The ISME Journal (2011) 5, 165–168

www.nature.com/ismej

COMMENTARY

Unraveling the viral tapestry (from inside the capsid out)

Shawn W Polson, Steven W Wilhelm and K Eric Wommack

The ISME Journal (2011) **5,** 165–168; doi:10.1038/ismej.2010.81; published online 17 June 2010

Introduction

The field of viral ecology has long endeavored to devise and adapt methodologies to peer beyond the visible and elucidate the roles of viruses in the environment. Much has been learned regarding the dynamics of viral assemblages and the significant role viruses have in biogeochemical cycles (Brussaard *et al.*, 2008). Despite these advances, detailed understanding of biological processes behind ecosystem-scale effects of viral infection has remained largely obscured, with research relegated to a handful of available host–virus culture systems. Increasingly affordable DNA sequencing has provided a route to assess the genetic diversity of viruses in the environment. Current research seeks to apply genomic technologies to address knowledge gaps in environmental virology, but obstacles presented by the unique biology of viruses must be addressed to understand the context and significance of viral genome and metagenome sequence data.

Coincident with the twentieth anniversary of the publication that launched the field (Bergh *et al.*, 1989), viral ecologists from around the world met in 2009 for a workshop of the Scientific Committee for Oceanographic Research (SCOR) Working Group on the Role of Viruses in Marine Ecosystems (http://scor-viral-ecology.dbi.udel.edu) and at a session entitled 'From Direct Counts to Metagenomics: Two Decades of Discovery in Aquatic Viral Ecology' at the 109th General Meeting of the American Society for Microbiology (ASM; http://www.asm.org). These meetings covered a broad range of topics relevant to environmental virology, however, the impact of metagenomics emerged as a major topic. Highlighted are important issues for viral metagenomics raised during a round-table discussion (SCOR) and through various abstracts presented at both forums.

Viral metagenomics: how did we get here?

Viral research formed the foundation of genomic biology, with the first whole genome sequence (WGS) being that of bacteriophage MS2 (Fiers *et al.*, 1976). During the subsequent 20 years numerous viral genomes were sequenced, laying groundwork for the advent of organismal genomics and an unfortunate, but short, decline in viral genomics. Bacterial genomics actually resulted in the production of ample phage sequence through the common, yet unintended, sampling of integrated prophage genomes within bacterial WGS. Putative prophage regions within bacterial WGS typically comprise genomic segments containing a high proportion of genes showing little or no homology to known sequences—a harbinger to the vast pool of unknown genes seen in today's viral metagenomic investigations. The past decade has seen a renaissance in viral genomics with the traditional paradigm of sequencing genomes of currently known viruses giving way to host-based viral genomics, where a host's viruses are isolated for the defined purpose of genome sequencing. This approach, exemplified by the recent advances in mycobacteriophage genomics (Hatfull *et al.*, 2008), provides a wealth of knowledge about the viral metaproteome and has led to practical applications, such as the mycobacterial recombineering system allowing genetic manipulation of these difficult to transform organisms (GF Hatfull, SCOR meeting website).

To date, reliance on cultivation prior to obtaining viral WGS has limited our view of viral genetic diversity; however, single viral particle sequencing may ultimately alleviate this limitation. A further complication is the lack of a universally shared genetic marker among viruses, akin to SSU rRNA gene in cellular organisms, upon which to base phylogenetic studies of viral diversity. The advent of metagenomic analysis, taken with its caveats, has finally provided a means to explore the enormous genetic potential within natural viral assemblages (Figure 1).

Considerations

The application of metagenomics to viruses has not been a straightforward process (Figure 2). Technical issues have forced decisions, potentially creating biases downstream. Investigators have formulated multiple strategies for addressing these issues, thus comparison of metagenomes across studies must take into account potential artifacts related to sample preparation.

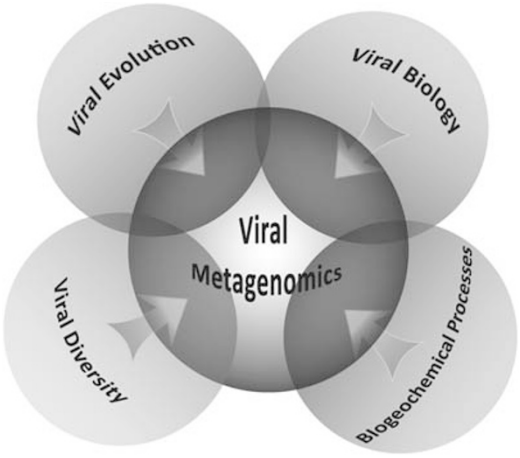

Figure 1 The promise of metagenomics for viral ecology. Owing to the lack of a universally shared genetic marker, connections between current understanding of viral evolution and viral diversity are poorly characterized and based largely on a subset of hallmark genes found across a few groups of viruses. Similarly, the small collection of cultivated strains and whole viral genome sequences means that there are few connections between viral diversity, the biology of viruses and the biogeochemical processes viral activities may impact. In contrast, such connections are substantially stronger for microorganisms owing to the larger collection of cultivated strains and whole genome sequences and the universally conserved SSU rRNA gene. Although metagenomic analyses of microbial communities will continue to strengthen these connections, metagenomics holds even greater promise for fundamental discoveries on the global diversity and evolutionary history of viruses, as well as predominant themes within viral biology and the impact of viral processes on global biogeochemical processes.

A key issue has been the significantly smaller amount of nucleic acid carried within a virus as compared with even the smallest prokaryotic cells. Even the larger genomes of some phycodnaviruses remain orders of magnitude smaller than those of their microalgal hosts. Therefore, preparation of samples for viral metagenomic investigations typically requires a combination of large samples and amplification of viral nucleic acids. Large sampling volumes have driven the need for expensive filtration regimes, such as tangential flow filtration, to concentrate virus particles into a workable volume. Even with large sample volumes, it is almost always necessary to perform some type of nucleic acid amplification. Amplification strategies have fallen into two types: linker/adapter (LA) and multiple displacement amplifications (MDA). Both strategies have been successfully applied to construct metagenomes, but each has potential drawbacks. Linker amplification is time consuming and requires relatively high sample concentration. MDA allows smaller samples to be analyzed, but questions have been raised regarding potential biases and artifacts (for example, preferential amplification of circular ssDNA and the generation of chimeric sequences) resulting from the application of MDA to mixed populations.

Although dsDNA viruses appear dominant in most environments, the current knowledgebase is deplete regarding the diversity/abundance of ssDNA and RNA viruses. To date, a single aquatic RNA virus metagenome exists in the literature (Culley *et al.*, 2006). Even less is known about ssDNA viruses in the oceans, where the sole metagenomic assessments have taken advantage the bias of MDA toward circular ssDNA amplification to bioinformatically-mine likely ssDNA virus sequences from metagenomes (K Rossario, SCOR meeting website; Angly *et al.*, 2006). SJ Williamson (SCOR meeting website) presented a method for co-purification of the dsDNA, ssDNA and RNA fractions from viral concentrates—a promising route for simultaneous investigation of viral diversity within these three genomic domains.

Bioinformatic analysis of metagenome sequence data also presents numerous virus-specific challenges. Most stem from the poor knowledgebase of viral proteins. Even among long-sequenced genomes, numerous gene products remain functionally obscure. This is compounded in environmental samples where novelty abounds. Often the most abundant predicted open reading frames (ORFs) in viral metagenomes have no homologs in sequence databases. In contrast, well-known genes that have been used in phylogenetic investigations of viral diversity, such as T4 major capsid protein, T7-like DNA polymerase or terminase can be relatively rare (SW Polson, abstracts 109th ASM meeting; S Jamindar and KE Wommack, SCOR meeting website).

Even among known viral genes, methods for functional assignment are often missing. Terms for common viral proteins are largely absent from the Gene Ontology terms, SEED subsystems and other databases used for annotation of microbial genomes. Existing metagenome annotation pipelines, such as MG-RAST (Meyer *et al.*, 2008), rely on these microbial-centric databases, unavoidably missing viral genes with known functions. Viral genes are often quite divergent from cellular homologs, causing additional annotations to be overlooked due to reliance on similarity thresholds defined for gene discovery within cellular organisms. Nevertheless, the viral-centric ACLAME database (A CLAssification of Mobile genetic Elements) (Leplae, 2004) and the Phage Proteomic Tree (Rohwer and Edwards, 2002) have proven to be extremely valuable in the analysis of viral metagenome data. On-going software development efforts such as the Viral Informatics Resource for Metagenome Exploration (http://virome.dbi.udel.edu) and the Phage Annotation Tools and Methods (http://www.phantome.org) should narrow the gap for extracting meaningful biological information from viral metagenomes.

Assessment of true viral diversity has been another difficult point. At the SCOR meeting, Mya Breitbart pointed out that each new viral metagenome presents a large number of novel genes hinting at a huge unknown diversity, however direct comparison of metagenome sequences from disparate locations often indicates significant overlap. We

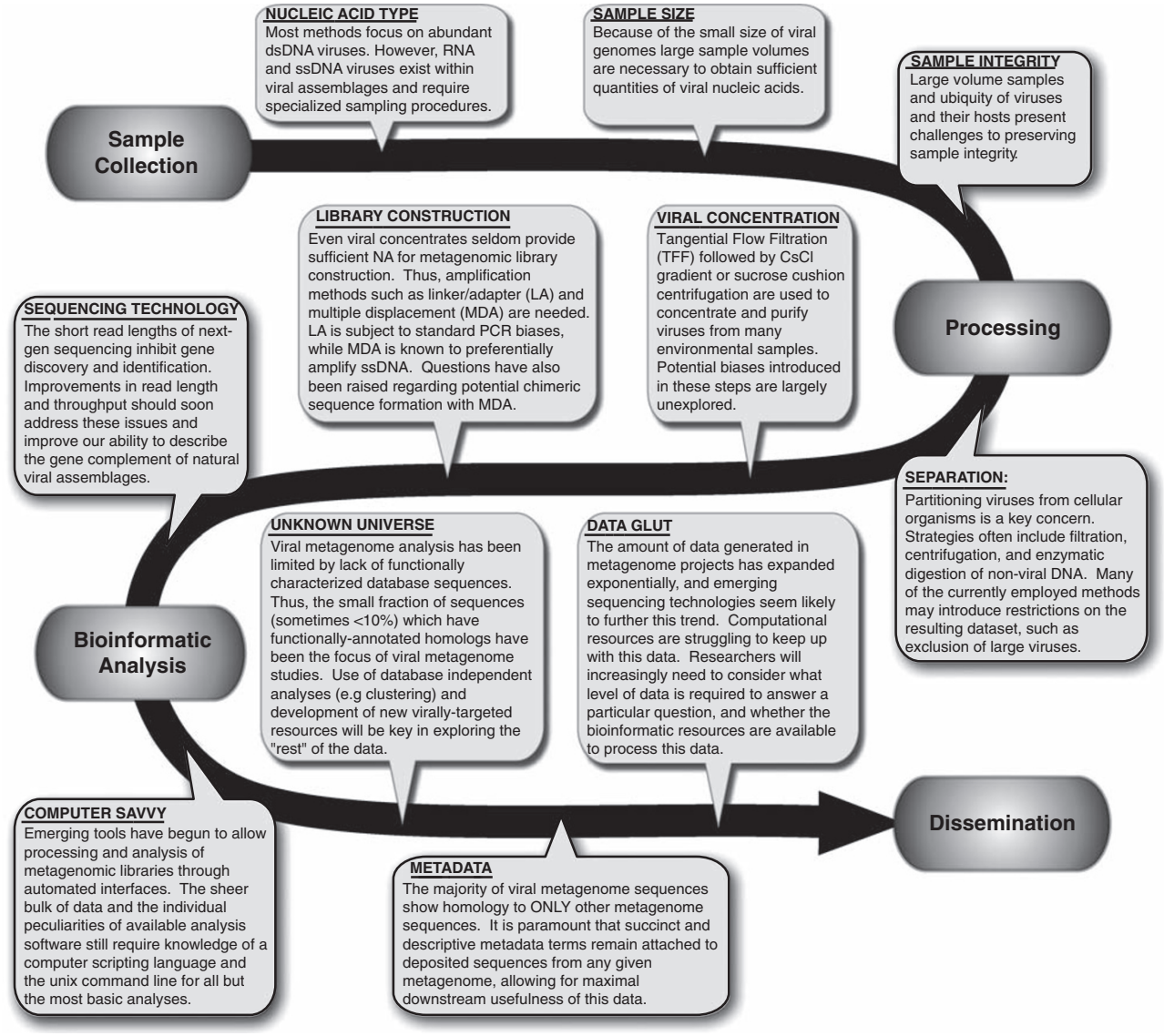

Figure 2 Issues, strategies and biases in viral metagenomics. The biology of viruses imposes numerous unique considerations during the process of planning, constructing and analyzing a viral metagenome. Detailed are various points of consideration during a viral metagenome workflow.

are presented with the ironic dichotomy of knowing little about viral genes, but seeing the same genes everywhere. The lack of a universal phylogenetic marker, poorly defined viral taxonomy, and the high likelihood of gene transfer events means that assessment of viral diversity remains a challenging issue. Moreover, the presence of unique sequence variants (the rare biosphere) raise questions concerning the functionality of genes that are rare: are these genes associated with functional viruses of low fitness, or are they evolutionary dead ends within the larger viral genomic pool?

Going forward

The Broad Institute has been charged with sequencing 200 viral genomes and 50 viral metagenomes as part of the Marine Microbiology Initiative of the Gordon and Betty Moore Foundation. This flood of sequence data should serve as a catalyst for novel viral ecology research. New approaches are also on the horizon, such as the prospect of single virus sequencing. At the SCOR and ASM meetings, L.A. Zeigler and S.J. Williamson reported successful isolation and genomic amplification of average sized phage genomes from single viral particles isolated by flow cytometry, while W.H. Wilson demonstrated application of a similar cytometry and amplification approach to obtain sequence information from large genome algal virus particles. Future development of approaches for sequencing single virus particles, akin to single-cell genomics, will provide an exquisite complement to metagenomic sequencing by providing oft-missing genomic context for common environmental ORFs.

Metagenomics has revolutionized the study of viral assemblages. Developing research is certain to provide a wealth of new data to the field in the coming years. However, it is important to remember that metagenomic data is not a replacement, but rather a starting point for future ecological studies. This wealth of gene sequence data will empower researchers to take new insights back to the bench (or field) and convert insight into new understanding. Genomic biology is only as good as the knowledge underpinning its databases, and a thorough understanding of basic viral biology is requisite to move forward. As viral biologists it is very easy to set our gaze tightly on the sub-micron realm; however, it is vital that as we explore the viral genosphere we continually strive to translate this data into real insights regarding virus-host interactions, and ultimately toward the mechanistic basis of global biogeochemical cycles.

SW Polson is at Department of Plant and Soil Sciences, University of Delaware, Newark, DE, USA and Delaware Biotechnology Institute, University of Delaware, Newark, DE, USA; SW Wilhelm is at Department of Microbiology, The University of Tennessee, Knoxville, TN, USA and KE Wommack is at Department of Plant and Soil Sciences, University of Delaware, Newark, DE, USA and Delaware Biotechnology Institute, University of Delaware, Newark, DE, USA.
E-mail: wommack@dbi.udel.edu

References

Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C *et al.* (2006). The marine viromes of four oceanic regions. *PLOS Biol* **4**: 2121–2131.

Bergh O, Borsheim KY, Bratbak G, Heldal M. (1989). High abundance of viruses found in aquatic environments. *Nature* **340**: 467–468.

Brussaard CPD, Wilhelm SW, Thingstad F, Weinbauer MG, Bratbak G, Heldal M *et al.* (2008). Global-scale processes with a nanoscale drive: the role of marine viruses. *ISME J* **2**: 575–578.

Culley A, Lang A, Suttle CA. (2006). Metagenomic analysis of coastal RNA virus communities. *Science* **312**: 1795–1798.

Fiers W, Contreras R, Duerinck F, Haegeman G, Iserentant D, Merregaert J *et al.* (1976). Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* **260**: 500–507.

Hatfull GF, Cresawn SG, Hendrix RW. (2008). Comparative genomics of the mycobacteriophages: insights into bacteriophage evolution. *Res Microbiol* **159**: 332–339.

Leplae R. (2004). ACLAME: A CLAssification of Mobile genetic Elements. *Nucleic Acids Res* **32**: 45D–449.

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M *et al.* (2008). The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**: 386.

Rohwer F, Edwards R. (2002). The phage proteomic tree: a genome-based taxonomy for phage. *J Bacteriol* **184**: 4529–4535.