Supporting Information

Hurwitz et al. 10.1073/pnas.1319778111

SI Methods

Virome, Metadata, and K-mer Analysis Details. ***Constructing viromes.*** As previously described, viromes were constructed from 32 seawater samples in the Pacific Ocean that varied by depth, season, proximity to shore, geographic distance, and oxygen concentration (Table S3 and Fig. S2) (1, 2). Briefly, seawater samples were prefiltered using a 150-μm grade A glass microfiber filter to enrich for bacterial and viral-sized particles in the filtrate (Fig. S2). The filtrate was then passed through a subsequent 0.22-μm filter to enrich for viral-sized particles only. Viral-sized particles were then concentrated from the remaining filtrate using $FeCl_3$ precipitation and purified by DNase and CsCl. Three comparison viromes were also constructed from a single sample at Scripps Pier (SIO) using the following concentration and purification protocols: (*i*) tangential flow filtration (TFF) and DNase and CsCl, (*ii*) $FeCl_3$ precipitation and DNase only, and (*iii*) $FeCl_3$ precipitation and DNase and sucrose as previously described (1). Following DNA extraction using Wizard PCR DNA Purification Resin and Minicolumns as previously described (3), viral DNA was randomly size sheared and amplified using linker amplification (LA) as described previously (4) in preparation for sequencing. Samples were sequenced using GS FLX Titanium sequencing chemistry on a 454 Genome Sequencer. Multiple samples were multiplexed on a single sequencing run by adding sample specific molecular barcodes during adapter ligation in the library preparation step.

Data preparation.

Quality control of reads. Following sequencing, reads were passed through a quality filter to remove low-quality reads that were less than 2 SDs from the mean length and quality or contained an "N" anywhere in the read (Fig. S2). To remove technical replicates, high-quality reads were then passed through cd-hit-454 (using default parameters, version 4.5.5) (5). Next, reads were binned by sample based on their molecular barcode, and the barcode and platform-specific adapter were trimmed to produce the 32 viromes used in this analysis. Custom quality filtering code was written in Perl and shell script to implement this protocol as a pipeline on a compute cluster running PBSPro (screenpipe.tar).

K-mer analyses. In each k-mer analysis below (both for removing contaminating reads and for pairwise all-vs.-all comparison of viromes), suffix arrays were created using mkvtree from the vmatch package version 2.1.5 (www.vmatch.de) 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.

Removal of high and low abundance contaminating reads. To remove other potential sequencing errors, a k-mer–based approach was applied to compare reads from each virome to a suffix array from the same virome (Figs. S2 and S3). The main concept was that rare reads, wherein the mode value of k-mers in the read is <1 in the same virome (Fig. S3), are likely to be contaminants (1, 6) and therefore were removed before further analysis. Moreover, portions of the read with k-mers that appear >1,000 [more than 10× the average read coverage by contig assembly analysis (2)] in the suffix array for the same virome are likely to be either sequencing artifacts or highly conserved protein domains that may distort the overall abundance of that read. As a result, these high-abundance k-mers were masked out with Ns in the read and trimmed from the beginning or end of the read. Further, any reads that were less than 100 bp after masking and trimming were removed. Given this k-mer filtering approach, we are able to detect reads with aberrant k-mers (from sequencing artifact or in highly conserved protein domains) that could confound an analysis of the relative abundance of reads between metagenomes.

Pairwise all-vs.-all analysis of viromes. After reads with aberrant k-mers were removed (both low and extremely high abundance k-mers), 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 $(i = 1, \ldots, 32)$ and (for $j = 1, \ldots, 32$)].

Data analysis pipeline.

Matrix of read counts based on reads with shared k-mers. Following the k-mer–based pairwise analysis of viromes, read mode tables were created to capture the abundance of each read between viromes (number of times virome *i* read appears in virome *j*; Fig. S3). 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 (Fig. S3). This analysis resulted in a single abundance value (k-mer mode) for each read in virome *i* compared with virome *j* (Fig. S3). The mode tables were then used to construct a 32 × 32 matrix of shared read counts $(y_{i,j})$ between pairwise combinations of viromes. Reads in virome *i* were considered to be shared with virome *j*, if the mode k-mer value for the read was >2. Each shared read in the mode table increments the total shared read $(y_{i,j})$ count for the pair of viromes by 1.

Creating matrices of covariates. Matrices of relational covariates were created for each metadata type in Table S3, specifically geographic region, season, proximity to shore, depth, and oxygen concentration. For each cell in the matrix comparing two viromes, if two samples have exactly the same value, then they are coded $(x_{ij} = 1)$ for being the same; otherwise, they are coded as $(x_{ij} = 0)$. In the case of oxygen concentration, which is a continuous value, high and low oxygen values were established using a cutoff of 0.06 mL/L (low oxygen values are indicated in bold in Table S3). Viromes were then coded $(x_{ij} = 1)$ for being the same if they were both high or low oxygen. Network analyses were performed according to the detailed methods provided later.

Runtime analyses and comparison with other methods. Given the dramatic decrease in sequencing costs with next-generation sequencing technologies, rapid and scalable methods to analyze large-scale genomic and metagenomic datasets are fundamental. Beyond the biological conclusions outlined here, this article also offers a method for dealing with the computational complexity of modern datasets. Specifically, our k-mer method is 57× faster than an all-vs.-all blast comparison of the same dataset (Table S1). Extrapolation to the scale of datasets currently being generated (~80,000,000 high-quality reads on a next-generation sequencing platform), the k-mer analysis would complete in 6 h compared with 1.5 wk for an all-vs.-all blast analysis on a 32-core high performance compute cluster (Table S2). Moreover, the k-mer method is comparable in computing time to other scalable heuristic clustering algorithms such as usearch and cdhit (7, 8) (Table S1). Although run times are similar, the k-mer method uses the entire dataset to compute read abundance across metagenomes (with mode k-mer abundance reported for the read; Fig. S3), whereas such abundance data are lost with clustering methods because the fast heuristics find only the top few hits resulting in presence/absence data (7, 8). Thus, the k-mer method provides comparable run times but more comprehensive analysis of metagenomes.

Statistical Modeling Details. ***Model structure.*** The relational data that we are modeling, in its simplest form, is a proportion $y_{i,j}/n_{i,j}$, where $y_{i,j}$ represents the number of reads in common (reading in one direction) over the total possible number of counts $n_{i,j}$ (reading in one direction). Although a natural approach to deal with these type of data are through binomial distribution, we consider an alternative by taking the natural log of the proportion within the following regression setting

$$\begin{aligned}\log(y_{i,j}/n_{i,j}) &= \boldsymbol{\beta}'\boldsymbol{x}_{i,j} + \delta_{i,j}\\ \log(y_{i,j}) - \log(n_{i,j}) &= \boldsymbol{\beta}'\boldsymbol{x}_{i,j} + \delta_{i,j}\\ \log(y_{i,j}) &= \log(n_{i,j}) + \boldsymbol{\beta}'\boldsymbol{x}_{i,j} + \delta_{i,j},\end{aligned}$$

where $\boldsymbol{x}_{i,j}$ is a vector of covariates (see the main text and *SI Methods* for information on their construction), and $\delta_{i,j}$ is an error term. Moving the term $\log(n_{i,j})$ over to the right side of the regression makes that term an offset (9). However, as our data actually consist of reading in two directions and averaging those results, we have

$$\log(\bar{y}_{i,j}) = \log(\bar{n}_{i,j}) + \boldsymbol{\beta}'\boldsymbol{x}_{i,j} + \delta_{i,j}.$$

Finally, we allow for more generality by not forcing the coefficient of $\log(\bar{n}_{i,j})$ to be equal to 1

$$\log(\bar{y}_{i,j}) = \gamma\log(\bar{n}_{i,j}) + \boldsymbol{\beta}'\boldsymbol{x}_{i,j} + \delta_{i,j}. \qquad \textbf{[S1]}$$

As the data are nondirected ($\bar{y}_{i,j} \equiv \bar{y}_{j,i}$ and $\bar{n}_{i,j} \equiv \bar{n}_{j,i}$), we only consider the data for $i < j$. To accommodate the potential dependencies that arise in nondirected relational data, consider the following decomposition of $\delta_{i,j}$ into the following mean zero random effects (based on refs. 10–12):

$$\begin{aligned}&\delta_{i,j} = a_i + a_j + \boldsymbol{z}_i'\boldsymbol{z}_j + \epsilon_{i,j}\\ &a_i \overset{\text{iid}}{\sim} \text{normal}(0, \sigma_a^2)\\ &\boldsymbol{z}_i \overset{\text{iid}}{\sim} \text{multivariate normal}_{k=2}\left(\mathbf{0}, \begin{bmatrix}\sigma_{z_1}^2 & 0\\ 0 & \sigma_{z_2}^2\end{bmatrix}\right)\\ &\epsilon_{i,j} \overset{\text{iid}}{\sim} \text{normal}(0, \sigma_\epsilon^2).\end{aligned} \qquad \textbf{[S2]}$$

Our network modeling framework, via the 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 *ii*, $\boldsymbol{z}_i'\boldsymbol{z}_j$ is measure of distance between viromes *i* and *j*. Each virome's position ($\boldsymbol{z}_i$) may be visualized in a *k*-dimensional latent space *Z* (after a Procrustes' transformation; see below), where virome *i* and virome *j* are considered similar if they are close in that space. For ease of visualization we consider a 2D space ($k = 2$; considered a 1D space) (13).

This modeling approach generalizes stand-alone techniques based on multidimensional scaling [e.g., principle components analysis (PCoA) and nonmetric multidimensional scaling (nMDS)] by embedding an ordination metric $\boldsymbol{z}_i'\boldsymbol{z}_j$ into a single inferential model that additionally accounts for the activity level for each virome in the network and covariates of interest (10–13). Accounting for this dependence structure (expressed by the statistical moments below) allows for appropriate quantification of uncertainty for parameters of interest, including the regression coefficients. The importance of single modeling and inferential framework is highlighted in ref. 11.

In particular, the error term leads to the following first moment:

$$E(\delta_{i,j}) = 0,$$

which implies the following first moments for each of the observations:

$$E\left[\log(\bar{y}_{i,j})\right] = \log(\bar{n}_{i,j}) + \boldsymbol{\beta}'\boldsymbol{x}_{i,j}.$$

The following are the nonzero second and third moments for the errors (as well as the observations). Note that even though our data are dyadic, the third moment is not equal to zero

$$\begin{aligned}&E(\delta_{i,j}^2) \equiv E(\delta_{i,j}\delta_{j,i}) \equiv E(\delta_{j,i}^2) = 2\sigma_a^2 + \sigma_\epsilon^2 + \sigma_{Z_1}^4 + \sigma_{Z_2}^4\\ &E(\delta_{i,j}\delta_{i,k}) = E(\delta_{i,j}\delta_{k,j}) = E(\delta_{i,j}\delta_{k,i}) = \sigma_a^2\\ &E(\delta_{i,j}\delta_{j,k}\delta_{k,i}) = \sigma_{Z_1}^6 + \sigma_{Z_2}^6.\end{aligned}$$

To estimate the parameters in the model, a Bayesian inferential approach was considered using the R statistical software (14) and gbme.R obtained from ref. 10 (www.stat.washington.edu/hoff/Code/hoff_2005_jasa). 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.

Procrustes transformation. Although the inner products ($\boldsymbol{z}_i'\boldsymbol{z}_j$) are identifiable, individually the random effects ($\boldsymbol{z}_i$) corresponding to the latent space are unidentifiable without constraints (10, 11, 15). Specifically, the inner products are invariant to rotation and reflection of the latent space *Z*. To circumvent this unidentifiability for visualization, for each *s*th Markov chain Monte Carlo (MCMC) scan, a Procrustes transformation is applied to rotate the space to a common orientation.

Goodness-of-fit. In Table S4, we report the Akaike and Bayesian information criteria (AIC and BIC) to compare the network mixed model (based on Eqs. **S1** and **S2**) to a simpler standard regression model

$$\begin{aligned}&\log(\bar{y}_{i,j}) = \log(\bar{n}_{i,j}) + \boldsymbol{\beta}'\boldsymbol{x}_{i,j} + \epsilon_{i,j}\\ &\epsilon_{i,j} \overset{\text{iid}}{\sim} \text{normal}(0, \sigma_\epsilon^2).\end{aligned} \qquad \textbf{[S3]}$$

Specifically, we considered the following criteria:

$$\begin{aligned}\text{AIC} &= -2\log p\left(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{Bayes}\right) + 2k\\ \text{BIC} &= -2\log p\left(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{Bayes}\right) + k\log(N),\end{aligned}$$

where *k* is the number of parameters, and *N* is the number of data points. The log likelihoods were evaluated at the means of the posterior distributions $[\hat{\theta}_{Bayes} = E(\theta|y)]$ (16, 17); please see ref. 18 for an overview of model selection for mixed models. In considering the effective number of parameters for the random effects portion of the network model, additional number of parameters compared with the standard regression model, we used both an optimistic (o) lower bound, consisting of three additional parameters ($\sigma_a^2, \sigma_{z1}^2, \sigma_{z2}^2$), and a pessimistic (p) upper bound where all of the random effects were treated as fixed (s_i, $\boldsymbol{z}_i$, $\forall_i = \{1, \ldots, A = \text{number of nodes}\}$; leading to $3 \times A$ additional parameters).

Based on the AIC values, for both the optimistic and pessimistic effective number of parameters, the network model is preferred to the standard regression model. The same holds true for the BIC values, except for one case (the pessimistic effective number of parameters for LineP open ocean data). Overall these results suggest that the network structure is an important consideration when modeling relational metagenomics data.

1. 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.
2. 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.
3. Henn MR, et al. (2010) Analysis of high-throughput sequencing and annotation strategies for phage genomes. *PLoS ONE* 5(2):e9083.
4. Duhaime MB, Deng L, Poulos BT, Sullivan MB (2012) Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: A rigorous assessment and optimization of the linker amplification method. *Environ Microbiol* 14(9):2526–2537.
5. Niu B, Fu L, Sun S, Li W (2010) Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinformatics* 11:187.
6. Zhang Q, Pell J, Canino-Koning R, Howe AC, Brown CT (2013) These are not the k-mers you are looking for: Efficient online k-mer counting using a probabilistic data structure. arXiv:1309.2975.
7. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19):2460–2461.
8. Huang Y, Niu B, Gao Y, Fu L, Li W (2010) CD-HIT Suite: A web server for clustering and comparing biological sequences. *Bioinformatics* 26(5):680–682.
9. 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.
10. Chiu GS, Westveld AH (2011) A statistical social network model for consumption data in trophic food webs. *Stat Methodol* 17(4432):139–160.
11. Gelman A, Carlin JB, Stern HS, Dunson DB, Rubin DB (2014) *Bayesian Data Analysis* (CRC, Boca Raton, FL).
12. Hoff PD, Raftery AE, Handcock MS (2002) Latent space approaches to social network analysis. *J Am Stat Assoc* 97(460):1090–1098.
13. Hoff PD (2005) Bilinear mixed-effects models for dyadic data. *J Am Stat Assoc* 100(469):286–295.
14. Hoff PD (2007) Model averaging and dimension selection for the singular value decomposition. *J Am Stat Assoc* 102(478):674–685.
15. Kass RE, Raftery AE (1995) Bayes factors. *J Am Stat Assoc* 90(430):773–795.
16. McCullagh P, Nelder JA (1989) *Generalized Linear Models* (CRC, Boca Raton, FL).
17. Muller S, Scealy JL, Welsh AH (2013) Model selection in linear mixed models. *Stat Sci* 28(2):135–167.
18. R Core Team (2013) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna).

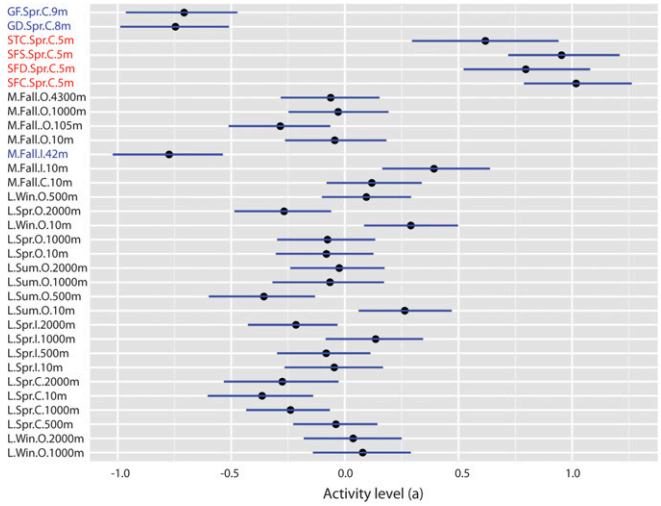

Fig. S1. Activity levels for viromes in the POV dataset. Dots are the median of the posteriors and bars are the 95% credible intervals for the activity levels of each virome. Viromes labeled in red are the four most active in the network (i.e., share the most reads across the network), and those labeled in blue are the three least active. Virome names follow the convention for the POV dataset as described by Hurwitz and Sullivan (1). Briefly, the initial letter indicates location (G, Great Barrier Reef; S, Scripps Pier; M, MBARI; L, LineP), season is indicated after the first period (Spr, spring; Sum, summer; Fall, fall; Win, winter), proximity to shore is indicated after the second period (C, coastal; I, intermediate; O, open ocean), and depth in meters is indicated after the third period.

1. 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.

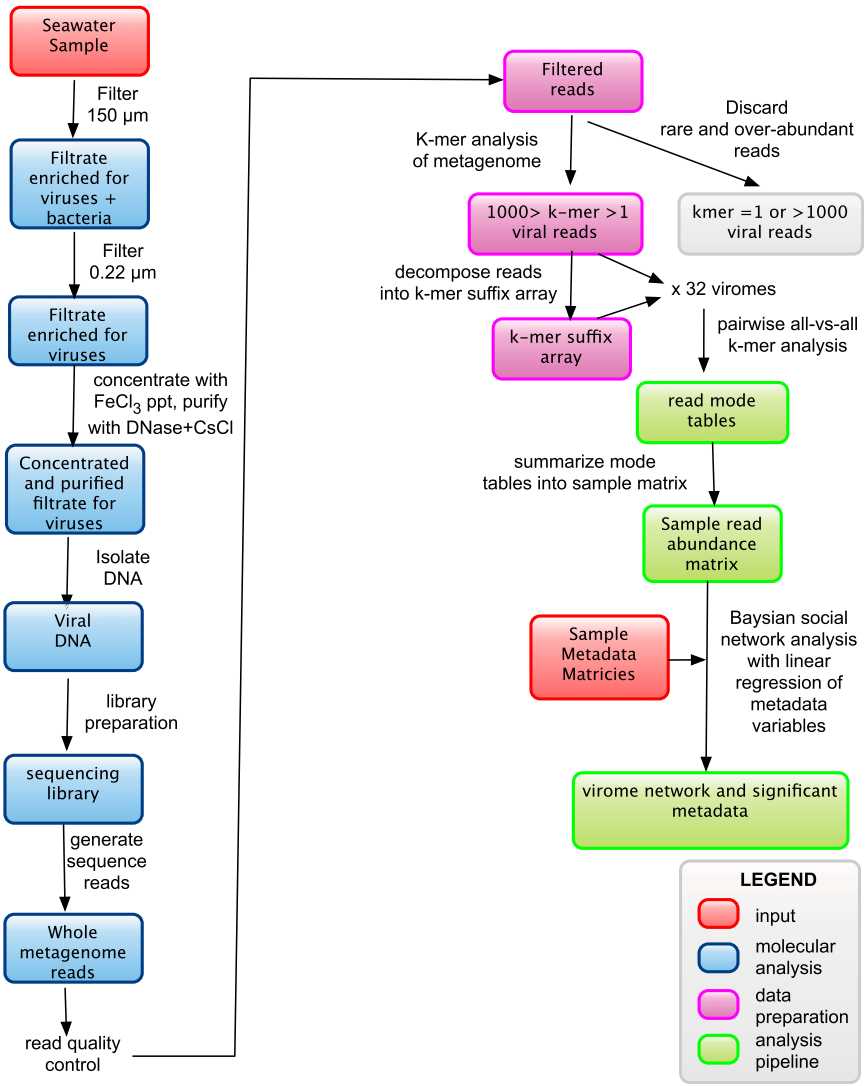

Fig. S2. Overview of the methods for molecular analysis, data preparation, and data analysis pipeline.

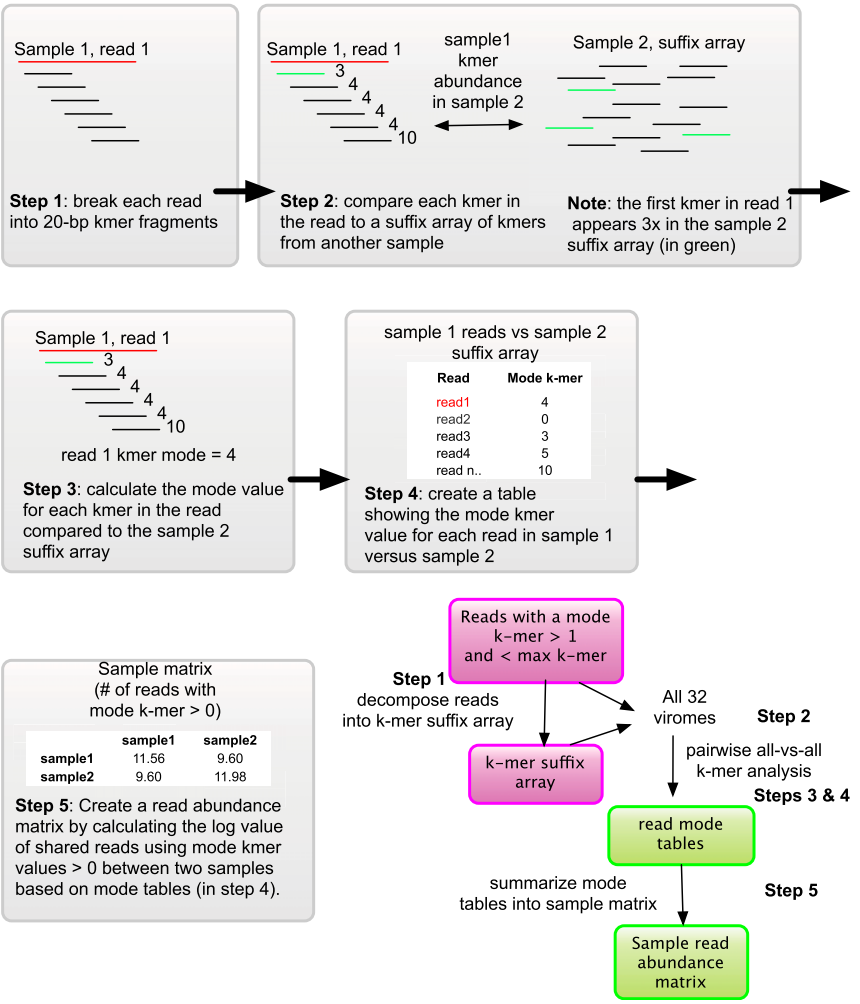

Fig. S3. Detailed overview of k-mer analyses and creation of the sample read abundance matrix.

Table S1. Benchmark comparison of pairwise sequence analysis methods for Pacific Ocean viromes

Method used	Total reads	CPU (s)	Runtime (reads/s)
All-vs.-all blast	6,120,792	2,268,000	3
k-mer analysis (used here)	6,120,792	39,600	155
Usearch (uclust)	6,120,792	38,520	159

Usearch version 7.0.1090 was run with the following parameters: cluster_ smallmem -id 0.95′. Given that usearch failed at 49% runtime due to a 4-GB memory constraint in the open source version, total runtime was calculated based on the partial runtime. All-vs.-all blast was run using NCBI Blast version 2.2.22 with the following parameters : a 12 -p blastn -v 10 -b 10 -e 1e-3.

Table S2. Compute time on a cluster for all-vs.-all blast analysis vs. k-mer analysis of viromes

Method used	Total reads	CPU hours	CPU days	CPU days on an HPC cluster (32 cores)	Weeks on an HPC cluster (32 cores)	Runtime
All-vs.-all blast	6,120,792	630.0	26.3	0.8	0.0	Actual
	80,000,000	8233.8	343.1	10.7	1.53	Calculated
k-mer analysis (used here)	6,120,792	11.0	0.46	0.0	0.0	Actual
	80,000,000	143.8	5.99	0.2	0.0	Calculated

Table S3. Thirty-two POV samples and metadata

Sample	No. reads	Geographic region	Depth (m)	Season	Proximity to shore	Oxygen (mL/L)
GD.Spr.C.8m	116,855	GBR	8	Spring	Coastal	NA
GF.Spr.C.9m	82,739	GBR	9	Spring	Coastal	NA
L.Spr.C.10m	107,244	LineP	10	Spring	Coastal	7.20
L.Spr.I.10m	92,415	LineP	10	Spring	Intermediate	6.79
L.Spr.O.10m	75,036	LineP	10	Spring	Open ocean	7.03
L.Sum.O.10m	165,256	LineP	10	Summer	Open ocean	6.22
L.Win.O.10m	192,685	LineP	10	Winter	Open ocean	6.93
L.Spr.C.500m	136,876	LineP	500	Spring	Coastal	**0.59**
L.Spr.I.500m	58,108	LineP	500	Spring	Intermediate	0.94
L.Sum.O.500m	42,118	LineP	500	Summer	Open ocean	0.89
L.Win.O.500m	167,616	LineP	500	Winter	Open ocean	0.82
L.Spr.C.1000m	97,126	LineP	1,000	Spring	Coastal	**0.19**
L.Spr.I.1000m	122,565	LineP	1,000	Spring	Intermediate	**0.21**
L.Spr.O.1000m	101,179	LineP	1,000	Spring	Open ocean	**0.32**
L.Sum.O.1000m	70,596	LineP	1,000	Summer	Open ocean	**0.38**
L.Win.O.1000m	147,537	LineP	1,000	Winter	Open ocean	**0.37**
L.Spr.C.1300m	98,478	LineP	1,300	Spring	Coastal	**0.35**
L.Spr.I.2000m	49,914	LineP	2,000	Spring	Intermediate	1.34
L.Spr.O.2000m	55,332	LineP	2,000	Spring	Open ocean	1.24
L.Sum.O.2000m	68,516	LineP	2,000	Summer	Open ocean	1.23
L.Win.O.2000m	125,896	LineP	2,000	Winter	Open ocean	1.31
M.Fall.C.10m	303,519	MBARI	10	Fall	Coastal	4.01
M.Fall.I.10m	321,754	MBARI	10	Fall	Intermediate	3.95
M.Fall.O.10m	203,238	MBARI	10	Fall	Open ocean	3.66
M.Fall.I.42m	31,528	MBARI	42	Fall	Intermediate	3.90
M.Fall.O.105m	156,509	MBARI	105	Fall	Open ocean	3.95
M.Fall.O.1000m	225,833	MBARI	1,000	Fall	Open ocean	**0.32**
M.Fall.O.4300m	144,588	MBARI	4,300	Fall	Open ocean	2.25
SFC.Spr.C.5m	487,339	SIO	5	Spring	Coastal	NA
SFD.Spr.C.5m	645,463	SIO	5	Spring	Coastal	NA
SFS.Spr.C.5m	504,826	SIO	5	Spring	Coastal	NA
STC.Spr.C.5m	821,404	SIO	5	Spring	Coastal	NA

Low oxygen samples (<0.60 mL/L) are denoted in bold. NA, not applicable.

Table S4. AIC and BIC to compare the network mixed model to a simpler standard regression model

Network dataset/model	$\log p(y \mid \hat{\theta}_{Bayes})$	k	AIC	BIC
Full dataset (32 samples)				
Network	1.687	11 (o)	18.626	64.898
		104 (p)	204.626	642.110
Standard regression	−400.085	8	816.171	849.823
LineP open ocean (11 samples)				
Network	24.066	9 (o)	−30.132	−12.066
		39 (p)	29.868	108.154
Standard regression	−27.374	6	66.748	78.792
LineP spring transect (11 samples)				
Network	36.755	9 (o)	−55.511	−37.445
		39 (p)	4.489	82.775
Standard regression	−31.357	6	74.714	86.758

Optimistic and pessimistic effective numbers of parameters are represented by (o) and (p), respectively. Noticeably smaller AIC and BIC values suggest a better fit.