Illuminating structural proteins in viral "dark matter" with metaproteomics

Jennifer R. Brum[a,1,2], J. Cesar Ignacio-Espinoza[b,1,3], Eun-Hae Kim[c,1,4], Gareth Trubl[c,2], Robert M. Jones[c,5], Simon Roux[a,2], Nathan C. VerBerkmoes[d,6], Virginia I. Rich[c,2,7], and Matthew B. Sullivan[a,b,c,2,7]

[a]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721; [b]Department of Molecular and Cellular Biology, University of Arizona, Tucson, AZ 85721; [c]Department of Soil, Water and Environmental Science, University of Arizona, Tucson, AZ 85721; and [d]Chemical Science Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831

Edited by Edward F. DeLong, University of Hawaii at Manoa, Honolulu, HI, and approved January 21, 2016 (received for review December 21, 2015)

Viruses are ecologically important, yet environmental virology is limited by dominance of unannotated genomic sequences representing taxonomic and functional "viral dark matter." Although recent analytical advances are rapidly improving taxonomic annotations, identifying functional dark matter remains problematic. Here, we apply paired metaproteomics and dsDNA-targeted metagenomics to identify 1,875 virion-associated proteins from the ocean. Over one-half of these proteins were newly functionally annotated and represent abundant and widespread viral metagenome-derived protein clusters (PCs). One primarily unannotated PC dominated the dataset, but structural modeling and genomic context identified this PC as a previously unidentified capsid protein from multiple uncultivated tailed virus families. Furthermore, four of the five most abundant PCs in the metaproteome represent capsid proteins containing the HK97-like protein fold previously found in many viruses that infect all three domains of life. The dominance of these proteins within our dataset, as well as their global distribution throughout the world's oceans and seas, supports prior hypotheses that this HK97-like protein fold is the most abundant biological structure on Earth. Together, these culture-independent analyses improve virion-associated protein annotations, facilitate the investigation of proteins within natural viral communities, and offer a high-throughput means of illuminating functional viral dark matter.

viruses | marine | proteins

Microorganisms are central to the Earth's ecosystem function (1), and it is becoming increasingly evident that viruses substantially influence microbially driven processes through mortality and manipulation of metabolism via viral-encoded metabolic genes (reviewed in ref. 2), including those involved in photosynthesis (3) and most of central carbon metabolism (4). However, holistic understanding of marine viruses has been limited in part by the dominance of "unknown" genomic sequences encountered when surveying viral communities in nature.

This "viral dark matter" in metagenomes manifests as an inability to obtain functional or taxonomic annotations for most (63–93%) of surveyed sequence space (5), as well as an inability to taxonomically annotate the vast majority (>99%) of viral populations observed in nature (6). Emerging approaches, such as comparison of metagenomes using shared *k*-mers (7), protein clusters (PCs) (8), and viral populations (6), enable ecological inferences without annotation (reviewed in ref. 9), but further conclusions are hindered by most viral PCs and populations remaining unknown. Taxonomic viral dark matter occurs due to limited representation of viruses in reference databases—86% of 1,531 sequenced genomes of bacterial and archaeal viruses were isolated from only 3 of 61 known host phyla (10). Some progress is being made using traditional isolation and genome-sequencing techniques to obtain reference genomes for both abundant (11, 12) and rare, but ubiquitous (13), marine viruses. However, identifying viral genomic information within microbial genomic datasets and using genome- and network-based analytics to classify these previously unidentified sequences is already rapidly increasing the number of available and classified viral reference genome sequences (10). With the emerging deluge of novel and diverse single-cell genomic datasets that contain viruses (14, 15), such methods are likely to uncover viruses for all known phyla in short order, which should presumably greatly illuminate taxonomic viral dark matter.

In contrast, high-throughput advances to resolve our understanding of functional viral dark matter are lagging. Examination of viral genomic sequence space organized into PCs based on similarity has revealed that the global virosphere (the catalog of genes encoded by viruses) is now well sampled in the upper oceans (6) and likely contains less than 3.9 million proteins (16). Although the abundance of viral PCs is becoming well understood, the functions of these PCs remain poorly characterized.

Significance

Marine viruses are abundant and have substantial ecosystem impacts, yet their study is hampered by the dominance of unannotated viral genes. Here, we use metaproteomics and metagenomics to examine virion-associated proteins in marine viral communities, providing tentative functions for 677,000 viral genomic sequences and the majority of previously unknown virion-associated proteins in these samples. The five most abundant protein groups comprised 67% of the metaproteomes and were tentatively identified as capsid proteins of predominantly unknown viruses, all of which putatively contain a protein fold that may be the most abundant biological structure on Earth. This methodological approach is thus shown to be a powerful way to increase our knowledge of the most numerous biological entities on the planet.

Author contributions: E.-H.K., N.C.V., V.I.R., and M.B.S. designed research; E.-H.K., R.M.J., and N.C.V. performed research; S.R. and N.C.V. contributed new reagents/analytic tools; J.R.B., J.C.I.-E., E.-H.K., G.T., R.M.J., and S.R. analyzed data; and J.R.B., J.C.I.-E., E.-H.K., V.I.R., and M.B.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: Sequences, assemblies, annotation, and processed data such as alignments, similarity matrixes, and network files have been deposited in iVirus, mirrors.iplantcollaborative.org/browse/iplant/home/shared/iVirus/TOV_4_metaproteomes. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE Partner Repository, www.ebi.ac.uk/pride/archive/ (identifier PXD000938).

[1]J.R.B., J.C.I.-E., and E.-H.K. contributed equally to this work.

[2]Present addresses: Department of Microbiology, The Ohio State University, Columbus, OH 43210; and Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, OH 43210.

[3]Present address: Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089.

[4]Present address: Roche Tissue Diagnostics, Oro Valley, AZ 85755.

[5]Present address: Cold Regions Research and Engineering Laboratory, Hanover, NH 03755.

[6]Present address: Border Biomedical Research Center, Department of Biological Sciences, The University of Texas at El Paso, El Paso, TX 79968.

[7]To whom correspondence may be addressed. Email: virginia.isabel.rich@gmail.com or mbsulli@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1525139113/-/DCSupplemental.

A promising approach to annotate portions of functional viral dark matter could be to elucidate which predicted proteins encode viral structural components. Computationally, artificial neural networks have been used to predict viral capsid and tail proteins from metagenomic data, which has been validated through in vivo expression and visualization of four putative viral structural genes (17). Experimentally, divergent structural proteins from cultivated viral isolates have been annotated using mass spectrometry (MS)-based proteomics (13, 18–20). Metaproteomics has now emerged as a powerful tool to investigate microbial communities (21, 22), and here we apply this approach to marine viral communities to identify virion-associated proteins and facilitate annotation of the structural components of viral dark matter, generating new insights regarding the structural proteins in natural viral communities.

Results and Discussion

Metaproteomic Datasets for Investigating Wild Marine Viruses. High-throughput experimental MS-based proteomics was applied to four purified marine viral communities from the Mediterranean Sea, Indian Ocean, and Atlantic Ocean (Table S1) collected through the *Tara* Oceans Expedition (23). After using several experimental approaches to generate metaproteomes (Table S2; see experimental overview in Fig. S1), we selected the sample preparation method that minimized keratin contamination and autotryptic peptides [filter-aided sample preparation 2 (FASP2)] and the mass spectrometer that produced the most peptide spectra (LTQ Orbitrap Velos Pro). We then evaluated three analytical search pipelines to compare these MS-derived peptide spectra against assembled contigs from their paired dsDNA viral metagenomes included in the *Tara* Oceans Viromes (TOV) dataset (6) (Fig. S1). Among these pipelines, TPP with X! Tandem enabled the identification of the most spectra, nonredundant proteins (i.e., the distinct nonidentical proteins those spectra represent), and PCs (defined as groups of proteins with 60% similarity across 80% coverage; Table S3). Furthermore, 26% of the total spectra were only identified using the TPP with X! Tandem pipeline, and only 8% of total spectra were not identified using this pipeline (Fig. S2*A*). Finally, the distribution of annotated spectra within the viral functional and taxonomic categories was highly similar among all three pipelines (Fig. S2*B*; Morisita's Index of 1.0 for each pairwise comparison). We thus generated the Quantitative Dataset consisting of the peptide spectral abundances and annotations obtained only from the FASP2 sample preparation method, the LTQ Orbitrap Velos Pro mass spectrometer, and the TPP with X! Tandem pipeline to quantitatively investigate viral protein abundances (Fig. S1).

The Quantitative Dataset consisted of 15,270 spectra representing 697 nonredundant proteins in 296 PCs (Table S3; Dataset S1). The majority (74% of spectral counts) of proteins in this dataset facilitated annotation of previously unannotated virion-associated proteins (i.e., "newly annotated"; Fig. 1). Taxonomically, 24% of the proteins were annotated as belonging to tailed phages (myoviruses, podoviruses, and siphoviruses; Fig. 1). However, there were very few tail proteins in the dataset; among the proteins with previous functional annotations, the majority (23%) were identified as capsid proteins and <1% were identified as tail proteins (Fig. 1), resulting in ∼100-fold more capsid than tail proteins. Two prior proteomic studies of marine phage isolates show that, although all ORFs annotated as tail proteins were detected in the proteomes of myoviruses infecting *Synechococcus* and *Prochlorococcus* (24), five of the nine putative tail proteins were not detected in *Cellulophaga* siphoviruses (13). This suggests that, even in isolates, MS-based proteomic methods may miss tail proteins—presumably due to loss during phage isolation or deficiencies in sample preparation method (i.e., inefficient digestion with trypsin due to limited K/R residues in these specific proteins or excessive digestion due to having too many K/R residues). In this complex community case using metaproteomics, lower conservation of tail proteins relative to capsids may also hamper their identification through annotation using reference databases (see discussion regarding conservation of viral-associated proteins below).

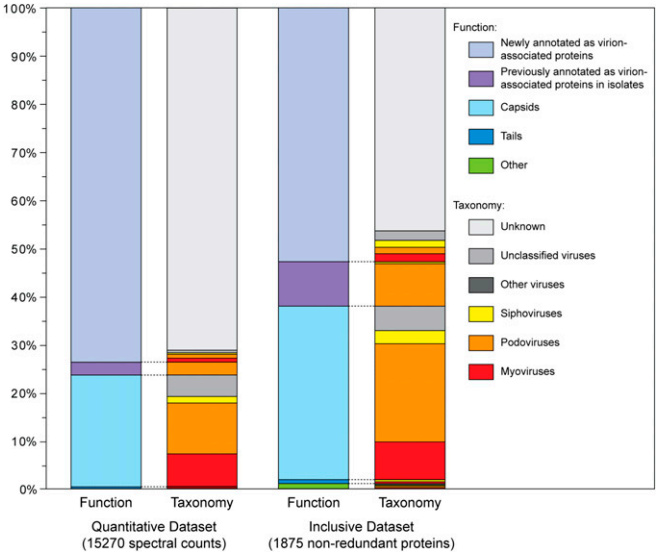

Fig. 1. Virion metaproteomics helps annotate previously unknown viral proteins. Functional annotations and their associated taxonomic annotations (linked by dashed lines) are presented for the Quantitative Dataset based on protein spectral abundances generated from one method and one analytical search pipeline, as well as the Inclusive Dataset that includes all proteins identified using all methods and analytical search pipelines combined. Annotations are based on the top BLASTP match (e value < 0.001) against the viral RefSeq database (full annotation details in Dataset S1). The "Capsids" category includes proteins annotated as head–tail connectors, necks, and portals, whereas the "Other" functional category includes scaffolding proteins and enzymes such as proteases. Hypothetical proteins in genomes of viral isolates are functionally annotated as "Newly annotated" but have a taxonomic annotation.

Collectively, experimentation with two sample preparation methods, three mass spectrometers, and three analytical search pipelines, generated additional peptide spectra beyond the Quantitative Dataset (Fig. S1). Due to the methodological differences, these data could not be combined quantitatively; however, they did provide expanded identification of virion-associated proteins in the four marine viral communities because not all methods identified the same proteins. The resulting Inclusive Dataset (see overview in Fig. S1) contained 1,875 nonredundant proteins grouped into 574 PCs (Table S4), which is ∼2.7- and ∼1.9-fold more proteins and PCs, respectively, than the Quantitative Dataset. Of these proteins, most (991 nonredundant proteins; 53% of the Inclusive Dataset) were again newly identified as virion-associated proteins (Fig. 1), providing functional annotation to 677,376 previously unannotated viral metagenomic reads from these samples, identified here as "structural" based on similarity to peptide spectra using the three analytical search pipelines. The metaproteomes included 176 proteins (9% of the Inclusive Dataset) previously seen in viral isolate experimental proteomes and identified as "viral-associated" or structural (e.g., ref. 13) (Fig. 1). In addition, the metaproteomes provided annotation for 84 previously unannotated hypothetical proteins in viral isolate genomes (4% of the Inclusive Dataset; Fig. 1).

To further examine the utility of metaproteomic analyses in natural viral samples, we first investigated whether the metaproteomes included proteins within the dominant PCs from the paired viral metagenomes. Of the 200 most abundant PCs in the viral metagenomes of each sample, 9% (72 of 800 PCs total) were experimentally detected in the metaproteomic Inclusive Dataset, including 47 PCs that had no prior functional annotation (Fig. 2). We next examined TOV-generated viral populations (i.e., contigs grouped based on similarity of ≥80% of their genes at ≥95% nucleotide identity) (6) for the presence of PCs detected in the metaproteomes. This showed that the metaproteomic PCs in the Inclusive Dataset were detected in viral populations from the

paired viral metagenomes that spanned a large range of population abundances—identifying proteins in the most abundant viral populations, as well as rare populations (Fig. 3*A*). Applying these same analyses to all 5,476 viral populations detected in the larger, globally distributed TOV dataset (6) revealed that metaproteome-detected PCs were found in populations spanning a large range of abundances across as many as 36 of the 43 samples (Fig. 3*B*). Together, this combined information (Figs. 1–3) suggests that metaproteomics is a powerful approach to inform annotation of previously unknown genomic content as structural genes in both isolates and variably abundant populations in natural viral communities.

Dominant Protein Clusters in Viral Metaproteomes. Within the Quantitative Dataset, one PC (CAM_CRCL_773, previously identified in the Global Ocean Sampling expedition, Pacific Ocean Viromes, and TOV datasets) (5, 6, 25) was by far the most abundant, representing 57.5% of spectral counts (Fig. 4*A*). Given this PC's dominance, we applied network analysis to the 400 protein members of this PC in the Inclusive Dataset, which showed two clearly separated groups divergent by ~30% amino acid identity (Fig. 4*B*). Within this PC, only 10 of the 400 constituent proteins were previously annotated (as capsid proteins of siphoviruses JD024 and D3112 that infect *Pseudomonas*), which represented only 1.6% of the PC's spectral counts derived from the Quantitative Dataset (Fig. 4*B* and Dataset S1). This PC thus included the majority (79%) of the previously unannotated spectra in the Quantitative Dataset (Fig. 1). In silico structural modeling of representative sequences from this PC suggested both groups represent major capsid proteins from phages similar to one another (the lambdoid phages HK97, ref. 26, and BPP-1, ref. 27; Fig. 4 *C* and *D*); however, these best fits were relatively weak (template modeling scores, TM scores, lower than the accepted cutoff of 0.5) (28). Thus, this dominant PC appears to be a major capsid protein of previously unexplored marine viruses.

The next four most abundant PCs in the Quantitative Dataset contained a total of 9.8% of the spectral counts (Fig. 4*A*) and were predominantly annotated as capsid proteins by sequence similarity (Dataset S1) and structural modeling (Fig. S3) of their total ORFs present within the Inclusive Dataset. The most abundant of these four PCs, CAM_CRCL_625, was a T4-like major capsid protein by consensus annotation of the PC's component ORFs (29) and also by structural modeling (30). Moving in order of decreasing spectral abundance, PCs CAM_CRCL_14716 and TARA_183056 were both functionally and taxonomically unannotated by sequence similarity; however, by structural modeling, both had best fits to a capsid protein of cyanophage Syn5 (31), although the TM score for the latter PC was below the recommended cutoff of 0.5 (28). Finally, PC TARA_207964 was annotated as a capsid protein from phage HMO-2011 (which infects Ca. *Puniceispirillum marinum* of the SAR116 clade) (11) by similarity, but was annotated as the major capsid protein of cyanophage P-SSP7 (32) by structural modeling, likely because there is currently no reference structure available in the modeling database for phage HMO-2011. Collectively, this combination of ORF annotation and structural modeling thus suggested that, of the top five most abundant PCs (which comprised approximately two-thirds of the spectra in the Quantitative Dataset), at least four were capsid proteins. This is consistent with the dominance of capsids in the annotated portion of the metaproteomes (Fig. 1), and with our understanding of virion structural proteins usually being dominated by capsid proteins in proteomes of viral isolates (13, 24).

We next sought to examine the global-scale distribution of these five most abundant metaproteome-detected PCs, by examining their presence in previously-identified TOV viral populations (6).

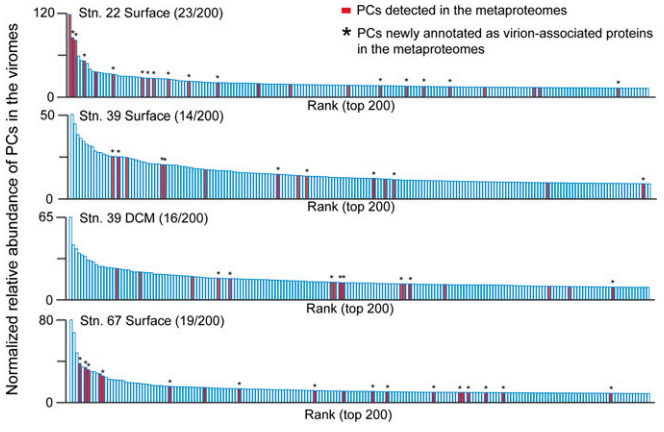

Fig. 2. Viral metagenomes place metaproteomics data into broader community context. Rank abundance of the 200 most abundant PCs in each sample's viral metagenome (by normalized metagenome read counts). PCs detected in the metaproteomic Inclusive Dataset are identified in red, and those that were thereby newly annotated as virion-associated proteins are denoted with asterisks. The number of metaproteome-detected PCs is reported in parentheses for each sample.

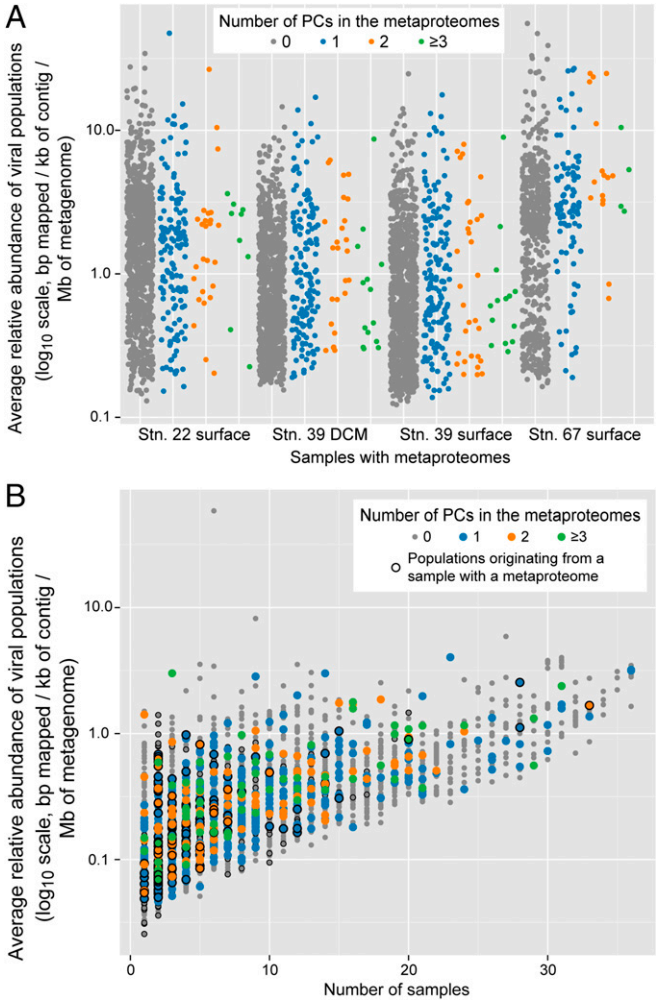

Fig. 3. Metaproteome-detected PCs are found in viral populations with a range of abundances and spatial distributions. Viral populations are from the TOV dataset (6). (*A*) Abundance of viral populations containing metaproteome-detected PCs (using the Inclusive Dataset) in the four samples with metaproteomes. (*B*) Abundance of viral populations versus the number of samples in which they are detected, with populations containing metaproteome-detected PCs (using the Inclusive Dataset) indicated. Populations originating from a sample with a metaproteome are also indicated, with "originating" defined as that population having the maximum coverage in one of those four samples from which metaproteomes were generated.

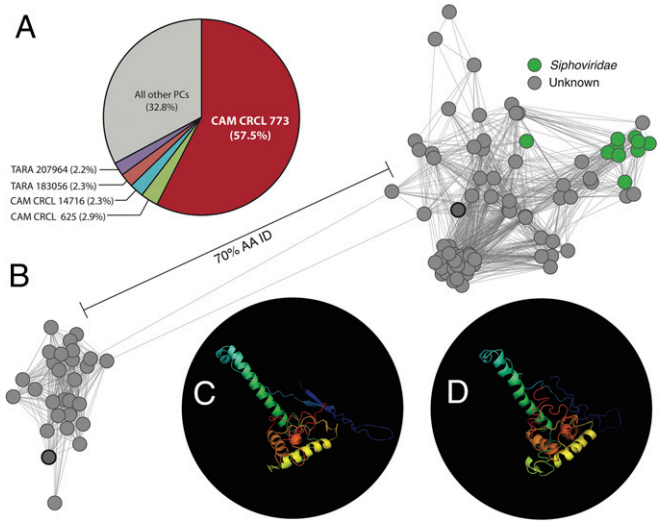

Fig. 4. Investigating a dominant unknown protein cluster. (*A*) Percentage of spectral counts within each PC from the Quantitative Dataset, focusing on the five most abundant PCs. (*B*) Network diagram showing amino acid sequence diversity and taxonomic affiliation for the 400 protein members in the Inclusive Dataset comprising the most abundant PC, CAM_CRCL_773. Line thickness (edge weights) correspond to amino acid identity, calculated as the number of identical residues within the alignment. Proteins used for structural modeling (*C* and *D*) are outlined in thick black. Taxonomic affiliation based on PC annotation is indicated by color. (*C*) Representative structural model for the group of amino acid sequences on the *Left* of the network diagram using the I-TASSER prediction server. The best-fit template was major capsid protein 2FS3 (C score, −4.81; 24% identity; TM score, 0.10) of Enterobacteria phage HK97, which infects *Escherichia coli*. (*D*) Representative structural model for the group of amino acid sequences on the *Right* of the network diagram using the I-TASSER prediction server. The best-fit template was major capsid protein 3J4U (C score, −2.16; 24% identity; TM score, 0.43) of the HK97-like *Bordetella* phage BPP-1.

The dominant metaproteome-detected PC (CAM_CRCL_773) was present in a total of 93 viral populations collectively found in every TOV sample across seven oceans and seas (Fig. 5). In contrast, the four next most abundant PCs were present in substantially fewer populations and showed somewhat more restricted geographic distributions. One PC (TARA_183056) was found in 10 populations that were present in every oceanic region examined except the Southern Ocean. Two PCs (CAM_CRCL_625 and TARA_207964) were found in 5 and 11 viral populations, respectively, predominantly present only in the Indian and Atlantic Oceans, and the Mediterranean and Red Seas. Finally, one PC (CAM_CRCL_14716) was present in only one viral population that showed the most geographic restriction, with the highest abundance from the Indian Ocean, where two of the four metaproteomic samples were collected, but low or nonexistent abundance in the remaining locations. Thus, the five most abundant PCs in the four metaproteomes from three stations are present in viral populations with both widespread and regionally restricted distributions.

Conservation of Virion-Associated Proteins. Conservation of structural similarity in viral capsid proteins, even in the absence of nucleotide sequence similarities, has long been recognized (33, 34). It is thus notable that the model-predicted structural similarities of the five most abundant PCs in the Quantitative Dataset (Fig. 4*A*) are to capsid proteins that all contain the HK97-like fold, including siphophage HK97, HK97-like phage BPP-1, myophage T4, podophage Syn5, and siphophage P-SSP7 (Fig. 4 *C* and *D* and Fig. S3) (27, 30, 31, 34). This HK97-like capsid protein fold has been found in viruses infecting organisms from all three domains of life (35) and is suggested to be the most abundant biological structure on Earth, based on the high abundance of total viruses (e.g., refs. 30, 34, and 36). The data presented here support that assertion: not only do the most abundant PCs in the metaproteomes (representing 67% of the Quantitative Dataset; Fig. 4) seem to contain this protein fold, four out of five of these PCs also appear widely distributed in the upper oceans as shown in our analysis of the TOV viral populations (Fig. 5).

To further investigate conservation in virion-associated proteins, selective constraints of the PCs from the Inclusive Dataset were examined using the ratio of nonsynonymous to synonymous polymorphisms (pN/pS), which has proven powerful for analysis of microbial metagenomic datasets (37, 38). Average pN/pS ratios for PCs in the metaproteome were significantly lower than those determined for all viral metagenome-derived PCs (0.67 vs. 0.84; $P < 0.001$, Mann–Whitney U test; Fig. 6). For comparison, viral metagenome PCs previously annotated as capsids also had relatively low pN/pS ratios (average, 0.48), whereas ratios for annotated tail proteins were higher (average, 0.69). Together, this information suggests stronger overall negative selection for virion-associated

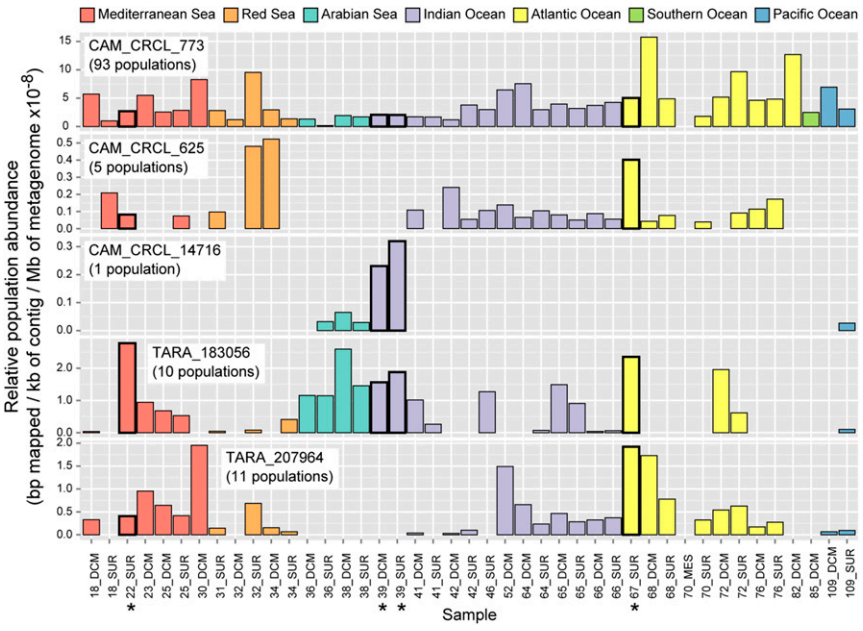

Fig. 5. Global distributions of viral populations containing abundant metaproteome PCs. Relative abundance of all TOV viral populations that contain the five most abundant metaproteome-detected PCs (one PC per panel, in order of decreasing abundance; Fig. 4*A*; one sample per bar with the seven oceans and seas distinguished by color). The total number of TOV viral populations containing each PC is reported in parentheses after the PC name. Sample names include *Tara* Oceans station number and depth category (DCM, deep chlorophyll maximum; Mes, mesopelagic; SUR, surface). Samples used for metaproteomics in this study are indicated by asterisks and have bold black outlines on the bar graphs. Note differing *y*-axes scales. See Fig. S4 for a map of all station locations.

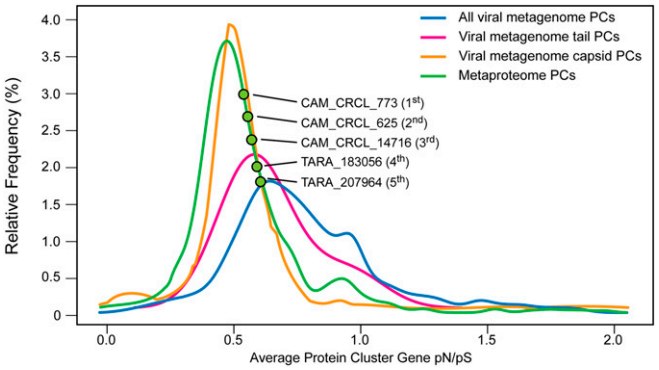

Fig. 6. Selective constraints in viral-associated proteins. Distribution of average pN/pS ratios for all viral metagenome PCs (blue), viral metagenome PCs with consensus annotations as either tails (pink) or capsids (orange), and experimentally detected PCs in the metaproteome (green; from the Inclusive Dataset). Green dots highlight the position of the five most abundant PCs in the Quantitative Dataset, with their order of abundance indicated in parentheses (Fig. 4*A*).

proteins (i.e., increased maintenance of their gene sequences), especially capsid proteins, relative to other viral genome-encoded proteins. This is analogous to previous observations of conservation in housekeeping genes in microorganisms (e.g., ref. 38) and underscores the importance of capsid protein structure maintenance to virion fitness.

Genomic Context for Experimentally Detected Viral Proteins. Genomic context frequently improves gene-specific functional and taxonomic interpretations. We thus examined the genomic context of the five most abundant metaproteome-detected PCs via their five longest associated contigs per PC in the TOV dataset (Fig. 7 and Dataset S2). The most abundant PC (CAM_CRCL_773) was present in contigs where few (24–29%) ORFs were annotated and also showed no taxonomic consensus, the latter of which is consistent with >99% of TOV viral populations (6). However, this genomic context did show that CAM_CRCL_773 was present within a genomic region containing ORFs encoding for a tail fiber, baseplate, and a terminase, as well as three additional unannotated PCs that were also detected in the metaproteome. Within these contigs, the presence of two tail genes and the significant similarities to tailed virus genes for the majority (90–100%) of annotated ORFs indicates that this dominant PC may belong to previously-unidentified *Caudovirales*.

In contrast, the second most abundant PC (CAM_CRCL_625) was present in contigs that were predominantly taxonomically annotated (58–100% of their ORFs), mainly as genes of *Myoviridae* infecting highly abundant hosts such as *Pelagibacter*, *Synechococcus*, and *Prochlorococcus* (Fig. 7 and Dataset S2). This PC was again found within a genomic region containing multiple tail and capsid proteins and two terminase subunits. Collectively, this genomic context combined with the sequence-based and structural modeling-based annotations (above) provides strong evidence that CAM_CRCL_625 is a capsid protein of myoviruses.

The third and fourth most abundant PCs (CAM_CRCL_14716 and TARA_183056) were found in predominantly unannotated contigs (11–29% of ORFs annotated; Fig. 7; Dataset S2). The former (CAM_CRCL_14716) was present in only one TOV contig, consistent with its more restricted geographic distribution (Fig. 5). Although the annotations present in both of these PCs' contigs did not allow taxonomic consensus to be reached, each PC occurred within genomic regions containing other metaproteome-detected PCs. Furthermore, the genomic context for TARA_183056 included a terminase gene as well tail fiber genes, suggesting it may belong to another unidentified *Caudovirales*.

Finally, the fifth most abundant PC (TARA_207964) was present in predominantly annotated contigs (57–78% annotated ORFs) in which the consensus taxonomy (56–91%) was podophage HMO-2011, a phage infecting a SAR116 bacterium (11) (Fig. 7 and Dataset S2). This matches this PC's annotation reported above via its component metagenomic ORFs. This PC was also present in a well-annotated genomic region that included a metaproteome-detected PC (TARA_40991) annotated as a portal protein, supporting the annotation of this PC (TARA_207964) as capsid protein of podophage HMO-2011.

Conclusions

In summary, this study establishes environmental metaproteomics as a high-throughput strategy for shedding light on viral dark matter in two ways: (*i*) defining formerly unannotated proteins as structural, and (*ii*) revealing which of these proteins are most abundant thereby focusing further inquiry (e.g., structural modeling). The 1,875 viral proteins observed in these metaproteomes allowed us to newly annotate 991 proteins as primarily structural. Surprisingly, the majority (67%) of the metaproteomic spectra were derived from just five environmentally dominant PCs. With a combination of sequence- and structural modeling-based annotation, these PCs are now predominantly identified as putative capsid proteins of tailed viruses containing the most abundant biological structure on Earth, the HK97-like protein fold. Furthermore, analysis of metaproteomic PCs facilitated understanding of increased selective pressures on genes encoding virion-associated proteins (e.g., capsids). Although this study focused on dsDNA viruses, the approach is generalizable to ssDNA and RNA viruses,

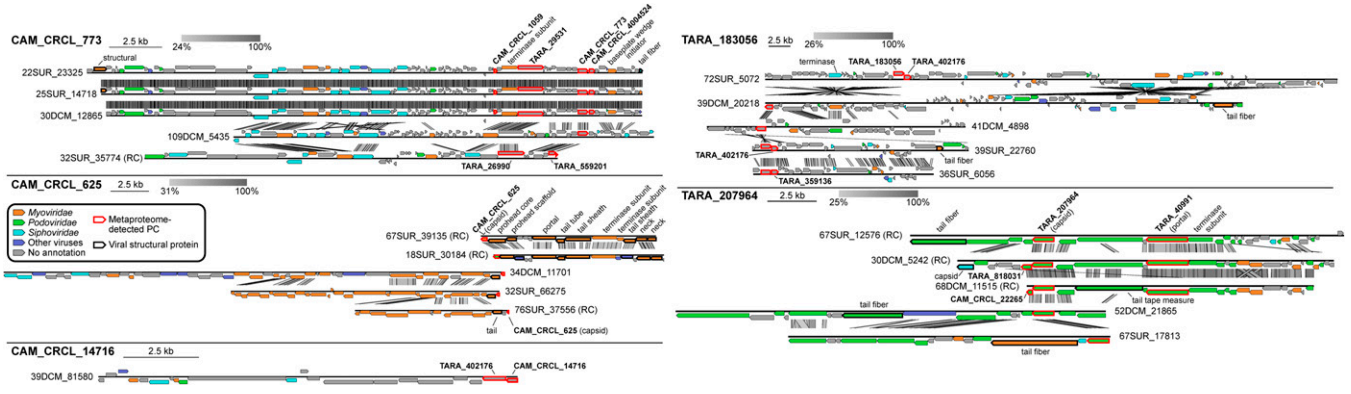

Fig. 7. Genomic context for abundant metaproteome-derived PCs. Genome maps are presented for the most abundant TOV contigs that contain the five most abundant metaproteome-detected PCs in the Quantitative Dataset (Fig. 4*A*). Boxes represent ORFs within contigs, with metaproteome-detected PCs outlined in thick red lines and other viral structural proteins in thick black lines. Contig names include *Tara* Oceans station number, depth (DCM, deep chlorophyll maximum; SUR, surface), and a unique numeral identifier; RC indicates reverse complement of the contig. Broad taxonomic annotations are indicated by the color of the boxes, and selected functional annotations are presented; see Dataset S2 for full annotations.

which currently require generation of separate metagenomes. Thus, this large-scale annotation strategy and the findings presented here will help guide the experimentation needed to refine structural annotations and offer glimpses of the viral metagenomic dark matter that obfuscates our understanding of the most abundant biological entities on Earth: viruses.

Methods

A detailed description of all metaproteomic, metagenomic, and bioinformatic procedures is provided in *SI Methods*.

ACKNOWLEDGMENTS. We thank Bonnie Poulos for preparing viral concentrates, Genoscope for viral metagenomic sequencing, members of Tucson Marine Phage Lab for comments on the manuscript, and University Information Technology Services Research Computing Group and the Arizona Research Laboratories Biotechnology Computing for High-Performance Computing Cluster access and support. We thank Kristen Corrier and Manesh Shah of University of Tennessee/Oak Ridge National Laboratory for efforts in filter-aided sample preparation (FASP) preparation of viral samples and MS analyses, and aspects of proteome informatics, respectively. The four viral concentrates were collected as part of exceptional commitment by scientists and sponsors who made the *Tara* Oceans expedition possible [full list in Brum et al. (6)]. Funding specific to this project was provided by a Ford Foundation Postdoctoral Fellowship (to E.-H.K.), the Gordon and Betty Moore Foundation through Grants GBMF2631 and GBMF3790 (to M.B.S.), and a grant to the UA Ecosystem Genomics Institute through the UA Technology and Research Initiative Fund and the Water, Environmental and Energy Solutions Initiative (to M.B.S. and V.I.R.). This article is contribution 35 of the *Tara* Oceans Expedition 2009–2012.

1. Falkowski PG, Fenchel T, Delong EF (2008) The microbial engines that drive Earth's biogeochemical cycles. *Science* 320(5879):1034–1039.
2. Suttle CA (2007) Marine viruses–major players in the global ecosystem. *Nat Rev Microbiol* 5(10):801–812.
3. Sullivan MB, et al. (2006) Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biol* 4(8):e234.
4. Hurwitz BL, Hallam SJ, Sullivan MB (2013) Metabolic reprogramming by viruses in the sunlit and dark ocean. *Genome Biol* 14(11):R123.
5. Hurwitz BL, Sullivan MB (2013) The Pacific Ocean virome (POV): A marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLoS One* 8(2):e57355.
6. Brum JR, et al. (2015) Patterns and ecological drivers of ocean viral communities. *Science* 348(6237):1261498.
7. Hurwitz BL, Westveld AH, Brum JR, Sullivan MB (2014) Modeling ecological drivers in marine viral communities using comparative metagenomics and network analyses. *Proc Natl Acad Sci USA* 111(29):10714–10719.
8. Hurwitz BL, Brum JR, Sullivan MB (2015) Depth-stratified functional and taxonomic niche specialization in the "core" and "flexible" Pacific Ocean Virome. *ISME J* 9(2):472–484.
9. Brum JR, Sullivan MB (2015) Rising to the challenge: Accelerated pace of discovery transforms marine virology. *Nat Rev Microbiol* 13(3):147–159.
10. Roux S, Hallam SJ, Woyke T, Sullivan MB (2015) Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *eLife* 4:e08490.
11. Kang I, Oh H-M, Kang D, Cho J-C (2013) Genome of a SAR116 bacteriophage shows the prevalence of this phage type in the oceans. *Proc Natl Acad Sci USA* 110(30):12343–12348.
12. Zhao Y, et al. (2013) Abundant SAR11 viruses in the ocean. *Nature* 494(7437):357–360.
13. Holmfeldt K, et al. (2013) Twelve previously unknown phage genera are ubiquitous in global oceans. *Proc Natl Acad Sci USA* 110(31):12798–12803.
14. Labonté JM, et al. (2015) Single-cell genomics-based analysis of virus-host interactions in marine surface bacterioplankton. *ISME J* 9(11):2386–2399.
15. Roux S, et al. (2014) Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-genomics. *eLife* 3:e03125.
16. Ignacio-Espinoza JC, Solonenko SA, Sullivan MB (2013) The global virome: Not as big as we thought? *Curr Opin Virol* 3(5):566–571.
17. Seguritan V, et al. (2012) Artificial neural networks trained to detect viral and phage structural proteins. *PLoS Comput Biol* 8(8):e1002657.
18. Lavigne R, Ceyssens P-J, Robben J (2009) Phage proteomics: Applications of mass spectrometry. *Bacteriophages: Methods and Protocols, Volume 2: Molecular and Applied Aspects*, eds Clokie MRJ, Kropinski AM (Humana, New York), pp 239–251.
19. Allen MJ, Howard JA, Lilley KS, Wilson WH (2008) Proteomic analysis of the EhV-86 virion. *Proteome Sci* 6:11.
20. Sullivan MB, et al. (2009) The genome and structural proteome of an ocean siphovirus: A new window into the cyanobacterial "mobilome." *Environ Microbiol* 11(11):2935–2951.
21. Hettich RL, Sharma R, Chourey K, Giannone RJ (2012) Microbial metaproteomics: Identifying the repertoire of proteins that microorganisms use to compete and cooperate in complex environmental communities. *Curr Opin Microbiol* 15(3):373–380.
22. VerBerkmoes NC, Denef VJ, Hettich RL, Banfield JF (2009) Systems biology: Functional analysis of natural microbial consortia using community proteomics. *Nat Rev Microbiol* 7(3):196–205.
23. Karsenti E, et al.; Tara Oceans Consortium (2011) A holistic approach to marine ecosystems biology. *PLoS Biol* 9(10):e1001177.
24. Sullivan MB, et al. (2010) Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. *Environ Microbiol* 12(11):3035–3056.
25. Yooseph S, et al. (2007) The Sorcerer II Global Ocean Sampling expedition: Expanding the universe of protein families. *PLoS Biol* 5(3):e16.
26. Gan L, et al. (2006) Capsid conformational sampling in HK97 maturation visualized by X-ray crystallography and cryo-EM. *Structure* 14(11):1655–1665.
27. Zhang X, et al. (2013) A new topology of the HK97-like fold revealed in *Bordetella* bacteriophage by cryoEM at 3.5 A resolution. *eLife* 2:e01299.
28. Yang J, et al. (2015) The I-TASSER Suite: Protein structure and function prediction. *Nat Methods* 12(1):7–8.
29. Tétart F, et al. (2001) Phylogeny of the major head and tail genes of the wide-ranging T4-type bacteriophages. *J Bacteriol* 183(1):358–366.
30. Fokine A, et al. (2005) Structural and functional similarities between the capsid proteins of bacteriophages T4 and HK97 point to a common ancestry. *Proc Natl Acad Sci USA* 102(20):7163–7168.
31. Gipson P, et al. (2014) Protruding knob-like proteins violate local symmetries in an icosahedral marine virus. *Nat Commun* 5:4278.
32. Liu X, et al. (2010) Structural changes in a marine podovirus associated with release of its genome into *Prochlorococcus*. *Nat Struct Mol Biol* 17(7):830–836.
33. Bamford DH, Grimes JM, Stuart DI (2005) What does structure tell us about virus evolution? *Curr Opin Struct Biol* 15(6):655–663.
34. Veesler D, Cambillau C (2011) A common evolutionary origin for tailed-bacteriophage functional modules and bacterial machineries. *Microbiol Mol Biol Rev* 75(3):423–433.
35. Pietilä MK, et al. (2013) Structure of the archaeal head-tailed virus HSTV-1 completes the HK97 fold story. *Proc Natl Acad Sci USA* 110(26):10604–10609.
36. Morais MC, et al. (2005) Conservation of the capsid structure in tailed dsDNA bacteriophages: The pseudoatomic structure of ϕ29. *Mol Cell* 18(2):149–159.
37. Simmons SL, et al. (2008) Population genomic analysis of strain variation in *Leptospirillum* group II bacteria involved in acid mine drainage formation. *PLoS Biol* 6(7):e177.
38. Schloissnig S, et al. (2013) Genomic variation landscape of the human gut microbiome. *Nature* 493(7430):45–50.
39. Pesant S, et al.; Tara Oceans Consortium Coordinators (2015) Open science resources for the discovery and analysis of Tara Oceans data. *Sci Data* 2:150023.
40. John SG, et al. (2011) A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environ Microbiol Rep* 3(2):195–202.
41. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F (2009) Laboratory procedures to generate viral metagenomes. *Nat Protoc* 4(4):470–483.
42. Patel A, et al. (2007) Virus and prokaryote enumeration from planktonic aquatic environments by epifluorescence microscopy with SYBR Green I. *Nat Protoc* 2(2):269–276.
43. Hurwitz BL, Deng L, Poulos BT, Sullivan MB (2013) Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. *Environ Microbiol* 15(5):1428–1440.
44. Luo R, et al. (2012) SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *Gigascience* 1(1):18.
45. Kultima JR, et al. (2012) MOCAT: A metagenomics assembly and gene prediction toolkit. *PLoS One* 7(10):e47656.
46. Hyatt D, et al. (2010) Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119.
47. Edgar RC (2004) MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113.
48. Lee WP, et al. (2014) MOSAIK: A hash-based algorithm for accurate next-generation sequencing short-read mapping. *PLoS ONE* 9(3):e90581.
49. Shannon P, et al. (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res* 13(11):2498–2504.
50. Wiśniewski JR, Zougman A, Nagaraj N, Mann M (2009) Universal sample preparation method for proteome analysis. *Nat Methods* 6(5):359–362.
51. Washburn MP, Wolters D, Yates JR, 3rd (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* 19(3):242–247.
52. Eng JK, McCormack AL, Yates JR (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 5(11):976–989.
53. Tabb DL, McDonald WH, Yates JR, 3rd (2002) DTASelect and Contrast: Tools for assembling and comparing protein identifications from shotgun proteomics. *J Proteome Res* 1(1):21–26.
54. Käll L, Canterbury JD, Weston J, Noble WS, MacCoss MJ (2007) Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods* 4(11):923–925.
55. Deutsch EW, et al. (2010) A guided tour of the Trans-Proteomic Pipeline. *Proteomics* 10(6):1150–1159.
56. Craig R, Beavis RC (2004) TANDEM: Matching proteins with tandem mass spectra. *Bioinformatics* 20(9):1466–1467.
57. Kessner D, Chambers M, Burke R, Agus D, Mallick P (2008) ProteoWizard: Open source software for rapid proteomics tools development. *Bioinformatics* 24(21):2534–2536.
58. Mondav R, et al. (2014) Discovery of a novel methanogen prevalent in thawing permafrost. *Nat Commun* 5:3212.
59. Vizcaíno JA, et al. (2014) ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol* 32(3):223–226.
60. Zhang Y (2008) I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* 9:40.
61. Wilkinson L (2012) Exact and approximate area-proportional circular Venn and Euler diagrams. *IEEE Trans Vis Comput Graph* 18(2):321–331.
62. R Core Team (2012) R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna).
63. Schlitzer R (2011) Ocean Data View. Version 4.4.4. Available at odv.awi.de, 2011.