Viruses as Winners in the Game of Life

Ana Georgina Cobián Güemes,[1] Merry Youle,[2] Vito Adrian Cantú,[3] Ben Felts,[4] James Nulton,[4] and Forest Rohwer[1]

[1]Department of Biology, San Diego State University, San Diego, California 92182; email: frohwer@gmail.com

[2]Rainbow Rock, Captain Cook, Hawaii 96704

[3]Computational Sciences Research Center, San Diego State University, San Diego, California 92182

[4]Department of Mathematics and Statistics, San Diego State University, San Diego, California 92182

Annu. Rev. Virol. 2016. 3:197–214

The *Annual Review of Virology* is online at virology.annualreviews.org

This article's doi: 10.1146/annurev-virology-100114-054952

Keywords

phage diversity, phage abundance, phage ecology, metagenomics, biogeography, viruses

Abstract

Viruses are the most abundant and the most diverse life form. In this meta-analysis we estimate that there are 4.80×10^{31} phages on Earth. Further, 97% of viruses are in soil and sediment—two underinvestigated biomes that combined account for only ~2.5% of publicly available viral metagenomes. The majority of the most abundant viral sequences from all biomes are novel. Our analysis drawing on all publicly available viral metagenomes observed a mere 257,698 viral genotypes on Earth—an unrealistically low number—which attests to the current paucity of viral metagenomic data. Further advances in viral ecology and diversity call for a shift of attention to previously ignored major biomes and careful application of verified methods for viral metagenomic analysis.

J.B.S. Haldane observed, "God has an inordinate fondness for stars and beetles."
To this we would add, "... and viruses."

Phage: a virus that infects a prokaryote (*Bacteria* and *Archaea*)

VLP: virus-like particle

Virion: the intercellular transport form of a virus, typically comprising the chromosome(s) enclosed within a protein capsid

Microbe: a prokaryote, i.e., an archaeon or a bacterium

HOW MANY VIRUSES?

Until the 1970s, viruses were of import only insofar as they were found to cause disease in us, our domesticates, and other eukaryotes of economic value. Bacteriophages—viruses that infect bacteria—were proving themselves to be eminently useful as model systems for molecular biology research, but were still thought to be of little significance to the functioning of the biosphere. There simply could not be enough environmental phages to matter, given the then measurable number of potential prokaryotic hosts counted by culturing techniques. Because typically only ~1% of the *Bacteria* were culturable by methods available at that time, bacterial populations were routinely underestimated by two orders of magnitude: the "great plate count anomaly" (1, p. 324). In the late 1970s, the reported environmental bacterial populations jumped 100-fold or more with the development of improved direct counting methods employing epifluorescence microscopy (2). In 1979, Torrella & Morita (3) concentrated particles larger than 0.2 μm from Oregon bay water by filtration and made direct counts of virus-like particles (VLPs) by transmission electron microscopy. Their counts, only about 10^4 mL^{-1} or one VLP per microbial cell, overlooked VLPs <0.2 μm, the majority of phages. This omission was corrected in 1989 when Bergh et al. (4) centrifuged water samples directly onto grids for viewing by transmission electron microscopy. They reported direct virion counts up to 1.5×10^7 mL^{-1}, an order of magnitude greater than the number of their hosts. Moreover, based on several reasonable assumptions, they predicted that in marine environments, as much as one-third of the bacterial population suffered phage attack daily—hardly insignificant.

This was only the beginning, as abundant microbes, and even more numerous phages, were then found in many environments, including inhospitable locations such as glacial ice (5), bubbling acidic hot springs (6), deep-sea vents (7), and deep sediments (8). Based on subsequent methodological developments, viruses are now recognized for what they are: the winners in the game of life. They are the most abundant and the most diverse life form, and are of great import for ecology and evolution.

A Global Census

How many viruses are there on Earth, and where are they located? One would expect to find the most viruses in the habitats with the most host organisms. The most abundant cellular organisms are the prokaryotes, with an estimated 4.15×10^{30} cells, the majority of which are found in soil, subseafloor sediments, and marine waters (**Table 1**) (9–12). The number of microbial eukaryotes, including algae, is comparatively small, in the range of only 10^3 to 10^4 mL^{-1} in seawater and other planktonic environments (13).

Taking a global viral census poses methodological challenges. Current methods for direct counting of VLPs commonly combine nucleic acid stains with epifluorescence microscopy or flow cytometry. These methods were developed for aquatic environments, and their application to soil and sediment is problematic. Additional steps are required to detach prokaryotes and VLPs from particles and surfaces, and successful methods are specific to particular situations. Both may be obscured by opaque particles, and background fluorescence can interfere. The marine environment remains the most extensively sampled and best-studied biome. Other intensively studied locations, such as the human gut and freshwater, make a relatively small contribution to

Table 1 Estimated number of virus-like particles on Earth

Biome	Number of prokaryotic cells	Median virus-to-microbe ratio	Virus-like particles per biome	Percentage of total virus-like particles
Marine	1.01×10^{29}	12.76	1.29×10^{30}	2.6828
Freshwater	1.26×10^{26}	14	1.76×10^{27}	0.0037
Other aquatic	2.44×10^{27}	30	7.32×10^{28}	0.1524
Sediment	3.80×10^{30}	11	4.18×10^{31}	87.0131
Soil	2.50×10^{29}	19.5	4.88×10^{30}	10.1481
Human-associated	2.80×10^{23}	0.1	2.80×10^{22}	0.0000
Other host–associated	Unknown	25	Unknown	Unknown
Total	**4.15×10^{30}**	**12**	**4.80×10^{31}**	

Virus-to-microbe ratios (**Supplemental Table 2**) were extracted from 53 previous studies (**Supplemental Table 1**) and used to calculate virus-like particles in each biome. Numbers of prokaryotic cells in marine, freshwater, other aquatic, sediment, and soil biomes are from Reference 9. Number of cells per human is from Reference 20; human population is from Reference 89. Median virus-to-microbe ratio for human biome is from Reference 19.

the global total. Soils and the extensive subsurface sediments warrant more attention, as their contribution is expected to be major.

Supplemental Material

Marine sediment contains up to 10^5 times more organic matter than the water column above, supporting bacterial densities about 10^3 times greater (11). Summed globally, the total number of prokaryotes in subseafloor sediment (2.9×10^{29}) is roughly equal to the estimates of Whitman et al. (9) for the total number of prokaryotes in seawater (1.2×10^{29}) and in soil (2.6×10^{29}). Cell densities alone would predict good hunting for the phage, as the probability of nonspecific phage-host contact would be $\sim 10^3$ times greater on average than in the water column above. Indeed, VLP counts of 10^9 mL^{-1} in sediment have been reported, three orders of magnitude greater than that observed in the overlying water column at that depth (14). Similarly, the large numbers of prokaryotes in soil, the rhizosphere, and the rhizosheath, and their associated viruses, remain largely terra incognita despite their accessibility and their importance to agriculture.

VLPs outnumber their hosts by a ratio of approximately ten to one in various oceanic locations (15), and similar ratios have been observed in some other biomes (**Figure 1**). It is not understood what drives this consistency, nor the observed variations. On this basis, earlier extrapolations from the prokaryote abundances in the major biomes yielded an estimated 1.2×10^{30} in open ocean, 2.6×10^{30} in soil, 3.5×10^{31} in oceanic subsurface, and $0.25–2.5 \times 10^{31}$ in terrestrial subsurfaces, consistent with the rule-of-thumb estimate of 10^{31} viruses on Earth (16). Here we revisited this question by combining the prokaryote populations drawn primarily from the classic 1998 paper by Whitman et al. (9) with the median measured virus-to-microbe ratios (VMRs) from 53 previously reported studies (**Supplemental Table 1**). In addition to the major biomes, we included some specialized environments of particular interest, such as niches associated with humans and other hosts. From a global perspective, the majority of prokaryotes inhabit soil, seawater, and sediment. Thus, we expected that these biomes would account for the majority of viruses.

This approach yielded 1.29×10^{30} VLPs in marine waters, 4.18×10^{31} in sediment, and 4.88×10^{30} in soil (**Table 1**). Adding in other lesser contributors brought the global total to 4.80×10^{31} VLPs (details in **Supplemental Table 2**). This confirms and slightly augments the oft-quoted number of 10^{31}. Given that the vast majority of cellular entities are prokaryotes, this number represents the global phage population. Significantly, sediment and soil combined accounted for 97% of the global total. In sum, phages are the most abundant life form, exceeding the runner-up,

VMR: virus-to-microbe ratio

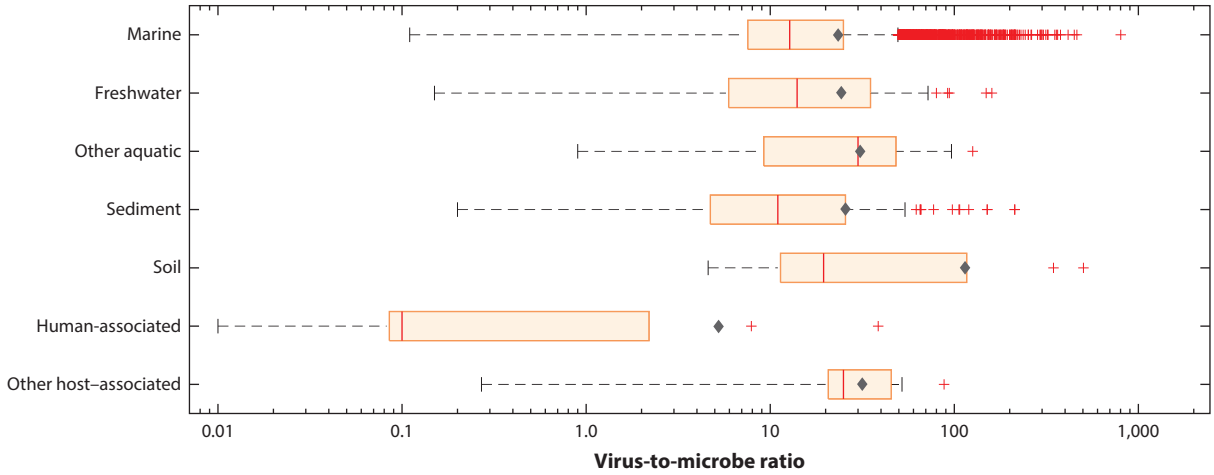

Figure 1

Virus-to-microbe ratios for major biomes. Box plots of biome virus-to-microbe ratios drawn from 53 previous studies (see **Supplemental Methods**). The means are indicated with gray diamonds. The medians are indicated with red lines.

Supplemental Material

the prokaryotes, by more than an order of magnitude. Populations of this magnitude dictate that phage evolution is driven by selection, with the contribution of drift being insignificant (17).

Some caveats are worth mentioning. To the best of our knowledge, there are only two published reports of VMRs for human-associated communities. One reported VMRs of 38.6 and 7.9 for gum-associated mucus and the adjacent milieu, respectively (18), and the other reported a mean VMR of 0.129 and a median VMR of 0.1 in fecal matter (19). Because the vast majority of human-associated microbes and VLPs are in the distal gut (20, 21), the fecal VMR was used to calculate the total VLPs for the human-associated biome. Admittedly, that fecal VMR, being two orders of magnitude lower than the VMRs typical of most other biomes, was suspect. However, combining it with the recently revised estimate of 3.9×10^{13} human-associated prokaryotes (20) yields 3.9×10^{12} human-associated VLPs, in good agreement with the previous estimate of 3.0×10^{12} (21). Unraveling the phage-host dynamics in the human-associated microbiome will require further investigation.

Several factors may result in underestimations of phage abundance. (*a*) Equating the number of VLPs with the number of phages overlooks those residing as prophages in bacterial genomes—not an insignificant number (see below). (*b*) Some stains used to enumerate microbial cells and VLPs by epifluorescence microscopy, such as SYBR Green, are relatively insensitive to single-stranded DNA and RNA, and thus overlook many small viruses. A more representative assay can be made using SYBR Gold (22). (*c*) VLP counts for soil and sediment are biased by reduced extraction efficiencies, and visualization and identification of VLPs can be compromised by particulate matter in samples of fecal matter, soil, sediment, etc. (*d*) Some VLP isolation methods for epifluorescence microscopy include filtration through 0.2-μm filters, which miss larger virions as well as virions stuck to particulate matter. (*e*) The 0.02-μm grids used for transmission electron microscopy viewing miss the smallest virions.

Prophage: a phage chromosome that resides within a host cell (lysogen) without immediately engaging in lytic replication

Conversely, are these VLPs really viruses? In some environments, some VLPs may be gene transfer agents, i.e., packaged cellular genes masquerading as tailed phage particles (23). Although gene transfer agents are produced by roseobacters, a group that may account for more than 25% of the prokaryotes in some marine environments (23), it remains unknown whether gene transfer

agents make a significant contribution to marine VLP counts. Also unknown is what fraction of the VLPs in various environments are infectious, and what fraction are defective upon release or subsequently inactivated by UV irradiation, physical damage, or enzymatic attack.

Femtogram (fg): 10^{-15} g

Petagram (Pg): 10^{15} g

Phages Matter

How much do 10^{31} viruses matter? One measure of this is to consider the matter they contain. The mass of a typical phage virion containing 50 kbp of DNA packaged inside an icosahedral capsid is calculated to be 0.0823 femtograms (fg); of this, 0.054 fg is DNA and 0.0283 fg is the protein capsid (A. Luque, personal communication). For comparison, each *Prochlorococcus* cell, a particularly small autotrophic marine bacterium, has a mass of 300 fg. Based on this virion mass, the 4.80×10^{31} VLPs on Earth have a total mass of 3.95×10^{15} g, or 3.95 petagrams (Pg). The stoichiometry of carbon, nitrogen, and phosphorus (C:N:P) in this typical virion is 20:6:1 (24), which partitions that mass as approximately 0.06 fg C, 0.02 fg N, and 0.0075 fg P. Calculation based on these data suggests that Earth's VLPs represent roughly 2.9 Pg C, i.e., two orders of magnitude less than the 350–500 Pg total C previously estimated for Earth's prokaryotes (9). Compared to microbial cells, virions are enriched in both nitrogen and phosphorus, with estimated global totals of 0.96 Pg N and 0.36 Pg P. As a result of this enrichment, >5% of the total marine dissolved organic phosphorus and nitrogen pools is estimated to reside in virions in some locations (24).

It is now widely recognized that microbes are of tremendous importance for global biogeochemical processes, and consequently, that which controls the microbes—the phages—runs the world. Prokaryotes are subject to predation by both heterotrophic protists and phages. Grazing by protist predators is size selective, whereas lysis by phages is strain specific. Moreover, in the oceans, grazing moves the organic carbon and nutrients to higher trophic levels, whereas lysis routes these components instead through the viral shunt as dissolved organic matter. This dissolved organic matter feeds the heterotrophic microbial community, thereby increasing net primary productivity and slowing the movement of carbon to the deep ocean (15, 25–29). Because of the stoichiometric mismatch between virions and their host cells, after lysis, a disproportionate amount of the phosphorus is found in the progeny virions, leaving the cellular debris that feeds the heterotrophs depleted in phosphorus (24). An estimated 10^{28} marine bacteria are lysed daily, including ~30% of the cyanobacteria and ~60% of the heterotrophic bacteria (28, 30). This selective, strain-specific lysis profoundly impacts prokaryote community diversity, increasing both richness and evenness (15, 25, 31–33).

Phages also add a horizontal dimension to microbial evolution by mediating the transfer of genes between cells (34). They nab useful metabolic genes from their hosts, maintain them, evolve them further to suit their own needs, and then sometimes return the new version to the microbial gene pool (35–38). On an ecosystem level, the phage community encodes environment-specific repertoires of microbial metabolic genes (39).

Through lysogeny, phage genes maintain a continual presence in microbial communities. Identification of prophages in microbial genomes is challenging, and numerous bioinformatics methods have been developed (40–44). Estimates of the percentage of prokaryotes in various biomes that carry one or more prophages have ranged between 0% and 100% (45). Approximately 82% of the sequenced prokaryotic genomes available as of 2015 are predicted to contain at least one prophage (K. McNair, personal communication). Resident prophages often account for the differences between strains within a bacterial species (46). Prophage-encoded exotoxin genes cause many notable human diseases, including cholera, diphtheria, and enterohemorrhagic diarrhea (47) as well as diseases that plague our agriculture and aquaculture.

Siphophage: a member of the family *Siphoviridae*

OTU: operational taxonomic unit

We do not yet understand the factors influencing the prevalence of lysogeny in any biome. Siphophages have long been associated with the temperate lifestyle, although lysogeny is not confined to that family. This group was observed to dominate the community in the Southern Ocean, marine sediment, desert, hypersaline ponds, and human fecal samples, accounting for 44% of the total in sediment (48–53). However, whether a temperate phage will follow the lytic or lysogenic pathway is decided at the start of each infection in response to host and environmental factors. The common interpretation based primarily on studies of coliphage λ posits that host abundance, as sensed by multiplicity of infection, determines which pathway is followed; low multiplicity of infection favors lytic replication, whereas high multiplicity of infection favors lysogeny (54). A recent analysis of phage communities on coral reefs suggests more complex dynamics (55).

HOW MANY DIFFERENT VIRUSES?

Ever since the discovery of phages, it has been evident that there are different phages capable of killing different bacteria. However, 100 years later we still do not know the extent of this diversity, its biogeography, or its dynamic role in ecosystem function. Whereas the diversity of all cellular life has been probed using universal genes such as the small subunit ribosomal RNA gene, the polyphyletic viruses have no gene in common, thus precluding a comprehensive PCR-based survey of viral diversity (56, 57). Therefore, other methods had to be devised.

Two approaches based on data available in 2003 both yielded an estimated 100 million phage species (58). (Phage species is an operational term used to denote different phages as defined by particular criteria.) One approach conservatively assumed that 10 different phage species infect each of the estimated 10 million microbial species—thus, it is estimated that there are 100 million different phages. This alone does not measure genetic diversity, because two phages denoted as different species by this criterion might differ in only one or two key proteins that determine host range. Similarly, the swapping of structural gene modules can yield a new species that differs in virion morphology without increasing global genetic diversity, whereas two phages indistinguishable by morphology and host range can differ significantly in other genome modules. The second approach compared all the sequenced phage open reading frames (ORFs) in GenBank at that time using BLAST and clustered the ORFs using an e-value of 10^{-4}. From this data the nonparametric estimator Chao1 (59) predicted that 2×10^9 phage ORFs remained to be discovered. Assuming 50 ORFs per phage genome with 50% of those being novel, calculations based on the Chao1 value predicted 100 million different phages.

A decade later, armed with more metagenomic data and new analysis tools, Sullivan and colleagues (60) presented an analysis based on protein clustering that reduced the estimated total number of phage ORFs from 2×10^9 to only 3.9×10^6, a demotion of almost three orders of magnitude. The debate continues.

Signature Genes

Although there is no universal phage gene, diversity within phage groups can be assessed using signature genes shared by all group members (61). For instance, the capsid portal gene (*g20*) is conserved among many of the large cyanomyophages that inhabit marine and freshwater environments. Several studies using *g20* reported rich community diversity, typically 100 or more operational taxonomic units (OTUs) and including clades for which there are no cultured isolates (62, 63). Attempts to correlate variations in community composition with depth, host abundance, season, and geographic distance yielded inconsistent results. Some OTUs demonstrated consistent seasonal variation, whereas others persisted in moderate abundance from year to year (64).

Sampling across a north-south Atlantic Ocean transect found similar cyanomyophages to be widely distributed with no apparent geographical segregation (63), and others were present in both marine and freshwater environments (65). However, some diversity within this group eluded these surveys; of 39 cyanophage isolates from the Gulf of Mexico, only 63% carried detectable *g20* sequences (66).

Cyanophage: a phage that infects cyanobacteria

Myophage: a member of the family *Myoviridae*

Podophage: a member of the family *Podoviridae*

The T4 superfamily (*Myophage*) was similarly surveyed using their major capsid protein (*gp23*). In aquatic environments, these are primarily T4-like cyanomyophages. The results here echoed the same trends: diversity (more than 100 OTUs) exceeding that represented in cultured isolates (67), seasonal succession patterns, and long-term persistence of some OTUs (64, 68). In some cases persistent OTUs were also the most abundant (>4% relative abundance), contradicting predictions of both the bank model of viral community structure and the classic kill-the-winner dynamic (68–70).

These analyses based on *g20* and *gp23* are limited to the myophages (predominantly cyanomyophages) and miss the podophages and the abundant siphophages. A more inclusive assessment of cyanophage diversity, including myophages and podophages, used *psbA*, the gene that encodes the D1 protein of oxygenic photosystem II (71). Phage-encoded sequences clustered by both phage family and by host, separate from the host *psbA* genes, and included clusters with no cultured isolates. Multiple clusters coexisted in some locales, whereas some clusters extended over vast geographic distances. Moreover, one cyanopodovirus subcluster was found to be globally distributed based on its DNA polymerase gene (*pol*) (72).

Other signature genes can provide a more complete picture of marine phage diversity. For example, the phosphate-starvation gene *phoH* is present in nearly 40% of all marine phages compared with 4% of nonmarine phages, reflecting the scarcity of phosphate in the marine environment (38). It is not restricted to a single viral family and is found in phages that infect heterotrophs as well as autotrophs, even in viruses of photosynthetic green algae. A *phoH*-based marine survey found, yet again, that most of the environmental diversity was not represented in cultured isolates; that *phoH* homologs are widely distributed, with most clusters represented in multiple oceanic regions; and that marine phage community composition varies with depth and geographical location. A subsequent survey of the *phoH* genes in Sargasso Sea phage communities at depths of 0 to 1,000 m over a 2-year period identified 3,619 OTUs (97% identical) and provided new insights into community dynamics (73). Approximately 96% of those OTUs were rare, each accounting for <0.01% of the total sequences, and more than 50% of the sequences were from five abundant OTUs. The presence of a few abundant OTUs (one to four in any particular sample) and many rare ones is consistent with the bank model of phage community structure. However, whereas that model predicts the cycling of phages between the two groups over time, here the rare OTUs remained rare, and the most abundant OTUs persisted through seasons and years.

To date, signature genes have been developed for only some viral families. They are strikingly lacking for the siphophages, the family that includes many temperate phages and that dominates both metagenomic data sets and cultured isolates. In even the best cases, signature genes fail to capture the full richness present in natural communities. However, they have provided insights into the global distribution of specific phage genes. The DNA polymerase conserved in the T7-like podophages and restricted to that group was found in multiple biomes (74). Moreover, identical or nearly identical 533-bp segments were recovered from different biomes, indicating that phages, or at least phage genes, have moved between biomes in recent evolutionary time. Similarly, sequences from algal virus DNA polymerase genes that were >98% identical were found from the northern Pacific Ocean to Antarctica (75). That phages from freshwater, sediment, and soil are able to infect marine prokaryotic communities suggests that phages can move successfully between biomes (76).

These and similar observations suggest that the global viral diversity may be less than previously estimated based on the diversity in individual biomes.

Virome: a viral metagenome

Phage Metagenomics

Expansion of the field of view from consideration of signature genes to assessment of phage community diversity was made possible by the development of viral metagenomics. Earlier sequencing methods that required cloning of phage DNA had often encountered problems. Many phages carry genes that are lethal to the cloning host cells, or their DNA contains modified bases that block cloning. An alternative method, linker-amplified shotgun sequencing, was first used to assess near-shore marine communities (49). Sample preparation included passage through a 0.16-μm tangential flow filter, purification of VLPs by CsCl gradient centrifugation, and subsequent DNA extraction. This procedure did not recover large viruses (e.g., algal phycodnaviruses) or RNA viruses. This initial marine survey found that more than 65% of the viral sequences were novel. Four years later, viromes prepared using next-generation sequencing from four oceanic regions contained >90% unknowns (77). Even 15 or more years later, despite the increased number of sequenced viral genomes in the public databases, 60–99% of the sequences in viromes from diverse biomes are still unknowns (16, 52, 53, 78, 79). More than 99% of viral genetic diversity remains to be explored (16).

Even though most sequences recovered are unknowns, bioinformatics methods can provide insights into community structure. Metagenomic reads are assembled into contigs in silico. The more diverse the community, the lower the probability of sequencing two overlapping fragments from the same genome—thus, shorter contigs and more unassembled singleton reads. Plots of contig spectra (number of contigs versus contig length) were best represented by power law–based mathematical models and provided estimates of both the richness and evenness of the sampled community (49). Application of this approach to near-shore marine communities estimated 374 to 7,114 genotypes present. Of these, the most abundant genotype represented only 2–3% of the total community, and only three genotypes contributed more than 1% of the reads. Subsequent metagenomic surveys of diverse environments reported genotypes numbering in the hundreds to tens of thousands (reviewed in 80; data in 6, 50, 77, 80, 81).

Virome Meta-Analysis

Supplemental Material

We have developed a new bioinformatics method, FRAP (fragment recruitment, assembly, purification) (**Figure 2**; **Supplemental Methods**) and used it to assess global viral diversity by analyzing 1,623 publicly available viromes (**Supplemental Table 3**). To create a reference library of the observed viral genotypes on Earth, all reads from each of the viromes were assembled separately using SPAdes (82). Comparative evaluations of assembler performance on metagenomic data sets had ranked SPAdes among the best based on contig accuracy (83). We assumed that each ≥1-kbp contig represented a partial viral genotype that was relatively abundant in that biome. On rare occasions, more than one nonoverlapping contig might have been recovered from the same genotype. All of these ≥1-kbp contigs were merged into a sequence pool. In addition, the 2,669 sequenced phage genomes (as of December 2015) and 67 archaeal virus genomes (as of February 2016) in the NCBI Viral Genomes database were added to the pool, along with an additional 123 bacteriophage genomes from the Broad Institute Marine Phage Sequencing Project. Subsequent dereplication of the pool by CD-HIT (84) at 98% identity yielded 2,267,978 unique viral contigs plus genomes that constitute the reference library.

Given this reference library, the fragment recruitment step could then retrieve the matching reads from any virome. In principle this FRAP method could be used to retrieve and thus purify

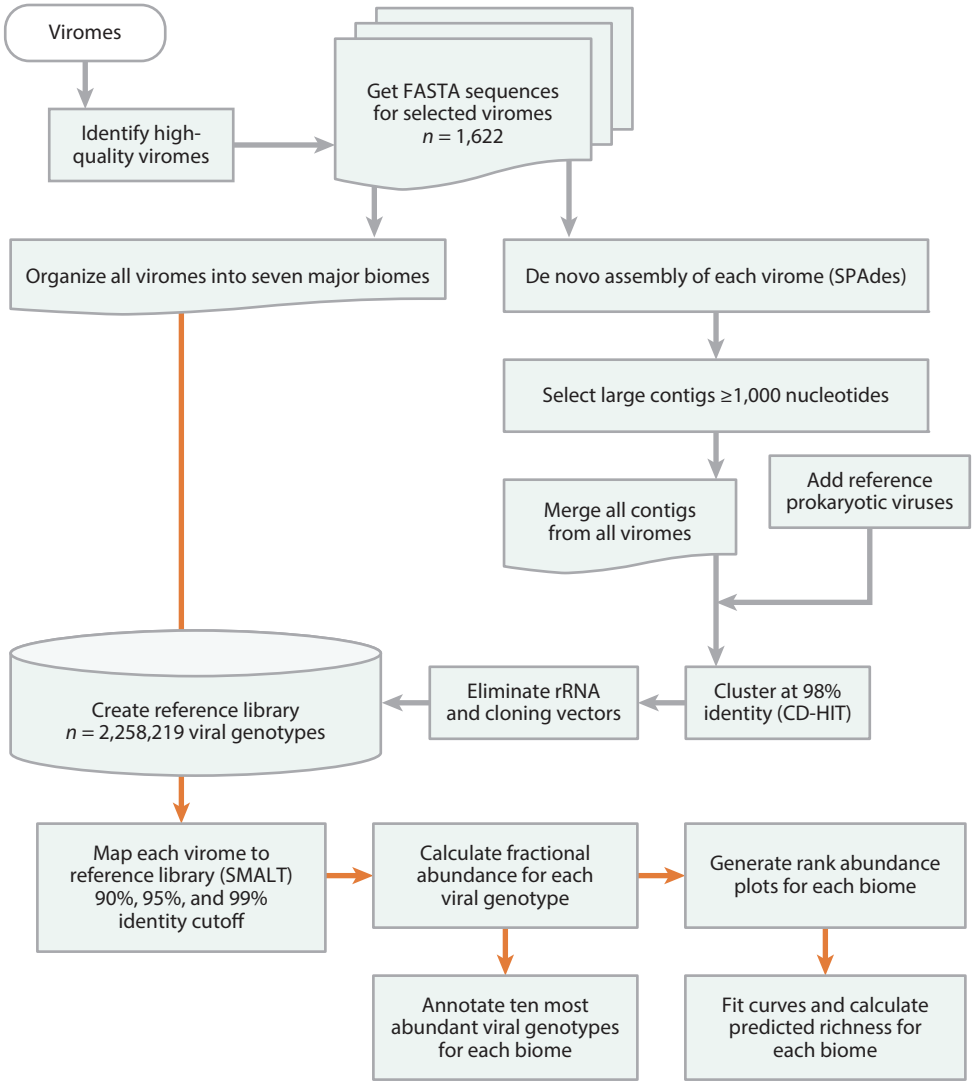

Figure 2

Bioinformatics pipeline for the FRAP (fragment recruitment, assembly, purification) method (see **Supplemental Methods**).

sequences that are only minor components of any data set. For our analysis, all reads from all viromes were mapped to the reference library at 90%, 95%, and 99% identity, and the normalized number of hits to each contig was tallied for each biome (**Table 2**; **Supplemental Methods**). This yielded the fractional abundance in each biome of every contig that was also present in the reference library (**Figure 3**). The 10 most abundant viral contigs in each case occupied the same rank for all three mapping identities, evidencing the robustness of the method.

Supplemental Material

Mapping at 99% identity provided the most conservative estimate of the number of identified viral contigs and was used to calculate the number of viral genotypes potentially encoded within each biome (**Figure 3**). Here we summed the lengths of all observed viral contigs and divided

Table 2 Virome metrics

Biome	Number of viromes	Number of reads	Percentage of reads assembled into ≥1-kbp contigs	Percentage of reads mapped to reference library		
				90% identity	95% identity	99% identity
Marine	192	56,676,517[1]	11.83	38.1	29.4	20.4
Freshwater	48	11,519,523	19.51	15.7	11.4	8.1
Other aquatic	19	3,342,537	14.04	40.2	35.8	27.3
Sediment	21	15,729,082	10.00	54.7	51.7	44.2
Soil	9	2,459,152	15.54	41.9	36.6	29.0
Human-associated	1,158	481,172,486	25.16	30.3	27.2	12.4
Other host–associated	167	34,600,192	3.81	34.6	31.1	25.3
Other	7	396,889	28.22	44.3	36.8	17.4
Total	**1,621**	**605,896,378**	**16.01**	**31.8**	**28.1**	**14.7**

[1]For the Tara Oceans Virome data set, only 1% of the reads were used. Including all would increase the number of reads from the marine biome to 1,688,798,702.

by the assumed average phage genome size of 50 kbp. The result represents the observed coding capacity expressed as the number of viral genotypes. The most viral genotypes were observed in the marine and human-associated biomes, and the least in soil. This reflects the size of the data sets and their relative contributions to the reference library: 192 viromes for the marine biome, 1,158 viromes for the human-associated biome, and 9 viromes for the soil biome. We developed two methods to calculate the predicted number of viral genotypes (richness) of each biome based on the observed rank abundance plots. The curve fit method draws on the information of the most abundant ranks; the tail fit method utilizes the information of the less abundant ranks. A power law model provides the best fit for the rank abundance curves (**Supplemental Table 4**). Based on this model and the number of VLPs for each biome, we calculated the number of predicted viral genotypes for each biome (**Supplemental Methods**). This method, in which we fit the power law model directly to each rank abundance plot, is limited to capturing only the most abundant ranks because the data do not allow valid ranking for the rare types. Consequently, we also implemented a tail fit method that addresses the rare types directly (**Supplemental Methods**).

Supplemental Material

The congruence of the results from the two methods correlated with the amount of virome data available from each biome. For the marine biome, for which we had 192 viromes, both predictions differed by approximately 10%, whereas the data from only 9 soil viromes yielded richness estimates that differed by more than an order of magnitude. Moreover, accurate estimation of the total number of viral genotypes awaits collection of more data from the soil and sediment biomes, which together house 97% of the VLPs on Earth.

How many different phages on Earth? Providing a direct answer to this question remains challenging, in part because we do not know whether phages are provincial or cosmopolitan. If a biome is sampled in two geographic locations, A and B, and the number of phage genotypes present in each is estimated, is the richness of the phage community in that biome equal to A + B, or is it significantly less? Likewise, if the number of phage genotypes in each biome is known, are we justified in summing them to calculate the global virome? Other studies have addressed this by assessing the fraction of genotypes shared between biomes or geographical regions. A

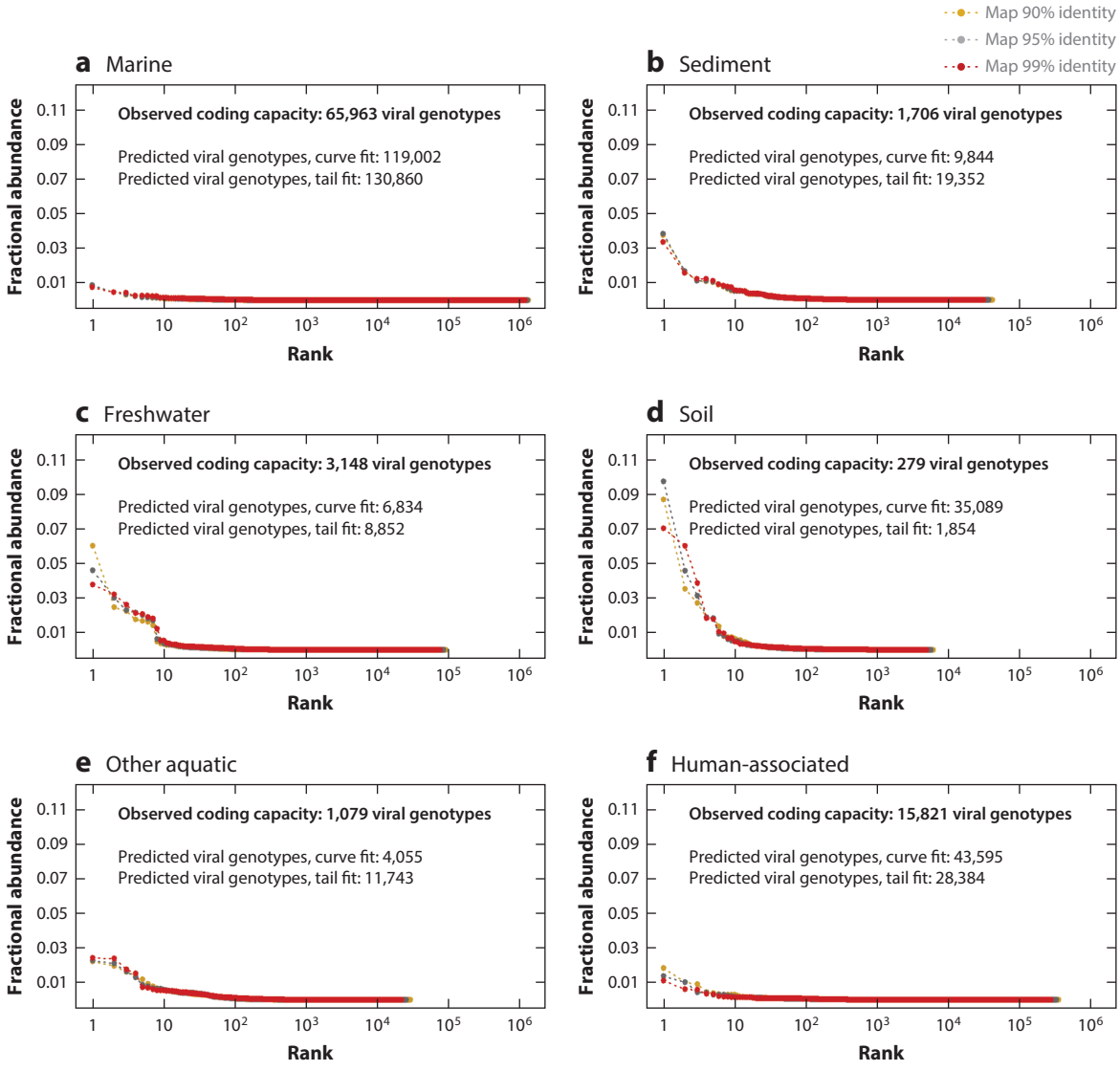

Figure 3

Viral rank abundance curves for (*a*) marine, (*b*) sediment, (*c*) freshwater, (*d*) soil, (*e*) other aquatic, and (*f*) human-associated biomes. Reads in these six major biomes were mapped to the genotypes present in the reference library at 90%, 95%, and 99% identity. In each case, the relative abundance of all recovered viral genotypes is shown as well as the observed coding capacity expressed as the potential number of different 50-kbp phage genomes and the predicted viral genotypes using both the curve fit and tail fit methods.

three-pronged analysis of viromes from four oceanic regions (the Pacific Ocean off the coast of British Columbia, the Arctic Ocean, the Gulf of Mexico, and the Sargasso Sea) found that a large fraction of the phage community is cosmopolitan; that is, they are found in two, three, or even four of the surveyed regions (77). Within this global distribution, the phage communities showed regionalization in that community members, including the cosmopolitan and the most

abundant, shifted in relative abundance from region to region. Assembly of reads from each region separately yielded a total of ~150,000 genotypes, whereas coassembly reduced the total to only 57,600 different genotypes. Similarly, a recent study of hypersaline ponds on three continents found that community composition varied with the level of salinity but that communities in environments with the same salinity are genetically connected across the globe (53). Conversely, phage communities present in three soil biomes shared essentially no genotypes (85).

The cosmopolitan range of individual phage genes or gene modules is another factor that complicates determination of the ecological or geographical range of phage genotypes. Studies described earlier that found nearly identical sequences of signature genes to be globally distributed indicated a global phage gene pool, at least within the marine biome. However, even within that biome, this does not distinguish between the global travels of phage genes by horizontal gene transfer and the presence of the same phage genotypes in geographically remote locations.

Supplemental Material

Here we used the total length of the unique DNA sequences observed in our meta-analysis to estimate the number of possible viral genotypes on Earth (**Supplemental Methods**). In brief, we started with all available viromes and all sequenced viral genomes, and then summed the lengths of all the dereplicated contigs (98% identity). This sum was our proxy for all viral DNA currently available in the databases. Division of this sum by the assumed 50-kbp average phage genome length (**Supplemental Table 4**) indicated that the observed viral DNA is sufficient to encode 257,698 different viruses.

This estimate is undoubtedly low. The sequenced samples so far represent only a minute fraction of the total viral-encoded information present in every biome. For example, including all 2.16×10^9 reads from the massive Tara Oceans Viromes data set (average read length of ~101 bp) would provide 2×10^8 kbp of marine viral genomic sequence. To put this in perspective, 1.29×10^{30} marine VLPs with an average genome of 50 kbp would contain 6.5×10^{31} kbp of DNA.

Which are the most abundant phage genes? To explore this, we annotated the ten most abundant contigs in each biome (**Supplemental Table 5**). The majority of them had no significant similarity to sequences in the NCBI database. Those that showed significant similarity from the soil biome included three replication proteins from a *Staphylococcus aureus* plasmid and one mobC and relaxase; those from sediment included an *S. aureus* plasmid replication protein, two uncultured virus clone sequences, and a major capsid protein from an uncultured member of *Gokushovirinae*; those from the marine biome included four podovirus polymerases, an *Escherichia coli* excinuclease, and four miscellaneous genes; those from the freshwater biome included two myophage major capsid genes and one nonstructural gene from a picorna-like virus; those from the human-associated biome included three from uncultured bacterial plasmids, a replication protein, and a mobilization protein.

Again, some caveats need to be addressed. Our results are unavoidably skewed due to the biased distribution of the current viromes. Of the 1,623 viromes, 71% were from human-associated communities, 10% from communities associated with other animals or plants, and 11% from marine environments (**Table 2**). Thus 92% of the viromes explored only 3% of the VLPs on Earth, and only 1.8% investigated the two biomes with 97% of the VLPs—soil and sediment (**Table 1**). Some biomes remain virtually unsampled. Even in the marine biome, only a very limited geographic territory and a restricted range of environmental parameters have been sampled. Despite the surging interest in the human microbiome, published counts for human-associated prokaryotic cells and VLPs are strikingly lacking.

The publicly available viromes are further compromised by poor methods. Of those viromes, 132 (7.5%) were omitted from our library because they were mislabeled microbial metagenomes or showed obvious contamination with human or microbial sequences. Some contained abundant φX174 sequences due to either the amplification bias of multiple displacement amplification (86, 87) or, when sequencing was performed on the Illumina platform, the failure to remove the Illumina PhiX quality control sequences prior to submission to the public databases. In the sediment and soil biomes, the highly abundant fragments from *S. aureus* plasmids could be from phages or gene transfer agents, or from bacterial contaminants of viromes that result from inadequate viral purification during sample preparation. Many sources of error can be avoided and more quantitative data obtained by carefully adhering to current best practices (86–88). Still needed is the comparable development of methods for RNA viruses and single-stranded DNA phages, as well as VLP purification procedures that do not exclude the largest viruses.

The future for FRAP. Given a high-quality reference library, FRAP can be used to get rid of the crap—i.e., to fish out matching sequences from a metagenomic data set, even when they constitute only a small percentage of the reads. This can potentially eliminate some sample purification steps or enable selective retrieval of different components from a mixed sample. However, FRAP's utility depends on the completeness and quality of the reference library. Library development, in turn, calls for clean sample preparation methods, sequencing of more base pairs, and longer read lengths (such as the 10- to 15-kbp lengths now possible with PacBio SMRT sequencing).

Recent advances in phage ecology have heightened our awareness that we live in a phage world. Much of that world remains a terra incognita. For the careful researcher equipped with today's technology, the opportunities for discovery are vast.

SUMMARY POINTS

1. Phages are the winners: the most numerous and genetically diverse life form on Earth. The estimated 4.80×10^{31} VLPs on Earth comprise at least 257,698 different viral genotypes.
2. We have barely begun to explore viral diversity. Viral metagenomic studies have focused on a few biomes and have ignored soil and sediment—the two biomes that combined contain 97% of the global viral population. Sampling has been sparse at best, and numerous environments remain terra incognita.
3. Global phage diversity far exceeds that represented by cultured isolates.
4. One cannot fully understand the ecology or evolution of any ecosystem without including the phages.
5. Our current knowledge of the fractional abundances of viruses in each biome is limited, and we need better strategies to describe the population structure of the viruses in each biome.
6. Further work is needed before we even know how much we still do not know, i.e., how far we are from understanding the viral dark matter of the biosphere.

FUTURE ISSUES

1. Viral metagenomic sampling has been narrowly focused on selected regions of the marine environment and on host-associated communities, primarily human-associated communities. Most of the globe and most biomes await exploration. Initial surveys of the human-associated phage community uncovered anomalous dynamics that await resolution.
2. Viruses that have chromosomes of RNA or single-stranded DNA are significant components of some communities but have been generally ignored. Correcting this requires new inclusive methods for both their direct VLP counts and their metagenomics.
3. Some phages and some phage genes are cosmopolitan, whereas others appear to be geographically or ecologically restricted. The question remains: Do all phages share the same global gene pool, with varying levels of access?
4. Genomic analysis of both phages and their hosts indicates that lysogeny is commonplace. Awaiting further exploration are the prevalence of temperate phages in various environments, the lysis-lysogeny decision, and the impact of lysogeny on the ecology and evolution of both phages and their hosts.
5. Phages are the greatest reservoir of unexplored genetic diversity on Earth. Current estimates of the number of different proteins encoded by phages vary widely. After more than a decade of phage metagenomics, the majority of phage sequences remain novel. Sequencing technology has advanced rapidly, and now enhanced bioinformatics methods are essential to analyze the mushrooming metagenomic data.
6. Assignment of function to phage-encoded proteins based on sequence homology is limited by the rapid rate of phage evolution. Other approaches are called for in order to translate environmental metagenomic data into the metabolic potential of phage communities.
7. The species concept is not directly applicable to viruses. An alternative, generally accepted metric is needed to facilitate discussion of viral diversity, ecology, and evolution.
8. In the current era of the microbiome, researchers are actively investigating the roles of microbes in processes including human health, ecosystem functioning, and global biogeochemical cycles. Now, a century after the discovery of phage, exploration of the role of phage in these and other activities is overdue.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

All calculations were made on Rob Edwards's lab cluster at San Diego State University, which is supported by National Science Foundation grant DBI-0850356 for computational resources. We are thankful to the members of the Biomath group at San Diego State University for critical discussion of this work; to Cynthia Silveira for early access to marine VMRs; to Rizki Wulandari, Jan Janouškovec, Nate Robinett, and Ben Knowles for suggestions and comments; and to Evelien

Adriaenssens for data sharing. This work was partially supported by the National Council on Science and Technology (CONACyT), Mexico.

LITERATURE CITED

1. Staley JT, Konopka A. 1985. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu. Rev. Microbiol.* 39:321–46
2. Hobbie JE, Daley RJ, Jasper S. 1977. Use of Nuclepore filters for counting bacteria by fluorescence microscopy. *Appl. Environ. Microbiol.* 33:1225–28
3. Torrella F, Morita RY. 1979. Evidence by electron micrographs for a high incidence of bacteriophage particles in the waters of Yaquina Bay, Oregon: ecological and taxonomical implications. *Appl. Environ. Microbiol.* 37:774–78
4. Bergh Ø, Børsheim KY, Bratbak G, Heldal M. 1989. High abundance of viruses found in aquatic environments. *Nature* 340:467–68
5. Anesio AM, Mindl B, Laybourn-Parry J, Hodson AJ, Sattler B. 2007. Viral dynamics in cryoconite holes on a high Arctic glacier (Svalbard). *J. Geophys. Res.* 112:G04S31
6. Bolduc B, Wirth JF, Mazurie A, Young MJ. 2015. Viral assemblage composition in Yellowstone acidic hot springs assessed by network analysis. *ISME J.* 9:2162–77
7. Ortmann AC, Suttle CA. 2005. High abundances of viruses in a deep-sea hydrothermal vent system indicates viral mediated microbial mortality. *Deep-Sea Res. I* 52:1515–27
8. Engelhardt T, Kallmeyer J, Cypionka H, Engelen B. 2014. High virus-to-cell ratios indicate ongoing production of viruses in deep subsurface sediments. *ISME J.* 8:1503–9
9. Whitman WB, Coleman DC, Wiebe WJ. 1998. Prokaryotes: the unseen majority. *PNAS* 95:6578–83
10. Williamson KE, Radosevich M, Wommack KE. 2005. Abundance and diversity of viruses in six Delaware soils. *Appl. Environ. Microbiol.* 71:3119
11. Kallmeyer J, Pockalny R, Adhikari RR, Smith DC, D'Hondt S. 2012. Global distribution of microbial abundance and biomass in subseafloor sediment. *PNAS* 109:16213–16
12. DeLong EF. 2003. Oceans of *Archaea*. *ASM News* 69:503–11
13. Rocke E, Pachiadaki MG, Cobban A, Kujawinski EB, Edgcomb VP. 2015. Protist community grazing on prokaryotic prey in deep ocean water masses. *PLOS ONE* 10:e012450
14. Danovaro R, Serresi M. 2000. Viral density and virus-to-bacterium ratio in deep-sea sediments of the Eastern Mediterranean. *Appl. Environ. Microbiol.* 66:1857
15. Wommack KE, Colwell RR. 2000. Virioplankton: viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev.* 64:69–114
16. Mokili JL, Rohwer F, Dutilh BE. 2012. Metagenomics and future perspectives in virus discovery. *Curr. Opin. Virol.* 2:63–77
17. Kimura M. 1962. On the probability of fixation of mutant genes in a population. *Genetics* 47:713
18. Barr JJ, Auro R, Furlan M, Whiteson KL, Erb ML, et al. 2013. Bacteriophage adhering to mucus provide a non-host-derived immunity. *PNAS* 110:10771–76
19. Kim MS, Park EJ, Roh SW, Bae JW. 2011. Diversity and abundance of single-stranded DNA viruses in human feces. *Appl. Environ. Microbiol.* 77:8062–70
20. Sender R, Fuchs S, Milo R. 2016. Revised estimates for the number of human and bacteria cells in the body. bioRxiv 036103. doi: 10.1101/036103
21. Haynes M, Rohwer F. 2011. The human virome. In *Metagenomics of the Human Body*, ed. KE Nelson, pp. 63–77. New York: Springer
22. Tuma RS, Beaudet MP, Jin X, Jones LJ, Cheung CY, et al. 1999. Characterization of SYBR Gold nucleic acid gel stain: a dye optimized for use with 300-nm ultraviolet transilluminators. *Anal. Biochem.* 268:278–88
23. Lang AS, Zhaxybayeva O, Beatty JT. 2012. Gene transfer agents: phage-like elements of genetic exchange. *Nat. Rev. Microbiol.* 10:472–82
24. Jover LF, Effler TC, Buchan A, Wilhelm SW, Weitz JS. 2014. The elemental composition of virus particles: implications for marine biogeochemical cycles. *Nat. Rev. Microbiol.* 12:519–28
25. Fuhrman JA. 1999. Marine viruses and their biogeochemical and ecological effects. *Nature* 399:541–48

26. Weinbauer MG. 2004. Ecology of prokaryotic viruses. *FEMS Microbiol. Rev.* 28:127–81
27. Weitz JS, Stock CA, Wilhelm SW, Bourouiba L, Coleman ML, et al. 2015. A multitrophic model to quantify the effects of marine viruses on microbial food webs and ecosystem processes. *ISME J.* 9:1352–64
28. Proctor LM, Fuhrman JA. 1990. Viral mortality of marine bacteria and cyanobacteria. *Nature* 343:60–62
29. Suttle CA. 2005. Viruses in the sea. *Nature* 437:356–61
30. Suttle CA. 2007. Marine viruses—major players in the global ecosystem. *Nat. Rev. Microbiol.* 5:801–12
31. Thingstad TF. 2000. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnol. Oceanogr.* 45:1320–28
32. Thingstad TF, Pree B, Giske J, Våge S. 2015. What difference does it make if viruses are strain-, rather than species-specific? *Front. Microbiol.* 6:320
33. Sandaa RA, Gómez-Consarnau L, Pinhassi J, Riemann L, Malits A, et al. 2009. Viral control of bacterial biodiversity—evidence from a nutrient-enriched marine mesocosm experiment. *Environ. Microbiol.* 11:2585–97
34. Paul JH. 1999. Microbial gene transfer: an ecological perspective. *J. Mol. Microbiol. Biotechnol.* 1:45–50
35. Frank JA, Lorimer D, Youle M, Witte P, Craig T, et al. 2013. Structure and function of a cyanophage-encoded peptide deformylase. *ISME J.* 7:1150–60
36. Lindell D, Sullivan MB, Johnson ZI, Tolonen AC, Rohwer F, Chisholm SW. 2004. Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *PNAS* 101:11013–18
37. Sullivan MB, Lindell D, Lee JA, Thompson LR, Bielawski JP, Chisholm SW. 2006. Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLOS Biol.* 4:e234
38. Goldsmith DB, Crosti G, Dwivedi B, McDaniel LD, Varsani A, et al. 2011. Development of *phoH* as a novel signature gene for assessing marine phage diversity. *Appl. Environ. Microbiol.* 77:7730–39
39. Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, et al. 2008. Functional metagenomic profiling of nine biomes. *Nature* 452:629–32
40. Bose M, Barber RD. 2006. Prophage Finder: a prophage loci prediction tool for prokaryotic genome sequences. *In Silico Biol.* 6:223–27
41. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. 2011. PHAST: a fast phage search tool. *Nucleic Acids Res.* 39(Suppl. 2):W347–52
42. McNair K, Bailey BA, Edwards RA. 2012. PHACTS, a computational approach to classifying the lifestyle of phages. *Bioinformatics* 28:614–18
43. Akhter S, Aziz RK, Edwards RA. 2012. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Res.* 40:e126
44. Fouts DE. 2006. Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res.* 34:5839–51
45. Paul JH. 2008. Prophages in marine bacteria: dangerous molecular time bombs or the key to survival in the seas? *ISME J.* 2:579–89
46. Canchaya C, Fournous G, Chibani-Chennoufi S, Dillmann ML, Brüssow H. 2003. Phage as agents of lateral gene transfer. *Curr. Opin. Microbiol.* 6:417–24
47. Casas V, Miyake J, Balsley H, Roark J, Telles S, et al. 2006. Widespread occurrence of phage-encoded exotoxin genes in terrestrial and aquatic environments in Southern California. *FEMS Microbiol. Lett.* 261:141–49
48. Breitbart M, Felts B, Kelley S, Mahaffy JM, Nulton J, et al. 2004. Diversity and population structure of a near-shore marine-sediment viral community. *Proc. R. Soc. B* 271:565
49. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, et al. 2002. Genomic analysis of uncultured marine viral communities. *PNAS* 99:14250–55
50. Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, et al. 2003. Metagenomic analyses of an uncultured viral community from human feces. *J. Bacteriol.* 185:6220–23
51. Brum JR, Hurwitz BL, Schofield O, Ducklow HW, Sullivan MB. 2015. Seasonal time bombs: dominant temperate viruses affect Southern Ocean microbial dynamics. *ISME J.* 10:437–49
52. Adriaenssens EM, Van Zyl L, De Maayer P, Rubagotti E, Rybicki E, et al. 2015. Metagenomic analysis of the viral community in Namib Desert hypoliths. *Environ. Microbiol.* 17:480–95

53. Roux S, Enault F, Ravet V, Colombet J, Bettarel Y, et al. 2016. Analysis of metagenomic data reveals common features of halophilic viral communities across continents. *Environ. Microbiol.* 18:889–903
54. Herskowitz I, Hagen D. 1980. The lysis-lysogeny decision of phage λ: explicit programming and responsiveness. *Annu. Rev. Genet.* 14:399–445
55. Knowles B, Silveira CB, Bailey BA, Barott K, Cantu VA, et al. 2016. Lytic to temperate switching of viral communities. *Nature* 531:466–70
56. Rohwer F, Edwards R. 2002. The phage proteomic tree: a genome-based taxonomy for phage. *J. Bacteriol.* 184:4529–35
57. Dwivedi B, Schmieder R, Goldsmith DB, Edwards RA, Breitbart M. 2012. PhiSiGns: an online tool to identify signature genes in phages and design PCR primers for examining phage diversity. *BMC Bioinform.* 13:37
58. Rohwer F. 2003. Global phage diversity. *Cell* 113:141
59. Chao A. 1984. Nonparametric estimation of the number of classes in a population. *Scand. J. Stat.* 11:265–70
60. Ignacio-Espinoza JC, Solonenko SA, Sullivan MB. 2013. The global virome: not as big as we thought? *Curr. Opin. Virol.* 3:566–71
61. Adriaenssens EM, Cowan DA. 2014. Using signature genes as tools to assess environmental viral ecology and diversity. *Appl. Environ. Microbiol.* 80:4470–80
62. Zhong Y, Chen F, Wilhelm SW, Poorvin L, Hodson RE. 2002. Phylogenetic diversity of marine cyanophage isolates and natural virus communities as revealed by sequences of viral capsid assembly protein gene *g20*. *Appl. Environ. Microbiol.* 68:1576
63. Jameson E, Mann NH, Joint I, Sambles C, Mühling M. 2011. The diversity of cyanomyovirus populations along a north–south Atlantic Ocean transect. *ISME J.* 5:1713–21
64. Chow CET, Fuhrman JA. 2012. Seasonality and monthly dynamics of marine myovirus communities. *Environ. Microbiol.* 14:2171–83
65. Dorigo U, Jacquet S, Humbert JF. 2004. Cyanophage diversity, inferred from *g20* gene analyses, in the largest natural lake in France, Lake Bourget. *Appl. Environ. Microbiol.* 70:1017
66. McDaniel LD, delaRosa M, Paul JH. 2006. Temperate and lytic cyanophages from the Gulf of Mexico. *J. Mar. Biol. Assoc. U.K.* 86:517–27
67. Comeau AM, Krisch HM. 2008. The capsid of the T4 phage superfamily: the evolution, diversity, and structure of some of the most prevalent proteins in the biosphere. *Mol. Biol. Evol.* 25:1321–32
68. Pagarete A, Chow CE, Johannessen T, Fuhrman J, Thingstad T, Sandaa R. 2013. Strong seasonality and interannual recurrence in marine myovirus communities. *Appl. Environ. Microbiol.* 79:6253–59
69. Thingstad T, Lignell R. 1997. Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. *Aquat. Microb. Ecol.* 13:19–27
70. Breitbart M, Rohwer F. 2005. Here a virus, there a virus, everywhere the same virus? *Trends Microbiol.* 13:278–84
71. Chenard C, Suttle C. 2008. Phylogenetic diversity of sequences of cyanophage photosynthetic gene *psbA* in marine and freshwaters. *Appl. Environ. Microbiol.* 74:5317–24
72. Huang S, Wilhelm SW, Jiao N, Chen F. 2010. Ubiquitous cyanobacterial podoviruses in the global oceans unveiled through viral DNA polymerase gene sequences. *ISME J.* 4:1243–51
73. Goldsmith DB, Parsons RJ, Beyene D, Salamon P, Breitbart M. 2015. Deep sequencing of the viral *phoH* gene reveals temporal variation, depth-specific composition, and persistent dominance of the same viral *phoH* genes in the Sargasso Sea. *PeerJ* 3:e997
74. Breitbart M, Miyake JH, Rohwer F. 2004. Global distribution of nearly identical phage encoded DNA sequences. *FEMS Microbiol. Lett.* 236:249–56
75. Short CM, Suttle CA. 2005. Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. *Appl. Environ. Microbiol.* 71:480–86
76. Sano E, Carlson S, Wegley L, Rohwer F. 2004. Movement of viruses between biomes. *Appl. Environ. Microbiol.* 70:5842
77. Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, et al. 2006. The marine viromes of four oceanic regions. *PLOS Biol.* 4:e368
78. Brum JR, Sullivan MB. 2015. Rising to the challenge: Accelerated pace of discovery transforms marine virology. *Nat. Rev. Microbiol.* 13:147–59

79. Watkins SC, Kuehnle N, Ruggeri CA, Malki K, Bruder K, et al. 2015. Assessment of a metaviromic dataset generated from nearshore Lake Michigan. *Mar. Freshw. Res.* doi: 10.1071/MF15172
80. Youle M, Haynes M, Rohwer F. 2012. Scratching the surface of biology's dark matter. In *Viruses: Essential Agents of Life*, ed. G Witzany, pp. 61–81. Dordrecht, Neth.: Springer
81. Tseng CH, Chiang PW, Shiah FK, Chen YL, Liou JR, et al. 2013. Microbial and viral metagenomes of a subtropical freshwater reservoir subject to climatic disturbances. *ISME J.* 7:2374–86
82. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19:455–77
83. García-López R, Vázquez-Castellanos JF, Moya A. 2015. Fragmentation and coverage variation in viral metagenome assemblies, and their effect in diversity calculations. *Front. Bioeng. Biotechnol.* 3:141
84. Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28:3150–52
85. Fierer N, Breitbart M, Nulton J, Salamon P, Lozupone C, et al. 2007. Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Appl. Environ. Microbiol.* 73:7059–66
86. Kim KH, Bae JW. 2011. Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Appl. Environ. Microbiol.* 77:7663–68
87. Duhaime MB, Sullivan MB. 2012. Ocean viruses: rigorously evaluating the metagenomic sample-to-sequence pipeline. *Virology* 434:181–86
88. 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.* 14:2526–37
89. United Nations. 2016. *Population and Vital Statistics Report: Statistical Papers*, Ser. A, Vol. LXVIII. New York: United Nations. **http://unstats.un.org/unsd/demographic/products/vitstats/Sets/Series_A_2016.pdf**