial DsrA and phage-host pairs from**  
943 **the Guaymas Basin hydrothermal environment.** The phage and bacterial *dsrA* encoding proteins from  
944 the metagenomes studied in this project were aligned with reference sequences. The phylogenetic tree was  
945 reconstructed by IQ-TREE v1.6.9 with settings as described in the methods. Branches with over 90%  
946 UFBoot bootstrap values were labeled with closed circles. Phage *dsrA* genes are labeled in red. The phage-  
947 host gene pairs (linked with dash lines) were labeled accordingly in the tree. The IMG metagenome IDs of

948 Guaymas Basin samples are 3300001683 and 3300003086.

949

950 **Supplementary Figure 11. Phylogenetic tree of phage and bacterial DsrA and phage-host pairs from**  
951 **Chesapeake Bay.** The phage and bacterial *dsrA* encoding proteins from the metagenomes studied in this  
952 project were aligned with reference sequences. The phylogenetic tree was reconstructed by IQ-TREE  
953 v1.6.9 with settings as described in the methods. Branches with over 90% UFBoot bootstrap values were  
954 labeled with closed circles. Phage *dsrA* genes are labeled in red. The phage-host gene pairs (linked with  
955 dash lines) were labeled accordingly in the tree. IMG metagenome IDs are: 3300010370, 3300010354,  
956 3300010299, 3300010318, 3300010297, 3300010300, and 3300010296.

957

958 **Supplementary Figure 12. Heatmap of amino acid identities between phage and bacteria *dsrA* and**  
959 ***soxYZ* genes.** This diagram contains the comparisons of (a) SUP05 Clade 1 and Clade 2 phage and bacterial  
960 *dsrA* for Lau Basin hydrothermal environments, (b) Betaproteobacteria Clade, Methylophilales-like Clade,  
961 and Gammaproteobacteria Clade phage and bacterial *soxYZ* for freshwater lake and Tara Ocean  
962 environments, (c) SUP05 Clade 1 and Clade 2 phage and bacterial *dsrA* for Guaymas Basin hydrothermal  
963 environments, (d) Chesapeake Bay Clade Pair 1 and 2 phage and bacterial *dsrA* for Chesapeake Bay  
964 environments. The corresponding phylogenetic trees of individual subpanels could be found in  
965 Supplementary Figures 8, 9, 10, and 11. Blank cell indicates no amino acid identity within this pair due to  
966 the short sequences/no sequence overlap.

967

968 **Supplementary Table 1.** Details of vMAGs used in this study. Metadata was recovered from IMG/VR.

969

970 **Supplementary Table 2.** Protein annotations generated using VIBRANT for each vMAG.

971

972 **Supplementary Table 3.** The phage and bacterial *dsrA* gene abundance percentage and phage *dsrA* gene  
973 expression table.

974

975 **Supplementary Table 4.** The phage and bacterial *soxYZ* gene abundance percentage.

976

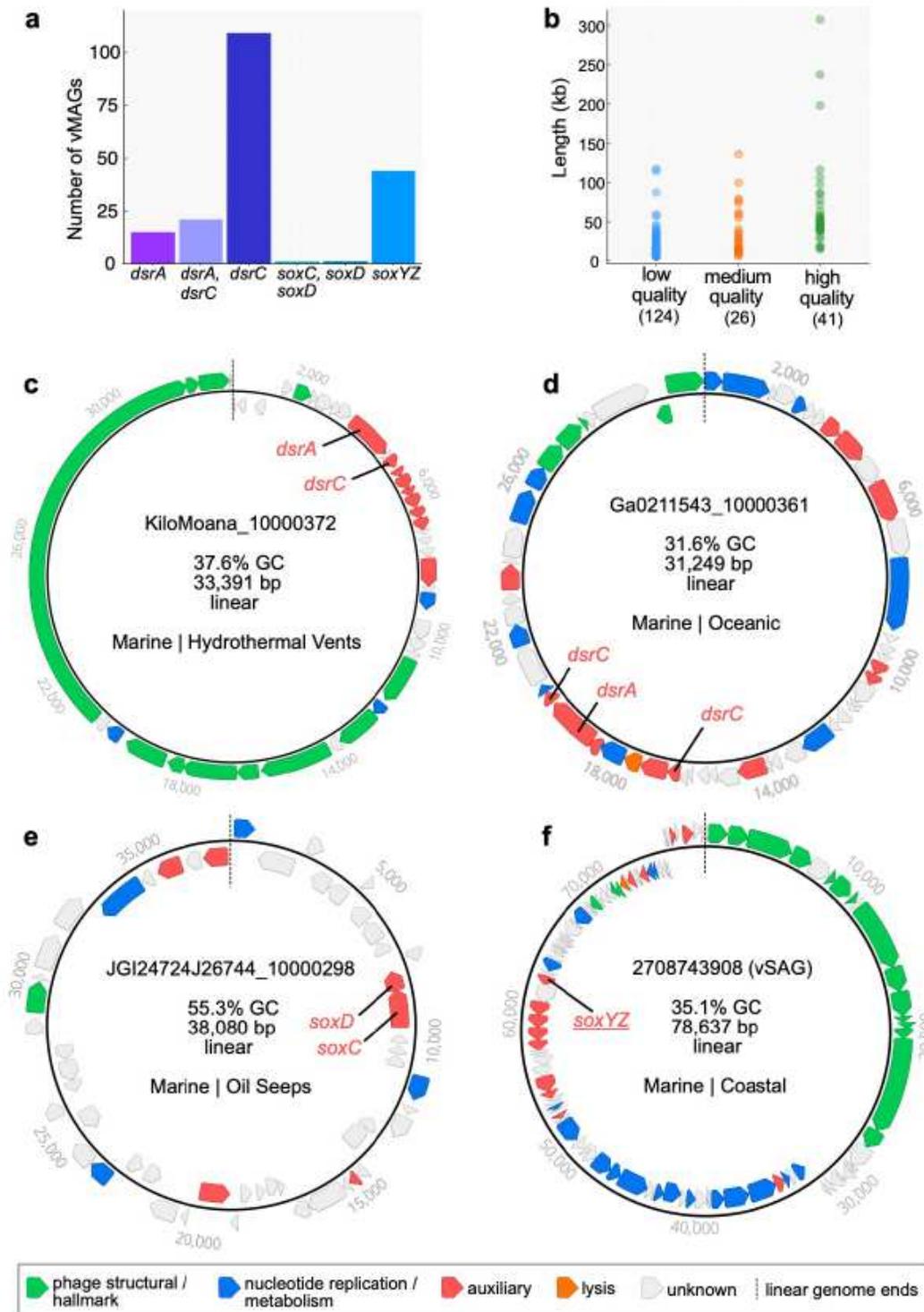
977 **Supplementary Table 5.** Validation of vMAG sequences as true phage identifications.

978

979 **Supplementary Table 6.** Phage AMG pairs and *dN/dS* calculations respective to Supplementary Figure 6.

980

# Figures



**Figure 1**

Dataset summary statistics and representative genome organization diagrams of vMAGs. a The number of vMAGs, 191 total, encoding single or multiple DSM AMG. b Estimated vMAG genome qualities as a function of scaffold lengths. vMAGs encoding c *dsrA* and *dsrC*, d *dsrA* and two *dsrC*, e *soxC* and *soxD*,

and f soxYZ. For c, d, e and f linear vMAG scaffolds are visualized as circular with the endpoints indicated by dashed lines, and predicted open reading frames are colored according to VIBRANT annotation functions.

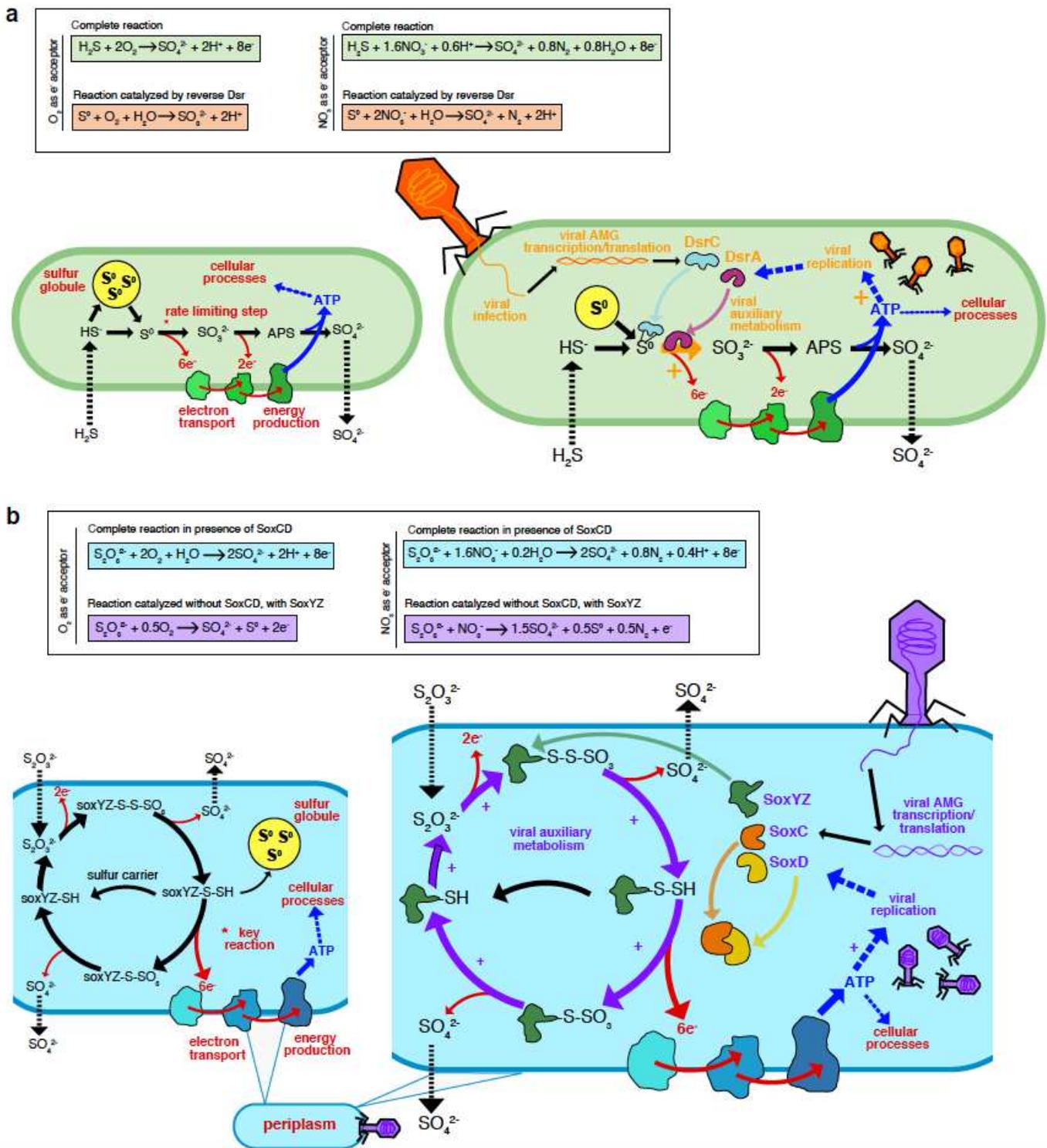
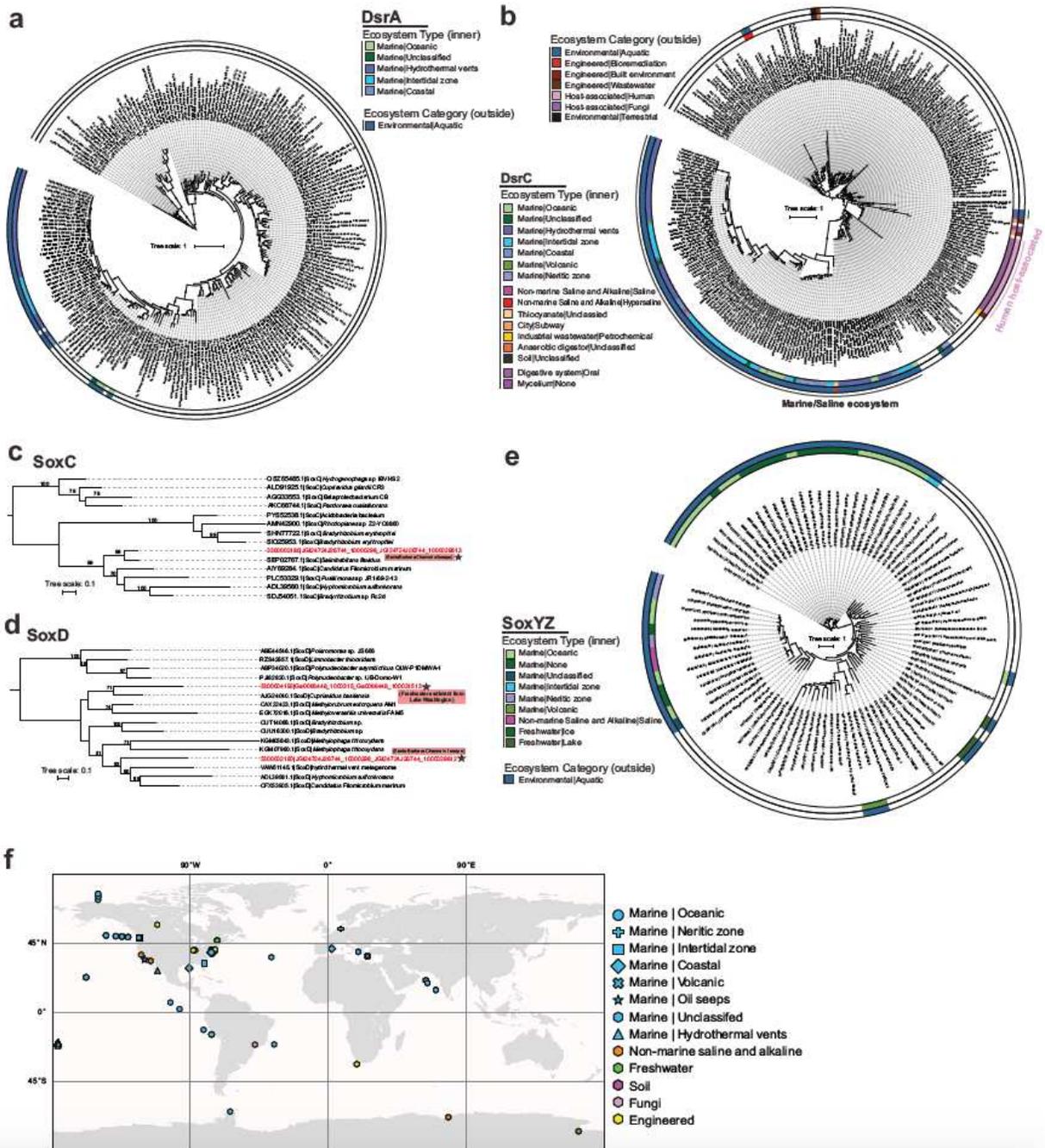


Figure 2

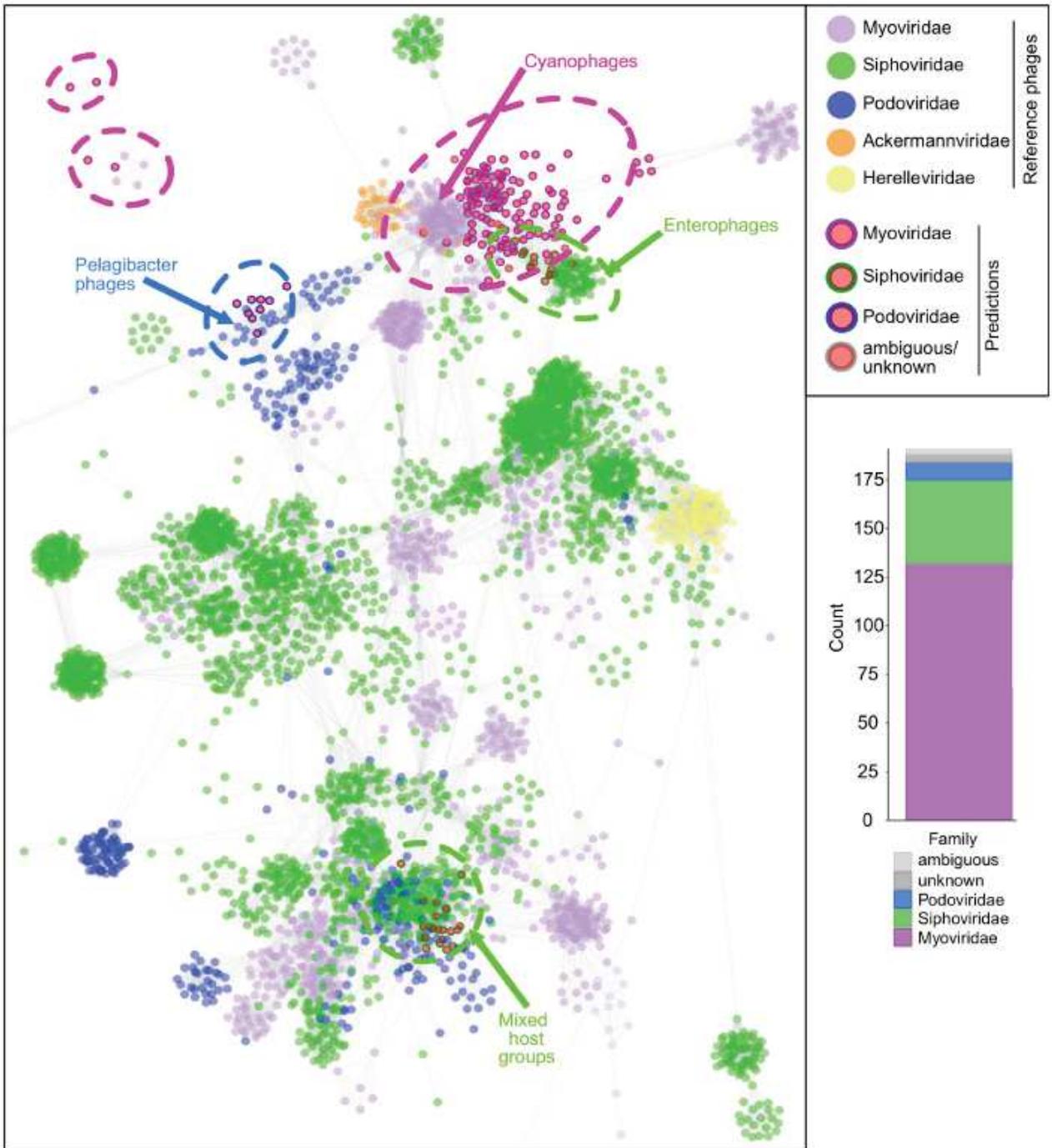
Conceptual diagrams of viral DsrA, DsrC, SoxC, SoxD and SoxYZ auxiliary metabolism. a Microbial dissimilatory oxidation of hydrogen sulfide and stored inorganic sulfur. The resulting production of ATP

utilized for cellular processes and growth and the pathway's rate limiting step is indicated with an asterisk (top). Viral infection and manipulation of sulfur oxidation by encoded DsrA or DsrC to augment the pathway's rate limiting step and increase energy yield towards viral replication (bottom). b Microbial dissimilatory oxidation of thiosulfate or storage of inorganic sulfur in the periplasm. The resulting production of ATP is utilized for cellular processes and the pathway's key energy yielding reaction indicated with an asterisk (top). Viral infection and manipulation of thiosulfate oxidation by encoded SoxC, SoxD or SoxYZ to augment the entire pathway zone and the key energy yielding step to increase energy yield towards viral replication (bottom). For a and b cellular processes are shown in red, sulfur oxidation pathway is shown in black, energy flow is shown in blue, and viral processes are shown in orange (a) or purple (b).



**Figure 3**

Phylogenetic tree of AMG proteins and distribution of phage genomes (on a world map). a, b Phylogenetic trees of phage DsrA and DsrC c, d, e SoxC, SoxD, SoxYZ. Ultrafast bootstrap (UFBoot) support values (> 50%) are labelled on the nodes. c, d Phage gene encoded protein sequences are labeled with stars and their environmental origin information is labeled accordingly. f World map showing distribution of phage genomes that contain the sulfur-related AMGs. Studies on human systems are excluded from the map.



**Figure 4**

Taxonomic assignment of vMAGs and protein network clustering with reference phages. In the protein network each dot represents a single vMAG (circles with outlines) or reference phage (circles without outlines), and dots are connected by lines respective to shared protein content. Genomes (i.e., dots) having more similarities will be visualized by closer proximity and more connections. Cluster annotations depicted by dotted lines were approximated manually. vMAG taxonomy was colored according to predictions by a custom reference database and script, shown by bar chart insert.

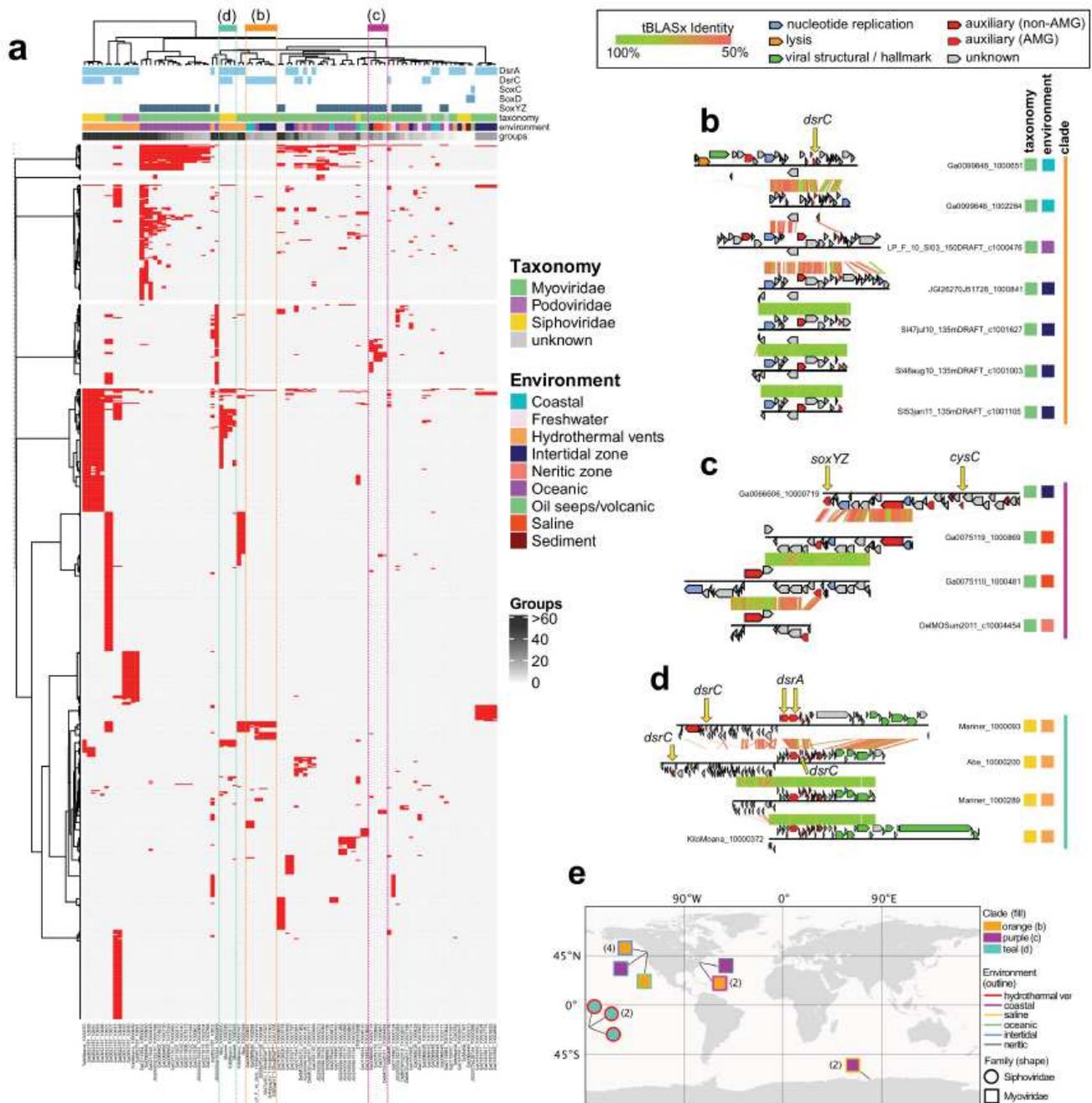
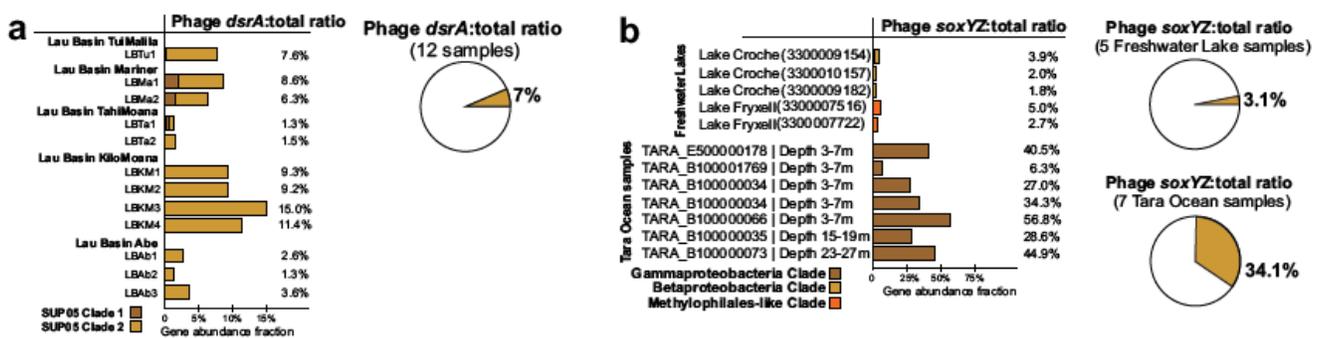


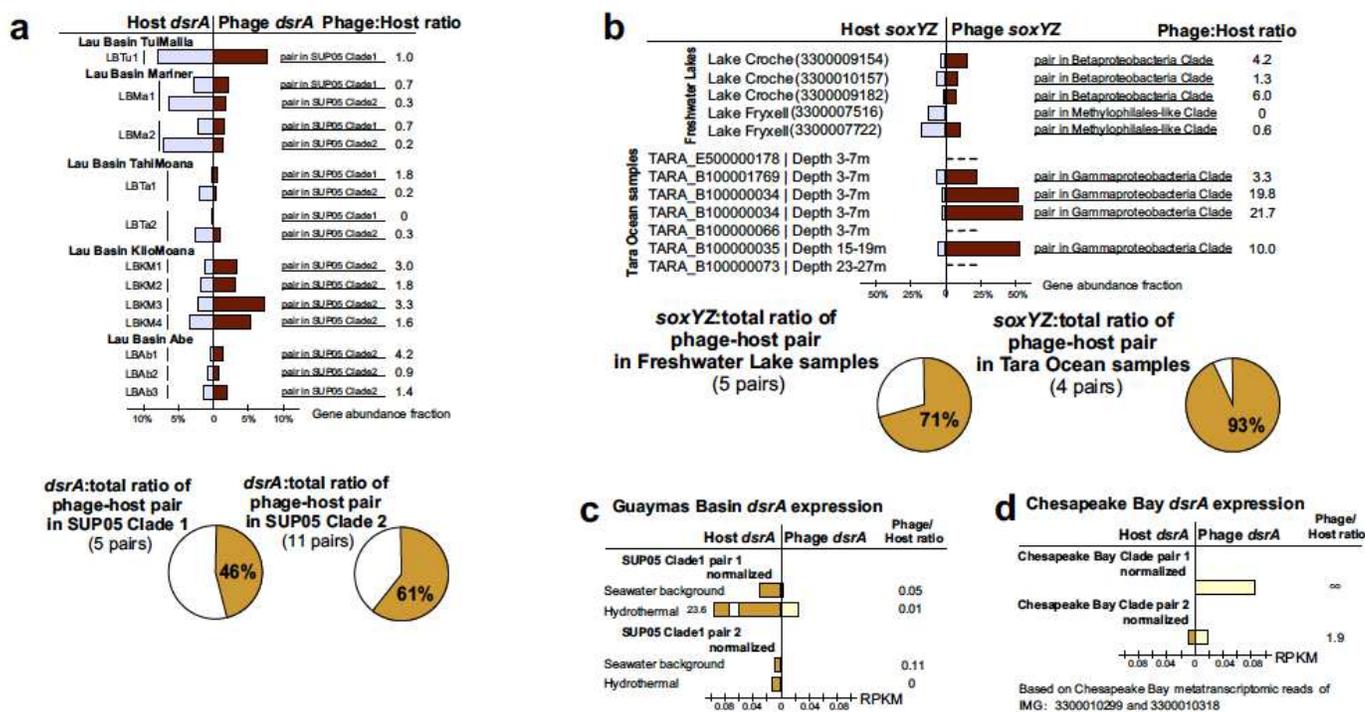
Figure 5

vMAG protein grouping and genome alignments. a vMAG hierarchical protein grouping where each row represents a single protein group (887 total) and each column represents a single vMAG (94 total). Metadata for encoded AMGs, estimated taxonomy, source environment and number of protein groups per vMAG is shown. Clades respective of b, c and d are depicted by colored dotted lines. Genome alignments of b seven divergent Myoviridae vMAGs encoding *dsrC* from diverse environments, c four divergent Myoviridae vMAGs encoding *soxYZ* from diverse environments, and d four divergent Siphoviridae vMAGs encoding *dsrA* and *dsrC* from hydrothermal environments. For the genome alignments, each black line represents a single genome and arrows represent predicted proteins which are colored according to VIBRANT annotations; genomes are connected by lines representing tBLASTx similarity. e Map of geographic distribution of 15 vMAGs depicted in b, c and d, annotated with respective clade, source environment and taxonomic family.



**Figure 6**

Phage to total *dsrA* and *soxYZ* gene coverage ratios. a Viral *dsrA* to total (viral and bacterial *dsrA* gene together) gene coverage ratios. The contribution of viral *dsrA* genes from different SUP05 Gammaproteobacteria clades is shown in different colors. The average viral *dsrA*:total ratio was calculated from 12 samples. b Viral *soxYZ* to total gene coverage ratios. The contribution of viral *soxYZ* genes from three different clades is shown in different colors. Genes from Freshwater Lake and Tara Ocean samples were compared separately, and the average viral *soxYZ*:total ratios were calculated and compared separately as for Freshwater Lake and Tara Ocean samples.



**Figure 7**

Phage to host *dsrA* and *SoxYZ* gene coverage ratios and *dsrA* gene expression comparison between phage and host pairs. a Phage *dsrA* to total gene coverage ratios of each phage-host pair. Average phage *dsrA*:total ratios of phage-host pairs in SUP05 Clade 1 and Clade 2 were calculated by 5 and 11 pairs of genes, respectively. b Phage *soxYZ* to total gene coverage ratios of each phage-host pair. The contribution of phage *soxYZ* genes from three different clades is shown in different colors. Average phage *dsrA*:total ratios of phage-host pairs in Freshwater Lakes and Tara Ocean were calculated separately. c Phage to host *dsrA* gene expression comparison in Guaymas Basin metatranscriptomes. The same database was used for mapping both hydrothermal and background metatranscriptomic datasets d Phage to host *dsrA* gene expression comparison in Chesapeake Bay metatranscriptomes. The same database was used for mapping all Chesapeake Bay metatranscriptomic datasets. Gene expression levels are shown in RPKM normalized by gene sequence depth and gene length.

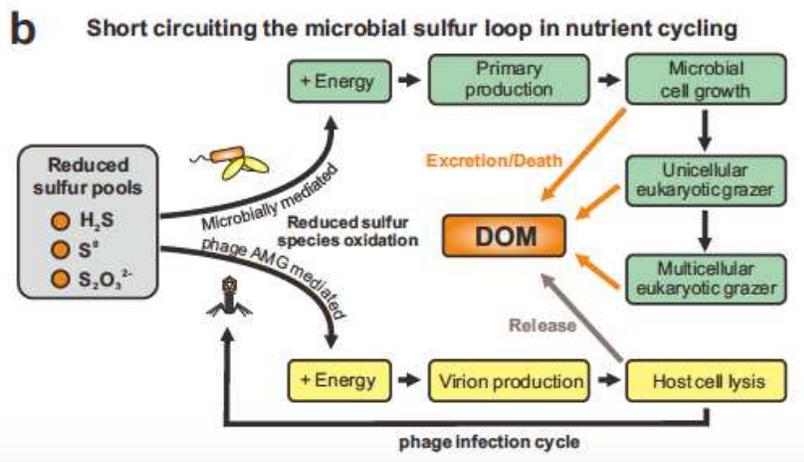
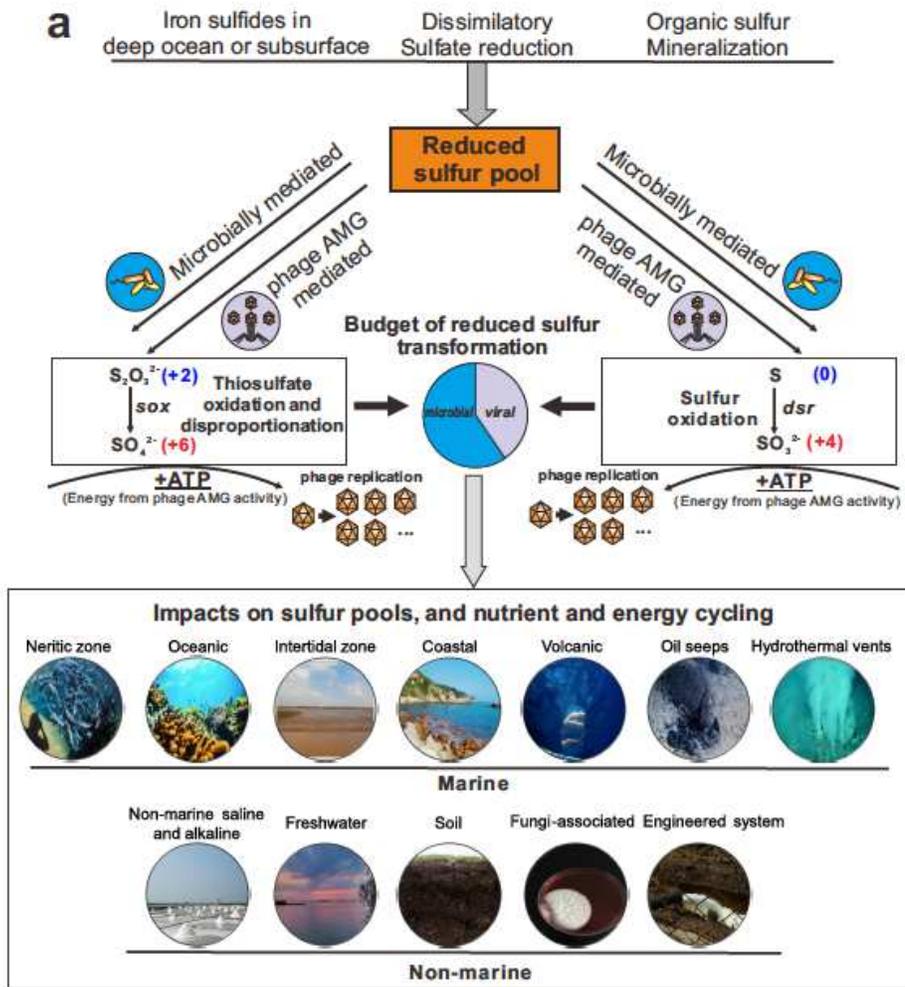


Figure 8

Conceptual figure indicating the ecology and function of AMG in sulfur metabolisms. a DSM AMG effect on the budget of reduced sulfur transformation. b Diagram of virus-mediated metabolism short circuiting the microbial sulfur loop in nutrient cycling.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [NatCommIMGVRsoxdsrSupplementaryFigures.pdf](#)
- [NatCommIMGVRsoxdsrSupplementaryTables.xlsx](#)