AEM
Journals.ASM.org

New Approaches Indicate Constant Viral Diversity despite Shifts in Assemblage Structure in an Australian Hypersaline Lake

Joanne B. Emerson,[a]* Brian C. Thomas,[a] Karen Andrade,[b] Karla B. Heidelberg,[c] Jillian F. Banfield[a,b]

Department of Earth and Planetary Science, University of California, Berkeley, Berkeley, California, USA[a]; Department of Environmental Science, Policy, and Management, University of California, Berkeley, Berkeley, California, USA[b]; Department of Biological Sciences, University of Southern California, Los Angeles, California, USA[c]

It is widely stated that viruses represent the most significant source of biodiversity on Earth, yet characterizing the diversity of viral assemblages in natural systems remains difficult. Viral diversity studies are challenging because viruses lack universally present, phylogenetically informative genes. Here, we developed an approach to estimate viral diversity using a series of functional and novel conserved genes. This approach provides direct estimates of viral assemblage diversity while retaining resolution at the level of individual viral populations in a natural system. We characterized viral assemblages in eight samples from hypersaline Lake Tyrrell (LT), Victoria, Australia, using 39,636 viral contigs. We defined viral operational taxonomic units (OTUs) in two ways. First, we used genes with three different functional predictions that were abundantly represented in the data set. Second, we clustered proteins of unknown function based on sequence similarity, and we chose genes represented by three clusters with numerous members to define OTUs. In combination, diversity metrics indicated between 412 and 735 sampled populations, and the number of populations remained relatively constant across samples. We determined the relative representation of each viral OTU in each sample and found that viral assemblage structures correlate with salinity and solution chemistry. LT viral assemblages were near-replicates from the same site sampled a few days apart but differed significantly on other spatial and temporal scales. The OTU definition approach proposed here paves the way for metagenomics-based analyses of viral assemblages using ecological models previously applied to bacteria and archaea.

Viruses are abundant and ubiquitous, and they influence nutrient cycling, host evolution, and community structure (1). Despite an appreciation for their enormous diversity, viruses are the least well characterized biological entities (2). All-inclusive molecular surveys commonly applied to microbial systems are not possible for viruses, owing to the lack of a universal marker gene, so other techniques have been employed to characterize the diversity and dynamics of viral assemblages. For example, viral counts and pulsed-field gel electrophoresis (PFGE) have demonstrated seasonal shifts in viral abundance and genome size diversity in Chesapeake Bay sediments and surface waters (3, 4). The amplification of target genes known to be conserved within specific viral groups has shown, for example, that single-stranded DNA (ssDNA) viral diversity generally changes on a time scale of months in marine systems (5), that cyanophage genetic diversity varied over 3 years in marine coastal waters (6), and that marine myoviral assemblages near the coast of California exhibited seasonal dynamics (7). However, the inherently high conservation of some genes chosen for such surveys may limit the extent to which ecologically relevant parameters, such as host range or habitat, can be correlated with viral biogeography and dynamics (8). In addition, the biggest limitations of such studies are amplification biases that may preclude accurate representation of natural abundances (9) and the fact that these analyses can only be used to estimate the diversity of relatively small groups of known viruses.

Statistical modeling of assembly success, developed for analyses of Sanger viral metagenomic data (10, 11), has been used to estimate the structure and alpha- and beta-diversity of viral assemblages from metagenomic data. However, given that the algorithm uses a relatively computationally intensive overlap-layout-consensus assembly approach, it is not designed to handle the large data sets generated from new sequencing technologies (particularly Illumina). The models also require average genome size as input, which can be estimated from another model (12) but cannot be directly assessed. Most importantly, without a means to more directly estimate viral assemblage diversity, it is impossible to validate the models. For the relatively small portion of viral metagenomic data with similarity to sequences in public databases (2), BLAST searches have been used to infer taxonomy and/or generate functional predictions, which can then be used to compare samples, estimate richness, and/or estimate functional diversity (13, 14). However, many such studies rely on analyses of short reads that can potentially yield inaccurate BLAST results (15, 16), and our previous work showed that BLAST searches from single reads can result in false-positive identification of viral types (17). One recently reported approach, based on protein clustering, offers the potential for quantitative comparisons of viral functional diversity from metagenomic data (18, 19). However, since this approach is based on protein sequences, it cannot resolve populations at the nucleotide level, so a complementary approach with resolution at the level of viral operational taxonomic units (OTUs) is necessary.

Hypersaline environments are ideal model systems for studying viral assemblages, because geochemical conditions remain rel-

Received 11 June 2013 Accepted 23 August 2013
Published ahead of print 30 August 2013
Address correspondence to Joanne B. Emerson, joanne.emerson@colorado.edu.
* Present address: Joanne B. Emerson, Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Boulder, Colorado, USA.
Supplemental material for this article may be found at http://dx.doi.org/10.1128/AEM.01946-13.

doi:10.1128/AEM.01946-13

TABLE 1 Description of Lake Tyrrell (Victoria, Australia) samples

Sample	Sample collection date (mo/day/yr)	Sample collection time	Site[a]	TDS[b] (% [by wt])	Temp (°C)	pH	Sequencing type	Sequencing (Mb)	Viral richness (no. of viral populations per sample)	% abundance of the most abundant population	No. of populations at 0.1% abundance or higher
2007At1	1/23/2007	15:00	A	31	22	7.23	Illumina PE	356	502	2.3	294
2007At2	1/25/2007	15:00	A	31	28	7.09	Illumina PE	845	589	1.8	308
2009B[c]	1/5/2009	7:21	B	24	18	6.86	Illumina PE	2,162	652	3.8	236
		12:37		26	30	7.13					
		18:00		27	36	7.02					
2010Bt1	1/7/2010	7:45	B	32	20	7.23	454-Ti	248	562	3.4	314
2010Bt2	1/7/2010	20:00	B	36	32	7.25	Illumina PE	425	628	2.4	294
2010Bt3	1/8/2010	8:00	B	34	21	7.2	454-Ti	239	588	2.4	336
2010Bt4	1/10/2010	0:36	B	32	33	7.16	Illumina PE	1,039	672	3.4	296
2010A	1/10/2010	12:50	A	35	37	7.05	Illumina PE	1,103	663	2.1	312

[a] Sites A and B are isolated pools ~300 m apart.
[b] TDS, total dissolved solids (salinity).
[c] 2009B is a combination of three samples collected on the same day, pooled after DNA extraction.

atively constant, and the community is strictly microbial, eliminating complex interactions with higher (metazoan) trophic levels. Many studies of haloviral isolates and assemblages have been conducted previously (reviewed in references 20 to 23), showing, for example, a decrease in the diversity and an increase in the abundance of viruses with increasing salinity (24, 25). Previously, we tracked 35 complete and near-complete virus and virus-like genomes in hypersaline Lake Tyrrell (LT), Victoria, Australia, and we showed that most populations were stable over days but dynamic over years (17). Here, we comprehensively investigated the diversity, genetic composition, and dynamics of LT viral assemblages (hundreds of populations, relative to 35 in our previous analysis of the same samples). We developed new methods for characterizing and comparing viral assemblages of unknown phylogeny and genetic composition, including the identification of common genes to represent OTUs, and we investigated the influence of environmental parameters on LT viral assemblage composition.

MATERIALS AND METHODS

LT sample collection and DNA extraction, sequencing, and assembly. As described previously, eight viral concentrates were recovered from 10-liter surface water samples collected from hypersaline Lake Tyrrell (LT), Victoria, Australia, from two sites (A and B) between 2007 and 2010 (Table 1 and reference 17). Sampling, DNA extraction, and sequencing methods were described previously (17, 26–28). Briefly, water samples were sequentially filtered through 3.0-, 0.8-, and 0.1-μm polyethersulfone filters, and filtrates that had been put through 0.1-μm filters were concentrated through tangential flow filtration for virome recovery.

In this study, we use the same eight viral concentrate libraries and assemblies from our previous publication (17), and we expand our analyses to include all contigs > 500 bp (39,636 contigs, relative to 35 contigs analyzed previously). Assembly parameters and assembly verification were discussed in detail in reference 17, but briefly, 454-sequenced samples were assembled with Newbler (29) (samples 2010Bt1 and 2010Bt4), and two assemblers were tested on each of the Illumina-sequenced samples, with the best assembly chosen for each sample (briefly, the best assembly retained a balance between a large number of long contigs and relatively uniform coverage depth across contigs, as described in detail in reference 17). Samples 2007At1, 2007At2, 2010Bt2, 2010Bt4, and 2010A were assembled with ABySS (30), and sample 2009B was assembled with Velvet (31). Because fragment recruitment to all assemblies was used for each sample (see below), differences in assembly parameters should not significantly affect the results. Unless otherwise noted, all analyses beyond assembly, beginning with pulling contigs > 500 bp, are new and specific to this study. Sequencing reads were uploaded to GenBank previously (BioProject identification [ID] PRJNA81851), and the 30,525 gene sequences from the current analysis (described shortly) are available through MG-RAST (ID 4529512.3) (32).

Statistical analyses. Unless otherwise noted, statistical analyses were performed in the R programming language, and the software functions mentioned for each analysis (i.e., anosim, hclust, bioenv, mantel, and rarefy) are part of the vegan package (33). Unless otherwise mentioned, all protein and nucleic acid sequence-based clustering was performed using UCLUST version 4.2.66 (34), with identity for proteins set at 0.4 and nucleic acids set at 0.95.

Generation of reference gene sequences and estimates of viral abundance. In order to generate reference gene sequences for fragment recruitment, we used Prodigal (35) to predict genes on all contigs > 500 bp. To increase the possibility that genes present in the system would be detected through fragment recruitment in downstream analyses, we removed all predicted genes < 300 bp. We recognize that small viral genes would be excluded from this analysis. We then made the gene sequences nonredundant by clustering at 95% nucleotide identity, yielding 30,525 genes. For abundance estimates, reads from each sample were recruited to these reference gene sequences, and the number of reads recruited to a given gene in a given sample was retained as the "mapping count." In accordance with the cutoff used for clustering, fragment recruitment was carried out using gsMapper (29) with a minimum of 95% nucleotide identity and an overlap length of at least 70 bp. We chose the 95% nucleotide identity threshold for clustering and fragment recruitment to be representative of a viral population, based on tests across 70, 80, 90, and 95% nucleotide identity (see supplemental material). However, we acknowledge that, without isolates, we cannot accurately define a nucleotide threshold for a viral population, and the threshold likely differs across genes and viral groups. Therefore, we use the terms "population" and "OTU" loosely to mean a group of viruses that we expect to be genetically similar, based on high (95%) nucleotide sequence identity across a single gene.

Functional predictions and selection of signature gene groups. We used InterProScan (36) to predict functional domains in predicted protein sequences. Additionally, all predicted protein sequences were clustered at 40% amino acid identity, under the assumption that each cluster would be likely to contain proteins of similar function. It is generally accepted that ~30% amino acid identity tends to yield robust protein alignments and indicate similar protein function (37), and as one specific example, ~20 to 40% amino acid identity suggests similar function in reoviruses (38). Using this as a guideline, we chose 40% as an amino acid

identity threshold, which is on the conservative end of these previous reports. However, it should be noted that no reoviruses—or members of any previously recognized viral group—were detected in the LT system, and we acknowledge that this cutoff does not guarantee similar function. In fact, 0.3% of our clusters contained mixtures of predicted functions, and it is likely that more mixtures exist in clusters of unknown function.

In lieu of a universal marker gene, seven sets of viral genes (seven "signature gene groups") were chosen for analyses of diversity and dynamics, with sequences in each signature gene group representing individual OTUs (viral populations). Specific descriptions of each signature gene group are given in the next two paragraphs, but in general, although no gene is present in each viral population, we consider sequence variants within each signature gene group to be representative of individual populations. We chose seven signature gene groups under the assumption that the combination of these groups would (i) maximize the number of viruses included in community analyses and (ii) minimize biases associated with each signature gene group. In order to accomplish these goals, we did not require all signature gene groups to be functionally or phylogenetically cohesive.

The first signature gene group includes all 30,525 genes from the LT metagenomes, and as such, it is the best signature gene group for maximizing the number of viruses included. However, because there is no limit to the number of genes from a given genome that can be present in the 30,525 genes ("all genes") signature gene group, this group contains multiple genes per genome (up to whole genomes in some cases). The second signature gene group includes all LT genes annotated as any type of "methyltransferase," including amino acid, nucleic acid, and other methyltransferases. This group was chosen because methyltransferase is the most common annotation in the data set and is therefore likely to represent a large number of viruses. Methyltransferase is also the most common annotation in the seven complete viral genomes previously reported from LT (17). However, we know from our previous work that some LT viral genomes contain multiple methyltransferases, whereas others contain no methyltransferases (there were two copies in two genomes, one copy in two genomes, and no copies in three genomes, with an average of 6/7 methyltransferases per genome), so this group will sample some viruses more than once and others not at all. The third and fourth signature gene groups (concanavalin A-like glucanases/lectins, henceforth referred to as "glucanases," and "major capsid proteins") were similarly defined based on common annotation and relatively high abundance and because they are functions likely to be associated with viruses (glucanases are predicted to be involved in host cell binding and entry [39], while major capsid proteins are structural components of the virion).

The final three signature gene groups were chosen as the largest clusters of unknown function (clusters 261, 667, and 1435), grouped by amino acid identity. Together with the 30,525-gene group, these three groups were chosen in order to ensure that our OTU-based analyses were not biased to include only viral genes with representatives in public databases. Within each signature gene group, a threshold of 95% nucleotide similarity was used to define OTUs, according to the clustering and fragment recruitment approaches described above. The presence/absence and relative abundance of each OTU were calculated for each sample through fragment recruitment as described above.

Hierarchical clustering. Several normalizations and transformations were tested in order to determine the most reasonable parameters for generating distance matrices (see the supplemental material). Based on the results of these tests, mapping counts were normalized by gene length (dividing the average length of all genes by a given gene length and using that as a multiplier for all mapping counts for that gene) and by the sequencing effort (dividing the average number of reads in all samples by the number of reads in a given sample and using that as a multiplier for all mapping counts for that sample). For each signature gene group, normalized mapping counts for each OTU and sample are provided in the supplemental material as Tables S1 to S7. We generated Bray-Curtis dissimilarity matrices from the normalized mapping counts and used ANOSIM (40) with 1,000 permutations to test for statistically significant differences between sample groups. Samples were grouped for ANOSIM analysis by year and location, i.e., four groups, group 2007A (two samples), 2009B (one sample), 2010B (four samples), and 2010A (one sample). Hierarchical clusters were generated for each signature gene group, using the program MeV (41) with a Pearson correlation and average linkage clustering.

Diversity measurements. Using the normalized mapping counts, we calculated Shannon's diversity index (42), Simpson's diversity index (43), richness (number of OTUs detected), and Pielou's evenness (44) for LT samples, and we ranked samples for each index and signature gene group (Fig. 1). We also developed richness and dominance estimates based on our data. For one estimate, we divided the total number of genes (30,525 genes) by the average number of predicted genes (74 genes) in the seven viral genomes from our previous work (17). To generate the richness values in Table 1, we multiplied the number of methyltransferases in each sample by 7/6, based on the representation of methyltransferases in the seven previously reported genomes. That calculation was also applied to all methyltransferases in the data set to predict total richness across samples. Those calculations were not applied to any of the other signature gene groups because of the relatively low representation (or nondetection) of the other groups in the seven sequenced LT genomes, preventing normalization according to genomic representation. On the basis of the relative abundance of each methyltransferase gene within each sample, we also estimated the percent abundance of the most dominant OTU and the number of populations at 0.1% abundance or higher in each sample (Table 1).

Correlations with environmental data. We correlated LT viral assemblage structures with environmental parameters (temperature, total dissolved solids, pH, and solution chemistry; see Table 1 and previously reported geochemical data [17]). 2009B is a pool of DNA from three samples collected on the same day, so we used the average of three measurements for each environmental variable for that sample. Using normalized mapping counts for each viral OTU, we used the bioenv function (33) to do the following: (i) calculate a Bray-Curtis community dissimilarity matrix, (ii) select all possible combinations of up to 6 subsets of the 13 environmental variables, (iii) calculate a Euclidean distance matrix for each subset of environmental variables across samples, and (iv) use Spearman correlations to identify the subset of environmental variables with the best rank correlation with the viral community dissimilarity matrix. To generate *P* values, we tested 1,000 permutations of the distance matrix for environmental data, using the mantel function (33).

RESULTS

Annotation and protein cluster analyses. Methyltransferases were the most abundant annotation in the data set, and glucanases, predicted to be involved in archaeal host cell recognition (39), were also relatively abundant (Fig. 2). Integrases, which are common markers for temperate viruses, were relatively rare (0.02% of LT viral concentrate reads mapped to integrases). Of the 11 largest protein clusters, 10 are of unknown function (Table 2). However, only one large cluster had no match in the GenBank nr database. Nine had highly significant BLAST hits to proteins of unknown function predicted from other hypersaline systems, and one is a putative terminase with significant BLAST hits to terminases from five different haloviruses from Spain (45).

We also compared our protein clusters to existing protein clusters reported in an analysis of Pacific Ocean Virome (POV) data, which included protein clusters from Global Ocean Sampling (GOS) data (46, 47), proteins from complete phage genomes, and new clusters from POV metagenomes (18). Because the UCLUST algorithm was updated between our LT analyses and this comparison of LT and POV protein clusters, we used the updated version (v6) (34) for the comparison. There were originally 4,238,638 clusters in the POV data set (18). To make the POV data compa-

		All Genes	Meth.	Gluc.	Capsids	C667	C261	C1435
2007At1	Shannon's	8.596	5.288	4.411	2.289	3.615	3.152	3.146
2007At2		8.960	5.511	4.491	2.638	3.673	3.737	3.178
2009B		8.979	5.566	4.548	2.901	3.675	3.766	3.575
2010Bt1		8.983	5.590	4.554	3.040	3.677	3.791	3.617
2010Bt2		9.054	5.609	4.962	3.065	3.681	3.802	3.677
2010Bt3		9.131	5.669	5.090	3.114	3.707	3.817	3.709
2010Bt4		9.144	5.716	5.110	3.124	3.748	3.825	3.724
2010A		9.365	5.749	5.163	3.556	3.785	3.836	3.753
	Simpson's	0.9995	0.9906	0.9654	0.8004	0.9556	0.9344	0.9196
		0.9997	0.9935	0.9698	0.8496	0.9611	0.9656	0.9503
		0.9997	0.9936	0.9723	0.8819	0.9633	0.9708	0.9637
		0.9997	0.9938	0.9732	0.9247	0.9652	0.9712	0.9648
		0.9998	0.9944	0.9989	0.9268	0.9667	0.9715	0.969
		0.9998	0.9946	0.9903	0.9296	0.9668	0.9723	0.9698
		0.9998	0.9948	0.9905	0.9298	0.9673	0.9738	0.9701
		0.9998	0.9953	0.9908	0.9603	0.9675	0.9745	0.9705
	Pielou's Evenness	0.8489	0.834	0.7816	0.559	0.8481	0.7607	0.7795
		0.8956	0.890	0.7822	0.697	0.8754	0.9125	0.8525
		0.9001	0.897	0.7907	0.711	0.8809	0.9243	0.8946
		0.9008	0.900	0.8185	0.760	0.8833	0.9248	0.9096
		0.9068	0.903	0.8780	0.761	0.8853	0.9267	0.9139
		0.9085	0.905	0.8839	0.773	0.8909	0.9285	0.9165
		0.9163	0.908	0.9040	0.809	0.8976	0.9293	0.9174
		0.9198	0.917	0.9305	0.849	0.9028	0.9296	0.9261
	Richness	18101	434	207	41	59	57	50
		20900	487	259	47	60	59	53
		21284	510	281	51	64	60	55
		22621	510	285	56	65	62	57
		23618	546	293	60	69	62	57
		24991	566	317	61	69	62	59
		25819	575	339	64	70	62	60
		26403	583	358	66	71	63	60

FIG 1 Lake Tyrrell (LT) (located in Victoria, Australia) viral diversity index calculations and sample rankings. Viral signature gene groups are shown in columns, and indices are shown in four horizontal blocks. The viral signature gene groups are all genes, methylases (Meth.), glucanases (Gluc.), major capsid proteins (capsids), and protein clusters 667 (C667), 261 (C261), and 1435 (C1435). There is one calculation per sample (eight samples) per index (four indices) per viral gene group (seven groups), and these calculations are the numbers in the table. Richness is a measure of the number of OTUs detected, and the other indices are measured according to the references provided in the text. Colors indicate specific samples (legend on the left), ranked in ascending vertical order within each index and signature gene group.

rable to ours, we first removed any protein sequences less than 100 amino acids in length, which reduced the number of clusters to 1,912,551. We then reclustered the sequences at 40% amino acid identity, resulting in 637,619 POV clusters for our analysis. We then clustered these POV clusters with our 14,927 LT protein clusters (for consistency, these LT clusters were regenerated from the original 30,525 LT protein sequences, using UCLUST v6). In total, we generated 651,107 clusters, representing an overlap of only 1,439 clusters between the two data sets and an addition of 13,488 new clusters from LT, meaning that 90% of the LT protein clusters are novel.

Viral assemblages across LT samples. We found significant correlations between the structures of LT viral assemblages and environmental factors, particularly salinity and potassium concentration (Table 3). Comparisons of the viral assemblages via hierarchical clustering resulted in the same sample groupings, regardless of which signature gene group was chosen to define the OTUs (Fig. 3; see Fig. S1 in the supplemental material). Samples from the same site and year consistently grouped together and separately from samples collected during different years and/or from the other site (of the two sites). Specifically, the two samples from the 2007 2-day time series (2007At1 and 2007At2) grouped together, as did the four samples from the 2010 4-day series (2010Bt1 to 2010Bt4), and samples 2009B and 2010A grouped separately from the other two groups. These groups were supported by ANOSIM analyses, which revealed high within-group similarity ($r = 0.93$ to 1) and highly significant P values (<0.006) for all signature gene groups, except for cluster 261 ($r = 0.5$; P value, 0.06). Within the 2010 4-day time series at site B, the samples grouped in different orders, depending on the signature gene group analyzed, so any patterns in viral assemblage structures on the time scale of days could not be resolved.

The methyltransferase hierarchical cluster is representative of hierarchical clustering analyses across signature gene groups and is shown in Fig. 3 (clustering of other signature gene groups is shown in Fig. S1 in the supplemental material). In addition to relationships among viral assemblages across samples, which are shown in the tree topology at the top of Fig. 3, the presence/absence and relative abundance of individual viral OTUs can be inferred from each horizontal row of the heat map. Importantly, this hierarchical cluster is a visual representation of viral assemblage structure that retains resolution of the behavior of individual populations. While overarching differences at the viral assemblage level could not be resolved on the time scale of days, as described above, dynamics are clearly visible at the level of individual populations on all temporal and spatial scales included in this study (days to years, two sites separated by ~300 m).

Diversity of LT viral assemblages. We calculated four univariate diversity indices for each of the seven LT viral signature gene groups and ranked samples for each index and gene group (Fig. 1). Sample rankings were highly variable across both gene groups and indices, and for each index, the range of values for each gene group was different, particularly for Shannon's diversity index (42). Few patterns were discernible, though sample 2007At1 was consis-

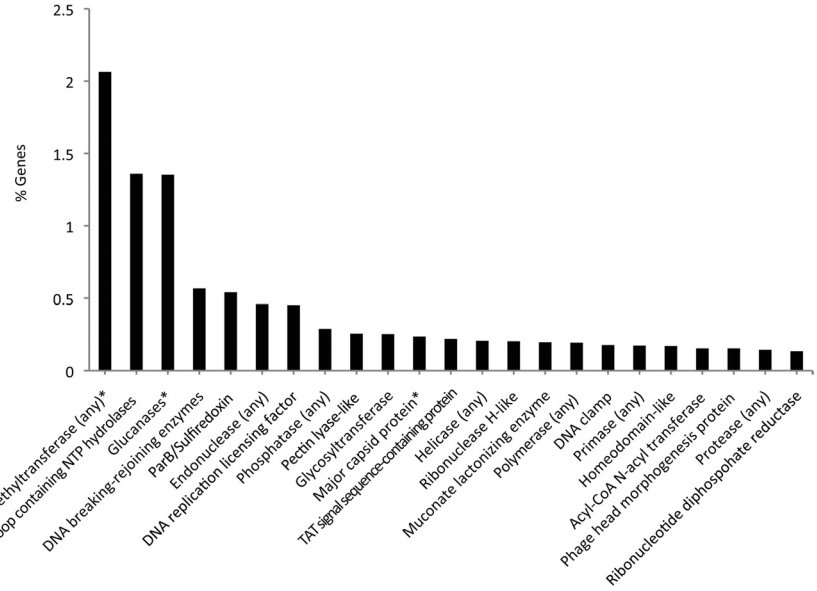

FIG 2 Abundance of predicted functions in Lake Tyrrell (Victoria, Australia) viral concentrates. The most common predicted functions in Lake Tyrrell are shown on the *x* axis in order of their abundance. The percentage of total genes (all 30,525 genes, including genes with no annotation) is on the *y* axis. Asterisks along the *x* axis indicate viral signature gene groups. NTP, nucleoside triphosphate; acyl-CoA, acyl coenzyme A.

tently the least rich and sample 2009B was often the least diverse and least even, despite being among the most rich. However, for the glucanase group, sample 2009B was among the most diverse and most even, so the significance of these trends is unclear.

Richness and dominance calculations developed specifically for this study, as described in Materials and Methods, are in Table 1. Richness estimates ranged from 502 to 672 viral populations per sample, and estimates of dominance (the number of OTUs at 0.1% abundance or higher) ranged from 236 to 336 per sample. For the entire LT viral community (all samples), the all-genes signature gene group calculation predicted 412 viral populations, and the methyltransferase signature gene group predicted 735 populations. Together with the richness data from Table 1, this suggests that 68 to 91% of the total viral richness that we were able to capture across the eight samples was present in each sample. We also used rarefaction curves from the methyltransferase signature gene group (Fig. 4), the all-genes group (see Fig. S2a in the supplemental material), and across protein clusters (Fig. S2b) to assess the amount of diversity that we were able to capture across LT samples. Curves for all samples, except for 2009B, become nearly flat as they approach an asymptote, suggesting that we have sampled most of the diversity in the LT system. 2009B is the only pooled sample, representing three samples collected on the same day, which could explain the difference in its rarefaction curve. It should be noted that these analyses do not account for reads that may belong to genes that were not assembled, so they are estimates of the amount of diversity captured for moderate- to high-abundance viral populations. Interestingly, despite the population and assemblage dynamics just described, the diversity remained relatively constant across samples (between 434 and 583 methyltransferase genes per sample, Fig. 1 and 4; also see the richness values in Table 1).

TABLE 2 Largest LT protein clusters (40% amino acid identity) and BLAST hits

Cluster[a]	No. of proteins	Top BLAST hit	E value[b]	No. of halovirus hits[b]
667*	71	eHP-36 halovirus	2.00E−28	2
261*	63	eHP-32 halovirus	1.00E−39	6
1435*	60	eHP-36 halovirus	6.00E−77	1
214	50	eHP-36 halovirus	2.00E−40	1
342	49	eHP-36 halovirus	9.00E−72	1
271	45	eHP-36 halovirus	8.00E−83	5
369	42	Uncultured halovirus	4.00E−42	3
356	41	None	N/A	N/A
157	41	eHP-32 halovirus, terminase	3.00E−71	5
210	41	eHP-36 halovirus	3.00E−68	5
24	40	eHP-36 halovirus	3.00E−80	5

[a] Clusters with an asterisk are also LT viral signature gene groups.
[b] N/A, not applicable.

DISCUSSION

In this study, we developed new metagenomic assembly-based techniques for comparing viral assemblages, and we used these

TABLE 3 Correlations between LT viral community structure and environmental factors

Viral gene group	Environmental factors[a]	Spearman's correlation coefficient	*P* value
All 30,525 genes	TDS, pH, K, F	0.6716	0.068
Methyltransferases	TDS, K, Mg, F, Br, SO_4	0.7252	0.038
Glucanases	TDS, temp, pH, K, F	0.5534	0.052
Capsid proteins	F, Br, SO_4	0.8079	0.095

[a] Subset of factors with the most significant correlations in bioenv analysis. TDS, total dissolved solids (% by weight); temp, temperature (°C); SO_4, sulfate concentration. The chemical symbols (K, F, Mg, etc.) indicate the concentration of that ion in solution.

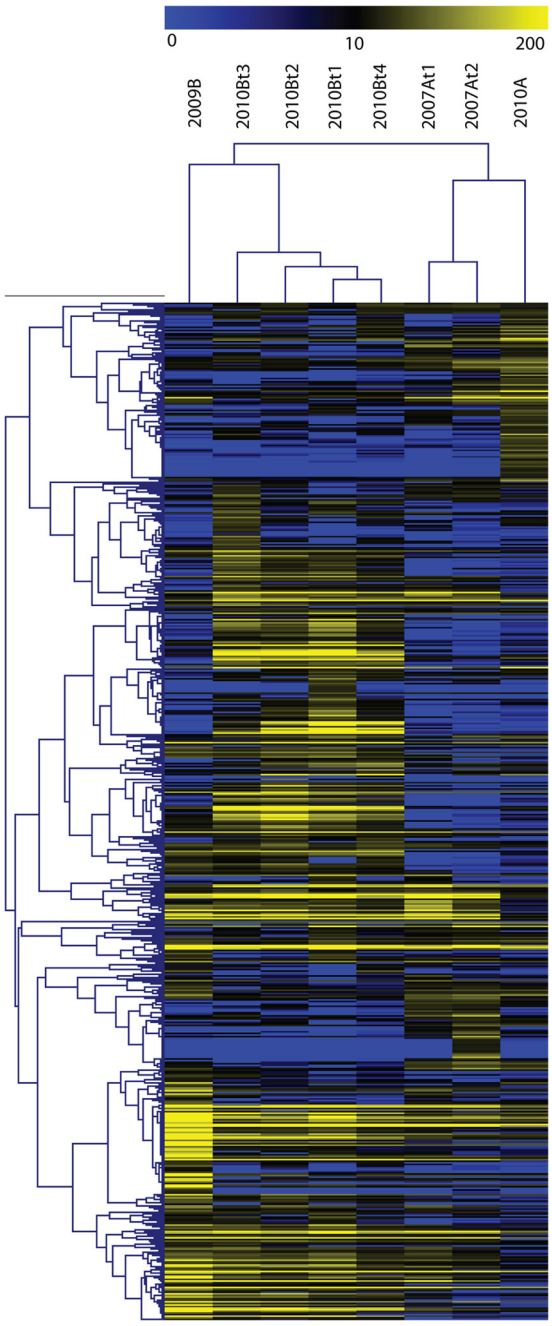

FIG 3 Hierarchical clustering of Lake Tyrrell (Victoria, Australia) viral methyltransferase OTUs. Hierarchical clustering analysis (Pearson correlation, average linkage clustering) of the relative abundances (normalized mapping counts; scale bar at top) of LT viral methyltransferase OTUs. The samples are clustered across the top, and methyltransferase gene sequences (OTUs) are clustered on the left.

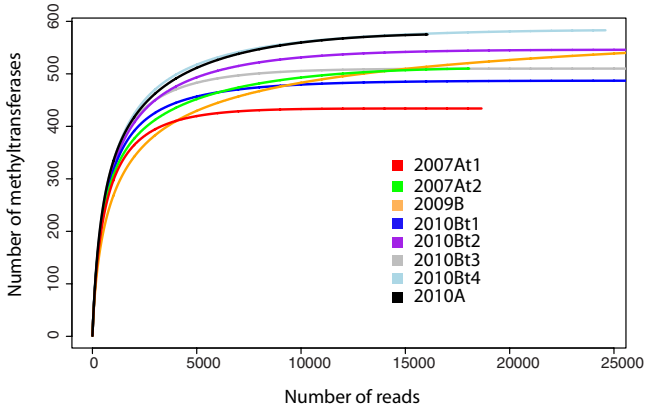

FIG 4 Methyltransferase rarefaction curves. Rarefaction curves were generated, using the function "rarefy" from the vegan package in the R programming language (33), using a matrix of normalized mapping counts from the methyltransferase gene group as input (converted to integers, as required by the program). For each sample, we display the cumulative number of new methyltransferase genes identified, as reads were selected at random from the matrix.

techniques to do the following: (i) provide direct, high-resolution estimates of viral assemblage diversity in a natural system, (ii) provide a means of resolving and visualizing viral assemblage and population dynamics simultaneously, (iii) correlate viral assemblage dynamics with environmental parameters, and (iv) identify viral proteins of unknown function that are likely to be adapted to hypersaline systems. The techniques that we describe for estimating diversity and comparing viromes can easily be applied to other deeply sequenced viral metagenomes by identifying signature gene groups from assembled data, based on common functional annotation and protein clustering, and then counting OTUs in each sample (e.g., through fragment recruitment to gene sequences). Matrices of counts of OTUs across samples can be used for a variety of ecological analyses, including those presented in this study. We have shown that these analyses work for a reasonably diverse system that contains hundreds of viral populations. Given that there is already precedent for scaling metagenomic assembly-based approaches to microbial systems of increasing complexity (48, 49), we suspect that our approaches can be applied to deeply sequenced viromes across a range of ecosystems.

We identified OTUs directly from our metagenomic data, avoiding reliance on public databases and on models that are designed for much lower sequencing throughput (10). This enabled the characterization of the diversity and dynamics of viral assemblages with simultaneous resolution of individual population behavior. This is a significant advance, as these analyses are typically decoupled. For example, viral assemblage diversity could be estimated from models or through PFGE and compared across samples, but the behavior of individual populations would be lost. Alternatively, a specific population or group of viruses could be followed through PCR amplification across samples, but the diversity and dynamics of the rest of the assemblage would be unknown.

De novo analyses of virome functional diversity have recently been reported through protein clusters (18, 19), and we provide a means for increasing the resolution of these analyses. In previous viral protein cluster analyses, each cluster represented a single OTU (a single entity for counting), with abundances for that OTU drawn from the number of proteins contained in the cluster. In this study, each sequence recruited to a cluster is a separate OTU. We achieve this by using the corresponding gene sequence for each protein sequence and then measuring the abundance of each gene sequence (OTU) through read recruitment. In addition to protein clustering, read clustering has been used previously as a measure of virome diversity (16). Relative to read clustering, our

analyses add organization into genes and OTUs to give more context to the measurements, and we add longer reference sequence lengths (>300 bp, as opposed to 100 bp) for more robust clustering and read recruitment.

We appreciate, particularly for all the genes and methyltransferase signature gene groups, that we may be counting an individual (in this case, a viral genome) more than once while using ecological techniques that assume that each individual will be counted only once. Where possible, we have accounted for this by normalizing by the average number of genes (all-genes group) or copies of methyltransferases per genome. Similarly, we acknowledge that we will not be able to account for all individuals in all signature gene analyses because not all signature genes will be present in all viruses. We have attempted to account for this by choosing a variety of signature gene groups and comparing results across groups. Presumably, our most robust results are those that are consistent across signature gene groups and are therefore likely to represent most of the viral assemblage.

We acknowledge that metagenomic assembly algorithms are not perfect, and our assemblies could contain errors, even after the careful manual assembly curation that we reported previously for these samples (17, 50, 51). However, we stress a number of points in support of this approach. (i) Metagenomic libraries were prepared without multiple-displacement amplification, reducing biases in the sequencing data prior to assembly (9, 52). (ii) Because no reference haloviral genomes were detected in LT metagenomes, a *de novo* method for characterizing these assemblages was required. (iii) Our assemblies were carefully manually curated (17), and similar manual curation methods have resulted in significant metagenomic assembly of bacteria and archaea in a variety of systems (i.e., genomes have been accurately reconstructed, based on coverage and the presence of single-copy marker genes, among other measures) (48, 49, 53–56). (iv) Longer reads (and, by inference, contigs) allow for more reliable BLAST searches and annotation (15), and the annotation supports the identification of virus-like genes (Fig. 2). (v) Any assembly errors will be the same across samples, because the assemblies were not directly compared (reads were recruited to a nonredundant set of genes collated from all assemblies), and because we are "splitting" the assemblies into individual genes, any contig-scale assembly errors should be minimized.

It should be noted that we cannot rule out the possibility of a small amount of contamination from plasmids and other free DNA in our viral concentrate libraries, due to the lack of a DNase treatment prior to virion lysis (a DNase treatment was attempted but resulted in complete degradation of all DNA [17]). Through comparisons to 16S rRNA gene sequences and known plasmid sequences, along with searches for plasmid genes in our libraries, we demonstrated previously that our libraries are dominated by viruses and that any such contamination is likely to be minimal (17). All of our viral signature gene groups are consistent with viral sequences, and two groups (glucanases and major capsid proteins) were chosen specifically because they are likely to contain viral sequences exclusively.

Diversity and dynamics of LT viral assemblages. Persistence of viral populations over days but variation over the 3-year study period and between isolated pools separated by ~300 m is consistent with our previous report of dynamics in most virus and virus-like populations from the same LT samples on the same temporal and spatial scales (17). Together, these two studies indicate that the LT viral populations that were abundant enough to assemble into contigs > 500 bp (the cutoff for this study) tended to exhibit similar dynamics to populations that assembled into contigs > 10 kb (the cutoff for the previous study, presumed to represent the most abundant populations in these assemblages, given their significant assembly). We take this to mean that viral populations of both moderate and high abundance exhibit similar dynamics in the LT system.

The timescales on which LT viral assemblages tended to be stable (days) or dynamic (years) are generally consistent with previous studies of both viral and microbial communities in other systems. For example, marine myoviruses off the California coast exhibited both dynamics and persistence, with myoviral assemblages most similar during adjacent months (the shortest time scale in the study) (7). Similarly, at a nearby site, the dominant members of myoviral and bacterial assemblages tended to remain dominant on the time scale of days to weeks (57). However, interestingly, nearly all microbial OTUs found across 72 16S rRNA gene amplicon samples collected over 6 years in the western English Channel (and many from sites throughout the global oceans) were present in a single, very deeply sequenced sample, suggesting that most bacterial taxa are always present but shift in relative abundance (58, 59). This suggests that, while most LT viral populations appear absent (below detection limits) in at least some samples over 3 years, it is possible that they are actually present at low abundance. Spatially, our results are consistent with a study of 32 samples from two freshwater lakes, which revealed high within-lake similarity relative to between-lake similarity (60). We also observed similarities at the same site (across the four 2010B assemblages) and differences between sites (between the 2010A and 2010B assemblages).

Since viral diversity within LT remained relatively constant throughout the 3-year study (e.g., see the richness calculations in Table 1), we infer that viral population and assemblage dynamics occur over relatively short timescales (<3 years), while the total diversity of the system remains relatively constant over time and space. In model form, this would be conceptually similar to the constant diversity dynamics model (61) and other similar models (14, 62), which predict stability of viral and microbial populations with dynamics at the subpopulation (strain) level. However, we would change the scale to indicate constant diversity at the viral assemblage level with dynamics at the level of individual viral populations.

Although analyses of viral assemblage structure grouped samples collected at similar times from the same location (regardless of which of the seven LT viral signature gene groups was used), the four univariate diversity indices indicate that samples collected from the same location over days have somewhat different diversity levels. However, it is concerning that the diversity level rankings varied, based on which signature gene group was used for the analysis. This suggests that these widely used univariate diversity indices (designed originally for macroecological data and more recently applied to microbial ecological data [for example, reference 63]) may not be appropriate for comparing viral assemblages. This may be because the LT samples are from the same ecosystem and have relatively constant diversity, and/or because these indices were designed for use with phylogenetic marker genes, which are lacking in viruses. For viruses, we are necessarily restricted to functional genes, each of which potentially provides a different measure of functional diversity that does not necessarily

equate to phylogenetic diversity. To reduce potential biases associated with analyses based on functional genes, we included a normalization factor in our richness and dominance estimates, based on the relative representation of our OTU genes in sequenced genomes from LT. Although this correction factor can only be used in systems from which representative whole-genome sequences can be reconstructed, viral genome assembly is becoming more facile in a variety of systems through Illumina metagenomic sequencing (17, 64).

Functional predictions in LT viral assemblages. Although only ~15% of the 30,525 genes from LT could be assigned a function, and the most abundant functional annotation (methyltransferase) represented only ~2% of LT reads (Fig. 2), some inferences can be made from the functional annotation. The prevalence and diversity of glucanases in LT suggest that haloviruses, probably targeting the abundant archaea (75 to 95% of the microbial community across LT samples [65]), have evolved a variety of surface receptors for host cell recognition (39). The presence of a number of twin-arginine translocation (TAT) signal sequences, which likely serve to target proteins for secretion in an already folded state (66), suggests that haloviruses have evolved means of ensuring that proteins rapidly folded under high intracellular salt concentrations can still be embedded in or secreted from host membranes (e.g., in preparation for the generation of viral envelopes). The identification of relatively few capsid proteins suggests either that haloviral structural genes are poorly represented in public databases and/or that such proteins are highly conserved across diverse viral groups. In the case of the latter, we would expect to encounter stability in major capsid proteins across time and space in the LT system, which we did not see, so we hypothesize that many more haloviral structural genes remain to be discovered and therefore are likely to be good targets for future proteomic studies. Consistent with some other studies (for instance, reference 67), the relatively small number of integrases (less than 0.02% of LT reads) suggests that temperate viruses may not be abundant members of planktonic haloviral assemblages, though given the novelty of LT viruses, we cannot discount the possibility that novel integrase genes were not recognized or that temperate viruses could be abundant under conditions not sampled by this study.

Correlations between LT viral assemblages and environmental factors. Significant correlations with LT viral assemblage structures were identified for various combinations of environmental factors, with the most highly correlated factors generally differing by viral signature gene group (Table 3). Environmental factors may indirectly drive shifts in specific viral populations, presumably by influencing the relative abundances of their hosts (68, 69). It is also conceivable that some environmental factors could directly select for different viral groups, for example, by promoting or reducing viral decay (68, 70–72). Significant correlations with LT viral assemblage structure were associated with subtle shifts in both salinity (total dissolved solids [TDS]) and potassium concentrations for three viral gene groups (all 30,525 genes, methyltransferases, and glucanases), consistent with previous observations across more extreme differences in salinity (25). These observations likely reflect different adaptations of host populations to solution chemistry, which changes with the extent of evaporative concentration.

Conclusions. In this direct estimate of viral assemblage diversity in a natural system, we show that viral assemblages in hypersaline Lake Tyrrell, Victoria, Australia, are diverse, containing ~412 to 735 populations at moderate-to-high abundance. Although some LT viral populations were dynamic over days, viral assemblages were generally stable at the same site over days and dynamic over years, and viral assemblage diversity remained relatively constant throughout the study. Salinity was shown to correlate with viral assemblage structure, and we infer that salinity may be a driver of host population dynamics. The techniques that we describe for estimating diversity, comparing viromes, and determining potential environmental influences on viral assemblages should be broadly applicable to deeply sequenced viromes across ecosystems. In addition, we provide a means of counting viral OTUs across samples to generate a simple data matrix that can be used as the foundation for many ecological analyses, potentially linking environmental virologists to tools from the well-developed fields of macroecology and microbial ecology.

ACKNOWLEDGMENTS

Funding for this work was provided by National Science Foundation award 0626526 and Department of Energy award DE-FG02-07ER64505.

We thank Cheetham Salt Works (Victoria, Australia) for site access; John Moreau, Jochen Brocks, Eric Allen, and Mike Dyall-Smith for field assistance; Shannon Williamson and Doug Fadrosh for training J.B.E. in virus-related laboratory techniques; and members of the University of California (UC) Berkeley Dimensions of Biodiversity Distributed Graduate Seminar (National Science Foundation award 1050680) for helpful discussions. We also thank three anonymous reviewers for thoughtful comments that improved the manuscript.

REFERENCES

1. **Suttle CA.** 2007. Marine viruses—major players in the global ecosystem. Nat. Rev. Microbiol. **5:**801–812.
2. **Rosario K, Breitbart M.** 2011. Exploring the viral world through metagenomics. Curr. Opin. Virol. **1:**289–297.
3. **Helton RR, Wang K, Kan J, Powell DH, Wommack KE.** 2012. Interannual dynamics of viriobenthos abundance and morphological diversity in Chesapeake Bay sediments. FEMS Microbiol. Ecol. **79:**474–486.
4. **Wommack KE, Bench SR, Bhavsar J, Mead D, Hanson T, Clokie MRJ, Kropinski AM.** 2009. Isolation independent methods of characterizing phage communities 2: characterizing a metagenome. Methods Mol. Biol. **502:**279–289.
5. **Tucker KP, Parsons R, Symonds EM, Breitbart M.** 2011. Diversity and distribution of single-stranded DNA phages in the North Atlantic Ocean. ISME J. **5:**822–830.
6. **Marston MF, Sallee JL.** 2003. Genetic diversity and temporal variation in the cyanophage community infecting marine *Synechococcus* species in Rhode Island's coastal waters. Appl. Environ. Microbiol. **69:**4639–4647.
7. **Chow C-ET, Fuhrman JA.** 2012. Seasonality and monthly dynamics of marine myovirus communities. Environ. Microbiol. **14:**2171–2183.
8. **Sullivan MB, Coleman ML, Quinlivan V, Rosenkrantz JE, DeFrancesco AS, Tan G, Fu R, Lee JA, Waterbury JB, Bielawski JP, Chisholm SW.** 2008. Portal protein diversity and phage ecology. Environ. Microbiol. **10:**2810–2823.
9. **Duhaime MB, Sullivan MB.** 2012. Ocean viruses: rigorously evaluating the metagenomic sample-to-sequence pipeline. Virology **434:**181–186.
10. **Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, Salamon P, Felts B, Nulton J, Mahaffy J, Rohwer F.** 2005. PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. BMC Bioinformatics **6:**41. doi:10.1186/1471-2105-6-41.
11. **Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F.** 2002. Genomic analysis of uncultured marine viral communities. Proc. Natl. Acad. Sci. U. S. A. **99:**14250–14255.
12. **Angly FE, Willner D, Prieto-Davo A, Edwards RA, Schmieder R, Vega-Thurber R, Antonopoulos DA, Barott K, Cottrell MT, Desnues C, Dinsdale EA, Furlan M, Haynes M, Henn MR, Hu Y, Kirchman DL, McDole T, McPherson JD, Meyer F, Miller RM, Mundt E, Naviaux RK,**

Rodriguez-Mueller B, Stevens R, Wegley L, Zhang L, Zhu B, Rohwer F. 2009. The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. PLoS Comput. Biol. **5:**e1000593. doi:10.1371/journal.pcbi.1000593.

13. **Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, McDaniel L, Moran MA, Nelson KE, Nilsson C, Olson R, Paul J, Brito BR, Ruan Y, Swan BK, Stevens R, Valentine DL, Thurber RV, Wegley L, White BA, Rohwer F.** 2008. Functional metagenomic profiling of nine biomes. Nature **452:**629–632.
14. **Rodriguez-Brito B, Li L, Wegley L, Furlan M, Angly F, Breitbart M, Buchanan J, Desnues C, Dinsdale E, Edwards R, Felts B, Haynes M, Liu H, Lipson D, Mahaffy J, Martin-Cuadrado AB, Mira A, Nulton J, Pasic L, Rayhawk S, Rodriguez-Mueller J, Rodriguez-Valera F, Salamon P, Srinagesh S, Thingstad TF, Tran T, Thurber RV, Willner D, Youle M, Rohwer F.** 2010. Viral and microbial community dynamics in four aquatic environments. ISME J. **4:**739–751.
15. **Wommack KE, Bhavsar J, Ravel J.** 2008. Metagenomics: read length matters. Appl. Environ. Microbiol. **74:**1453–1463.
16. **Roux S, Enault F, Robin A, Ravet V, Personnic S, Theil S, Colombet J, Sime-Ngando T, Debroas D.** 2012. Assessing the diversity and specificity of two freshwater viral communities through metagenomics. PLoS One **7:**e33641. doi:10.1371/journal.pone.0033641.
17. **Emerson JB, Thomas BC, Andrade K, Allen EE, Heidelberg KB, Banfield JF.** 2012. Dynamic viral populations in hypersaline systems as revealed by metagenomic assembly. Appl. Environ. Microbiol. **78:**6309–6320.
18. **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:**e57355. doi:10.1371/journal.pone.0057355.
19. **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:**1428–1440.
20. **Dyall-Smith M, Tang S-L, Bath C.** 2003. Haloarchaeal viruses: how diverse are they? Res. Microbiol. **154:**309–313.
21. **Porter K, Russ BE, Dyall-Smith ML.** 2007. Virus-host interactions in salt lakes. Curr. Opin. Microbiol. **10:**418–424.
22. **Pina M, Bize A, Forterre P, Prangishvili D.** 2011. The archeoviruses. FEMS Microbiol. Rev. **6:**1035–1054.
23. **Santos F, Yarza P, Parro V, Meseguer I, Rossello-Mora R, Anton J.** 2012. Culture-independent approaches for studying viruses from hypersaline environments. Appl. Environ. Microbiol. **6:**1635–1643.
24. **Sandaa R-A, Foss Skjoldal E, Bratbak G.** 2003. Virioplankton community structure along a salinity gradient in a solar saltern. Extremophiles **7:**347–351.
25. **Bettarel Y, Bouvier T, Bouvier C, Carré C, Desnues A, Domaizon I, Jacquet S, Robin A, Sime-Ngando T.** 2011. Ecological traits of planktonic viruses and prokaryotes along a full-salinity gradient. FEMS Microbiol. Ecol. **76:**360–372.
26. **Heidelberg KB, Nelson WC, Holm JB, Eisenkolb N, Andrade K, Emerson JB.** 2013. Characterization of eukaryotic microbial diversity in hypersaline Lake Tyrrell, Australia. Front. Microbiol. **4:**115. doi:10.3389/fmicb.2013.00115.
27. **Narasingarao P, Podell S, Ugalde JA, Brochier-Armanet C, Emerson JB, Brocks JJ, Heidelberg KB, Banfield JF, Allen EE.** 2012. De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. ISME J. **6:**81–93.
28. **Podell S, Ugalde JA, Narasingarao P, Banfield JF, Heidelberg KB, Allen EE.** 2013. Assembly-driven community genomics of a hypersaline microbial ecosystem. PLoS One **8:**e61692. doi:10.1371/journal.pone.0061692.
29. **Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J-B, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM.** 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature **437:**376–380.
30. **Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I.** 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. **19:**1117–1123.
31. **Zerbino DR, Birney E.** 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. **18:**821–829.
32. **Meyer F, Paarmann D, D'Souza M, Olson R, Glass E, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards R.** 2008. The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics **9:**386. doi:10.1186/1471-2105-9-386.
33. **Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H.** 2011. vegan: Community Ecology Package. R package version 2.0-2. http://CRAN.R-project.org/package=vegan.
34. **Edgar RC.** 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics **19:**2460–2461.
35. **Hyatt D, Chen G-L, LoCascio P, Land M, Larimer F, Hauser L.** 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics **11:**119. doi:10.1186/1471-2105-11-119.
36. **Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R.** 2005. InterProScan: protein domains identifier. Nucleic Acids Res. **33:**W116–W120.
37. **Feng DF, Johnson MS, Doolittle WF.** 1985. Aligning amino acid sequences: comparison of commonly used methods. J. Mol. Evol. **21:**112–125.
38. **Mohd Jaafar F, Attoui H, Mertens PPC, de Micco P, de Lamballerie X.** 2005. Structural organization of an encephalitic human isolate of Banna virus (genus Seadornavirus, family Reoviridae). J. Gen. Virol. **86:**1147–1157.
39. **Gorlas A, Koonin EV, Bienvenu N, Prieur D, Geslin C.** 2012. TPV1, the first virus isolated from the hyperthermophilic genus Thermococcus. Environ. Microbiol. **14:**503–516.
40. **Clarke KR.** 1993. Non-parametric multivariate analyses of changes in community structure. Aust. J. Ecol. **18:**117–143.
41. **Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharon V, Howe EA, Li J, Thiagarajan M, White JA, Quackenbush J.** 2006. TM4 microarray software suite. Methods Enzymol. **411:**134–193.
42. **Shannon CE, Weaver W.** 1964. The mathematical theory of communication. University of Illinois Press, Urbana, IL.
43. **Simpson EH.** 1949. Measurement of diversity. Nature **163:**688.
44. **Pielou EC.** 1966. The measurement of diversity in different types of biological collections. J. Theor. Biol. **13:**131–144.
45. **Garcia-Heredia I, Martin-Cuadrado A-B, Mojica FJM, Santos F, Mira A, Anton J, Rodriguez-Valera F.** 2012. Reconstructing viral genomes from the environment using fosmid clones: the case of haloviruses. PLoS One **7:**e33802. doi:10.1371/journal.pone.0033802.
46. **Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W, Jaroszewski L, Cieplak P, Miller CS, Li H, Mashiyama ST, Joachimiak MP, van Belle C, Chandonia J-M, Soergel DA, Zhai Y, Natarajan K, Lee S, Raphael BJ, Bafna V, Friedman R, Brenner SE, Godzik A, Eisenberg D, Dixon JE, Taylor SS, Strausberg RL, Frazier M, Venter JC.** 2007. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol. **5:**e16. doi:10.1371/journal.pbio.0050016.
47. **Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers Y-H, Falcon LI, Souza V, Bonilla-Rosso G, Eguiarte LE, Karl DM, Sathyendranath S, Platt T, Bermingham E, Gallardo V, Tamayo-Castillo G, Ferrari MR, Strausberg RL, Nealson K, Friedman R, Frazier M, Venter JC.** 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol. **5:**e77. doi:10.1371/journal.pbio.0050077.
48. **Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF.** 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature **428:**37–43.
49. **Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC, Wilkins MJ, Hettich RL, Lipton MS, Williams KH, Long PE, Banfield JF.** 2012. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. Science **337:**1661–1665.
50. **Charuvaka A, Rangwala H.** 2011. Evaluation of short read metagenomic assembly. BMC Genomics **12**(Suppl 2)**:**S8. doi:10.1186/1471-2164-12-S2-S8.

51. **Nagarajan N, Pop M.** 2013. Sequence assembly demystified. Nat. Rev. Genet. **14:**157–167.
52. **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:**2526–2537.
53. **Baker BJ, Tyson GW, Webb RI, Flanagan J, Hugenholtz P, Allen EE, Banfield JF.** 2006. Lineages of acidophilic archaea revealed by community genomic analysis. Science **314:**1933–1935.
54. **Allen EE, Tyson GW, Whitaker RJ, Detter JC, Richardson PM, Banfield JF.** 2007. Genome dynamics in a natural archaeal population. Proc. Natl. Acad. Sci. U. S. A. **104:**1883–1888.
55. **Andersson AF, Banfield JF.** 2008. Virus population dynamics and acquired virus resistance in natural microbial communities. Science **320:** 1047–1050.
56. **Morowitz MJ, Denef VJ, Costello EK, Thomas BC, Poroyko V, Relman DA, Banfield JF.** 2011. Strain-resolved community genomic analysis of gut microbial colonization in a premature infant. Proc. Natl. Acad. Sci. U. S. A. **108:**1128–1133.
57. **Needham DM, Chow C-ET, Cram JA, Sachdeva R, Parada A, Fuhrman JA.** 2013. Short-term observations of marine bacterial and viral communities: patterns, connections and resilience. ISME J. **7:**1274–1285.
58. **Caporaso JG, Paszkiewicz K, Field D, Knight R, Gilbert JA.** 2012. The Western English Channel contains a persistent microbial seed bank. ISME J. **6:**1089–1093.
59. **Gibbons SM, Caporaso JG, Pirrung M, Field D, Knight R, Gilbert JA.** 2013. Evidence for a persistent microbial seed bank throughout the global ocean. Proc. Natl. Acad. Sci. U. S. A. **110:**4651–4655.
60. **Jones SE, Cadkin TA, Newton RJ, McMahon KD.** 2012. Spatial and temporal scales of aquatic bacterial beta diversity. Front. Microbiol. **3:**318. doi:10.3389/fmicb.2012.00318.
61. **Rodriguez-Valera F, Martin-Cuadrado A-B, Rodriguez-Brito B, Pasic L, Thingstad TF, Rohwer F, Mira A.** 2009. Explaining microbial population genomics through phage predation. Nat. Rev. Microbiol. **7:**828–836.
62. **Banfield JF, Young M.** 2009. Variety, the splice of life, in microbial communities. Science **326:**1198–1199.
63. **Shaw AK, Halpern AL, Beeson K, Tran B, Venter JC, Martiny JBH.** 2008. It's all relative: ranking the diversity of aquatic bacterial communities. Environ. Microbiol. **10:**2200–2210.
64. **Minot S, Grunberg S, Wu GD, Lewis JD, Bushman FD.** 2012. Hypervariable loci in the human gut virome. Proc. Natl. Acad. Sci. U. S. A. **109:**3962–3966.
65. **Emerson JB, Andrade K, Thomas BC, Norman A, Allen EE, Heidelberg KB, Banfield JF.** 2013. Virus-host and CRISPR dynamics in Archaea-dominated hypersaline Lake Tyrrell, Victoria, Australia. Archaea **2013:** 370871. doi:10.1155/2013/370871.
66. **Bolhuis A.** 2002. Protein transport in the halophilic archaeon Halobacterium sp. NRC-1: a major role for the twin-arginine translocation pathway? Microbiology **148:**3335–3346.
67. **Santos F, Yarza P, Parro V, Briones C, Antón J.** 2010. The metavirome of a hypersaline environment. Environ. Microbiol. **12:**2965–2976.
68. **Fuhrman JA.** 1999. Marine viruses and their biogeochemical and ecological effects. Nature **399:**541–548.
69. **Lozupone CA, Knight R.** 2007. Global patterns in bacterial diversity. Proc. Natl. Acad. Sci. U. S. A. **104:**11436–11440.
70. **Hewson I, Barbosa JG, Brown JM, Donelan RP, Eaglesham JB, Eggleston EM, LaBarre BA.** 2012. Temporal dynamics and decay of putatively allochthonous and autochthonous viral genotypes in contrasting freshwater lakes. Appl. Environ. Microbiol. **78:**6583–6591.
71. **Corinaldesi C, Dell'Anno A, Magagnini M, Danovaro R.** 2010. Viral decay and viral production rates in continental-shelf and deep-sea sediments of the Mediterranean Sea. FEMS Microbiol. Ecol. **72:**208–218.
72. **Wommack KE, Hill RT, Muller TA, Colwell RR.** 1996. Effects of sunlight on bacteriophage viability and structure. Appl. Environ. Microbiol. **62:** 1336–1341.