Ecogenomics of uncultivated globally abundant ocean viruses

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

[1] Department of Microbiology, The Ohio State University, Columbus, OH, USA
[2] Theoretical Biology and Bioinformatics, Utrecht University, Utrecht, The Netherlands.
[3] Centre for Molecular and Biomolecular Informatics, Radboud University Medical Centre, Nijmegen, The Netherlands.
[4] Department of Marine Biology, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
[5] Structural and Computational Biology, European Molecular Biology Laboratory, Heidelberg, Germany
[6] Department of Ecology and Evolutionary Biology, University of Michigan, MI, USA
[7] Division of Microbial Ecology, Department of Microbiology and Ecosystem Science, Research Network Chemistry Meets Microbiology, University of Vienna, Vienna, Austria
[8] Austrian Polar Research Institute, Vienna, Austria
[9] Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA
[10] Department of Marine Biology and Oceanography, Institut de Ciències del Mar (CSIC), Barcelona, Spain
[11] Institute of Marine Sciences (CNR-ISMAR), National Research Council, Venezia, Italy
[12] CEA - Institut de Génomique, GENOSCOPE, Evry, France
[13] PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany
[33] MARUM, Bremen University, Bremen, Germany
[15] Directors' Research European Molecular Biology Laboratory, Heidelberg, Germany
[16] Sorbonne Universités, UPMC Université Paris 06, CNRS, Laboratoire Adaptation et Diversité en Milieu Marin, Station Biologique de Roscoff, Roscoff, France
[17] Sorbonne Universités, UPMC Université Paris 06, CNRS, Laboratoire d'oceanographie de Villefranche (LOV), Observatoire Océanologique, Villefranche-sur-Mer, France
[18] Department of Oceanography, University of Hawaii, Honolulu, Hawaii, USA
[19] Mediterranean Institute of Advanced Studies, CSIC-UiB, Esporles, Mallorca, Spain
[20] King Abdullah University of Science and Technology (KAUST), Red Sea Research Center (RSRC), Thuwal, Saudi Arabia
[21] Max-Delbrück-Centre for Molecular Medicine, 13092 Berlin, Germany
[22] CNRS, UMR 8030, CP5706, Evry, France
[23] Université d'Evry, UMR 8030, CP5706, Evry, France
[24] Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, OH, USA

[†] *Tara* Oceans coordinators and affiliations are listed following the Acknowledgements.

[*] correspondence to mbsulli@gmail.com

CONTENTS

Dataset generation

With the goal to make this large dataset tractable for analysis, we first organized it into different levels of organization, namely contigs, populations, and Viral Clusters (VCs, Supplementary Figure 1). To this end, we first quality-controlled metagenomic shotgun sequencing reads and assembled them into contigs, then binned contigs into sets of contigs representing viral populations and finally grouped viral populations into clusters of related viral populations which we denote as VCs. In the following, we describe in detail the procedures performed to achieve these goals. All the raw data are available at ENA or IMG, with sample identifiers indicated in Supplementary Table 1. Processed data, including assembled contigs, populations definition and abundance, clusters definition and abundance, annotated viral contigs are available at http://mirrors.iplantcollaborative.org/browse/iplant/home/shared/iVirus.

Selection of binning parameters

To identify the optimal parameters for genome binning with Metabat[1], different sets of parameters were evaluated (all with minProb=80%, minBinned=40%, minCV=1, minContig=1,500, minContigByCorr=500, s=15,000 and fuzzy=1): 3 "sensitive" (i.e. p1, p2, p3 at 90%, 90% and 95% respectively) with a minimum correlation of 92%, 95% and 98%, and 3 "specific" (i.e. p1, p2, p3 at 95%, 90% and 95% respectively) with a minimum correlation of 92%, 95% and 98%. The binning efficiency was evaluated based on the identification of single copy marker genes from microbial genomes(COG0012; COG0049; COG0052; COG0048; COG0016; COG0018; COG0080; COG0088; COG0081; COG0087; COG0090; COG0085; COG0091; COG0092; COG0093; COG0094; COG0096; COG0097; COG0098; COG0099; COG0100; COG0102; COG0103; COG0124; COG0172; COG0184; COG0185; COG0186; COG0197; COG0200; COG0201; COG0215; COG0256; COG0522; COG0495; COG0525; COG0552, see ref. 2) and viral genomes (TerL[3]), looking for the set of parameters maximizing the number of viral bins generated (i.e. bins with only viral marker gene and no microbial marker gene), and minimizing the percentage of bins with multiple single-copy marker genes (i.e. including multiple genomes). Both these parameters were optimal with the "sensitive" options and a

minimum correlation of 98% (data not shown).

Identification of contigs and bins from eukaryotic viruses

While automatic tools exist to distinguish microbial and viral contigs in a mixed dataset, such as VirSorter[4] used in this study, no such tool is available to discriminate between viruses infecting eukaryotes versus viruses infecting bacteria and archaea. To this end, we chose to rely on a BLAST affiliation of the contig genes to the NCBI RefseqVirus database. The contigs considered as originating from eukaryotic viruses fulfilled one of these conservative conditions:

- >50% of genes displaying a best BLAST hit to an eukaryotic virus
- at least 1 gene with a best BLAST hit to an eukaryotic virus and no genes with a best BLAST hit against an archaeal or bacterial virus
- more than 10% of the genes or 3 genes (whichever was higher) with a best BLAST hit to an eukaryotic virus, and less than 25% of genes with a best BLAST hit to an archaeal and bacterial virus

Arguably, because of the similarity between genomes from archaeal or bacterial viruses and eukaryotic viruses, some of these contigs might still originate from archaeal or bacterial viruses. However, a best BLAST hit affiliation of all genes from archaeal and bacterial viruses in NCBI RefSeq database (v70, 05-26-2015), against all viruses (i.e. including eukaryotic viruses) revealed that only 0.13% of these genes had a best BLAST hit to an eukaryotic virus. Hence, given that some eukaryotic viruses are small enough to get through 0.22μm filters, any contig fulfilling the above conditions likely originated from an eukaryote virus, and was thus (conservatively) designated as a "eukaryotic virus".

The affiliation of viral genome bins (based on the affiliation of its contigs members) was performed as follows:

- if >10kb or >25% of the bin (whichever was lower) was affiliated to a eukaryotic virus, the bin was considered as a eukaryotic virus bin.
- the bin was considered as an archaeal or bacterial virus bin otherwise.

Identification of epi-, meso-, and bathypelagic viral populations

The coverage of viral populations (based on a mapping of QC'd reads to the contigs with Bowtie 2[5], using options --sensitive, -X 2000, and --non-deterministic, with all other parameters set to default) across the different samples revealed a strong separation between epi- and mesopelagic samples versus bathypelagic samples: 15,203 populations were detected (i.e. ≥75% of the genome covered) only in epi- and mesopelagic samples, 47 were detected only in bathypelagic samples, 19 were detected in both layers, but with a highest coverage in an epi- or mesopelagic sample, and 11 were detected in both layers, but with a highest coverage in a bathypelagic sample (Supplementary Table 2). Based on this uniqueness of deep-sea viruses, we decided to focus exclusively on epi- and mesopelagic viruses (i.e. contigs with a highest coverage in an epi- or mesopelagic sample) for this study, and focus on the bathypelagic viruses in a separate study. Thus, the final set of sequences analyzed in this study was composed of 298,383 contigs and 15,222 viral populations.

Level of reads mapping to viral contigs and populations

On average, 66.83% of epi- or mesopelagic reads per sample mapped back to an assembled contig (min: 33.75%, max 84.11%). Over these, reads mapping to contigs identified to originate from microbial or eukaryotic viruses corresponded to 12.47% on average for epipelagic (SRF / DCM / MIX) samples (1.49% - 47.89%), and 40.05% on average for mesopelagic (MES) samples (11.07% - 58.27%), where the lower biomass available (and lower concentration of viruses) likely explains this increased detection of non-viral signal. On average, for both epi- and mesopelagic samples, 40.95% of these mapped reads were linked to contigs too small (often <5kb[4]) to determine their origin (i.e. viral or

cellular genome) with confidence (range: 24.50% - 68.04%). These small contigs likely correspond to genomes with a coverage too low to lead to the assembly of a large genome fragment(s), either rare viruses or cellular genomes. Eventually, 46.26% of the mapped reads were linked to viral contigs for epipelagic samples (7.21% - 65.92%), and 20.42% for mesopelagic samples (5.25% - 47.14%). Among the reads mapped to any viral contig, an average of 20% were mapped to a contig associated with a viral population (the other viral contigs being too small to be confidently linked to any specific viral population identified). This constituted a two-fold improvement in terms of ratio of reads mapped to viral populations compared to the TOV dataset[6].

Viral cluster definition and affiliation

Additional taxonomic affiliation of NCBI Refseq viral genomes (v70) identified in GOV VCs

The following affiliation were added to the taxonomy downloaded from NCBI, based on manual curation of the corresponding genomes: Synechococcus phage S-SKS1: *Myoviridae*, Persicivirga phage P12024L: *Siphoviridae*, Persicivirga phage P12024S: *Siphoviridae*, Cyanophage KBS-P-1A: *Podoviridae*, Synechococcus phage S-CBP2: *Podoviridae*, Synechococcus phage S-CBP3: *Podoviridae*, Cyanophage SS120-1: *Podoviridae*, Cyanophage MED4-117: *Myoviridae*, Marinomonas phage P12026: *Siphoviridae*, Sulfitobacter phage pCB2047-C: *Podoviridae*, Vibrio phage pYD21-A: *Siphoviridae*, Paenibacillus phage phiIBB_Pl23: *Siphoviridae*, Bacillus phage Grass: *Myoviridae*, Deep-sea thermophilic phage D6E: *Myoviridae*, Geobacillus virus E2: *Siphoviridae*, Pseudomonas phage phiPto-bp6g: *Siphoviridae*, Lactobacillus johnsonii prophage Lj771: *Siphoviridae*, Enterococcus phage EF62phi: *Podoviridae*.

Complete and near-complete genome identification

Circular contigs (i.e. contigs with identical 5' and 3' ends) were considered as new complete genomes[7]. To identify linear contigs that might represent complete or near-complete genomes, the genome size information of each VC was first gathered (based on known isolates and circular contigs). If the standard deviation of this VC's estimated genome size was within 15% of the VC average genome size, i.e. if the genomes in that VC were not too variable in size (which was the case for 220 out of 245 VCs with two or more complete genomes), then all contigs in the VC longer than 80% of this VC average genome size were considered as near-complete genomes. Overall, 345 contigs were detected as complete because these were circular, and 425 were linear but within 20% of the expected genome size of the VC, and thus considered as near-complete.

Comparison of clustering methods for viral genome fragments and corresponding taxonomic level

Although there is currently no uniform standard to classify viruses into taxonomic groups based on genomic data[8], we sought to obtain nevertheless an overview of the phylogenomic diversity of the viral clusters we identified. Several methods have been previously proposed to classify new viral genomes into taxonomic groups. Although phylogenetic trees based on marker genes can be used for “fine-level” groupings (i.e. species level), the lack of universally conserved marker gene and high rate of horizontal gene transfer hampers the detection of larger groups (e.g. genus level). Instead, researches relied on pairwise genome comparisons to identify groups of viral genomes sharing a substantial portion of their genes as a reflection of their common ancestral origin.

A first way to estimate genome-to-genome similarities is by counting the number of proteins shared between two genomes relative to the total number of proteins in each genome. Based on reference genomes, thresholds were proposed of 40% of genes shared for two viruses to belong in the same genus, and 20% of genes shared to belong in the same subfamily[9]. Although these “hard” thresholds are operationally useful, they do not account for different rates of evolution and horizontal gene transfers between viral groups[3].

Another proposed method named "phage proteomic tree" uses first an all-vs-all BLAST comparison to calculate a global viral genome similarity matrix, and then applies a clustering algorithm (e.g. hierarchical clustering) to identify groups of similar genomes[10]. This method was successfully applied to new environmental viral genomes, but requires complete or near-complete genomes[11].

Finally, a network clustering of viral genomes based on shared content was also proposed[12]. Instead of representing relationships between viral genomes in a binary tree, this approach uses a reticulate representation (i.e. a network) to display the viral genome sequence space, and identify in this network "clusters" of viruses ("Viral Clusters" or VCs) sharing a significant fraction of their genes (i.e. sharing more genes that would be expected by chance). Because of this reticulate representation, this method can more accurately account for the mosaic nature of viral genomes. Moreover, it is also less sensitive to incomplete genomes as frequently seen in environmental samples, and is easily scalable to include tens of thousands of sequences[3].

Because of this scalability and better ability to classify mosaic and/or partial genomes, this latter method was used in this study. For corroboration, we compared out results with the phage proteomic tree and percentage of shared genes for the 756 complete and near-complete epi- and mesopelagic GOV genomes (Extended Data Fig. 2). Overall, ~75% of the VCs corresponded to monophyletic groups in the phage proteomic tree, confirming that both methods provide similar classifications. Most of the discrepancies were found in poorly resolved regions of phage genome sequence space as reflected by sporadic representation in the network or long branches in the proteomic tree, unlikely to be all resolved by any techniques until more data become available (Extended Data Figure 2). Examining the percentage of shared PCs between viruses, we found that viral contigs within a VC tend to share 20% to 50% of their genes (i.e. PCs). This was consistent with the observation that 54 of the 68 known viral genera with more than one genome are gathered in single VCs in this dataset (12 others are in 2 VCs, and only 2 are split in more than 2 VCs), and previous results showing that VCs gathered viruses from the same genus[3,12]. Unfortunately, the number of subfamilies defined by the International Committee on Taxonomy of Viruses (ICTV) is too low (only 6) to robustly evaluate the VC classification at this level. Hence, we used VCs as our larger taxonomic unit (Supplementary Figure 1), and considered it an operational taxonomic unit not associated with a specific taxonomic level.

Genome and contig map comparisons for VCs

Genome and contig map comparison were generated with Easyfig[13]. For each VC, all GOV complete and/or near-complete genomes are included if available. If the VC did not include any complete or near-complete genomes, all the longest contigs (within 80% of the largest sequence) were included. Non-GOV sequences were included when available (only the ones most closely related to a GOV contig were added to the plot when multiple non-GOV sequences are included in the VC).

Host prediction for viral contigs and VCs

Different methods have been proposed to predict virus-host pairings *in silico*. The most recent review of these methods[14] identified three main ways of predicting host(s) for a new virus: (i) similarity between viral and host genomes (as identified by BLAST), (ii) similarities between viral genome and CRISPR spacers in the host (strong, yet rarely identified signal), and (iii) similarities in nucleotide composition of host and viral genomes. To help interpret the host prediction results on GOV dataset, we first summarize here the results of the different benchmarks conducted for these methods, before discussing the results obtained on GOV sequences.

Host prediction method sensitivity and specificity

The first method (direct sequence similarity between virus and host genome) is the most accurate if the similarity region is long enough (at least 2kb). Unfortunately, this signal will originate from the presence of an integrated prophage in the host genome or several horizontally transfer genes, which

means that any virus-host prediction based on this method requires that (a) the virus has the capacity to integrate into or exchange genes with the host genome, (b) a host genome (or closely related one) was sequenced, and (c) the prophage or exchanged gene(s) region in the host genome was correctly assembled. Hence, this approach will only provide host information for a small fraction of the newly sequenced viruses.

The second method (CRISPR-based similarity) can also be very accurate when using exact or near-exact matches (up to 2 mismatches) between the CRISPR spacer and the viral genome[14]. This method requires the host to use a CRISPR-based defense, and the correct host sequence to be available in the reference database (i.e. a host having encountered the virus recently enough to have generated CRISPR spacers that still retain near-perfect identity to the viral genome), so will only be suitable for a small fraction of the new viruses.

Finally, the third method (nucleotide composition) is also able to accurately identify virus-host pairings, especially when comparing 4-mer frequency vectors between host and viral genomes[14]. Notably, correct host predictions at the order level could be obtained even after excluding the exact host species from the host database, and only retaining host sequences from the same genus[3]. This method will however provide a predicted host to the viral genome in every case, and thresholds on the similarity between host and viral 4-mer frequency vectors have to be applied to get accurate predictions[3]. Hence, viral genomes could still have no predicted host even when using this nucleotide composition approach, because the similarity between 4-mer frequency vectors was such that no confident predictions could be made.

Results of host prediction on GOV contigs

The three host prediction methods (BLAST similarity, CRISPR spacer matches, nucleotide composition) were applied to GOV viral contig associated with a viral population (24,353 contigs), to try to predict host at the phylum level (or class for Proteobacteria).

As could be expected, only a minor fraction of the viral contigs could be associated with a putative host with CRISPR matches (n=22), however these predictions were always consistent (i.e. when a viral contig was associated with different host genomes by CRISPR matches, they always corresponded to the same host group). A larger fraction of viral contigs could be linked to a host through BLAST matches (n=859). Among these, 34 contigs were linked to different host groups by these BLAST matches (i.e. a viral genome was similar to host genomes belonging to different phyla), and these predictions were discarded for the rest of the analysis. Finally, an even larger fraction of viral contigs were linked to a host through nucleotide composition (n=3,292). These results are provided in Supplementary Table 4.

When multiple predictions from different methods were obtained for the same viral contig, these were found to be consistent in most cases (>85%). This was consistent with the expected accuracy of such host prediction method[3,14]. Moreover, most of the conflicting cases correspond to host predictions relying on microbial metagenome contigs, which affiliation is not as certain as for complete genomes. Hence, these discrepancies are likely due to a combination of the few false-positive predictions generated through tetranucleotide frequencies similarity and affiliation issues for microbial metagenome contigs.

Host prediction at the VC level

Because of the sparseness and uncertainty of the host prediction at the individual contig or population level, this signal was studied at the VC level. For each VC, the number of host prediction for each host group was calculated, and compared to an expected number of host prediction obtained by chance based on the number of sequences in the VC, the number of host sequences of this group in the reference database, and the global rate of host prediction across all contigs. This is meant to control for false positive predictions that will be database-biased (i.e. more frequent for hosts more represented

in the database) and scale with VC size, and make sure that VC-host links are based on a significant numbers of predictions. Overall, a host prediction could be achieved for 392 VCs.

Function, taxonomic affiliation, and host prediction for new viral AMGs (*dsrC*, *soxYZ*, P-II, *amoC*)

Auxiliary Metabolic Genes (AMGs) are metabolic genes encoded by viruses and used to reprogram their host cell metabolism during infection[15,16]. Identifying AMGs is thus critical to predict the potential impact of viruses on geochemical cycles, however such AMG analysis from uncultivated virus presents multiple challenges, especially the verification of functionality of the AMG from the gene and predicted protein sequences only, and the association of the AMG with a specific host group and metabolic pathway. Here, we describe the different analyses conducted to address these two questions.

Motif detection and 3D structure predictions

Multiple alignments of the 4 new viral AMGs further analyzed were screened for the presence of conserved residues known to be important in the function of these proteins.

For *dsr*C, these include two conserved cysteine residues in the C-terminal part of the protein, identified as Cys-B and Cys-A[17]. The presence of both these residues is considered as required for a protein to be identified as bona fide DsrC, i.e. a "true DsrC" involved in the Dsr proteins-mediated sulfur oxidation or sulfite reduction reactions. Further groups of "DsrC-like" proteins were recently delineated by phylogenetic analysis and based on presence/absence of the two cysteine residues[17]. For example, sequences that only have the CysA residue are classified as TusE. While some members of the TusE group are known to be involved in persulfide-driven thiolation of tRNAs, it is unclear if they also play a role in the sulfur energy metabolism of sulfur-oxidizing bacteria[17]. In the GOV dataset, one AMG (DsrC-5) displayed both Cys-B and Cys-A and branched within the bona fide DsrC clade (Extended Data Fig. 5 and Supplementary Figure 2), while the other AMGs (DsrC-1 to 4 clades) lacked Cys-B and affiliated with the TusE group.

As no specific motifs have been defined to identify SoxY and SoxZ proteins, we relied on the conserved residues highlighted in their respective PFAM domains (PF13501 and PF08770). Screening the SoxYZ multiple alignments for these residues revealed that viral *soxYZ* genes encoded 23 of 24 conserved positions (Supplementary Figure 3). Notably, all the virus-encoded SoxYZ proteins included the motif KXX(X/-)GGC that is required for SoxYZ activity (the C residue is the active site of protein-bound sulfur oxidation[18]), and thus likely have a role in enhancing microbial sulfur oxidation.

In the case of P-II-like proteins, two conserved motifs, a C-terminal region signature, and a conserved uridylation site, have been described and manually curated (PROSITE documentation PDOC00439). Three of the four clades of viral P-II display the two motifs, and only one clade (P-II-3) seemed to lack the conserved uridylation site (Supplementary Information Figure 4). Because P-II structure has been determined through X-ray analysis, we could predict 3D structures of the P-II AMGs using I-TASSER[19](with default parameters). The models obtained were similar to the known P-II structure (Supplementary Information Figure 4) and their quality assessed with ProSA[20] comparable to the quality of experimentally determined structures for proteins of similar size in the Protein Data Bank (Supplementary Information Figure 4). Except for P-II-3 (whose function remains uncertain), we thus considered that viral P-II proteins in clades P-II-1, P-II-2 and P-II-4 are likely functioning as true P-II proteins.

For the AmoC protein, no specific functional motifs have been described, and some butane monooxygenase not involved in ammonia oxidation but closely related to bacterial AMO have been identified, which indicates that remote homology is not sufficient to identify a new gene as "true" AMO[21]. Here, the *amo*C AMG displays 94% Amino Acid Identity to AmoC proteins from known ammonia-oxidizer *Nitrosopumilus maritimus* SCM1 (Supplementary Figure 5, by comparison, butane mooxygenase displayed 37-39% Amino Acid Identity to AMO proteins[22]). Consistently, a prediction of

transmembrane helices in the viral and *Nitrosopumilus maritimus* AmoC identified the same transmembrane domains in both proteins (Supplementary Figure 5). Thus, we considered the *amo*C AMG as likely functioning as true AmoC involved in ammonia oxidation.

Selection pressure and signal of expression of new AMGs

Since conserved motifs and predicted 3D structures suggested that viral copies of *dsr*C, *sox*YZ and P-II genes had retained their "cellular" function, we next tried to verify that these genes were still functional and active once encoded by the virus.

To this end, we first looked for signs of purifying selection by calculating the ratio of non-synonymous to synonymous polymorphism rates (pN/pS) for each gene on the AMG-containing contigs, as in[23]. All but one AMG displayed signs of strong purifying selection (pN/pS ratios <0.3, Extended Data Table 1), the only exception being P-II-4, for which the signal of purifying selection was weaker (pN/pS of 0.66). Overall, except for P-II-4, this confirmed that the viral copies of these genes are still under strong negative selection and likely functional.

We then tried to identify transcript(s) of these genes in Tara Oceans metatranscriptomes, and could detect coverage of *dsr*C and *sox*YZ AMGs in 3 samples (metatranscriptomic reads mapping to the viral contigs at ≥95% ANI, Extended Data Table 1). Moreover, in the cases where an AMG was covered in a metatranscriptome, 1 to 53 other genes from the same contigs were also covered, suggesting that the transcripts identified are indeed originating from the same viral population.

Hence, combining the selective constraint and expression data, the only AMG which may have lost or evolved new functionality once transferred to the viral genome might be P-II-4 (no transcript detected and pN/pS higher than the other AMGs).

Taxonomic affiliation of AMG-containing contigs

As some of the AMG-containing contigs were not included in the VCs (because they were too short), a manual annotation was performed based on gene content. The 11 *dsrC*-containing contigs all displayed ≥50% of their genes affiliated by best BLAST hit to genomes from the T4-like superfamily (Extended Data Figure 5). The two longest contigs were either in VC_2, which is affiliated to the T4 phage superfamily (GOV_bin_3019_contig-100_4), or not clustered in any VC (GOV_bin_5582_contig-100_23). However, this latter contig, despite being unclustered, displayed a gene similar to a T4 baseplate protein (Extended Data Figure 5). Hence, we concluded that all the *dsrC*-containing contigs were likely originating from T4-like phages.

Similarly, *soxYZ*-containing contigs displayed ≥50% of genes affiliated by best BLAST hit to the T4 subfamily (or to another contig affiliated to the T4 subfamily for Tp1_25_DCM_0-0d2_scaffold13291_1), and the two contigs included in the VC analysis were included in VC_2, alongside other long contigs from the same viral populations (Extended Data Figure 6). Again, this consistent signal strongly suggested that *sox*YZ-containing contigs are originating from T4-like phages.

Conversely, the GOV contigs including P-II AMGs were more diverse: 6 contigs were long enough to be included in the VC analysis, and were found in 5 different VCs (Extended Data Figure 7). Moreover, one of the short P-II-containing contigs (GOV_bin_5834_contig-100_7) was associated with a viral population clustered in VC_12 (a *Siphoviridae* VC: *P12024virus*). Combined with the fact that one of the large P-II-containing contigs in a new VC displayed a *Siphoviridae* structural gene (Tp1_39_OMZ-0-0d2_scaffold996_1), this suggested that P-II AMGs are found in at least two different families of viruses (*Myoviridae* and *Siphoviridae*), and possibly in 6 different genera.

Finally, the *amo*C AMG was detected in a single contig (GOV_bin_4552_contig-100_2), affiliated to a new VC (VC_623). This contig displayed 16 predicted genes: 3 with a best BLAST hit against a known archaeovirus (including a major capsid protein), 1 similar to *amo*C, and 12 hypothetical proteins

(no similarity to NCBI nr or PFAM). Best BLAST hit affiliations of genes predicted on the other contigs in VC_623 are consistent with uncultivated archaeoviruses. Thus, we considered GOV_bin_4552_contig-100_2 as a genome fragment from a newly described archaeaovirus.

Identification of putative hosts of AMG-containing contigs

Finally, we tried to identify the hosts of viruses encoding *dsr*C, *sox*YZ, P-II, or *amo*C genes based on the host prediction and the AMG phylogenetic trees.

For *dsr*C, no host could be predicted from any of the viral contigs, and accordingly, all clades of viral *dsr*C were distinct from any reference and only grouped with other metagenomic sequences from *Tara* Oceans microbial metagenomes[24] (Extended Data Table 1 and Extended Data Figure 5). However, for DsrC-5, some of these microbial metagenomic contigs (the longest ones) could be affiliated further. Beyond DsrC, 9 of these contigs displayed other members of the *dsr* operon, including *dsr*A and *dsr*B which can used as phylogenetic markers for these sulfur-oxidizing micro-organisms[25]. These metagenomic sequences were thus inserted into a DsrAB reference tree[26] as outlined previously[27]. Briefly, *dsr*AB sequences were translated into amino acids, aligned to the reference alignment by using MAFFT[28], and added to the reference tree without changing its topology by using the Evolutionary Placement Algorithm in RAxML[29]. The metagenomic contigs all grouped near a recently described DsrAB from a BAC clone from the Mediterranean Sea identified as an uncultivated deep-branching phototrophic sulfur-oxidizing Gammaproteobacteria[30] (Supplementary Information Figure 6). Hence, this *dsr*C-5 AMG clade likely corresponded to viruses infecting these Gammaproteobacteria.

Of the 4 *sox*YZ-containing contigs, 2 had a predicted host in the Bacteroidetes and Alphaproteobacteria groups respectively (Extended Data Table 1). The Bacteroidetes host prediction is likely incorrect, as no bacteria in this phylum are known to be sulfur-oxidizers or carry *sox* genes, and this prediction was obtained from a signal based on nucleotide composition, which can have an error rate of 20% (when the host species is absent from the reference database). The host prediction to Alphaproteobacteria is more likely to be correct because it originates from a blastn sequence similarity, but this similarity is to a metagenomic contig which itself is affiliated with Alphaproteobacteria, so that this host prediction depends on the correct affiliation of the microbial contig. Finally, on the *sox*YZ phylogenetic tree, the viral versions of the gene cluster together at the root of a group including Betaproteobacteria, Gammaproteobacteria, and unclassified Proteobacteria. Thus, taken together, these different results suggest that *sox*YZ-containing contigs likely infect sulfur-oxidizing Proteobacteria, but could not reveal with more certainty the actual host group.

Consistently with their affiliation to multiple VCs, P-II AMG clades were all linked to different host groups. One sequence from each clade P-II-1 and P-II-3 was predicted as infecting a Bacteroidetes host based on nucleotide composition (Extended Data Table 1), and the P-II phylogenetic tree confirmed that these viruses likely infect Bacteroidetes (Extended Data Figure 7). The host of clade P-II-2 was more uncertain: no host prediction was available, its nearest neighbor in the phylogeny was a sequence from a Gammaproteobacteria Single-Amplified Genome (SAG), yet it was found in a deep-branching clade containing mostly Verrucomicrobia. For clade P-II-4, both the nearest neighbor in the phylogenetic tree and a host prediction based on nucleotide composition suggested these viruses infect Gammaproteobacteria hosts. Hence, P-II-containing viruses are likely infecting hosts from at least two different phyla (Bacteroidetes and Proteobacteria), and possibly a third one (Verrucomicrobia).

Finally, no host prediction was possible for the *amo*C-containing contig based on similarity to a prophage, CRISPR spacer, or tetranucleotide composition (Extended Data Table 1). The AmoC phylogeny suggested that the AMG is most closely related with sequences from Thaumarchaeota / marine group I.1a (Extended Data Figure 8). This is consistent with the affiliation of the viral contig as an archaeovirus based on best BLAST hits (see above), as archaeal and bacterial viruses are expected to harbor limited genetic similarity[31]. Hence, this *amo*C-containing virus is likely infecting an uncultivated ammonia-oxidizing Thaumarchaeota.

Supplementary Tables

Supplementary Table 1: List of viromes included in the GOV dataset. For each virome, the corresponding expedition, station number, and depth is indicated. *Tara* Oceans stations are prefixed with "Tara_" and Malaspina stations with an "M". Accession numbers are given for raw reads available in ENA (for *Tara* Oceans samples) and on JGI IMG (for *Malaspina* samples). Longhurst provinces and biomes are defined based on Longhurst[32] and environmental features are defined based on Environment Ontology (http://environmentontology.org/). The total number of reads and bp sequenced, as well as the number of bp mapped to viral contigs within and outside of populations are indicated. *Malaspina stations for which no water mass or basin data are available because these were not included in the previous study[33].

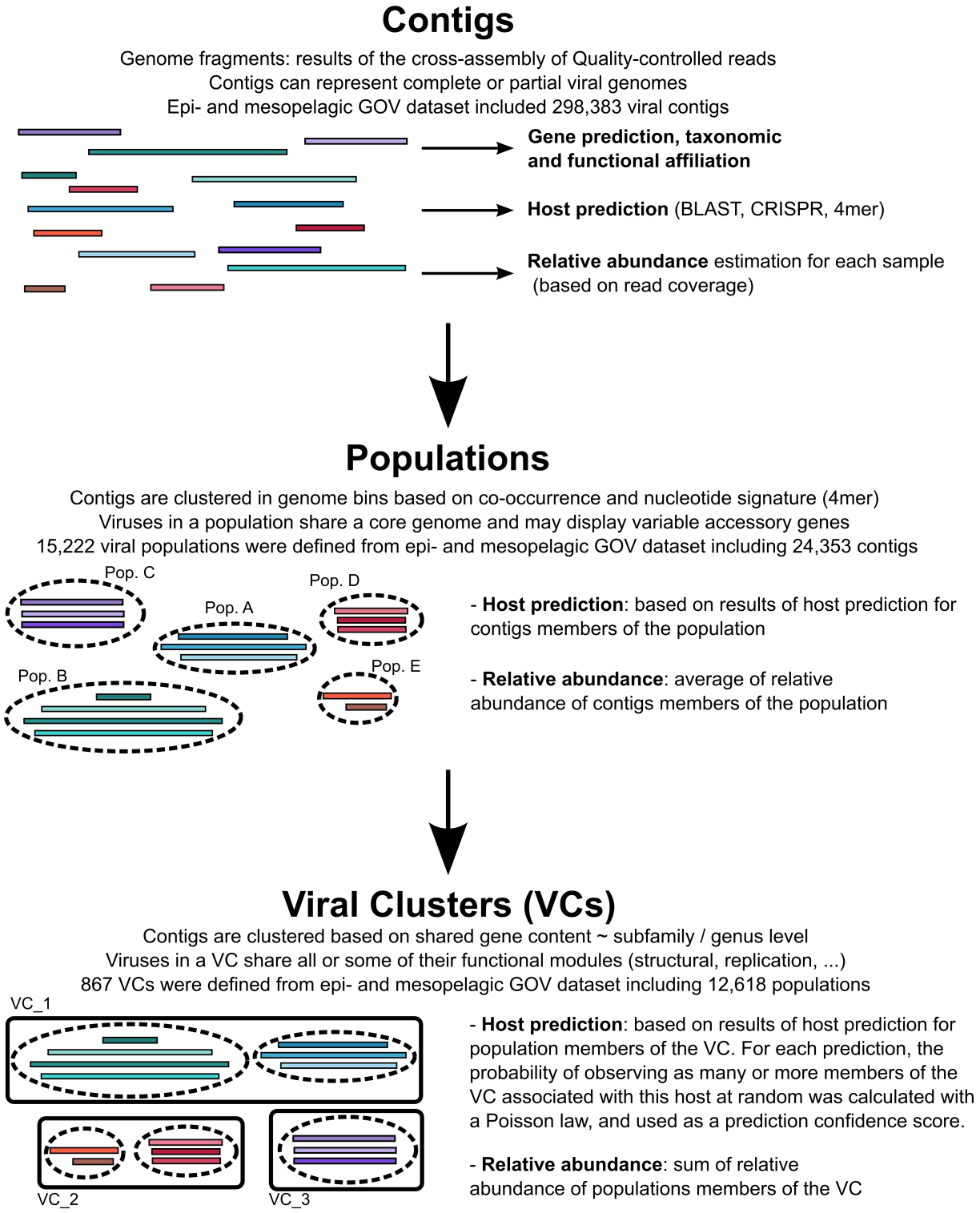
Supplementary Table 2: GOV viral population summary. The number of contigs, total length and length of the largest contig, type of assembly used, and highest normalized coverage across the GOV samples is indicated in the first tab. For populations already identified in the TOV dataset (contigs similar at 95% ANI on ≥50% of their length), the size of the TOV contig is noted. In the second tab, the normalized coverage (average coverage of the population contig(s) normalized by the total sequencing depth of the sample) is indicated as coverage / Gb of metagenome for all GOV samples.

Supplementary Table 3: Summary of Viral Clusters (VCs). The first tab lists, for each VC, the number of members (total, and by dataset, i.e. originating from RefSeq, environmental phages, VirSorter Curated Dataset, and GOV), alongside the affiliation of RefSeq members of the VCs (when available) at the family, subfamily, and genus levels. The second tab includes the cumulative normalized coverage of each VC in each sample (based on the coverage of populations members of the VC), as well as the sum of coverage for the 38 recurrently abundant VCs and all other VCs at the bottom.

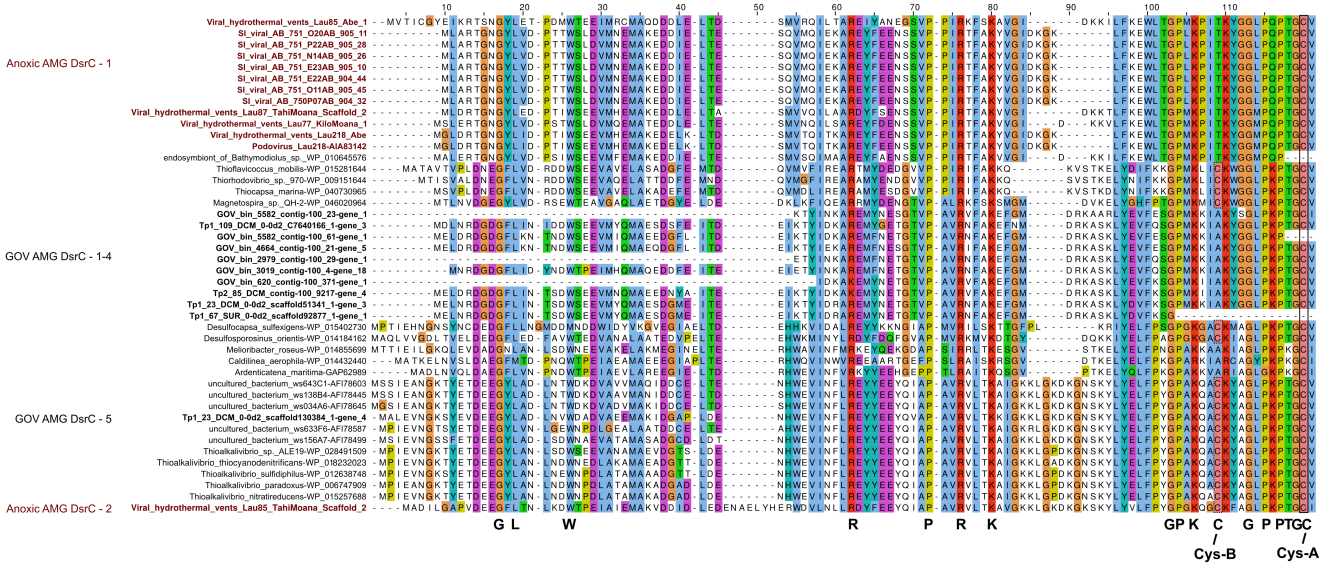
Supplementary Table 4: List of host prediction for GOV viral contigs associated with a population. For each prediction, the type of signal (blastn, CRISPR, tetranucleotide composition), the host sequence used for the prediction alongside its affiliation, and the strength of the prediction (length of the blastn match, number of mismatches in the CRISPR spacer, and distance between viral and host tetranucleotide frequencies vectors) is indicated.

Supplementary Table 5: List of PFAM domains detected in GOV viral contigs. For each PFAM domain, the number of genes detected in the GOV dataset is indicated, alongside the category of the domain (as in [34]). The category "other" category includes PFAM domains with vague descriptions, multiple functions, or regulatory functions.

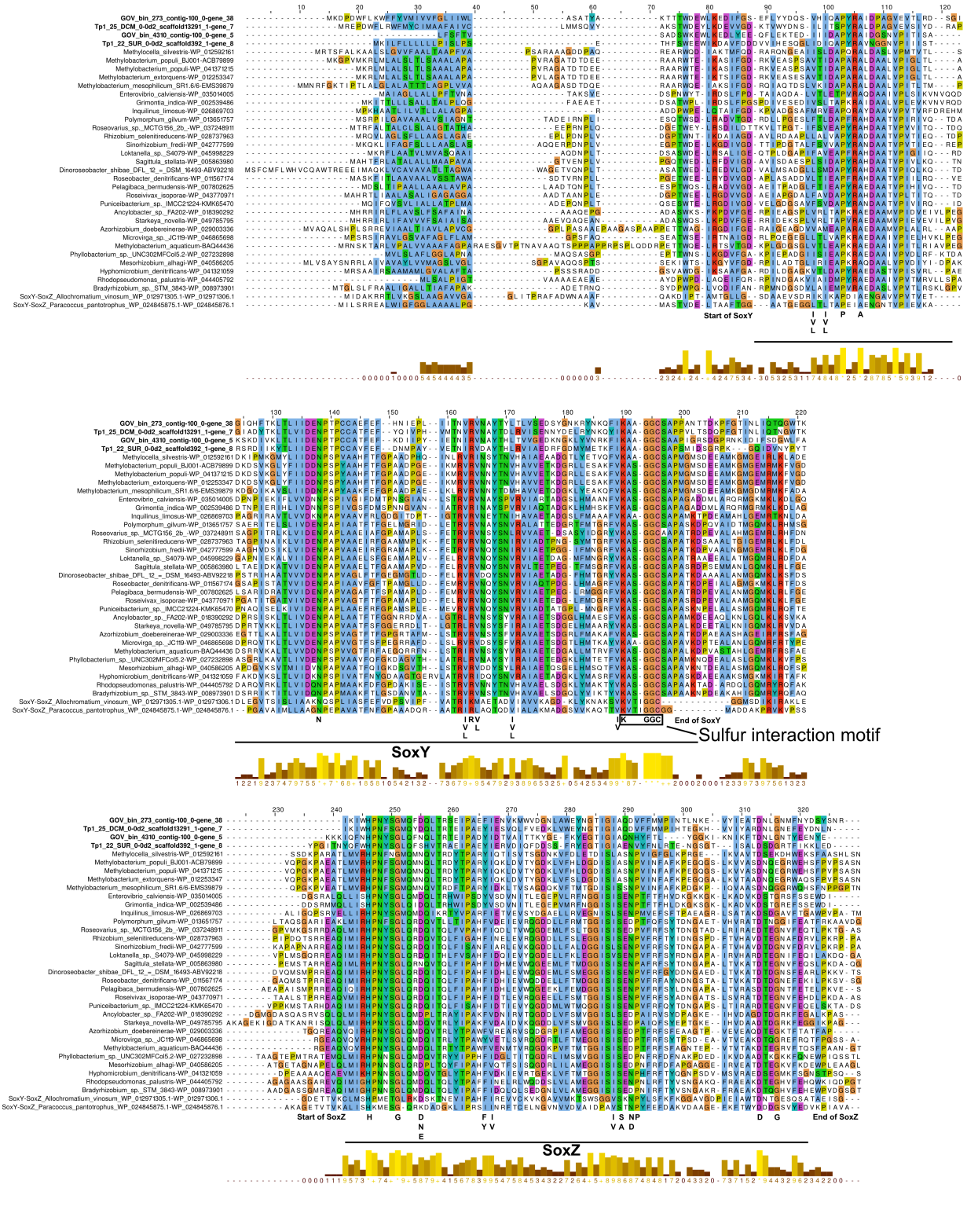
Supplementary Figures



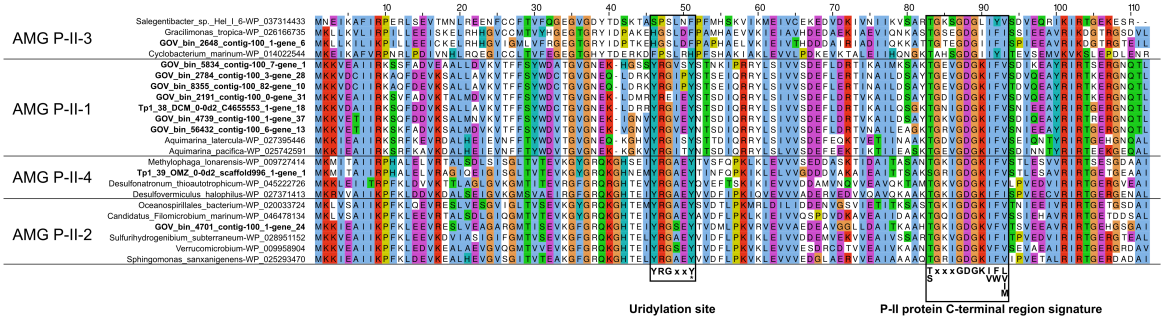
Supplementary Figure 1: Schematic of the different levels of organization used in this study. The base unit is the contig, i.e. assembled genome (fragment). These contigs are gathered (when available) in viral populations, a proxy for viral "species", through genome binning based on co-occurrence and similarity in nucleotide composition. A higher level of organization (VCs, subfamily ~ genus level) is achieved by clustering the contigs based on shared gene content.

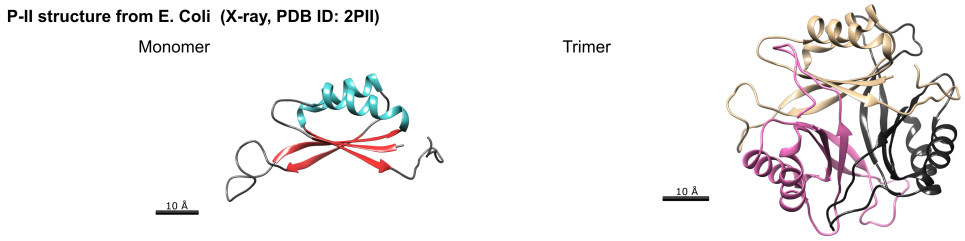


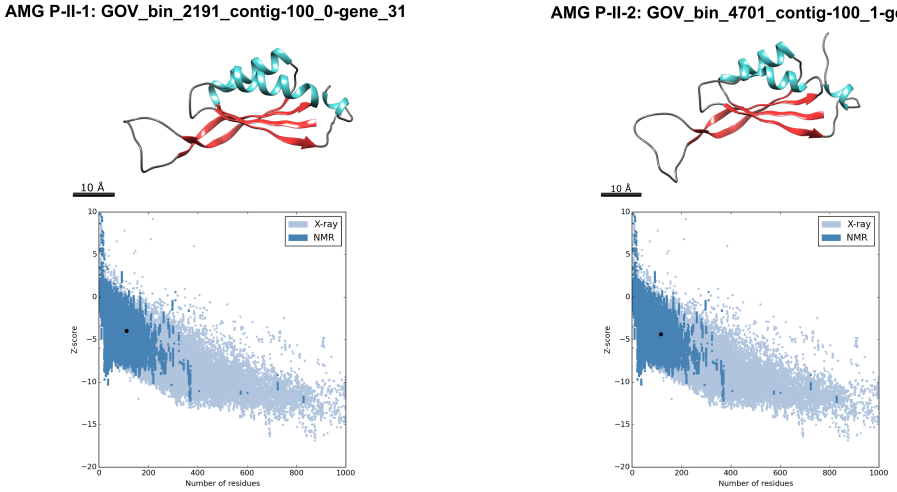
Supplementary Figure 2: Multiple alignment of *dsr*C protein sequences. Conserved residues are indicated below the alignment, and the two conserved C residues representing the active sites of "true *dsr*C" (Cys-B and Cys-A) are named as in[17]. Viral AMGs are highlighted in bold, with previously described anoxic SUP05 viruses sequences in red (from[35,36]) and epipelagic GOV sequences in black.



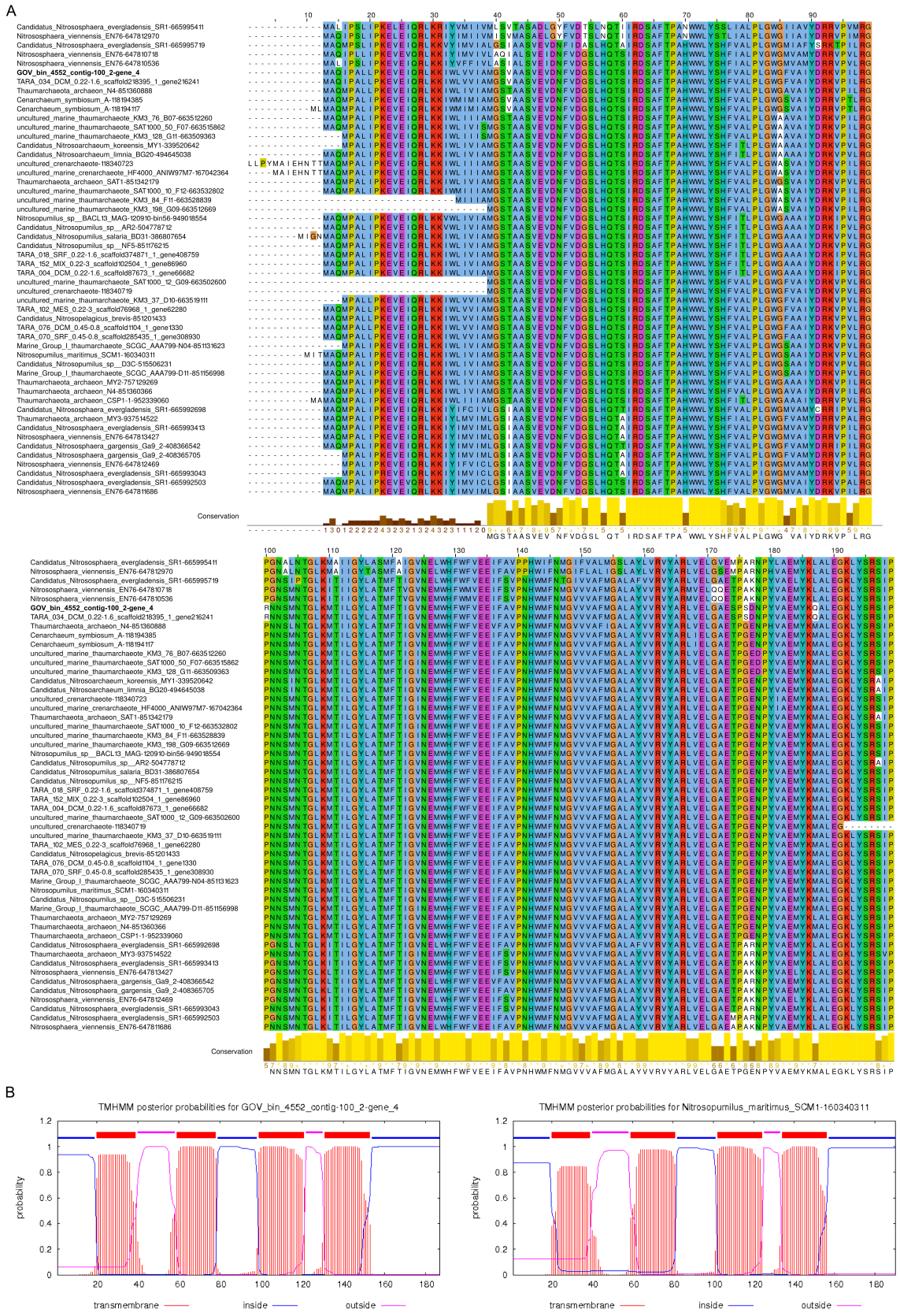
Supplementary Figure 3: Multiple alignment of *sox*YZ protein sequences. Conserved residues are indicated below the alignment for SoxY and SoxZ protein domains, based on the respective PFAM domains (PF13501 and PF08770). Viral AMGs are highlighted in bold.



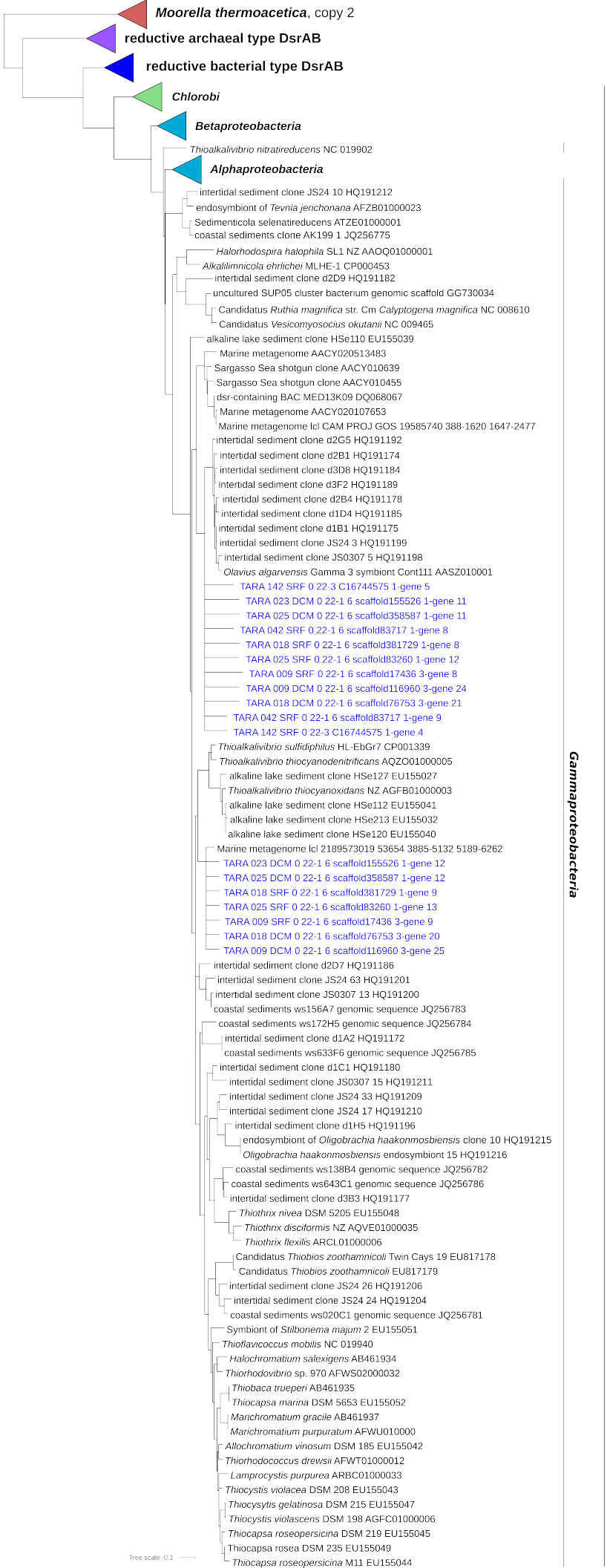




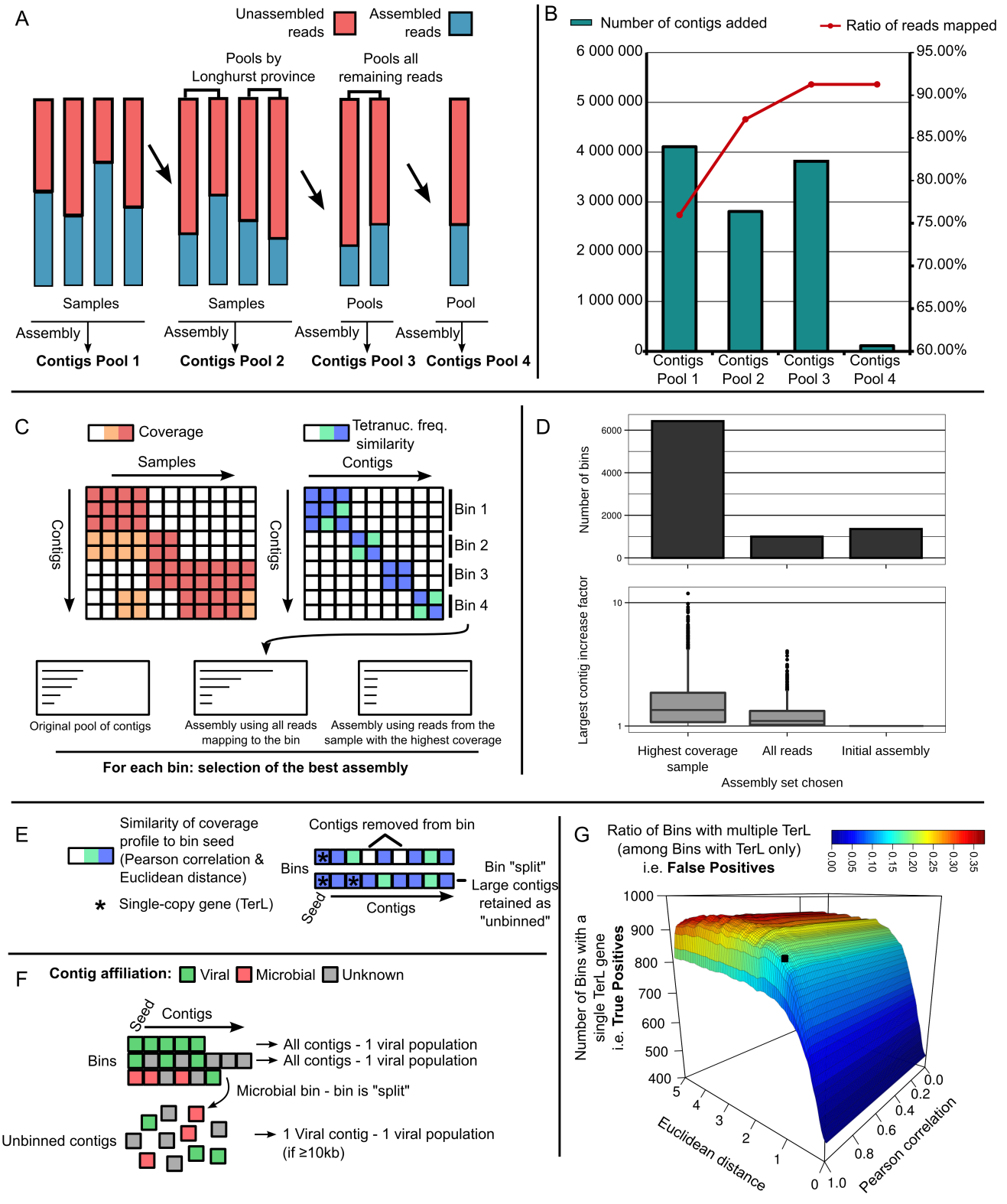
Supplementary Figure 4: Alignment (A) and predicted 3D structures (B) of P-II AMGs. Conserved motifs are indicated below the alignment (PROSITE: PDOC00439). P-II uridylation site is highlighted with a star. Characterized structure (from E. Coli) and predicted 3D conformations are colored according to secondary structures (alpha helix: blue, beta strand: red), except for the trimer structure of E. Coli P-II where each subunit is colored differently. For predicted structures, the model quality as assessed by ProSA[20] is indicated below the model. Viral AMGs are highlighted in bold.



Supplementary Figure 5: Alignment (A) and predicted transmembrane domain (B) of *amo*C AMGs. The viral sequence is highlighted in bold, and conserved residues are indicated below the alignment. Transmembrane domains were predicted with TMHMM[37] for the AMG *amo*C (left), and a reference *amo*C from the ammonia-oxidizing *Nitrosopumilus maritimus* SCM1 (right).



Supplementary Figure 6: Dissimilatory sulfite reductase (*dsr*AB) tree showing the phylogeny of oxidative bacterial type *dsr*AB. Sequences from Tara Ocean microbial metagenomes close to DsrC-5 AMG are indicated in blue and are affiliated with sulfur-oxidizing Gammaproteobacteria. Other phylogenetic groups and *dsr*AB families are collapsed and shown as triangles.



455 **Supplementary Figure 7: Overview and result of the cross assembly, binning, and viral contigs selection process. A.** Iterative assembly viromes. First, for each sample, reads were mapped to the set of contig generated through MOCAT[38]. Reads not assembled (i.e. not mapped to any contigs) were then used in another assembly, using Idba_ud[39]. Unmapped reads after this second round of sample-by-

sample assembly were then pooled by Longhurst province (i.e. all unmapped reads from all samples within one province), and cross-assembled with Idba_ud[40]. Finally, all unmapped reads after this third round of assembly were gathered and assembled with Idba_ud. **B.** Results of the iterative assembly process. For each assembly round, the number of contigs is displayed alongside the cumulated percentage of reads mapped to a contig. **C.** Overview of the binning process. Contigs generated through the iterative assembly were binned based on correlation between their abundance profile and similarities between their tetranucleotide frequency (using Metabat[1]). For each bin, two contig pools (beyond the initial set of contigs) were generated, assembling either all reads mapping to the contig pool, or only reads from the sample in which the bin had the highest coverage (both assemblies computed with Idba_ud). The set of contigs including the largest genome fragment was then kept for each bin. **D.** Results of the re-assembly of bins. For each type of bin assembly (highest coverage sample, all samples, or initial assembly) the number of bins for which this type was selected is indicated on top, with the distribution of increase in length of longest contig at the bottom. **E.** Bin refinement based on abundance profile similarities. For each bin, the abundance profile of each contig was compared to the abundance profile of the bin seed contig (largest contig), and contigs not well correlated to the bin seed were excluded. Bins still displaying multiple TerL gene (single-copy marker gene for viruses) after this bin refinement step were split. **F.** Bin affiliation and viral population definition. Bins were either affiliated as entirely viral and considered as single viral populations, or included non-viral contigs, in which case viral contigs in these bins were considered as "unbinned" and selected as viral population seed if ≥10kb **G.** Selection of thresholds for bin refinement based on abundance profile similarities. Thresholds to exclude contigs from bins based on Euclidean distance and Pearson correlation coefficient between contig abundance profile and bin seed profile were explored, looking for the best compromise between number of true positive (z-axis, number of bins with a single TerL) and number of false negative (in colors, number of bins with multiple TerL). The thresholds combination chosen is indicated with a black square.

References

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

2. Ciccarelli, F. D. *et al.* Toward automatic reconstruction of a highly resolved tree of life. *Science* **311,** 1283–7 (2006).

3. Roux, S., Hallam, S. J., Woyke, T. & Sullivan, M. B. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *Elife* **4,** 1–20 (2015).

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

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

6. Brum, J. *et al.* Patterns and ecological drivers of ocean viral communities. *Science* **348,** 1261498–1–10 (2015).

7. Roux, S., Tournayre, J., Mahul, A., Debroas, D. & Enault, F. Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinformatics* **15,** 1–12 (2014).

8. Krupovic, M. *et al.* Taxonomy of prokaryotic viruses: update from the ICTV bacterial and archaeal viruses subcommittee. *Arch. Virol.* **in press,** 0–4 (2016).

9. Lavigne, R., Seto, D., Mahadevan, P., Ackermann, H.-W. & Kropinski, A. M. Unifying classical and molecular taxonomic classification: analysis of the Podoviridae using BLASTP-based tools. *Res. Microbiol.* **159,** 406–14 (2008).

10. Rohwer, F. & Edwards, R. The Phage Proteomic Tree : a Genome-Based Taxonomy for Phage. *J. Bacteriol.* **184,** 4529–4535 (2002).

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

12. Lima-Mendez, G., Van Helden, J., Toussaint, A. & Leplae, R. Reticulate representation of evolutionary and functional relationships between phage genomes. *Mol. Biol. Evol.* **25,** 762–77 (2008).

13. Sullivan, M. J., Petty, N. K. & Beatson, S. A. Easyfig: a genome comparison visualizer. *Bioinformatics* **27,** 1009–10 (2011).

14. Edwards, R. A., McNair, K., Faust, K., Raes, J. & Dutilh, B. E. Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiol. Rev.* **in press,** (2015).

15. Thompson, L. R. *et al.* Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proc. Natl. Acad. Sci. U. S. A.* **108,** E757–64 (2011).

16. Breitbart, M., Thompson, L. R., Suttle, C. A. & Sullivan, M. B. Exploring the Vast Diversity of

Marine Viruses. *Oceanography* **20,** 135–139 (2007).

17. Venceslau, S. S., Stockdreher, Y., Dahl, C. & Pereira, I. A. C. The "bacterial heterodisulfide" DsrC is a key protein in dissimilatory sulfur metabolism. *Biochim. Biophys. Acta* **1837,** 1148–64 (2014).

18. Quentmeier, A. & Friedrich, C. G. The cysteine residue of the SoxY protein as the active site of protein-bound sulfur oxidation of Paracoccus pantotrophus GB17. *FEBS Lett.* **503,** 168–172 (2001).

19. Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* **5,** 725–38 (2010).

20. Wiederstein, M. & Sippl, M. J. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res.* **35,** W407–10 (2007).

21. Stahl, D. A. & de la Torre, J. R. Physiology and diversity of ammonia-oxidizing archaea. *Annu. Rev. Microbiol.* **66,** 83–101 (2012).

22. Sayavedra-Soto, L. A. *et al.* The membrane-associated monooxygenase in the butane-oxidizing Gram-positive bacterium Nocardioides sp. strain CF8 is a novel member of the AMO/PMO family. *Environ. Microbiol. Rep.* **3,** 390–6 (2011).

23. Schloissnig, S. *et al.* Genomic variation landscape of the human gut microbiome. *Nature* **493,** 45–50 (2013).

24. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348,** 1–10 (2015).

25. Loy, A. *et al.* Reverse dissimilatory sulfite reductase as phylogenetic marker for a subgroup of sulfur-oxidizing prokaryotes. *Environ. Microbiol.* **11,** 289–99 (2009).

26. Müller, A. L., Kjeldsen, K. U., Rattei, T., Pester, M. & Loy, A. Phylogenetic and environmental diversity of DsrAB-type dissimilatory (bi)sulfite reductases. *ISME J.* **9,** 1152–65 (2015).

27. Pelikan, C. *et al.* Diversity analysis of sulfite- and sulfate-reducing microorganisms by multiplex dsrA and dsrB amplicon sequencing using new primers and mock community-optimized bioinformatics. *Environ. Microbiol.* (2015). doi:10.1111/1462-2920.13139

28. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT : a novel method for rapid multiple sequence alignment based on fast Fourier transform. **30,** 3059–3066 (2002).

29. Berger, S. A. & Stamatakis, A. Aligning short reads to reference alignments and trees. *Bioinformatics* **27,** 2068–75 (2011).

30. Sabehi, G. *et al.* New insights into metabolic properties of marine bacteria encoding proteorhodopsins. *PLoS Biol.* **3,** e273 (2005).

31. Krupovic, M., Prangishvili, D., Hendrix, R. W. & Bamford, D. H. Genomics of bacterial and

archaeal viruses: dynamics within the prokaryotic virosphere. *Microbiol. Mol. Biol. Rev.* **75,** 610–35 (2011).

32. Longhurst, A. *Ecological geography of the sea*. (2007).

33. Salazar, G. *et al.* Global diversity and biogeography of deep-sea pelagic prokaryotes. *ISME J.* 1–13 (2015). doi:10.1038/ismej.2015.137

34. Hurwitz, B. L., Brum, J. R. & Sullivan, M. B. Depth-stratified functional and taxonomic niche specialization in the "core" and "flexible" Pacific Ocean Virome. *ISME J.* **9,** 472–84 (2015).

35. Anantharaman, K. *et al.* Sulfur Oxidation Genes in Diverse Deep-Sea Viruses. *Science* **344,** 757–760 (2014).

36. Roux, S. *et al.* Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta- genomics. *Elife* **3,** 1–20 (2014).

37. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305,** 567–80 (2001).

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

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

40. Dutilh, B. E. *et al.* A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* **5,** 1–11 (2014).