New dimensions of the virus world discovered through metagenomics

David M. Kristensen[1], Arcady R. Mushegian[1,2], Valerian V. Dolja[3] and Eugene V. Koonin[4]

[1] Stowers Institute of Medical Research, Kansas City, MO 64110, USA
[2] Department of Microbiology, Molecular Genetics and Immunology, Kansas University Medical Center, Kansas City, KS 66160, USA
[3] Department of Botany and Plant Pathology and Center for Genome Research and Biocomputing, Oregon State University, Corvallis, OR 97331, USA
[4] National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD 20894, USA

Metagenomic analysis of viruses suggests novel patterns of evolution, changes the existing ideas of the composition of the virus world and reveals novel groups of viruses and virus-like agents. The gene composition of the marine DNA virome is dramatically different from that of known bacteriophages. The virome is dominated by rare genes, many of which might be contained within virus-like entities such as gene transfer agents. Analysis of marine metagenomes thought to consist mostly of bacterial genes revealed a variety of sequences homologous to conserved genes of eukaryotic nucleocytoplasmic large DNA viruses, resulting in the discovery of diverse members of previously undersampled groups and suggesting the existence of new classes of virus-like agents. Unexpectedly, metagenomics of marine RNA viruses showed that representatives of only one superfamily of eukaryotic viruses, the picorna-like viruses, dominate the RNA virome.

The present state of viral metagenomics

Over the past two decades, the study of marine viruses using electron and fluorescent microscopy revealed an unexpected abundance of virus particles. At 10^6 to 10^9 particles per milliliter of sea water, viruses are the most abundant microbes in the sea and, most likely, in the entire biosphere [1,2]. Furthermore, they have emerged as crucial geochemical and ecological factors in marine ecosystems [3–5]. More recently, extensive data on the metagenomics of marine viruses have been reported [6–9]. Viral metagenomics is either pursued specifically by deep sequencing of environmental samples enriched for virus particles or emerge serendipitously through detection of virus-specific sequences in databases yielded by other metagenomic projects. The latter type of studies is mostly limited to known classes of viruses but the former has the potential to discover completely unknown viruses.

The gene repertoires of the putative marine viromes that were derived by sequencing double-stranded DNA (dsDNA) isolated from the fractions enriched for virus-like particles brought several major surprises and potential concerns. In particular, the estimates of the number of unique viral genotypes yielded breathtaking numbers of $>10^{30}$, making the marine viromes the most genetically diverse biological communities on earth [7,9]. The main and highly unexpected findings were that a substantial majority of the putative viral sequences were not significantly similar to any sequences in the current databases, and that those sequences that did have detectable homologs represented, primarily, various bacterial genes often having specific roles in central metabolism rather than distinct classes of genes commonly found in known bacteriophages or other viruses [6,10,11]. These remarkable findings suggest two possibilities that are not mutually exclusive. First, known viruses might not be representative of actual viromes, with the implication that marine viruses are the principal reservoir of new genes in the ocean. Second, the samples deemed to represent viromes might be, largely, not of viral origin and reflect contamination of the samples with non-viral DNA, which would indicate a serious shortcoming of the current metagenomic protocols.

Here, we apply several computational approaches to analyze the marine dsDNA viromes and show that, despite non-negligible contamination with bacterial genes, these sequences represent a collection that is markedly different in its statistical features from both prokaryotic and known viral genomes. Thus, there seems to be a realistic possibility that the actual marine viromes consist predominantly

Glossary

GOS: Global Ocean Survey, currently the most extensive collection of metagenomic data (mostly bacterial sequences) from a variety of locations in the Atlantic and the Pacific.
GTA: gene transfer agents, virus-like particles that consist of protein derived from defective prophages but incorporate host (bacterial or archaeal) DNA sequences.
NCLDV: nucleocytoplasmic large DNA viruses, an assemblage of six families of eukaryotic viruses that possess large DNA genomes (between approximately 100 Kb and 1 Mb), replicate mostly on the cytoplasm of infected cells but often go through a nuclear stage of reproduction as well, and are thought to be monophyletic on the basis of the conservation of a group of essential genes.
Pangenome: the entire set of genes found in a particular division of life forms (e.g. the pangenome of prokaryotes).
Picorna-like superfamily: a large assemblage of positive-strand RNA viruses of eukaryotes that infect animals, plants and unicellular organisms, and dominate marine RNA viromes.
Virome: the full compendium of viruses from a particular habitat.

Corresponding author: Koonin, E.V. (koonin@ncbi.nlm.nih.gov).

0966-842X/$ – see front matter © 2009 Published by Elsevier Ltd. doi:10.1016/j.tim.2009.11.003 Available online 26 November 2009

of virus-like particles that are different from well-characterized phages and might resemble gene transfer agents (GTAs). We further discuss the metagenomic studies of eukaryotic large nucleocytoplasmic DNA viruses (NCLDVs) and RNA viruses, and argue that, although still a young field, metagenomics is already revealing unexpected, yet fundamental features of the virus world.

The enigmatic marine dsDNA viromes

The most visible and provocative direction of viral metagenomics so far has been, beyond a doubt, the study of the totality of virus-like particles isolated from an environmental sample, mostly following the protocols developed for the DNA phages. Viral metagenomic sequences, or "viromes", were obtained by sequencing DNA extracted from partially purified environmental samples with a combination of filtration and density-dependent centrifugation in a cesium chloride gradient, a protocol that is expected to substantially enrich for virus-like particles [12,13]. Sequence similarity searches have shown that at least 50–60% and, more typically, close to 90% of the resulting DNA reads did not encode proteins that were significantly similar to others encoded in known genes of either virus or cellular origin [9,10]. This low percentage of sequences with detectable similarity to known viral proteins might indicate that many or perhaps most of the metagenomic sequences represent novel virus genes that have no matches in the databases because the true diversity of viruses has not been adequately sampled [7]. Alternatively (or additionally), despite all enrichment efforts, many of the reads could be of cellular rather than viral origin and have no matches in the databases because these genes belong to the poorly conserved fraction (sometimes referred to as the "gene cloud") of the pangenome of cellular organisms [14,15], and because the true diversity of cellular life has not been adequately sampled either. The use of short reads (until very recently, ~100 bp in most pyrosequencing protocols, compared with ~650 bp in traditional Sanger sequencing) exacerbates the problem, as evident from the fact that viral metagenomics efforts using Sanger sequencing yielded a fraction of reads with detectable homologs closer to 50% [6,16].

To obtain a better sense of the representation of viral and cellular genes in metagenomic samples, we reanalyzed data from some of the most extensive studies of this type. Specifically, we reanalyzed and compared DNA sequence data from the following sources: (i) a surface sample taken from the 5 m depth in the Sargasso Sea ("Sargasso-Total",

Box 1. The datasets used in this analysis

We reanalyzed data from a total microbial metagenome sampled from the surface of the Sargasso Sea—referred to in the text as ''Sargasso-Total'' (ST)—obtained from the J. Craig Venter Institute. A second dataset that we reanalyzed was a virus-enriched sample from the same habitat (although collected at 80 m deep rather than at 5 m as in the former) referred to as ''Sargasso-Viral'' (SV) and was obtained from the San Diego State University Center for Universal Microbial Sequencing. The ST DNA had been sequenced by a modified whole-genome shotgun approach [17], whereas the SV data had been obtained by 454 pyrosequencing [10]. Probably because of this, the ST dataset contains 15 times more sequences that are on average 6-fold longer than the reads in the SV sample. To reduce the effect of these differences in sample size and sequence length, we produced a fragmented version (''Sargasso-Total-fragmented'' or STf) of the ST data with the same number and length distribution of sequences as in SV. The fragmentation was performed by randomly choosing pieces of the ST sequences to match the size of each DNA fragment in SV (which was easily accomplished by using a simple custom perl script). The STf data refers to the average of three randomly generated replicates of this process.

We also examined three other oceanic virus-enriched samples—from the Arctic, British Columbia coast and Gulf of Mexico—available at the SDSU Center for Universal Microbial Sequencing [10] (Figure I). These datasets, also obtained by 454 pyrosequencing, are collectively referred to as ''Ocean-Viral'' (OV). To represent the genes of known phages, we used 27,254 protein-coding genes (1579 of which are our *de novo* predictions made using the GeneMarkS program [67]) from 323 complete dsDNA phage genomes available at NCBI as of November 2007 (''Phage-Total'' or PT). Because this sample contains 15 times fewer sequences that are on average 6-fold longer than the sequences in the SV dataset, we also produced a fragmented version of this dataset (''Phage-Total-fragmented'' or PTf). This was done, as for STf, by randomly choosing pieces of the much longer PT sequences to match exactly the size of each DNA fragment in SV.

For all samples, we compared the DNA sequences to the protein databases using BLASTX [63] and recorded the fraction of matching reads at *e*-values below 1e–5, 0.001, 0.01 and 0.1. Because the trends observed at each cut-off value were the same (data not shown), only the results for *e*-value <0.001 are reported. In the search for conserved orthologs, a simulated database size of 100,000,000 sequences was used for *e*-value calculation.

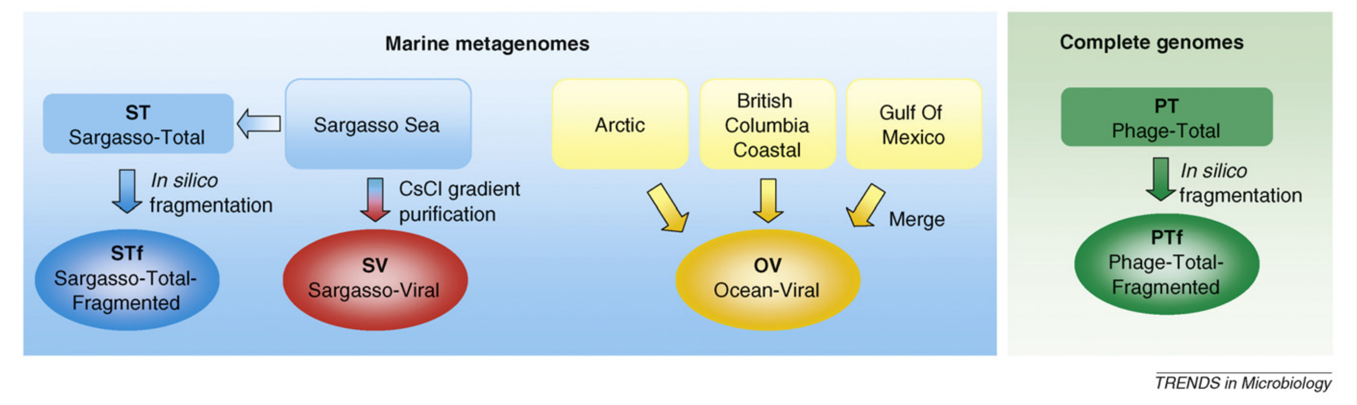


Figure I. A schematic representation of the metagenomic datasets.

ST) [17]; (ii) a virus-enriched sample also taken from the Sargasso Sea, albeit collected at 80 m depth ("Sargasso-Viral", SV) [10]; (iii) three marine samples from the Arctic, British Columbia Coast and Gulf of Mexico, respectively ("Ocean-Viral", OV) [10]; and (iv) a database containing protein-coding genes of known phages ("Phage-Total", PT). Both the SV and OV datasets contained many more short sequences than ST and PT. For this reason, we produced and analyzed fragmented and sampled versions of ST and PT that had a similar dataset size and length distribution to that of SV: "Sargasso-Total-fragmented" (STf) and "Phage-Total-fragmented" (PTf). Further description of samples and analysis protocols is provided in Box 1.

Our reanalysis shows that the virus-enriched SV is characterized by several distinctive properties compared with the other metagenomic samples, and, in particular, to the microbial STf dataset, which was shredded to resemble more closely the shorter sequence reads in SV. The fraction of sequences in SV that match a known protein in the non-redundant (NR) protein database is 10-fold lower than the respective fraction in STf, and more than 20-fold lower than in ST, PT or PTf (Figure 1a). The fraction of sequences that match the phage proteins that are known to be conserved in three or more phages (phage orthologous groups or POGs [18]; D.M. Kristensen, unpublished), as well as the fraction of POGs that were matched, are small in ST, with <4% of sequences matching 40% of POGs. These numbers are even smaller in the case of SV, with <2% of sequences matching only 12% of the POGs (Figure 1b). Nevertheless, compared with STf, the fraction of sequences matching POGs is 6-fold greater in SV. Thus, after the differences in the number and length of the sequences are taken into account, SV indeed appears substantially enriched in virus sequences. Furthermore, the fraction of sequences in SV that match a phage-specific POG (as opposed to POGs with low "phageness quotient", i.e. those that also include genes commonly found in bacterial genomes) is more than 10-fold greater than the respective fraction in STf (and even 2-fold greater than ST), although it is nearly 50-fold lower than in PTf (or PT). Thus, SV is strongly enriched for phage-specific genes although they account for only a small fraction of the total number of genes in the sample (<0.5%) and a relatively small fraction of phage-specific POGs (<10%). Conversely, sequences in SV match a significantly lower fraction of conserved cellular orthologs compared with STf (Figure 1c), and are far less likely to have a best match to known prokaryotic proteins (Figure 1d).

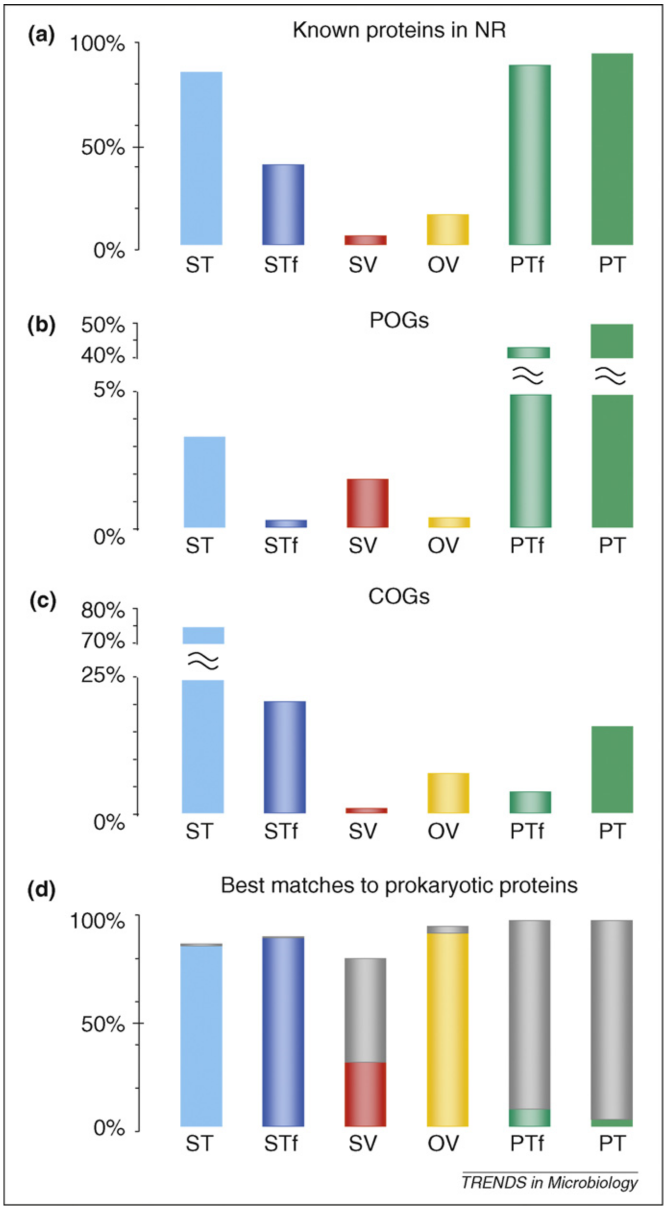
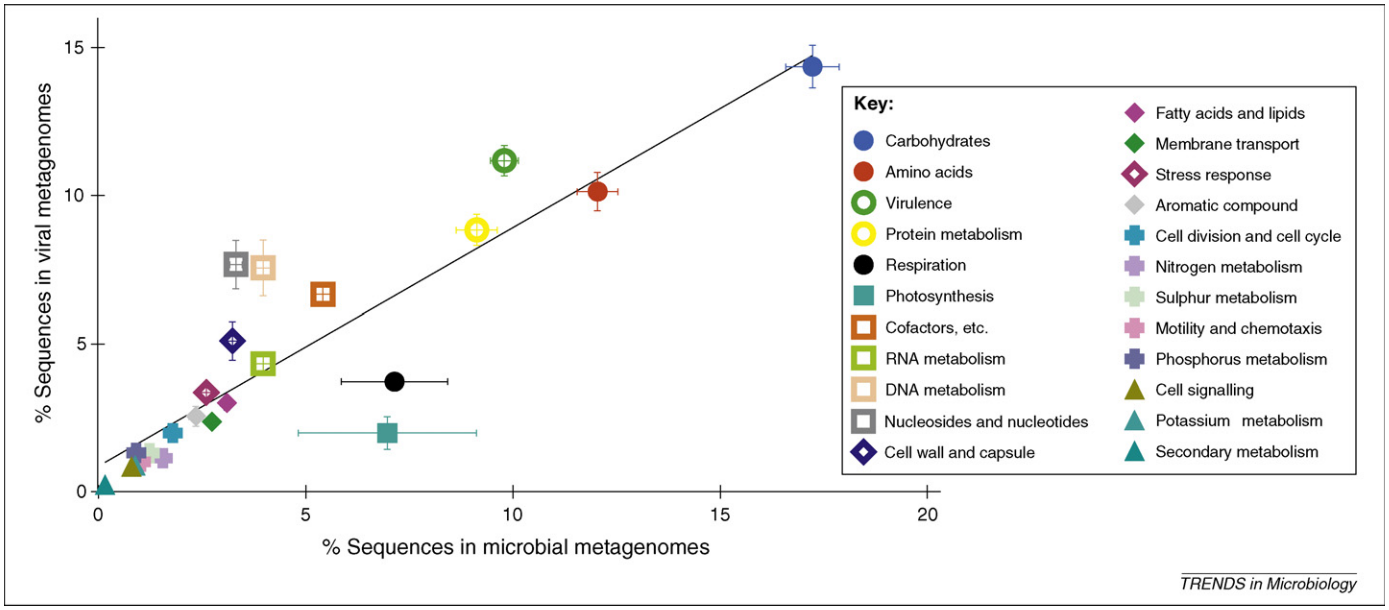


Figure 1. Properties of marine DNA viromes compared with non-viral marine metagenomic samples. For each sample (definitions and color scheme given in Box 1), the proportions with matches to **(a)** known proteins in the GenBank non-redundant (NR) database, **(b)** conserved phage orthologous groups (POGs), **(c)** conserved cellular orthologous groups (COGs) and **(d)** taxonomic classification of best matches (among known proteins in NR) as prokaryotes (color) or phages (gray) are shown.

The enrichment in viral proteins despite the presence of typical bacterial genes in SV is non-negligible as demonstrated, for instance, by the comparison of the content of sequences encoding ribosomal components between SV and STf. None of the known viral genomes encode any ribosomal proteins or ribosomal RNAs, and reliance on host ribosomes for genome expression was incorporated in most modern definitions of viruses [19]. Moreover, data from comparative genomics indicate that ribosomal protein genes are among the least prone to any form of horizontal gene transfer, despite a few demonstrable exceptions [20]. We observed that ~3% of the sequences in ST and ~0.6% of the sequences in STf showed significant similarity to ribosomal proteins. By contrast, matches to ribosomal proteins are observed in only 0.02% and 0.1% of the sequences in SV and OV (30-fold and 6-fold less than in STf, respectively). Many matches to rRNA genes were also observed in all samples. Thus, the analyzed virus metagenomes are depleted of cellular gene markers compared with non-viral metagenomes, but the contamination of SV and OV with genes of cellular origin is non-negligible.

Even more striking conclusions come from the recent functional profiling of diverse viromes [11]. We examined data from this study, namely, the mean percentages of sequences in 45 microbial and 42 viral metagenomes assigned to functional categories by comparison with functionally characterized proteins. There is a strong positive correlation between the distributions of gene functions in



TRENDS in Microbiology

Figure 2. Correlation between functional compositions of viral and cellular metagenomes. The plotted data consist of mean percentages of the sequences from 45 microbial and 42 viral metagenomes that match characterized proteins in major functional categories. Data obtained from Ref. [11].

the two metagenome types: the correlation coefficient is 0.88 for the raw data values and 0.91 for rank-ordered values (Figure 2). In both the cellular and viral metagenomes, the most common functional categories are unmistakably "cellular" such as carbohydrate metabolism and amino acid metabolism. This correlation might, in part, be explained by the choice of available functional categories, most of which encompass typical cellular rather than viral functions. However, some of the cellular categories, such as DNA replication, recombination, repair and nucleotide metabolism, are well represented in the known viral genomes [18], and several virus-specific categories were also included. Paradoxically, the enrichment for virus-like particles in the virome samples did not result in depletion of typical cellular functions relative to typical viral functions. In both the cellular and viral metagenomes, the virus-specific categories are in the lower end of the rank distribution (Figure 2) where they are intermingled with some seemingly virus-unrelated functions.

What's in a virus?

So what is the nature of the currently available marine virus metagenomic sequences? The comparisons described above show that, despite the low fraction of sequences with similarity to any known proteins, these viromes are significantly enriched for viral genes. Thus, a striking possibility is that these metagenomes are representative samples of the marine virus pangenome, whereas known phage genomes are not. Should this be the case, one will have to infer that viruses and cellular organisms (at least prokaryotes) each have a possibly overlapping "cloud" of poorly conserved genes [14]. This vast set of genes would be randomly sampled by typical marine viruses resulting in a representation that is not seen in the collection of known phages, perhaps because the latter are dominated by viruses infecting just two clades of bacteria (Proteobacteria and actinomycetes) or for other, still unknown, reasons.

It also seems possible that a substantial portion of CsCl-purified fraction is represented not by bona fide viruses but rather by GTAs or similar phage-like particles that apparently are derivatives of defective prophages [21,22]. These phage-like particles might contain a collection of bacterial and viral genes that are sampled randomly or at least with an unknown bias, which in any case does not result in the generation of typical phage genomes (where "high-phageness" genes dominate the conserved gene fraction). The GTAs were originally discovered in the alphaproteobacterium *Rhodobacter* but subsequently were identified in many diverse bacteria and archaea, and particularly in marine bacterioplankton [23,24]. The paucity of recognizable genes in the marine viromes implies that, if these samples consist largely of GTA-like agents, the GTA "genomes" are enriched not simply in bacterial as opposed to viral genes, but specifically in genes from the poorly conserved "cloud" of the prokaryotic gene universe as opposed to conserved genes. This conclusion is far from trivial because, although the cloud of poorly conserved genes in the biosphere is vast, the fraction of these genes in each individual cellular genome sequenced thus far is relatively small, on the order of 10–15% of all genes [14]. Thus, the possible enrichment of abundant GTA-like agents for "cloud" genes seems to imply a specific selective pressure, perhaps a purifying selection against GTAs including conserved genes or, in other words, against horizontal transfer of conserved genes. We have to emphasize that the hypothesis on the dominance of GTA-like particles encompassing poorly conserved prokaryotic genes in marine viromes is not about contamination with cellular material; on the contrary, GTAs and similar entities are bona fide components of viromes.

Should the material isolated by enriching for viral particles indeed consist predominantly of GTA-like entities, it would be fair to conclude that metagenomics is changing the existing ideas of the structure and dynamics of the prokaryotic world. In the new picture, the traffic of genetic information, in particular of rare, poorly conserved genes, in prokaryotes would be even much more extensive than previously conceived, and the role of specialized vehicles in this traffic would be decisive. This picture is compatible with the recent studies of cyanophages that revealed the major contribution of viruses to the dissemination of photosynthetic genes, and with the estimates (based on the available data on virus abundance in the sea) that suggest that, over the history of the biosphere, viruses have repeatedly sampled the entire diversity of bacterial genes [25,26].

These conclusions would also reveal a new face of the virus world by showing that the quantitatively dominant denizens of the virosphere (at least, in the marine environment) might not be familiar viruses encompassing "viral hallmark genes" (VHGs) involved in viral replication and morphogenesis [27,28] but rather GTAs and similar agents for which "genomes" consist (mostly) of packages of host genes. Should this be the case, the question of whether metagenomic studies were successful in isolating viromes becomes, more or less, obsolete, because the very boundary between viral and cellular genomes almost disappears. Of course, in the study of virus evolution, it still should be useful to distinguish "true" viruses capable of productive infection of a specific host from various types of virus-like agents incapable of replication.

A closer examination of the known part of the virosphere reveals agents that were isolated as viruses but seem to be GTA-like in their genomic composition. The primary examples are numerous viruses of hyperthermophilic archaea that possess very few VHGs and encompass mostly genes for small DNA-binding proteins with predicted regulatory functions (incidentally, the type of proteins that are likely to escape detection in metagenomic analyses owing to their small size and limited sequence conservation) and proteins without detectable homologs [29,30], and insect polydnaviruses that possess, primarily, genes derived from the host along with unique genes [31,32]. Remarkably, a recent study of polydnaviruses demonstrated their origin from regular viruses of the nudivirus group and showed that viral genes integrated in the host genome contribute to the biogenesis of polydnaviruses but are not included into their genomes [33]. Thus, in many respects, polydnaviruses can be viewed as eukaryotic counterparts of GTAs.

Before we embrace the radical reappraisal of the structure and dynamics of the microbial and virus worlds outlined above, it is worth emphasizing that the majority of sequences that comprise the purported viromes are of unknown origin. Although these metagenomic sequence collections definitely are enriched in phage-specific sequences, the apparent contribution of prokaryotic genes is far from negligible and seems to differ between viromes (e.g. it is considerably greater in OV compared with SV). Indeed, the single definitive criterion for the true "virusness" of metagenomic sequences is their presence in the complete, contiguous genomes of viruses that contain indisputable viral genes such as VHGs. This criterion was successfully fulfilled in the analysis of cyanophage genomes that encode proteins of photosystems I and II [26,34]. However, it remains to be addressed for other typical prokaryotic (and patently non-viral) genes such as those encoding ribosomal components, vitamin metabolism-related genes or genes related to cellular motility.

The serendipitous emergence of metagenomics of large nucleocytoplasmic DNA viruses of eukaryotes

The eukaryotic large NCLDVs are a large, probably monophyletic, assemblage of eukaryotic dsDNA viruses that includes six families with genome sizes between ~100 kb and ~1 Mb which infect diverse eukaryotic hosts (Table 1). These viral families share nine core genes that are represented in all sequenced NCLDV genomes, and approximately 30 additional genes that are found in the majority of viruses from the six families and are likely to have been present in the putative common ancestor of the NCLDVs [35,36].

We are unaware of any metagenomic study aimed at the characterization of NCLDVs and only one metagenomic study of any eukaryotic DNA viruses, which reported the discovery of numerous herpesvirus genes in coral habitats [37]. The marine viromes discussed in the preceding sections contain no detectable homologs of NCLDV genes. Unexpectedly, however, analysis of the total metagenomic sequences from the Sargasso Sea and Atlantic and Pacific Ocean samples that were obtained with methods supposed to enrich for bacteria revealed numerous sequences homologous to NCLDV genes [38,39]; in particular, highly conserved homologs of approximately 25% of the almost 1000 genes of the giant mimivirus [40], the prototype of one of the NCLDV families. Some of these sequences were also closely related to sequenced genes of giant algal viruses, and phylogenetic analysis supported clustering of these genes with mimivirus homologs [39], thus revealing the

Table 1. The NCLDV families and their representation in marine metagenomic samples

Virus family	Known host range	Genome size (kb)[a]	Number and percentage of NCLDV-related sequences [b]	
			Sargasso Sea metagenome	GOS metagenome
Poxviridae	Animals (vertebrates and insects)	134–360	0	0
Asfarviridae	Animals (mammals)	170	1 (0.4%)	20 (1%)
Iridoviridae	Animals (vertebrates and insects)	103–191	7 (3%)	74 (5%)
Ascoviridae	Animals (insects)	119–186	0	0
Mimiviridae	*Amoebozoa* (*Acanthamoeba*)	1181	50 (22%)	279 (20%)
Phycodnaviridae	*Chlorophyta* (green algae), *Haptophyta*, stramenopiles	155–407	160 (71%)	1032 (72%)

[a] Data from the NCBI genome database.

[b] The amino acid sequences of the following three core NCLDV proteins were used as queries to search the metagenomic nucleotide sequences using the TBLASTN program [63]: the major capsid protein (ortholog of vaccinia virus D13 protein), a transcription factor (ortholog of vaccinia virus A2 protein) and the disulfide oxidoreductase (ortholog of vaccinia virus E10 protein). Metagenomic sequences with the similarity to the NCLDV proteins above the chosen cut-off value (expectation value $<10^{-5}$) were pooled and searched against the non-redundant protein database (NCBI) using the BLASTX program [63], and the taxonomic affiliation of the most similar sequences were determined using BLAST taxonomy reports.

diversity of this family of the NCLDVs that so far was represented only by the mimivirus and the closely related mamavirus [41].

A systematic search for homologs of the viral family B of DNA polymerases in the GOS (Global Ocean Survey) metagenomic samples and subsequent phylogenetic analysis revealed multiple sequences that clustered in a phylogenetic tree with those from members of four NCLDV families [42]. The other two families, poxvirus and ascovirus, so far remain limited to animals in their host ranges.

The family B DNA polymerase alone might not be sufficient to unequivocally establish the NCLDV origin of a metagenomic sequence read, because homologous polymerases are also found in phages and in cellular life forms [43].To further characterize the representation of the NCLDVs in marine metagenomic samples, we performed BLAST searches of the Sargasso Sea and GOS metagenomic sequences using as queries representative sequences of all NCLDV families for three core proteins that are conserved in all NCLDVs but have no closely similar homologs outside this class of viruses (Table 1) [36]. The family affinities of the detected similar sequences were determined by reverse BLAST comparisons and a taxonomic breakdown of the results. We found that NCLDV families were represented remarkably consistently in both marine metagenomic samples, with the greatest abundance observed for phycodnaviruses, followed by mimiviruses and a much lower presence of apparent relatives of iridoviruses and asfarviruses (Table 1).

A notable result of the serendipitously emerging field of NCLDV metagenomics is the discovery of multiple and diverse sequences that are most closely related to mimivirus and asfarvirus genes [42] (Table 1). These families are so far represented by a single sequenced genome (asfarviruses) and two closely related genomes (mimiviruses), thus metagenomics is currently the only source of information on their diversity. Furthermore, marine metagenomes contain sequences that appear to be related to all known NCLDV families but are as distant from all of them as those families are from one another; these novel sequences might represent new families of NCLDVs (E.V. Koonin, unpublished).

Perhaps, even more striking observations were made through comparisons between the marine metagenomic samples and protein sequences encoded in the genome of the virophage, a novel virus with a 22-kb DNA genome that parasitizes on the giant mimivirus. The virophage genome encompasses 18 protein-coding genes, several of which are distantly related to genes involved in the replication of NCLDVs, others seem to be derived from plasmids or transposons, and two or three genes seem to have been recently acquired from the host mimivirus [41]. The GOS metagenomic sample contains numerous sequences of closely related homologs of approximately half of the virophage genes, including the predicted primase–helicase and DNA-packaging ATPase [41]. By contrast, no homologs of the virophage capsid protein were identified. The predicted virophage primase belongs to a family of putative primases–polymerases distantly related to the bacterial PolA and encoded in mobile elements present in the genomes of diverse bacteria [44]. These poorly characterized mobile elements also contain distant homologs of some of the other virophage genes. These observations emphasize the potential of the metagenomic data for the identification of novel classes of mobile elements that might have the capacity to give rise to new types of viruses with unexpected host ranges.

Where do the NCLDV genes in the marine metagenomes come from? The source samples were thought to be enriched for prokaryotic cells, thus there seem to be two probable sources of the NCLDV sequences: (i) picoeukaryotes such as certain algae for which cells are small enough to pass the 0.8 μm bacterial filters (e.g. *Micromonas*, which is the most abundant picoeukaryote [45], or *Ostreococcus*) and (ii) free giant viruses comparable in size to the mimivirus. Indeed, a considerable variety of sequences from picoeukaryotes was detected in the Sargasso Sea samples [46]. Regardless of which of these sources makes the main contribution, metagenomics of the NCLDVs reveals the diversity of the asfarvirus and mimivirus families, expands the host range of these families (e.g. by indicating that asfarviruses, so far isolated only from animals, probably reproduce in marine unicellular eukaryotes) and might lead to the discovery of new viral families. The study of the metagenomic homologs of virophage genes goes a step further by suggesting novel routes of viral evolution.

Metagenomics of RNA viruses: the unexpected dominance of the picornavirus-like superfamily

Compared with metagenomics of DNA viruses that benefited from immense flow of data both from virus-centered environmental projects [8,47] and from bacteria-oriented studies (that serendipitously yield sequences of large viruses, see above), RNA virus metagenomics is far more limited in scope. This disparity is largely a result of the formidable technical challenges presented by RNA metagenomics. Unlike DNA viruses, RNA viruses are small and barely detectable by epifluorescent microscopy, thus the relative prevalence of RNA viruses in the environment remains poorly characterized [48]. Sequencing RNA viral genomes extracted from the environment is not trivial either because RNA is fragile and needs to be reverse-transcribed into DNA. All these problems aside, RNA metagenomics has already made a major impact on our current understanding of virus evolution and global ecology.

RNA virus communities have been studied in several habitats, including clinical samples from mammals [49], but the coastal waters of British Columbia and Hawaii yielded most of the current data [50–52]. In contrast to DNA viromes which contain few sequences with detectable homologs (see above), up to ~40% of the reads in RNA viromes exhibited significant similarity to sequences of known RNA viruses [51]. This difference could stem from several factors, including the greater diversity of DNA viruses and a higher level of contamination of the DNA viromes by non-viral sequences (already discussed in this article). Another reason for this disparity could be the excess of prokaryotes over eukaryotes in marine environments. Indeed, up to 90% of the biomass of marine cellular organisms comes from prokaryotes, with unicellular eukaryotes being a distant second, and multicellular

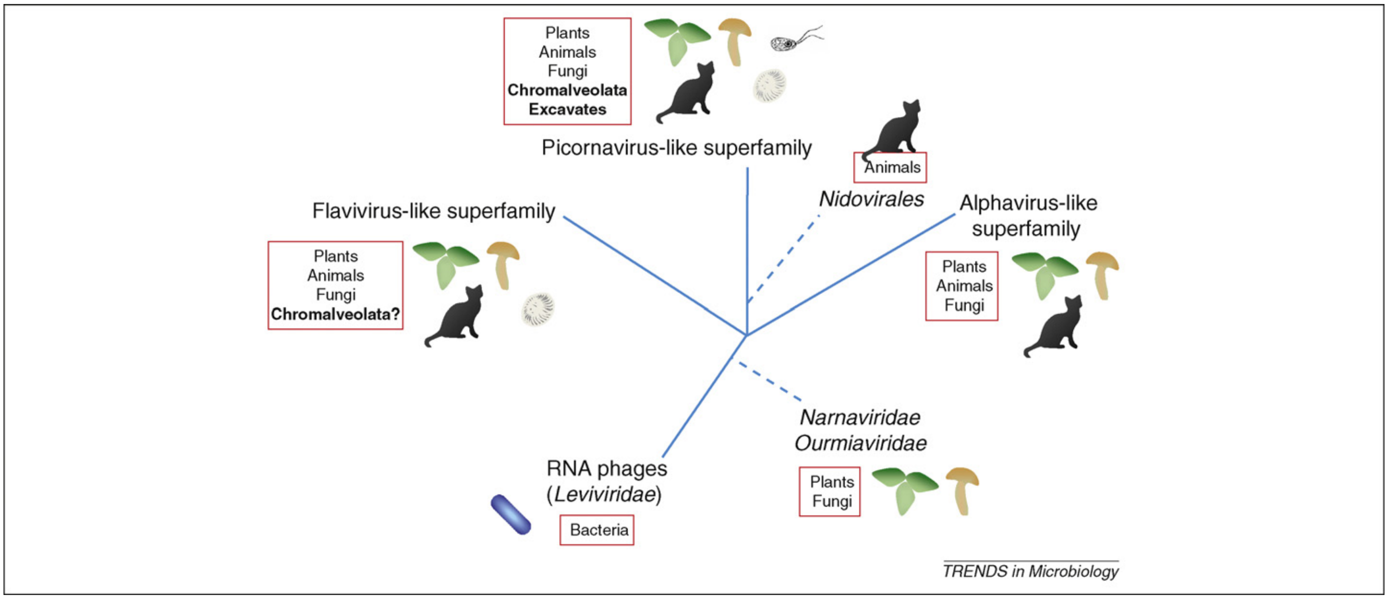


Figure 3. The diversity of positive-strand RNA viruses and their hosts. The schematic tree of the major groups of positive-strand RNA viruses is based on the analysis in Ref. [64], and the positions of *Nidovirales* [65] and *Narnaviridae/Ourmiaviridae* [66] are shown tentatively on the basis of subsequent studies. Host range is indicated for each branch. Data on viruses that infect unicellular eukaryotes (bold font) come primarily from metagenomic studies (see main text).

plants and animals together providing only a minor contribution [4]. For reasons we are only beginning to address, there are astronomical numbers and remarkable diversity of prokaryotic DNA viruses, but only a handful of RNA viruses that infect bacteria and none are known to infect archaea [27]. In line with this distribution, no RNA phages have been found so far in the marine RNA viromes [51].

Surprisingly, despite the enormous variety of positive-strand, negative-strand and double-strand RNA viruses in plants and animals, the known RNA virome of marine metagenomic samples overwhelmingly consists of a single evolutionary lineage of the positive-strand RNA viruses: the picornavirus-like superfamily. In one viral community from the coastal waters of British Columbia, picorna-like sequences comprised ~98% of all recovered virus-related sequences [51]. This result is corroborated by the results of the studies designed to specifically sample the marine picorna-like viruses, which demonstrated their persistent and widespread occurrence throughout the Pacific [50,52].

Whereas all previously recognized viruses of the order *Picornavirales* [53] infect either animals or plants, the novel marine viruses appear to infect unicellular species from the *Chromalveolata* supergroup including such diverse organisms as diatoms, a Raphidophyte, a Thraustrochytrid and a dinoflagellate (Figure 3). Phylogenetic analysis suggests that these viruses form several new virus families [54–57] which are rapidly populated with new members identified in parallel metagenomic studies [51,52,58].

The discovery of a previously unsuspected diversity of RNA viruses in the sea dramatically expanded our knowledge of the global virus ecology and prompted deeper inquiry into the origins and evolution of the picorna-like viruses. Recently, it was found that the phylogenies of picorna-like viruses and their hosts are radically incongruent. The most plausible explanation of this incongruency appears to be that, at least at the early stages of their evolution, picorna-like viruses did not coevolve with the hosts but diversified into five major evolutionary lineages before the radiation of the eukaryotic supergroups that subsequently “sampled” the viral diversity [59]. Furthermore, the core genes of picorna-like viruses were traced to ancestors from bacteria, bacterial retroelements and phages. Taken together, these findings suggest that picorna-like viruses originated during eukaryogenesis, probably in a ‘Big Bang’ manner, via mixing and matching genes of diverse prokaryotic origins [27,59].

The cause of the strong bias towards picorna-like viruses in marine RNA viral metagenomes studied thus far remains unknown. The richness of the marine community of picorna-like viruses is a stark contrast to the scarcity of other marine RNA viruses. So far, only one positive-strand RNA virus of the flavivirus-like superfamily and none of the alphavirus-like superfamily have been identified in environmental samples [51]. A trivial explanation of the picorna-like virus dominance could be the limited scale of the RNA metagenomic surveys that might result in a biased representation of the viral diversity in the environment. Alternatively, the representation of marine RNA viruses in the current metagenomic samples might be adequate, that is, unicellular eukaryotes (at least, in the sea) could be indeed primarily infected by picorna-like viruses. Such an affiliation of picorna-like RNA viruses with diverse unicellular eukaryotic hosts seems to imply that these viruses are the ancestral group from which the positive-strand RNA viruses of multicellular eukaryotes evolved [59]. More extensive sampling of marine RNA viromes has the potential to put this hypothesis to test and, perhaps, reveal the marine prototypes of all major evolutionary lineages of the animal and plant positive-strand RNA viruses.

Metagenomics updates the study of the evolution of eukaryotic positive-strand RNA viruses, but the origins of double-strand and negative-strand RNA viruses remain unexplained. So far, only a few double-strand, reovirus-like viruses and no negative-strand RNA viruses have been identified in unicellular eukaryotes, whereas a great variety of viruses of both classes are known to infect animals including marine invertebrates and vertebrates [48,51,60]. Both double-strand and negative-strand RNA viruses are also common in land plants where, however, neither class is nearly as diverse as positive-strand RNA viruses [61]. This pattern of global ecology is compatible with at least two possibilities. First, the double-strand and negative-strand RNA viruses could have emerged relatively recently, following the appearance of multicellular eukaryotes, from positive-strand or retroid RNA viruses, probably, on several independent occasions [62]. The second, perhaps, less probable possibility is that double-strand and negative-strand RNA viruses are as ancient as eukaryotes themselves but, for unknown reasons, they are at present either extinct or restricted to niches not yet sampled by metagenomics. Whatever the case, deeper examination of the marine RNA viromes will probably provide evidence towards this outstanding problem of virus evolution.

Concluding remarks and future directions

Recent advances in virus metagenomics demonstrate that viruses and virus-like elements are the most abundant biological entities on this planet and that their genomic diversity inferred from model systems is vastly undersampled. Moreover, analysis of the metagenomic data suggests a distinct possibility that the dominant forms in viromes could be qualitatively different from well-characterized viruses and might resemble GTAs. An implication of these findings is that highways of gene transfer between the virus world and the "stable" genomes of cellular life forms could be much wider than previously suspected. Virus metagenomics has yielded other unexpected findings as well, such as the discovery of a wealth of sequences apparently originating from eukaryotic DNA viruses (NCLDVs) in marine metagenomes supposedly dominated by prokaryotes, and the demonstration of the almost exclusive representation of picorna-like viruses in marine RNA viromes.

Despite these provocative findings, the study of virus diversity in the environment and in particular viral metagenomics clearly is in its infancy and is in need of improved technologies for sample preparation and sequence data analysis. Some of the probably most productive avenues for future research include: a thorough characterization of marine viromes (in which an important role will belong to proteomics); the study of new viromes from diverse and, in particular, extreme habitats; a thorough analysis of the role of viruses in ecology and geochemistry; and an estimation of the actual size of the virosphere. Furthermore, to adequately reconstruct the evolution of the entire virus world, the metagenomic approach has to be complemented by a census of viruses infecting all major lineages of prokaryotic and eukaryotic hosts. There is no doubt that, with the advent of new generations of sequencing and proteomic technologies, virus metagenomics has a bright future and will substantially contribute to an emerging new understanding of the genetic diversity of life.

References

1 Bergh, O. *et al.* (1989) High abundance of viruses found in aquatic environments. *Nature* 340, 467–468
2 Suttle, C.A. (2005) Viruses in the sea. *Nature* 437, 356–361
3 Fuhrman, J.A. (1999) Marine viruses and their biogeochemical and ecological effects. *Nature* 399, 541–548
4 Suttle, C.A. (2007) Marine viruses – major players in the global ecosystem. *Nat. Rev. Microbiol.* 5, 801–812
5 Haaber, J. and Middelboe, M. (2009) Viral lysis of *Phaeocystis pouchetii*: implications for algal population dynamics and heterotrophic C, N and P cycling. *ISME J.* 3, 430–441
6 Edwards, R.A. and Rohwer, F. (2005) Viral metagenomics. *Nat. Rev. Microbiol.* 3, 504–510
7 Breitbart, M. and Rohwer, F. (2005) Here a virus, there a virus, everywhere the same virus? *Trends Microbiol.* 13, 278–284
8 Rohwer, F. and Thurber, R.V. (2009) Viruses manipulate the marine environment. *Nature* 459, 207–212
9 Bench, S.R. *et al.* (2007) Metagenomic characterization of Chesapeake Bay virioplankton. *Appl. Environ. Microbiol.* 73, 7629–7641
10 Angly, F.E. *et al.* (2006) The marine viromes of four oceanic regions. *PLoS Biol.* 4, e368
11 Dinsdale, E.A. *et al.* (2008) Functional metagenomic profiling of nine biomes. *Nature* 452, 629–632
12 Breitbart, M. *et al.* (2002) Genomic analysis of uncultured marine viral communities. *Proc. Natl. Acad. Sci. U. S. A.* 99, 14250–14255
13 Thurber, R.V. *et al.* (2009) Laboratory procedures to generate viral metagenomes. *Nat. Protoc.* 4, 470–483
14 Koonin, E.V. and Wolf, Y.I. (2008) Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.* 36, 6688–6719
15 Lapierre, P. and Gogarten, J.P. (2009) Estimating the size of the bacterial pan-genome. *Trends Genet.* 25, 107–110
16 Wommack, K.E. *et al.* (2008) Metagenomics: read length matters. *Appl. Environ. Microbiol.* 74, 1453–1463
17 Venter, J.C. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304, 66–74
18 Liu, J. *et al.* (2006) Protein repertoire of double-stranded DNA bacteriophages. *Virus Res.* 117, 68–80
19 Raoult, D. and Forterre, P. (2008) Redefining viruses: lessons from Mimivirus. *Nat. Rev. Microbiol.* 6, 315–319
20 Sorek, R. *et al.* (2007) Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* 318, 1449–1452
21 Stanton, T.B. (2007) Prophage-like gene transfer agents-novel mechanisms of gene exchange for *Methanococcus*, *Desulfovibrio*, *Brachyspira*, and *Rhodobacter* species. *Anaerobe* 13, 43–49
22 Lang, A.S. and Beatty, J.T. (2007) Importance of widespread gene transfer agent genes in alpha-proteobacteria. *Trends Microbiol.* 15, 54–62
23 Biers, E.J. *et al.* (2008) Occurrence and expression of gene transfer agent genes in marine bacterioplankton. *Appl. Environ. Microbiol.* 74, 2933–2939
24 Paul, J.H. (2008) Prophages in marine bacteria: dangerous molecular time bombs or the key to survival in the seas? *ISME J.* 2, 579–589
25 Clokie, M.R. and Mann, N.H. (2006) Marine cyanophages and light. *Environ. Microbiol.* 8, 2074–2082
26 Sharon, I. *et al.* (2009) Photosystem I gene cassettes are present in marine virus genomes. *Nature* 461, 258–262
27 Koonin, E.V. *et al.* (2006) The ancient virus world and evolution of cells. *Biol. Direct* 1, 29
28 Koonin, E.V. *et al.* (2009) Compelling reasons why viruses are relevant for the origin of cells. *Nat. Rev. Microbiol.* 7, 615
29 Prangishvili, D. *et al.* (2006) Evolutionary genomics of archaeal viruses: unique viral genomes in the third domain of life. *Virus Res.* 117, 52–67
30 Prangishvili, D. *et al.* (2006) Viruses of the Archaea: a unifying view. *Nat. Rev. Microbiol.* 4, 837–848
31 Desjardins, C. *et al.* (2005) New evolutionary frontiers from unusual virus genomes. *Genome Biol.* 6, 212
32 Dupuy, C. *et al.* (2006) Unfolding the evolutionary story of polydnaviruses. *Virus Res.* 117, 81–89
33 Bezier, A. *et al.* (2009) Polydnaviruses of braconid wasps derive from an ancestral nudivirus. *Science* 323, 926–930
34 Sullivan, M.B. *et al.* (2006) Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biol.* 4, e234

35 Iyer, L.M. *et al.* (2001) Common origin of four diverse families of large eukaryotic DNA viruses. *J. Virol.* 75, 11720–11734

36 Iyer, L.M. *et al.* (2006) Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Res.* 117, 156–184

37 Vega Thurber, R.L. *et al.* (2008) Metagenomic analysis indicates that stressors induce production of herpes-like viruses in the coral *Porites compressa*. *Proc. Natl. Acad. Sci. U. S. A.* 105, 18413–18418

38 Ghedin, E. and Claverie, J.M. (2005) Mimivirus relatives in the Sargasso sea. *Virol. J.* 2, 62

39 Monier, A. *et al.* (2008) Marine mimivirus relatives are probably large algal viruses. *Virol. J.* 5, 12

40 Raoult, D. *et al.* (2004) The 1.2-megabase genome sequence of Mimivirus. *Science* 306, 1344–1350

41 La Scola, B. *et al.* (2008) The virophage as a unique parasite of the giant mimivirus. *Nature* 455, 100–104

42 Monier, A. *et al.* (2008) Taxonomic distribution of large DNA viruses in the sea. *Genome Biol.* 9, R106

43 Koonin, E.V. (2006) Temporal order of evolution of DNA replication systems inferred by comparison of cellular and viral DNA polymerases. *Biol. Direct* 1, 39

44 Iyer, L.M. *et al.* (2008) A new family of polymerases related to superfamily A DNA polymerases and T7-like DNA-dependent RNA polymerases. *Biol. Direct* 3, 39

45 Foulon, E. *et al.* (2008) Ecological niche partitioning in the picoplanktonic green alga *Micromonas pusilla*: evidence from environmental surveys using phylogenetic probes. *Environ. Microbiol.* 10, 2433–2443

46 Piganeau, G. *et al.* (2008) Picoeukaryotic sequences in the Sargasso Sea metagenome. *Genome Biol.* 9, R5

47 Schoenfeld, T. *et al.* (2008) Assembly of viral metagenomes from yellowstone hot springs. *Appl. Environ. Microbiol.* 74, 4164–4174

48 Lang, A.S. *et al.* (2009) RNA viruses in the sea. *FEMS Microbiol. Rev.* 33, 295–323

49 Delwart, E.L. (2007) Viral metagenomics. *Rev. Med. Virol.* 17, 115–131

50 Culley, A.I. *et al.* (2003) High diversity of unknown picorna-like viruses in the sea. *Nature* 424, 1054–1057

51 Culley, A.I. *et al.* (2006) Metagenomic analysis of coastal RNA virus communities. *Science* 312, 1795–1798

52 Culley, A.I. and Steward, G.F. (2007) New genera of RNA viruses in subtropical seawater, inferred from polymerase gene sequences. *Appl. Environ. Microbiol.* 73, 5937–5944

53 Le Gall, O. *et al.* (2008) Picornavirales, a proposed order of positive-sense single-stranded RNA viruses with a pseudo-T=3 virion architecture. *Arch. Virol.* 153, 715–727

54 Lang, A.S. *et al.* (2004) Genome sequence and characterization of a virus (HaRNAV) related to picorna-like viruses that infects the marine toxic bloom-forming alga *Heterosigma akashiwo*. *Virology* 320, 206–217

55 Takao, Y. *et al.* (2006) Complete nucleotide sequence and genome organization of a single-stranded RNA virus infecting the marine fungoid protist *Schizochytrium* sp. *J. Gen. Virol.* 87, 723–733

56 Tomaru, Y. *et al.* (2009) Isolation and characterization of a single-stranded RNA virus infecting the bloom-forming diatom *Chaetoceros socialis*. *Appl. Environ. Microbiol.* 75, 2375–2381

57 Nagasaki, K. *et al.* (2005) Comparison of genome sequences of single-stranded RNA viruses infecting the bivalve-killing dinoflagellate *Heterocapsa circularisquama*. *Appl. Environ. Microbiol.* 71, 8888–8894

58 Culley, A.I. *et al.* (2007) The complete genomes of three viruses assembled from shotgun libraries of marine RNA virus communities. *Virol. J.* 4, 69

59 Koonin, E.V. *et al.* (2008) The Big Bang of picorna-like virus evolution antedates the radiation of eukaryotic supergroups. *Nat. Rev. Microbiol.* 6, 925–939

60 Brussaard, C.P. *et al.* (2004) Discovery of a dsRNA virus infecting the marine photosynthetic protist *Micromonas pusilla*. *Virology* 319, 280–291

61 Hull, R. (2009) *Comparative Plant Virology*, Academic Press

62 Koonin, E.V. (1992) Evolution of double-stranded RNA viruses: a case for polyphyletic origin from different groups of positive-stranded RNA viruses. *Semin. Virol.* 3, 327–339

63 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402

64 Koonin, E.V. and Dolja, V.V. (1993) Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences. *Crit. Rev. Biochem. Mol. Biol.* 28, 375–430

65 Gorbalenya, A.E. *et al.* (2006) Nidovirales: evolving the largest RNA virus genome. *Virus Res.* 117, 17–37

66 Rastgou, M. *et al.* (2009) Molecular characterization of the plant virus genus *Ourmiavirus* and evidence of inter-kingdom reassortment of viral genome segments as its possible route of origin. *J. Gen. Virol.* 90, 2525–2535

67 Besemer, J. *et al.* (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.* 29, 2607–2618