Modeling ecological drivers in marine viral communities using comparative metagenomics and network analyses

Bonnie L. Hurwitz[a,1], Anton H. Westveld[b,c], Jennifer R. Brum[a], and Matthew B. Sullivan[a,1]

[a]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721; [b]Research School of Finance, Actuarial Studies and Applied Statistics, College of Business and Economics, Australian National University, Canberra, ACT 0200, Australia; and [c]Statistics Laboratory at the Bio5 Institute and Statistics Graduate Interdisciplinary Program, University of Arizona, Tucson, AZ 85721

Edited by David M. Karl, University of Hawaii, Honolulu, HI, and approved June 16, 2014 (received for review October 21, 2013)

Long-standing questions in marine viral ecology are centered on understanding how viral assemblages change along gradients in space and time. However, investigating these fundamental ecological questions has been challenging due to incomplete representation of naturally occurring viral diversity in single gene- or morphology-based studies and an inability to identify up to 90% of reads in viral metagenomes (viromes). Although protein clustering techniques provide a significant advance by helping organize this unknown metagenomic sequence space, they typically use only ~75% of the data and rely on assembly methods not yet tuned for naturally occurring sequence variation. Here, we introduce an annotation- and assembly-free strategy for comparative metagenomics that combines shared k-mer and social network analyses (regression modeling). This robust statistical framework enables visualization of complex sample networks and determination of ecological factors driving community structure. Application to 32 viromes from the Pacific Ocean Virome dataset identified clusters of samples broadly delineated by photic zone and revealed that geographic region, depth, and proximity to shore were significant predictors of community structure. Within subsets of this dataset, depth, season, and oxygen concentration were significant drivers of viral community structure at a single open ocean station, whereas variability along onshore–offshore transects was driven by oxygen concentration in an area with an oxygen minimum zone and not depth or proximity to shore, as might be expected. Together these results demonstrate that this highly scalable approach using complete metagenomic network-based comparisons can both test and generate hypotheses for ecological investigation of viral and microbial communities in nature.

virus | microbial ecology | Bayesian network

Microorganisms drive global biogeochemical cycles (1), with abundances and taxonomic composition tuned to spatiotemporally varying environmental conditions (2–5). Viruses then modulate these biogeochemical processes through mortality, horizontal gene transfer, and metabolic reprogramming (6). However, our understanding of how viral communities change in response to biological, physical, and chemical factors and host availability has been limited by technical challenges.

Most viruses in the ocean lack both cultivated representatives [85% of 1,100 sequenced phage genomes derive from only 3 of 45 bacterial phyla (7)] and a universally conserved marker gene (8); thus, metagenomics is commonly applied to characterize the ecology and evolution of viral assemblages. Problematically, however, our ability to investigate these assemblages via metagenomics remains limited by the lack of known viruses and viral proteins in biological sequence databases. The first viral metagenome (virome) used thousands of Sanger reads and found that 65% of sequences were unknown [i.e., no database match for reads >600 bp (9)]. Adoption of next-generation sequencing (NGS) technologies then generated hundreds of thousands of reads (averaging 102 bp in length) per virome and returned ~90% sequence novelty (10). This unknown problem has not been significantly improved on in subsequent oceanic virome studies regardless of sequencing platform (11). This novelty limits taxonomic and functional inferences about viral assemblages and makes comparative analyses that only use the known portion of these datasets minimally informative at best and completely misleading at worst. Additionally, the standard practice of comparing new datasets against large genomic databases is compute intensive and increasingly unfeasible given escalating scales in datasets and databases.

To circumvent similar issues, Yooseph et al. (12) clustered environmental reads with known proteins from available databases to define sequence similarity-based protein clusters (PCs) to analyze the first global ocean microbial metagenomic datasets. This PC approach helps to both organize the vastly unknown sequence space in metagenomes and identify abundant proteins in environmental datasets even where taxonomy and function are unknown. Application of this approach to viromes has also been fruitful and has led to (*i*) a dataset of 456K protein clusters (11), (*ii*) comparative estimates of viral community diversity across sites (11, 13), and (*iii*) an estimate that the global virome is three orders of magnitude less diverse than previously thought (14). Although a valuable approach for metagenomic data, particularly for viromes where functional and taxonomic information is

Significance

Microorganisms and their viruses are increasingly recognized as drivers of myriad ecosystem processes. However, our knowledge of their roles is limited by the inability of culture-dependent and culture-independent (e.g., metagenomics) methods to be fully implemented at scales relevant to the diversity found in nature. Here we combine advances in bioinformatics (shared k-mer analyses) and social networking (regression modeling) to develop an annotation- and assembly-free visualization and analytical strategy for comparative metagenomics that uses all the data in a unified statistical framework. Application to 32 Pacific Ocean viromes, the first large-scale quantitative viral metagenomic dataset, tested existing and generated further hypotheses about ecological drivers of viral community structure. Highly computationally scalable, this new approach enables diverse sequence-based large-scale comparative studies.

Author contributions: B.L.H. and A.H.W. designed research; B.L.H. and A.H.W. performed research; B.L.H. and A.H.W. contributed new reagents/analytic tools; B.L.H., A.H.W., and J.R.B. analyzed data; and B.L.H., A.H.W., J.R.B., 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: All methods and source code are fully documented and are freely available at http://code.google.com/p/tmpl/. Viromes are available in the iPlant Community Data folder (imicrobe/pov) at http://www.iplantcollaborative.org/.

[1]To whom correspondence may be addressed. Email: bhurwitz@E-mail.arizona.edu or mbsulli@E-mail.arizona.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1319778111/-/DCSupplemental.

especially limiting, there are drawbacks to the PC approach including (*i*) only mapping ~75% of the data (11) and (*ii*) a reliance on metagenomic assembly algorithms not yet optimized for handling sequence variation derived from sequencing artifact and real population heterogeneity (15).

Recently, k-mer–based approaches were introduced to facilitate genome annotation (16) and for whole genome comparison to identify relationships among organisms without assembly and synteny analysis (17). For larger-scale metagenomic datasets, this approach offers a computationally scalable option for direct comparisons. Specifically, this shared k-mer strategy enables a similarity metric and the ability to identify clusters of metagenomes to infer how microbial communities are affected by environmental factors (18). These are significant advances, but they suffer from the lack of a unified statistical framework for evaluating genetic predictors of community structure based on multiple ecological variables that can be dependent on one another.

Here, we introduce a strategy to comparatively evaluate complete metagenomes by combining a shared k-mer approach with social network analysis to place all data into a unified context. Expanding on prior k-mer-based metagenomic methods (17, 18), a model was used to determine the statistical significance of ecological variables in forming the network while also accounting for dependency among these variables. The resulting network allows for data-driven hypothesis testing and generation through the evaluation of k-mer–based virome proximity in network space and statistical evaluation of ecological variables that drive these relationships. Application of this approach to 32 Pacific Ocean viromes (POVs) reveals a high-level overview of shared sequence space between these viromes, investigates the environmental characteristics that drive variability in viral community structure, and identifies testable hypotheses regarding viral community dynamics. Finally, although demonstrated on viromes, this strategy can be efficiently implemented on many large-scale sequence datasets with broad uses from environmental to clinical applications.

Results

Similar to previous metagenomic studies of ocean viruses (9, 10, 19–21), the 6,000,000 read POV dataset was dominated by the unknown (<6% of reads matched known viruses) (22). To more holistically compare viral metagenomes in a computationally scalable way (57× faster than BLAST and comparable to heuristic clustering algorithms; Tables S1 and S2), a strategy was employed using read-level k-mer similarity analyses between viromes as input to a social network analysis (SNA) to model relationships between viromes and metadata using statistical regression methodologies (23–25) (for details, see *SI Methods*). These analyses resulted in (*i*) a unified comparative network of viromes based on sequence composition (Fig. 1) and (*ii*) a statistical measure of the effect of covariates (i.e., season, proximity to shore, and depth) on the network structure using Eq. **1** (Table 1). This technique was applied to the complete dataset and two subsampled datasets to examine broad-scale, temporal, and spatial patterns as follows: (*i*) all 32 Pacific Ocean samples (Fig. 1 *A* and *B*), (*ii*) open ocean, station P26 LineP samples that vary by season and depth (Fig. 1*C*), and (*iii*) spring LineP transect samples that vary by proximity to shore and depth (Fig. 1*D*).

Broad-Scale Patterns Across 32 Pacific Ocean Viral Communities. Visually, eight regions emerged from the full 32 POV network that broadly differed by photic zone (three photic vs. five aphotic; Fig. 1*A*). In the aphotic portion of the network, the first region contained three viromes from summer at LineP open ocean station P26 and the second region contained three spring LineP samples from deep samples of the transect. The third region in the aphotic region of the network contained viromes from all three biomes with deep samples and across seasons, whereas the fourth and fifth regions contained outlier LineP spring viromes sampled from the base of the oxygen minimum zone.

A sixth region, in the photic portion of the network, contained all four of the spring/summer surface ocean LineP samples regardless of whether they were coastal, intermediate, or open ocean stations, as well as one Monterey Bay (MBARI) virome sampled in fall from the deep chlorophyll maximum (DCM; 42 m) at an intermediate ocean transect site. The seventh region contained surface water viromes including four near-replicate viromes [a single viral-concentrate that was differentially concentrated or purified (13)], sampled in the spring from Scripps Pier, one MBARI fall coastal virome, and one LineP winter open ocean virome. Finally, an eighth region contained five viromes including three from the MBARI photic zone at intermediate and open ocean stations and two shallow samples from the Great Barrier Reef.

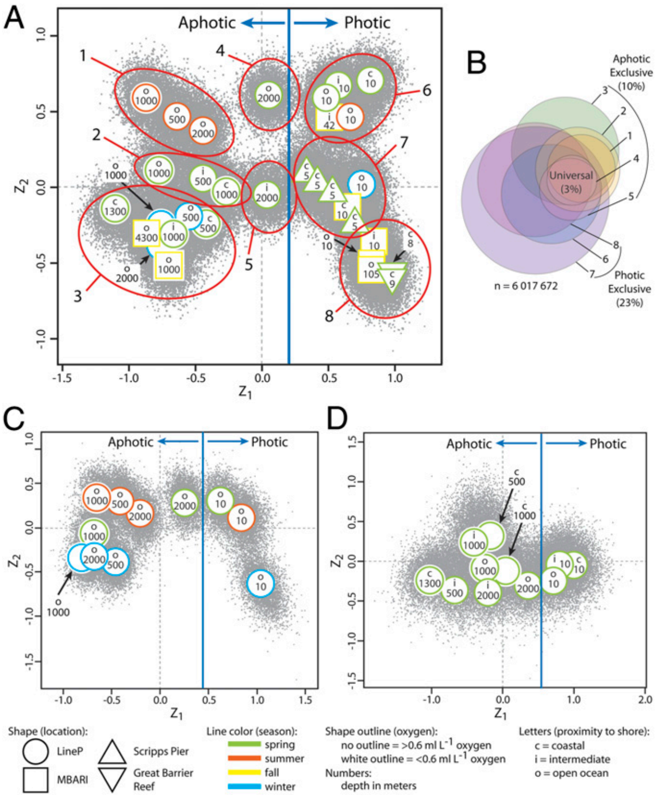

Fig. 1. Visualizing relationships between marine viral communities. (*A*) Social network of all POV samples with clusters circled in red. (*B*) Euler diagram (stress = 0.1027) depicting the portion of sequences shared by the eight clusters circled in *A* and the percent of sequences unique to the photic or aphotic zone. (*C*) Social network of all samples from LineP open ocean station P26. (*D*) Social network of all samples from the spring LineP transect. Dots in the social network graphs represent statistical samples taken from the marginal posterior distributions. Labels are placed at the posterior mean for each virome.

To complement these overall qualitative patterns (based on quantitative underpinnings), our unified network regression model was also used to evaluate ecological drivers of the observed network structure. Here, biogeographic region, depth, and proximity to shore were significant predictors of the overall POV network, but season was not (Table 1). Overall, only 3% of reads (n = 196,924) were universal to all samples within the POV network, whereas 23% and 10% of reads were exclusive to photic or aphotic parts of the network, respectively (Fig. 1*B*). The four near-replicate viromes from Scripps Pier contained the most activity (shared reads) in the network, whereas the shallow Great Barrier Reef viromes and one MBARI virome had the least (Fig. S1).

Finer-Scale Patterns Across Subnetworks. Given the complexity of the overall POV network, smaller refined subnetworks were examined to differentiate spatiotemporal features. First, analyses were focused on the most temporally well-resolved subsets of samples that included 11 viromes from the LineP open ocean

Table 1. Bayesian inference numerical summaries for social networks with selected covariates

Network dataset/Covariate	Parameter	Posterior median	Lower limit credible interval (2.5%)	Upper limit credible interval (97.5%)
Full dataset [32 samples (nodes)]				
log($\bar{n}_{i,j}$)	γ	0.63	**0.42**	**0.88**
Geographic region	β_1	0.14	**0.06**	**0.21**
Depth	β_2	0.12	**0.06**	**0.19**
Season	β_3	0.03	−0.02	0.08
Proximity to shore	β_4	0.12	**0.08**	**0.16**
Oxygen	β_5	−0.00	−0.21	0.08
LineP open ocean [11 samples (nodes)]				
log($\bar{n}_{i,j}$)	γ	0.828	**0.206**	**1.298**
Depth	β_1	0.224	**0.11**	**0.343**
Season	β_2	0.124	**0.032**	**0.257**
Oxygen	β_3	−0.336	**-0.589**	**-0.084**
LineP spring transect [11 samples (nodes)]				
log($\bar{n}_{i,j}$)	γ	6.737	**5.555**	**7.874**
Depth	β_1	0.117	−0.053	0.287
Proximity to shore	β_2	0.047	−0.063	0.147
Oxygen	β_3	0.867	**0.253**	**1.205**

Statistically significant covariates for each network are shown in bold. Covariates are considered significant if the upper and lower credible intervals (Baysian confidence intervals) do not overlap with zero. The covariate log($\bar{n}_{i,j}$) is an offset (although we do not restrict the coefficient to be equal to 1), which accounts for the fact that more shared read space may occur between two viromes if the either of the viromes is larger.

P26 station. Visually, again upper ocean, photic zone viromes were clearly separated from deep water, aphotic zone viromes, with seasonality leading to structure within these zones (Fig. 1*C*). Statistical regressions suggested that ecological drivers included depth, season, and oxygen concentration (Table 1).

A second subset of the data from LineP allowed focus on spatial variability from the coastal to open ocean samples collected on a springtime research cruise (Fig. 1*D*). Visually, again the photic and aphotic zone viromes were separated in shared k-mer space, but this time no strong patterns were observed with depth within these larger zones or with proximity to shore. Statistical analyses supported these qualitative observations, as only oxygen represented a structuring factor (Table 1).

LineP: A Case Study in Niche Specialization by Season. The power of the above analyses is the ability to visually represent viromes and define significant metadata factors to drive further investigation into underlying patterns. Given that reads have associated abundances (via the k-mer mode; *SI Methods* and Figs. S2 and S3), reads that are exclusive to specific viromes or parts of the network can be mined out of the underlying data. We demonstrate this by examining reads that are distinct by season (summer vs. winter) and photic zone (photic vs. aphotic) at open ocean station P26 at LineP in the Pacific Ocean based on Fig. 1*C* and Table 1. The exclusive read data demonstrate metabolic differences in parts of the network that likely derive from viral-encoded auxiliary metabolic genes as follows (Fig. 2), given the purity of these viromes (11, 13, 26).

The largest proportion of exclusive reads in all viromes, irrespective of photic zone or season, encodes genes related to nucleotide metabolism (Fig. 2). Given that viruses require nucleotides for replication, this result is not unexpected. When broadly comparing the photic zone and aphotic zone, aphotic viromes contain more overall metabolic functional capacity. In particular, genes related to the tricarboxylic acid (TCA) cycle, mannose and fructose metabolism, and electron transport chain (ETC) are more highly represented in aphotic viromes and are likely involved in energy production (26). These genes may be less represented in photic samples, given the capacity for viruses to encode and express photosynthetic genes (6) that allow them to derive energy for phage replication. Fatty acid metabolism may also be a source of energy production in phage in all seasons and photic zones, but most highly represented in summer aphotic viromes perhaps due to increased phage production in the summer and less energy derived from other sources. Interestingly, aphotic viromes and winter photic viromes contain genes related to cysteine and methionine metabolism, whose role is currently unknown but may be related to scavenging iron from Fe-S clusters in iron limited regions given that cysteine is important for Fe-S cluster biogenesis and degradation (27). Last, pyrimidine, purine, and glutathione metabolism may be important in winter aphotic viromes. Given that glutathione improves cold resistance in bacteria (28), viruses may help to provide protection to their infected hosts in the winter. These data suggest that viruses coevolve with their hosts and bolster host metabolism to improve host vitality for phage production given environmental selective pressures on the host.

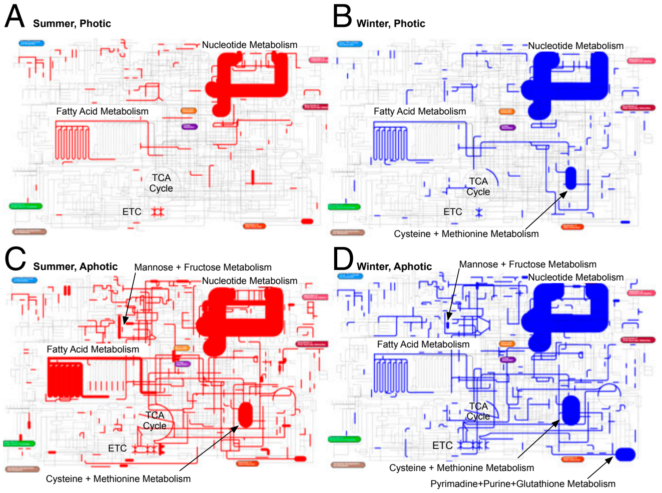

Fig. 2. Map of viral-encoded metabolic host genes from summer/winter and photic/aphotic at open ocean station P26 at LineP in the Pacific Ocean. The width of the lines corresponds to the normalized read abundance for viral encoded host genes in metabolic pathways from (*A*) summer photic (10 m) virome at LineP P26, (*B*) winter photic (10 m) virome at LineP P26, (*C*) summer aphotic (500, 1,000, and 2,000 m) viromes at LineP P26, and (*D*) winter aphotic (500, 1,000, and 2,000 m) viromes at LineP P26. ETC, electron transport chain. For map generation, see ref. 59.

Discussion

Microbes are now well recognized as critical players in ecosystems ranging from oceans and soils to humans and bioreactors. Their viruses are often as important, but methodological challenges have made it difficult to investigate even relatively fundamental questions such as how viral communities change over space and time. All currently used methods (e.g., morphology, viral genome fingerprinting, single-gene analyses, database-dependent annotation, and PC-based metagenomics) have issues that prevent extrapolating inferences to whole viral communities. The approach presented here that combines shared k-mer and social network analysis uses all of the reads, does not require assembly or database-dependent annotation, and includes a statistical framework (i.e., regression modeling) to evaluate ecological drivers of the resulting network structure. Application to 32 Pacific Ocean viromes allowed us to test hypotheses about how viral communities change over space and time, as well as generate new hypotheses where expectations were not met.

Broad-Scale Inferences About Pacific Ocean Viral Communities. Comparisons made across the entire POV dataset revealed that seasonality and oxygen were relatively unimportant in structuring viral communities, whereas geography, depth, and proximity to shore were significant. These first findings, that seasonality and oxygen do not structure POV communities, are likely a result of the overall dataset containing samples with many differing ecological features: e.g., half the dataset is associated with an oxygen minimum zone (LineP) and the rest have relatively invariant oxygen conditions. Thus, other features that strongly vary within the dataset overwhelm the more specific effects of season and oxygen (but see discussion of subnetworks below).

That geography is a strong driver of viral community structure is striking as ocean viruses have been posited to have extensive dispersion capability [inferred from global distribution of identical genetic marker sequences (29)] and a recent morphological survey of ocean viral communities found that geographic distance was not significant in explaining their variability across six oceans and seas (30). However, genome-wide variation likely far exceeds that of highly conserved marker genes and morphology-based metrics. A previous metagenomic study also suggested that four spatially diverse ocean viral communities were quite different (10), but the viromes were prepared in a nonquantitative manner (31), and only the known portion of these viromes (∼2% of the total reads with annotation to known phage) were analyzed to make this inference. Our observation of geographic variability in total viral metagenomes is consistent with variability of their dominant hosts, bacteria, which have geographic variability at the community level (18, 32), as well as within abundant phyla, including either large-scale genomic changes (e.g., in *Pelagibacter*) (33) or small-scale genomic changes more strongly localized to genomic islands (e.g., in cyanobacteria) (34).

Beyond geography, both depth and proximity to shore represent some of the strongest gradients available in the oceans, so it is not surprising they might also structure viral community composition across this larger sample set. Total viral particle counts broadly mirror those of prokaryotes across depth profiles in the oceans (35), which suggests that as microbial population abundances and structure change with depth, so too would their viruses. For example, microbial metagenomes from an open ocean depth profile show cyanophage abundance broadly mirrors that of their hosts (36). As well, depth-related variability in extracellular marine viral communities has been documented using viral genome fingerprinting (37, 38), and our findings support this, showing that depth is a clear driver of viral community structure. The latter driving factor, proximity to shore, is discussed with the subnetwork findings below.

Finer-Scale Evaluation of Ecological Factors That Structure Viral Communities. The decades of study along the LineP oceanographic transect (39) present an ideal backdrop for investigating temporal and spatial variability in viral communities. To focus on temporal variability, we examined a subnetwork of 11 viromes from February (winter), June (spring), and August (summer) at a single LineP station (open ocean station P26). This analysis revealed that depth, season, and oxygen were significant drivers of viral community structure in this subset of the data. In addition to the discussion of depth above, it is noteworthy that the LineP transect region is strongly stratified to the point of establishing one of the largest ocean interior oxygen minimum zones (40), so it is not surprising that viral community structure would significantly vary with depth and oxygen. That seasonality was also a driver is consistent with studies demonstrating annual cyclical changes in marine bacterial community structure (3, 41, 42). Although our single-year virome dataset does not permit inferences about year-to-year variation, similar annual repeatability has been observed in total viral abundance at the Bermuda Atlantic Time Series station (41), suggesting that annual repeatability of microbial hosts may lead to the same for their viral predators.

Additionally, the LineP transect is ideal for evaluating spatial changes in viral community structure along coastal to open ocean gradients. The strong vertical oxygen gradients along this transect (43) structured the viral community in the temporally focused subnetwork analysis above and also do so here in the spatial subnetwork analysis for a single season. Mechanistically, these strong gradients in oxygen are likely structuring LineP microbial populations as observed for total bacterial community composition (43) and dominant bacterial phyla [e.g., SUP05 and Marine Group A (MGA) (43)], which in turn structure their viral communities. These results are supported by studies of viral communities along strong oxygen gradients in stratified lakes using morphology or viral genome fingerprinting (37, 44), as well as a metagenomic investigation of viruses in a marine oxygen minimum zone off of Chile (45).

Notable outliers in our dataset include viromes from the base of the deep ocean oxycline (LineP spring 2,000-m viromes from the intermediate and open ocean). Distinct viral communities have been observed within oxyclines of marine and saline lake environments (44, 45), and thus these samples may represent viruses infecting bacteria adapted to dysoxic conditions (43). Alternatively, these deep oxycline viral communities could include surface water viruses that were entrapped on sinking particles and released at depth as a result of degradation, explaining their greater similarity to photic zone samples.

Additionally, although proximity to shore was a significant driver of viral community variability in the network with all samples, it was not significant when focusing solely on the LineP transect despite gradients in nutrients and productivity that occur along this transect (46). A lack of spatial variability in abundance of a specific bacterial phyla (MGA) along this transect has been observed (47), supporting our findings. Thus, the change in significance of proximity to shore as a structuring variable may be explained in the same fashion as for oxygen concentration. Specifically, the inclusion of coastal samples from MBARI, Scripps Pier, and the Great Barrier Reef may have been the primary drivers of this relationship in the full sample network, and their exclusion in the LineP transect network resulted in oxygen being the overwhelmingly dominant structuring variable, as was also noted for MGA distribution along this transect (47).

Last, we note that activity between viromes (shared reads) varies with sequencing effort. Four deeply sequenced near-replicate viromes from Scripps Pier showed the highest activity with other viromes (Fig. S1) likely due to greater representation of reads derived from the rare virosphere. Because the network is normalized for sequencing effort, this does not affect network structure, but is important when considering activity between viromes.

Analytical Advances. The approach outlined here provides a significant advance over alternative ecological methods for dimensionality reduction such as principle components analysis (PCoA) and nonmetric multidimensional scaling (nMDS; for details, see *SI Methods*). Broadly, the contrast lies in the fact that PCoA and nMDS are generally descriptive approaches (48), whereas the network approach outlined here provides a full inferential framework. Specifically, relational data methods are used to create a dependence structure in ordination space that includes random effects and as a result allows for the proper inference for regression coefficients (i.e., metadata). Or, in simple terms, the distances between viromes based on shared reads can be visually represented, while at the same time accounting for biological factors in a single statistical model. The importance of single modeling and inferential framework is highlighted in Chiu and Westveld (25).

This approach is also inherently different from other statistical frameworks [e.g., MaAsLin (49)] that identify associations between metadata and the abundance of operational taxonomic units (OTUs) or functions in metagenomic samples. Specifically, MaAsLin outputs a list of OTUs or functions that are significant given a metadata type. Given that the results are granular (by OTU or function) and only account for only one metadata type at a time, they cannot be combined. In contrast, our analytical framework (*i*) uses a model that enables simultaneous examination of shared sequence space between viromes in conjunction with multiple metadata types and (*ii*) requires no prior organizational bins (e.g., OTUs for MaAsLin), which is critical for viruses that lack a universal barcode gene for such taxonomic assignations. Both advances are fundamental for surveying complex viral communities to look for ecological drivers of community structure but also help broaden the toolkit available for other comparative metagenomic datasets (e.g., bacterial).

This approach may also prove to be important in other microbial analyses wherein taxonomic identification is less of a concern given rRNA sequence datasets. Specifically, we use entire metagenomes rather than a single gene (like 16S in bacteria) to assess the composition of microbial communities. The use of complete metagenomes is particularly important in cases where metagenomes may contain closely related species, indistinguishable on the level of the 16S gene alone, that have functional differences that make them distinct.

Further, because reads that represent significant patterns in the network can be mined out (see results LineP seasonal niche differentiation), this approach drives functional comparative metagenomic analyses. This approach is also important pragmatically in terms of runtime because fewer reads require extensive functional annotation. Moreover, the remaining unknown fraction of reads exclusive to a certain part of the network provides a starting point for future empirical analyses to understand the function of novel viral species. This approach is broadly applicable to metagenomes comprised of any microbe from viruses, to bacteria or fungi, and extends current approaches through the use of whole metagenomes and a comprehensive statistical framework.

Conclusions. Although marine microbes and their viruses are fundamental to Earth system function, the culture-independent metagenomic techniques used to study them present "big data" analytical challenges. The combination of shared k-mer and social network analysis presented here provides a powerful way to visualize and explore relationships between metagenomic samples and populations and statically evaluate the underlying factors that drive this variability. These methods are computationally tractable and widely applicable across sequence datasets and have the capacity to affect how data are stored, visualized, and analyzed, making use of big data analytics and the large-scale context that is now becoming available in metagenomic data repositories. These types of analyses and scales of data are needed to predictively model Earth's most abundant biological entities, viruses, and their predominant hosts, microorganisms.

Methods

Methods detailed below are further documented in *SI Methods*, Figs. S2 and S3, and Tables S1–S4. All source code is freely available at ref. 50.

Dataset. The 32-virome POV dataset (Table S3) (11) was examined to identify patterns of sequence similarity in viromes and determine the relationship between these patterns and depth, season, proximity to shore, geographic distance, and oxygen concentration. This dataset is a recently available public resource that leverages well-characterized sample-to-sequence preparation methods to generate quantitative viromes (13, 31). A full description of metadata associated with each virome and methods used to prepare the viromes and perform read quality control is included in *SI Methods* and Table S3. One additional filtering step was applied beyond the quality control steps for the POV dataset (11) that entailed removing reads with low abundance k-mers (k-mer = 1) in their own virome that were suspected of being contaminants (13) and reads with high-abundance k-mers (>1,000) that are likely to be either sequencing artifacts or highly conserved protein domains that may distort the overall abundance of that read.

k-mer Analyses. In the k-mer analysis below, suffix arrays were created using mkvtree from the vmatch package version 2.1.5 (51) using parameters (-pl -allout -v). Reads were compared with suffix arrays using vmatch's vmerstat (-minocc 1 -counts) to search for the frequency of 20-bp k-mers across the read. The k-mer size was set by examining the uniqueness ratio in the dataset (52). The k-mer value of 20 was chosen given that it represented an inflection point where k-mer hits moved from random to representative of the sequence content.

Pairwise All-vs.-All Analysis of Viromes. High-quality reads for each virome were compared with suffix arrays from all other viromes in a pairwise fashion (compute pipeline kmercompare.tar) to achieve an all-vs.-all analysis of the viromes [virome *i* vs. virome *j*, (for *i* = 1,. . .,32) and (for *j* = 1,. . .,32)]. The abundance for each read (in virome *i*) was calculated by finding the mode k-mer value for all k-mers in that read compared with the virome j suffix array (*SI Methods* and Figs. S2 and S3). This analysis resulted in a single abundance value (k-mer mode) for each read in virome *i* compared with virome *j*. The data were then normalized by averaging $(\overline{y}_{i,j})$ (shared read count) and $(\overline{n}_{i,j})$ (total read count) between virome *i* and virome *j*. Normalized shared read counts were used to construct a 32 × 32 matrix of viromes.

Network Analysis. To model the valued (nonbinary) nondirected data above, we consider the latent space approach outlined in Eq. **1** (23–25, 53). Our network modeling framework, via random effects, decomposes the statistical variation in the data to account for (*i*) the activity level (a_i) of each virome *i* (average amount of sequence space shared across the network for each virome *i*) and (*ii*) similarity (clustering) of shared sequence amount among viromes. For *i*), $z_i'z_j$ is measure of distance and similarity between viromes *i* and *j*. Each virome's position (z_i) may be visualized in a *k*-dimensional latent space Z (after a Procrustes' transformation to convert into a similar grid to compare) where virome *i* and virome *j* are considered similar if they are close in that space. For ease of visualization, we consider the case where $k=2$ (ref. 53 considered a 1D space).

Finally, we account for a set of relational covariates ($x_{ij}=1$ if similar, 0 if not) based on geographic region, season, proximity to shore, depth, and oxygen concentration using values in Table S3. In the case of oxygen concentration, which is a continuous value, high and low oxygen values were determined based on a cutoff of 0.06 mL/L.

$$\begin{aligned}
&\log\left(\overline{y}_{i,j}\right)=\alpha+\gamma\log\left(\overline{n}_{i,j}\right)+\beta' x_{ij}+a_i+a_j+z_i' z_j+\varepsilon_{ij}\\
&i<j,\\
&a_i \sim \text{identically distributed normal}\left(0,\sigma_a^2\right),\\
&z_{i,1} \sim \text{identically distributed normal}\left(0,\sigma_{z1}^2\right),\\
&z_{i,2} \sim \text{identically distributed normal}\left(0,\sigma_{z2}^2\right),\\
&\varepsilon_{ij} \sim \text{identically distributed normal}\left(0,\sigma_{\varepsilon}^2\right).
\end{aligned} \qquad \textbf{[1]}$$

To estimate the parameters in the model, a Bayesian inferential approach was considered using the R statistical software (54) and gbme.R obtained from refs. 24 and 55. For our analyses, empirical Bayes priors were considered (the default for the gbme.R). To examine the joint posterior distribution of the parameters, a Markov chain of 1,000,000 scans was constructed. The first 500,000 scans were removed for burn-in, and the chain was thinned by every 100th scan, leaving 5,000 samples.

Construction of Euler Diagrams Depicting Shared Read Content in Networks. Using data from the pairwise k-mer analysis described above, reads were detected that were unique or shared between subsets of viromes that visually clustered in the networks using a PERL script (get_section.pl). Reads were considered exclusive if they were present (mode k-mer ≥ 2) in two or more viromes in a cluster and absent (mode k-mer < 2) from viromes outside that cluster. For single virome clusters, all reads that were not shared with other viromes and present within that single virome at a k-mer abundance > 2 were considered exclusive. Reads that were present in a virome just once (k-mer = 1) were removed from the analysis given a higher probability of contamination (13) per the discussion above. The results were then used to compute an Euler diagram using the venneuler function (56) in the R statistical software (54).

Annotating Exclusive Reads. Exclusive reads per the method above for LineP summer photic viromes, summer aphotic viromes, winter photic viromes, and winter aphotic viromes were compared against the similarity matrix of proteins (SIMAP) released on August 20, 2013 (57) using BLASTX (58) to assign function as previously described (22). Briefly, these analyses were implemented using a custom data analysis pipeline written in Perl and bash shell and executed on a high-performance computer using PBSPro (blastpipeline_simap.tar). Hits were considered significant if they had an E value < 0.001, and only top hits were retained. Interpro ids in the SIMAP functional annotation were mapped to EC numbers using the swissprot_kegg_proteins_ec.csv as a mapping (59) (ipr_to_ec.pl). Read hit counts were normalized based on sequencing effort in the included viromes and converted into ipath2 format (create_ipath.pl) for visual representation in the ipath2 viewer (58).

ACKNOWLEDGMENTS. We thank Tucson Marine Phage Laboratory members for comments on the manuscript; D. D. Billheimer for statistics support; E. Allers, L. Deng, M. Dude-Duhaime, B. Poulos, J. Wright, K. Mitchell, C. Mewis, and A. Worden for logistical support, sample collection, and/or processing of viral concentrates; and University Information Technology Services Research Computing Group for high throughput computing access and support. Sequencing was provided by the US Department of Energy Joint Genome Institute Community Sequencing Program under the Office of Science of the US Department of Energy Contract DE-AC02-05CH11231 and the Gordon and Betty Moore Foundation Marine Microbial Initiative, whereas funding was from National Science Foundation Grants DBI-0850105 and OCE-0961947, Biosphere 2, BIO5, and Gordon and Betty Moore Foundation Grant GBMF3790 (to M.B.S.), as well as an Integrative Graduate Education Research Traineeship and National Science Foundation graduate research fellowship (to B.L.H.).

1. Falkowski PG, Fenchel T, Delong EF (2008) The microbial engines that drive Earth's biogeochemical cycles. *Science* 320(5879):1034–1039.
2. Caporaso JG, Paszkiewicz K, Field D, Knight R, Gilbert JA (2012) The Western English Channel contains a persistent microbial seed bank. *ISME J* 6(6):1089–1093.
3. Chow CE, Fuhrman JA (2012) Seasonality and monthly dynamics of marine myovirus communities. *Environ Microbiol* 14(8):2171–2183.
4. Fortunato CS, Herfort L, Zuber P, Baptista AM, Crump BC (2012) Spatial variability overwhelms seasonal patterns in bacterioplankton communities across a river to ocean gradient. *ISME J* 6(3):554–563.
5. Zaikova E, et al. (2010) Microbial community dynamics in a seasonally anoxic fjord: Saanich Inlet, British Columbia. *Environ Microbiol* 12(1):172–191.
6. Breitbart M (2012) Marine viruses: Truth or dare. *Annu Rev Mar Sci* 4:425–448.
7. Holmfeldt K, et al. (2013) Twelve previously unknown phage genera are ubiquitous in global oceans. *Proc Natl Acad Sci USA* 110(31):12798–12803.
8. Edwards RA, Rohwer F (2005) Viral metagenomics. *Nat Rev Microbiol* 3(6):504–510.
9. Breitbart M, et al. (2002) Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci USA* 99(22):14250–14255.
10. Angly FE, et al. (2006) The marine viromes of four oceanic regions. *PLoS Biol* 4(11): e368.
11. 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.
12. Yooseph S, et al. (2007) The Sorcerer II Global Ocean Sampling expedition: Expanding the universe of protein families. *PLoS Biol* 5(3):e16.
13. 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.
14. Ignacio-Espinoza JC, Solonenko SA, Sullivan MB (2013) The global virome: Not as big as we thought? *Curr Opin Virol* 3(5):566–571.
15. Degnan PH, Ochman H (2012) Illumina-based analysis of microbial community diversity. *ISME J* 6(1):183–194.
16. Edwards RA, et al. (2012) Real time metagenomics: Using k-mers to annotate metagenomes. *Bioinformatics* 28(24):3316–3317.
17. Song K, et al. (2013) Alignment-free sequence comparison based on next-generation sequencing reads. *J Comput Biol* 20(2):64–79.
18. Jiang B, et al. (2012) Comparison of metagenomic samples using sequence signatures. *BMC Genomics* 13:730.
19. Dinsdale EA, et al. (2008) Functional metagenomic profiling of nine biomes. *Nature* 452(7187):629–632.
20. Williamson SJ, et al. (2008) The Sorcerer II Global Ocean Sampling Expedition: Metagenomic characterization of viruses within aquatic microbial samples. *PLoS ONE* 3(1): e1456.
21. Bench SR, et al. (2007) Metagenomic characterization of Chesapeake Bay virioplankton. *Appl Environ Microbiol* 73(23):7629–7641.
22. Hoff PD, Raftery A, Handcock M (2002) Latent space approaches to social network analysis. *J Am Stat Assoc* 97(460):1090–1098.
23. Hoff PD (2005) Bilinear mixed-effects models for dyadic data. *J Am Stat Assoc* 100(469):286–295.
24. Chiu GS, Westveld AH (2011) A unifying approach for food webs, phylogeny, social networks, and statistics. *Proc Natl Acad Sci USA* 108(38):15881–15886.
25. Hurwitz BL, Hallam SJ, Sullivan MB (2013) Metabolic reprogramming by viruses in the sunlit and dark ocean. *Genome Biol* 14(11):R123.
26. Sharon I, et al. (2011) Comparative metagenomics of microbial traits within oceanic viral communities. *ISME J* 5(7):1178–1190.
27. Zhang J, Li Y, Chen W, Du GC, Chen J (2012) Glutathione improves the cold resistance of Lactobacillus sanfranciscensis by physiological regulation. *Food Microbiol* 31(2): 285–292.
28. Breitbart M, Rohwer F (2005) Here a virus, there a virus, everywhere the same virus? *Trends Microbiol* 13(6):278–284.
29. Brum JR, Schenck RO, Sullivan MB (2013) Global morphological analysis of marine viruses shows minimal regional variation and dominance of non-tailed viruses. *ISME J* 7(9):1738–1751.
30. Duhaime MB, Sullivan MB (2012) Ocean viruses: Rigorously evaluating the metagenomic sample-to-sequence pipeline. *Virology* 434(2):181–186.
31. Ladau J, et al. (2013) Global marine bacterial diversity peaks at high latitudes in winter. *ISME J* 7(9):1669–1677.
32. Brown MV, et al. (2012) Global biogeography of SAR11 marine bacteria. *Mol Syst Biol* 8:595.
33. Coleman ML, Chisholm SW (2007) Code and context: *Prochlorococcus* as a model for cross-scale biology. *Trends Microbiol* 15(9):398–407.
34. Weinbauer MG (2004) Ecology of prokaryotic viruses. *FEMS Microbiol Rev* 28(2): 127–181.
35. DeLong EF, et al. (2006) Community genomics among stratified microbial assemblages in the ocean's interior. *Science* 311(5760):496–503.
36. Jiang S, Fu W, Chu W, Fuhrman JA (2003) The vertical distribution and diversity of marine bacteriophage at a station off Southern California. *Microb Ecol* 45(4):399–410.
37. Steward G, Montiel JL, Azam F (2000) Genome size distributions indicate variability and similarities among marine viral assemblages from diverse environments. *Limnol Oceanogr* 45(8):1697–1706.
38. Freeland H (2007) A short history of ocean station papa and Line P. *Prog Oceanogr* 75(2): 120–125.
39. Whitney FA, Freeland HJ, Robert M (2007) Persistently declining oxygen levels in the interior waters of the eastern subarctic Pacific. *Prog Oceanogr* 75(2):179–199.
40. Parsons RJ, Breitbart M, Lomas MW, Carlson CA (2012) Ocean time-series reveals recurring seasonal patterns of virioplankton dynamics in the northwestern Sargasso Sea. *ISME J* 6(2):273–284.
41. Gilbert JA, et al. (2012) Defining seasonal marine microbial community dynamics. *ISME J* 6(2):298–308.
42. Wright JJ, Konwar KM, Hallam SJ (2012) Microbial ecology of expanding oxygen minimum zones. *Nat Rev Microbiol* 10(6):381–394.
43. Brum JR, Steward GF (2010) Morphological characterization of viruses in the stratified water column of alkaline, hypersaline Mono Lake. *Microb Ecol* 60(3):636–643.
44. Cassman N, et al. (2012) Oxygen minimum zones harbour novel viral communities with low diversity. *Environ Microbiol* 14(11):3043–3065.
45. Whitney F, Crawford WR, Harrison PJ (2005) Physical processes that enhance nutrient transport and primary productivity in the coastal and open ocean of the subarctic NE Pacific. *Deep Sea Res Part II Top Stud Oceanogr* 52(5):681–706.
46. Allers E, et al. (2013) Diversity and population structure of Marine Group A bacteria in the Northeast subarctic Pacific Ocean. *ISME J* 7(2):256–268.
47. Dinsdale EA, et al. (2013) Multivariate analysis of functional metagenomes. *Front Genet* 4: 41.
48. Huttenhower C (2014) MaAsLin: Multivariate analysis by linear models. Available at http://huttenhower.sph.harvard.edu/maaslin. Accessed December 30, 2013.
49. Hurwitz BL (2014) TMPL source code. Available at http://code.google.com/p/tmpl. Accessed December 30, 2013.
50. vmatch (2013) vmatch package version 2.1.5. Available at www.vmatch.de.
51. Kurtz S, Narechania A, Stein JC, Ware D (2008) A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics* 9:517.
52. Chiu GS, Westveld AH (2014) A statistical social network model for consumption data in trophic food webs. *Stat Methodol* 17(4432):139–160.
53. R Core Team (2012) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna).
54. Hoff P (2013) gbme.R. http://www.stat.washington.edu/hoff/Code/hoff_2005_jasa. Accessed on December 31, 2013.
55. Wilkinson L (2012) Exact and approximate area-proportional circular Venn and Euler diagrams. *IEEE Trans Vis Comput Graph* 18(2):321–331.
56. Rattei T, et al. (2006) SIMAP: the similarity matrix of proteins. *Nucleic Acids Res* 34(Database issue):D252–D256.
57. Yamada T, Letunic I, Okuda S, Kanehisa M, Bork P (2011) iPath2.0: Interactive pathway explorer. *Nucleic Acids Res* 39(Web Server issue):W412–W415.
58. De Ferrari L, Aitken S, van Hemert J, Goryanin I (2012) EnzML: Multi-label prediction of enzyme classes using InterPro signatures. *BMC Bioinformatics* 13:61.
59. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389–3402.