Available online at www.sciencedirect.com

SciVerse ScienceDirect

Current Opinion in Virology

The global virome: not as big as we thought?

J Cesar Ignacio-Espinoza[1], Sergei A Solonenko[2] and Matthew B Sullivan[1,2]

Viruses likely infect all organisms, serving to unknown extent as genetic vectors in complex networks of organisms. Environmental virologists have revealed that these abundant nanoscale entities are global players with critical roles in every ecosystem investigated. Curiously, novel genes dominate viral genomes and metagenomes, which has led to the suggestion that viruses represent the largest reservoir of unexplored genetic material on Earth with literature estimates, extrapolating from 14 mycobacteriophage genomes, suggesting that two *billion* phage-encoded ORFs remain to be discovered. Here we examine (meta)genomic data available in the decade since this provocative assertion, and use 'protein clusters' to evaluate whether sampling technologies have advanced to the point that we may be able to sample 'all' of viral diversity in nature.

Addresses
[1] Department of Molecular and Cellular Biology, University of Arizona, Tucson, AZ, USA
[2] Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA

Corresponding author: Sullivan, Matthew B (mbsulli@email.arizona.edu)

Current Opinion in Virology 2013, **3**:566–571

This review comes from a themed issue on **Virus evolution**

Edited by **Valerian V Dolja** and **Mart Krupovic**

For a complete overview see the Issue and the Editorial

Available online 26th July 2013

1879-6257/$ – see front matter,

http://dx.doi.org/10.1016/j.coviro.2013.07.004

Viral abundances and the Virus–Host Ratio: Viral-like particles from stain-based counts are routinely ~10-fold more abundant than microbes which suggests there are about 10^{31} viruses on Earth [1]. These particles are of great importance ecologically as they impact microbes, which fuel planet Earth [2], through mortality, horizontal gene transfer and modulating global biogeochemical cycles [3]. Evolutionarily, intriguing hypotheses now exist suggesting that some RNA viral groups may predate eukaryotic supergroups [4], while others may represent a fourth domain of life [5] and still others have contributed to the evolutionary trajectory of global photosystems [6]. Pragmatically, viruses in nature are incredibly hard to study with even basic, stain-based counts first, including 'fake viral particles' [7] and missing some viral types [8] that may be a substantive fraction of those in nature [9,10], and second, not linking viruses to their hosts (but see emerging methods [11–13]).

In spite of these 'count' challenges, host range assays and emerging theory also support the conclusion that multiple viruses infect any given bacterium [14], and genomics is beginning to delineate how different viruses infecting a particular bacterium might be. For example, one study [15••] showed that at least 12 viral *genera* infect *Cellulophaga baltica* strains and re-evaluates available genomes to show that this is also the case for non-marine heterotrophic hosts (*Escherichia coli*, *Pseudomonas aeruginosa*, *Staphylococcus aureus*, *Mycobacterium tuberculosis*), but not for marine cyanobacteria where only three cyanophage genera infect marine *Prochlorococcus* and *Synechococcus*. If more than one virus infecting a bacterium is constant throughout the microbial realm, and if there is not significant overlap of viruses from one host to another (an open question), then it is easy to suggest that viral genetic diversity should be high — perhaps balanced by smaller genome sizes to be on par with that of their bacterial hosts.

Again, problematically, estimating viral diversity is complicated due to the lack of a universal genetic marker. Estimates from metagenomes using assembly-based methods suggest that there are between 532 (Arctic Ocean) to 129 000 (British Columbia) viral types in 10–150 L of seawater [16]. These highly variable estimates makes global extrapolations challenging, but the underlying data also have issues (reviewed in Ref. [17]) that could drastically impact an estimate of viral richness that relates back to a single time point and sample.

Regardless, viruses are clearly abundant and their high rates of mutation (though little data for environmentally relevant viruses exist) leave us expecting that new genomic technologies will reveal extensive sequence diversity as the global virome is sampled. Such metagenomic data should greatly impact the ever-widening scope of viral evolutionary studies. As a discovery tool, metagenome-derived sequences can be phylogenetically analyzed using core gene sets, analogous to 16S analysis, to document novel clades in well-studied lineages [18]. Additionally, metagenomics can recover novel, full virus genome sequences [19–21] which provide opportunities to test spatio-temporal evolutionary theories by sampling naturally occurring communities over space and time [22].

Rampant mosaicism and the homogeneity of viral genomes: To counter this expectation of extensive sequence diversity, horizontal gene transfer (HGT) may homogenize viral diversity. In fact, *Caudovirales* evolutionary theory suggests that rampant mosaicism blurs taxonomic boundaries, a hypothesis predominantly derived from

observations in siphovirus genomes [23]. Broadly, phage genomes are thought to evolve through accumulations of HGT events involving transcriptionally autonomous genetic units (called 'morons') that introduce variation and only remain if offering a fitness advantage. Beyond 'morons' other mobile elements impact phage genomes ranging from promoter stem loops (PeSLs, transcriptionally autonomous with consensus-flanking sequences [24]) to more traditionally characterized elements such as homing endonucleases [25], introns [26], and transposons [27], and there are likely more to be discovered. Additionally, identical phage gene sequences occur globally and phage infection can cross biomes; this has led to the hypothesis that phages have high dispersal rates and access to a global gene pool [28].

With so many mechanisms seeking to homogenize genomes, it is hard to imagine that any vertical evolutionary signal remains in phage genomes. However, HGT has not blurred all evolutionary phage lineages. Again, among *Caudovirales*, T4-like phage genomes [29,30] and metagenomic contigs [31•] are predominantly syntenic with much variation captured in niche-defining hypervariable islands similar to those observed in microbes [32,33]. Further, phylogenomic approaches show that most (~80%) T4-like phage core genes are vertically inherited [34•], suggesting that at least some genes resist HGT. Outside T4-like phages, though not as formally tested, similar vertically inherited core gene sets are emerging for T7-like cyanophages [35], and seem probable for the nearly invariant, small ssDNA [19] and RNA [20] viral genomes assembled from metagenomes.

Finally, since delineating an ecologically and evolutionarily meaningful unit to 'count' is fundamental for studies in nature, virology could learn from parallel research efforts in microbes. Specifically, there has been significant interest in the role of HGT and whether it blurs microbial species boundaries. The current view is that ecological and genetic species can be defined (reviewed in Ref. [36••]) and arise when new alleles (mutations) or genes (HGT) sweep through a population in conjunction with ecological differentiation, where subsequent HGT events would help to maintain cohesion rather than disperse it [37]. Notably, there are many parallel and competing microbial species concepts, but such careful and meticulous empirical and theoretical work on microbes in nature is constrained predominantly to study of an r-selected, copiotroph (*Vibrio* sp.). Thus application of these evolutionary principles may not be straightforward for other microbes featuring different lifestyles. Similarly, variable lifestyles were recently observed in marine phages [12••], with evolutionary implications for rates of HGT being more important when comparing across viruses that span temperate and lytic lifestyles (due to increased accessibility of prophage genomic DNA).

The protein cluster as an organizational tool to explore viral sequence space

Viral metagenomes are dominated by the 'unknown', so a critical advance has been to organize unknown viral sequence space into protein clusters (PCs, [38••,39,40]), again taking the nod from our microbiology colleagues [41]. The PCs approach is valuable as a universal metric for comparing diversity of viral assemblages, and as a community resource will provide each first, a 'handle' to apply OTU-based ecological theory, independent of known function, using new and expanding community tools (e.g., QIIME, http://qiime.org/), and second, an informatics structure to allow propagation of functional data to novel 'unknown' proteins. Finally, PCs serve as a metric to estimate how well viral sequence space has been sampled given the genomic and metagenomic data that have been accumulated in the decade since Rohwer's estimate of two billion viral-encoded proteins [42•].

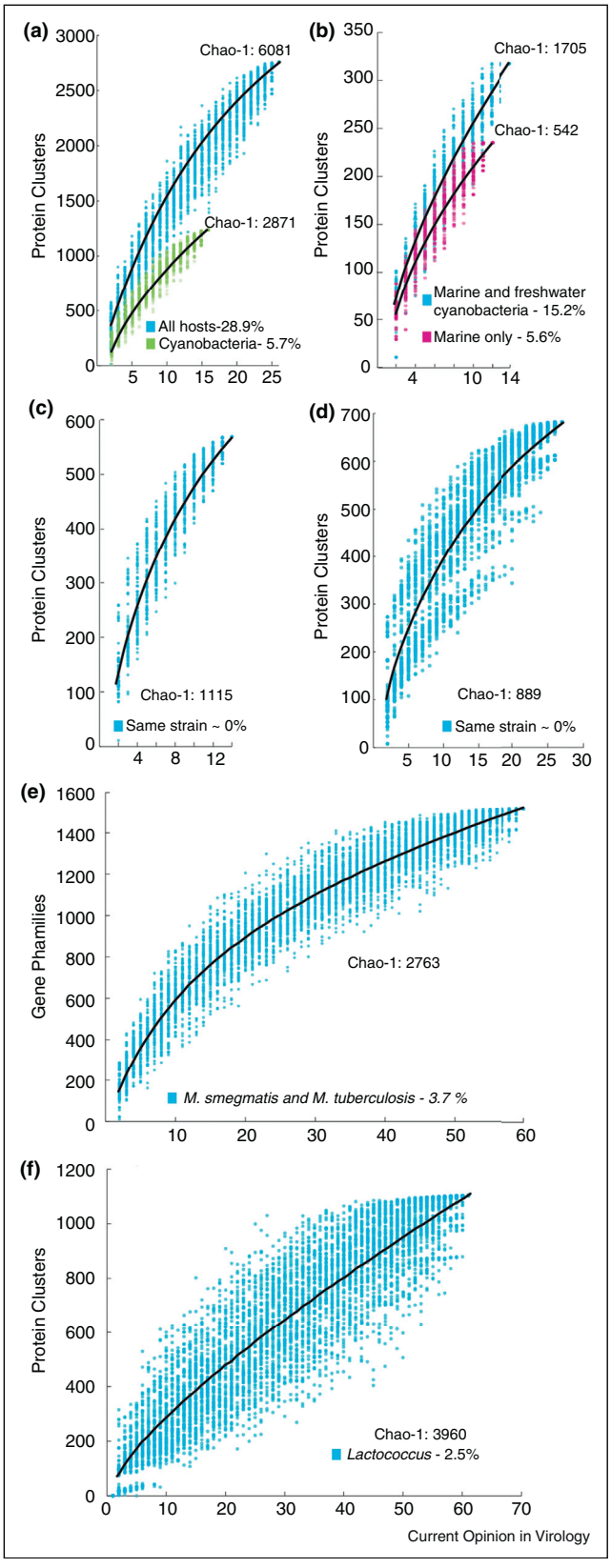
An alternative organization scheme to PCs is Phage Orthologous Groups [43]. POGs are grounded in the evolutionary concept of the conservation of ortholog function (as opposed to the relative evolutionary freedom of duplicated paralogs), and leverage phylogeny and heuristics to define sets of genes thought to share function across different organisms. These heuristics (e.g., reciprocal best blast hit) are required for scaling analyses to modern datasets and are relatively accurate compared to more sophisticated phylogenetic approaches [43]. However, the assumption that orthologs are more likely to share function than paralogs should be considered for each new research question [44].

Pragmatically, PCs and POGs are complementary approaches, with PCs perhaps most valuable for monitoring the expansion of sequence space and POGs best utilized to document phage-specific functional space expansions [45] that might be used to test 'phageness' of metagenomes — thus providing a new dimension in viral metagenomic analyses. Looking forward, as computational approaches scale, protein structure is a valuable trait in viral evolution and taxonomy [46] that will enable detection of functional similarity across evolutionarily distant lineages using information beyond the primary sequence used by PCs and POGs.

Using phage genomes to estimate global viral sequence diversity

Given that the most abundant hosts available in nature are bacteria, it is thought that the most abundant viruses infect bacteria-termed phages. Deeply genome-sequenced phage groups include the cyanophages (T4-likes [29], T7-likes [35], siphoviruses [47]), the mycobacteriophages (siphoviruses and myoviruses [48]), phages that infect *Staphylococcus* [49], *Pseudomonas* [50] and *Lactococcus* [51]. Here, by sequentially evaluating each

Figure 1



new gene in a genome against the genes already observed, we calculate 'collectors curves' or 'rarefaction curves' of PCs to document how well sampled the 'flexible' or 'pan' genome is in each phage group.

These analyses revealed a pan genome estimate that ranged from a few hundred to a few thousand PCs per phage–host system examined (Figure 1). While highly controversial, the number of microbial species is thought to be ~6 million globally [52], and we calculate here that the viral pan genome associated with any particular group of bacteria is a few hundred to a few thousand PCs. Assuming that viruses do not infect or share genes across host species, then the global virome should be on the order of 0.6–6 billion protein clusters — very similar to Rohwer's decade-old estimate of two billion proteins [42•]. A violation in either of those assumptions of host range and shared genes would reduce this total.

Using metagenomes to estimate global viral sequence diversity

Metagenomics is often used to survey nature beyond that can be observed through culture-based observations. The Pacific Ocean Virome (POV) dataset represents the largest viral metagenomic survey to date and is prepared in a quantitative manner using a highly efficient iron-chloride concentration method [53], as well as purification steps [40] and linker amplified library construction [54] that have well understood biases. The POV dataset is comprised 32 metagenomes obtained from four regions in the Pacific Ocean that represent a relatively wide range of ecological characteristics from the coastal-to-open and surface-to-deep oceans [38••]. Its value for ecological study is already apparent having helped enumerate the most abundant ocean viruses observed to date (pelagiphages [55]).

Notably, the initial description of the POV dataset (Figures 3 and 4 in [38••]) shows a non-linear relationship between sampling effort and PC accumulation, which suggests that sampling is approaching saturation in the Pacific Ocean pelagic viral sequence space (Figure 2). In fact, ~422 K PCs were observed in the POV dataset [38••], with extrapolation suggesting a total of 516 K to

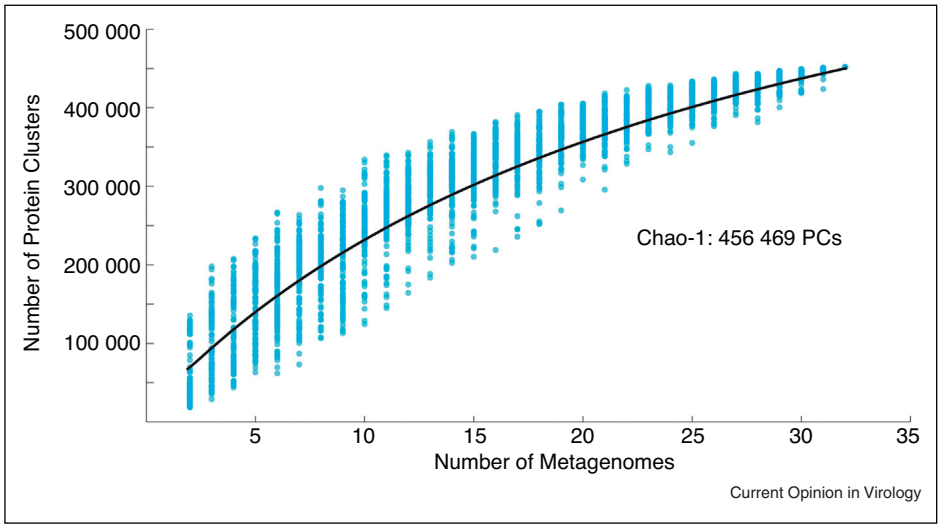
(Figure 1 Legend) Viral pan-genomes seem an endless source of novel genes. Genomic comparisons of A.T4-like Myoviruses (26 genomes), B. T7-like cyanopodoviruses (14 genomes), C. *Pseudomonas* phages (14 genomes), D. *Staphylococcus* phages (27 genomes), E. *Mycobacterium* phages (60 genomes) and F, *Lactococcus* phages (61 genomes). Each dot corresponds to the number of protein clusters when *k* members are sampled (from $k = 2$ to n). Only 1000 randomizations of the process are shown. Trend lines correspond to the average of 10 000 repetitions. Chao-1 is calculated as: observed + $[(\text{Singletons})^2]/(2 \times \text{Doubles})$. Microbial % in each panel corresponds to the 16S gene divergence of the host strains used for isolating the phages. The multiple trajectories in panel D reflect the fact that multiple phage groups are pooled in these analyses.

Table 1

Richness estimation from metagenomics. Four functions [58] were used to fit the accumulation curve observed in Figure 2

	Function	POV metagenome	
		Asymptote (95% CI)	*R*-Squared
Michaelis–Menten	$Y = (a \times x)/(1 + (b \times x))$	1 313 853 (963 153–3 232 913)	0.8181
Negative exponential	$Y = a \times (1 - \exp((-b) \times x)$	516 600 (508 400–524 700)	0.999
Rational	$Y = (a \times (b \times x))/(1 + (c \times x))$	691 957 (678 691–710 620)	0.9981
Hyperbolic	$Y = (a \times x)/(b + x)$	737 200 (680 400–793 900)	0.9894

Figure 2



Exploration of viral protein sequence space by means of PCs suggests near saturation. Protein cluster rarefaction curve of the number of protein clusters that metagenomic comparison totals calculated as described in Figure 1.

1.3 M PCs depending upon the model used for fitting the accumulation curve (Table 1). This suggests that between 32 and 82% of the possible pelagic Pacific Ocean virus proteins have been sampled. Further, the oceans are thought to harbor about 33% of the total microbial species on Earth [52]. Assuming the same is true for viruses then this suggests that there are likely to be no more than 3.9 M PCs (the 1.3 M maximum × 3-fold more species) in the global virome — or nearly three orders of magnitude less than the two billion previously estimated [42•].

Conclusions

There undoubtedly remains much diversity to be discovered in the world of viruses. However, the large discrepancy between genome-derived and metagenome-observed PC diversity requires consideration of its underlying causes. First, ascertainment bias is high as only a fraction of the phage–host systems and environments that occur in nature are sequenced. Even in the relatively well-sampled oceans, pelagic Pacific Ocean waters do not represent the diversity of possible ocean environments. As well, recent studies show that firstly, cultures do not represent the dominant morphotypes in the ocean [9], and secondly, roughly half of the viruses in the oceans (RNA viruses, [10]) and the larger-genome giant viruses [56] would not be captured using the methods that generated the POV dataset. Notably, RNA viruses are abundant, but their genomes appear small and of low diversity; in contrast, giant viruses are much less abundant, but their genomes are large and quite diverse. With so little data on these viral groups their contribution to the global virome is a big unknown. Second, the Chao1 non-parametric richness estimator is flawed when the shape of the 'rare tail' in the rank abundance curve is not well described — so much so that it is not reliable for comparisons across samples [57]. This is likely a significant problem for the genomic data, but less so for the metagenomic data examined here. Third, metagenomic PC observations are reliant upon open reading frame predictions, which are not very good in viruses (smaller ORFs, overlapping ORFs, smaller genome sizes limit algorithm 'training') and could lead to underestimations. These are unlikely to be order of magnitude underestimations but may be a factor. Fourth,

viral genomic content and host ranges may be more commonly shared across hosts than is currently recognized. This would decrease global virome estimates by an unknown factor, and new methods should enable culture-independent [11,13] and higher-throughput [12••] linkages to be made between viruses and hosts. Obtaining a better estimate of the global virome will clearly require filling knowledge gaps across all of these factors. However, it seems likely that viral sequence space, while large, is unlikely to approach the two billion genes estimated from 14 genomes a decade ago.

Acknowledgements

We thank Christine Schirmer and Ann Gregory for assistance with manuscript preparation and critical comments on the manuscript, as well as UITS Research Computing Group and the Arizona Research Laboratories Biotech Computing for high-performance computing access and support. Funding was provided by National Science Foundation (OCE-0961947), Biosphere 2, BIO5 and Gordon and Betty Moore Foundation grants to MBS, a Fulbright Scholarship to JCIE, and an NSF IGERT Comparative Genomics Training Grant award to SAS.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Suttle CA: **Viruses in the sea**. *Nature* 2005, **437**:356-361.
2. Falkowski PG, Fenchel T, DeLong EF: **The microbial engines that drive earth's biogeochemical cycles**. *Science* 2008, **320**:1034-1039.
3. Suttle CA: **Marine viruses — major players in the global ecosystem**. *Nat Rev Microbiol* 2007, **5**:801-812.
4. Koonin EV, Wolf YI, Nagasaki K, Dolja VV: **The Big Bang of picorna-like virus evolution antedates the radiation of eukaryotic supergroups**. *Nat Rev Microbiol* 2008, **6**:925-939.
5. Legendre M, Arslan D, Abergel C, Claverie J-M: **Genomics of Megavirus and the elusive fourth domain of Life**. *Commun Integr Biol* 2012, **5**:102-106.
6. Sullivan MB, Lindell D, Lee JA, Thompson LR, Bielawski JP, Chisholm SW: **Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts**. *PLoS Biol* 2006, **4**:e234.
7. Forterre P, Soler N, Krupovic M, Marguet E, Ackermann H-W: **Fake virus particles generated by fluorescence microscopy**. *Trends Microbiol* 2013, **21**:1-5.
8. Holmfeldt K, Odić D, Sullivan MB, Middelboe M, Riemann L: **Cultivated single-stranded DNA phages that infect marine Bacteroidetes prove difficult to detect with DNA-binding stains**. *Appl Environ Microbiol* 2012, **78**:892-894.
9. Brum JR, Schenck RO, Sullivan MB: **Global morphological analysis of marine viruses shows minimal regional variation and dominance of non-tailed viruses**. *ISME J* 2013 http://dx.doi.org/10.1038/ismej.2013.67.
10. Steward GF, Culley AI, Mueller JA, Wood-Charlson EM, Belcaid M, Poisson G: **Are we missing half of the viruses in the ocean?** *ISME J* 2013, **7**:672-679.
11. Allers E, Moraru C, Duhaime MB, Beneze E, Solonenko N, Barrero-Canosa J, Amann R, Sullivan MB: **Single-cell and population level viral infection dynamics revealed by phageFISH, a method to visualize intracellular and free viruses**. *Environ Microbiol* 2013 http://dx.doi.org/10.1111/1462-2920.12100.
12. •• Deng L, Gregory A, Yilmaz S, Poulos BT, Hugenholtz P, Sullivan MB: **Contrasting life strategies of viruses that infect photo- and heterotrophic bacteria, as revealed by viral tagging**. *MBio* 2012, **3**:e00373-12.

Environmental genomics has been largely centered on isolations. The authors present a high-throughput method that halves the need for isolations in genomic characterizations.

13. Tadmor AD, Ottesen EA, Leadbetter JR, Phillips R: **Probing individual environmental bacteria for viruses by using microfluidic digital PCR**. *Science* 2011, **333**:58-62.
14. Weitz JS, Poisot T, Meyer JR, Flores CO, Valverde S, Sullivan MB, Hochberg ME: **Phage–bacteria infection networks**. *Trends Microbiol* 2013, **21**:82-91.
15. •• Holmfeldt K, Solonenko N, Shah M, Corrier K, Riemann L, VerBerkmoes NC, Sullivan MB: **Twelve previously unknown phage genera are ubiquitous in global oceans**. *Proc Natl Acad Sci U S A* 2013 http://dx.doi.org/10.1073/pnas.1305956110. (in press).

Environmental genomics has centered on few phage host systems. These authors introduce a new and diverse ubiquitous phage–bacteria system in the pelagic ocean, as well as broach difficult questions about phage genomic boundaries in a taxonomic framework.

16. Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, Chan AM, Haynes M, Kelley S, Liu H: **The marine viromes of four oceanic regions**. *PLoS Biol* 2006, **4**:e368.
17. Duhaime MB, Sullivan MB: **Ocean viruses: rigorously evaluating the metagenomic sample-to-sequence pipeline**. *Virology* 2012, **434**:181-186.
18. Roux S, Faubladier M, Mahul A, Paulhe N, Bernard A, Debroas D, Enault F: **Metavir: a web server dedicated to virome analysis**. *Bioinformatics* 2011, **27**:3074-3075.
19. Roux S, Krupovic M, Poulet A, Debroas D, Enault F: **Evolution and diversity of the Microviridae viral family through a collection of 81 new complete genomes assembled from virome reads**. *PLoS ONE* 2012, **7**:e40418.
20. Culley AI, Lang AS, Suttle CA: **Metagenomic analysis of coastal RNA virus communities**. *Science* 2006, **312**:1795-1798.
21. Wu Q, Luo Y, Lu R, Lau N, Lai EC, Li W-X, Ding S-W: **Virus discovery by deep sequencing and assembly of virus-derived small silencing RNAs**. *Proc Natl Acad Sci U S A* 2010, **107**:1606-1611.
22. Emerson JB, Thomas BC, Andrade K, Allen EE, Heidelberg KB, Banfield JF: **Dynamic viral populations in hypersaline systems as revealed by metagenomic assembly**. *Appl Environ Microbiol* 2012, **78**:6309-6320.
23. Hendrix RW, Hatfull GF, Smith MCM: **Bacteriophages with tails: chasing their origins and evolution**. *Res Microbiol* 2003, **154**:253-257.
24. Arbiol C, Comeau AM, Kutateladze M, Adamia R, Krisch HM: **Mobile regulatory cassettes mediate modular shuffling in T4-type phage genomes**. *Genome Biol Evol* 2010, **2**:140-152.
25. Zeng Q, Bonocora RP, Shub DA: **A free-standing homing endonuclease targets an intron insertion site in the psbA gene of cyanophages**. *Curr Biol* 2009, **19**:218-222.
26. Landthaler M, Shub DA: **Unexpected abundance of self-splicing introns in the genome of bacteriophage Twort: introns in multiple genes, a single gene with three introns, and exon skipping by group I ribozymes**. *Proc Natl Acad Sci U S A* 1999, **96**:7005-7010.
27. Sullivan MB, Krastins B, Hughes JL, Kelly L, Chase M, Sarracino D, Chisholm SW: **The genome and structural proteome of an ocean siphovirus: a new window into the cyanobacterial 'mobilome'**. *Environ Microbiol* 2009, **11**:2935-2951.
28. Breitbart M, Rohwer F: **Here a virus, there a virus, everywhere the same virus?** *Trends Microbiol* 2005, **13**:278-284.
29. Sullivan MB, Huang KH, Ignacio-Espinoza JC, Berlin AM, Kelly L, Weigele PR, Defrancesco AS, Kern SE, Thompson LR, Young S *et al.*: **Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments**. *Environ Microbiol* 2010, **12**:3035-3056.

30. Millard AD, Zwirglmaier K, Downey MJ, Mann NH, Scanlan DJ: **Comparative genomics of marine cyanomyoviruses reveals the widespread occurrence of *Synechococcus* host genes localized to a hyperplastic region: implications for mechanisms of cyanophage evolution**. *Environ Microbiol* 2009, **11**:2370-2387.

31. • Comeau AM, Arbiol C, Krisch HM: **Gene network visualization and quantitative synteny analysis of more than 300 marine T4-like phage scaffolds from the GOS metagenome**. *Mol Biol Evol* 2010, **27**:1935-1944.
Isolations of T4-like phages have continually highlighted the remarkable synteny of their genomes. The authors of this work extend these observations to metagenomes noting that it is a hallmark of the T4-like family.

32. Coleman ML: **Genomic islands and the ecology and evolution of Prochlorococcus**. *Science* 2006, **311**:1768-1770.

33. Wilhelm LJ, Tripp HJ, Givan SA, Smith DP, Giovannoni SJ: **Natural variation in SAR11 marine bacterioplankton genomes inferred from metagenomic data**. *Biol Direct* 2007, **2**:27.

34. • Ignacio-Espinoza JC, Sullivan MB: **Phylogenomics of T4 cyanophages: lateral gene transfer in the 'core' and origins of host genes**. *Environ Microbiol* 2012, **14**:2113-2126.
This study provides a rare quantitative look at how the evolution of single genes in a phage genome compares to that inferred from the species tree.

35. Labrie SJ, Frois-Moniz K, Osburne MS, Kelly L, Roggensack SE, Sullivan MB, Gearin G, Zeng Q, Fitzgerald M, Henn MR *et al.*: **Genomes of marine cyanopodoviruses reveal multiple origins of diversity**. *Environ Microbiol* 2013, **15**:1356-1376.

36. •• Polz MF, Alm EJ, Hanage WP: **Horizontal gene transfer and the evolution of bacterial and archaeal population structure**. *Trends Genet* 2013, **29**:170-175.
These authors take on a contentious and challenging subject (microbial speciation) to help signal emerge from the noise. This review examines diverse datasets to help bring a unified perspective to the topic.

37. Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabó G, Polz MF, Alm EJ: **Population genomics of early events in the ecological differentiation of bacteria**. *Science* 2012, **336**:48-51.

38. •• Hurwitz BL, Sullivan MB: **The Pacific Ocean Virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology**. *PLoS ONE* 2013, **8**:e57355.
This paper introduces the Pacific Ocean Virome (POV) — the largest viral-fraction metagenomic dataset, and one that is prepared with well-characterized steps leading to a quantitative dataset for follow-up ecological studies.

39. Roux S, Enault F, Robin A, Ravet V, Personnic S, Theil S, Colombet J, Sime-Ngando T, Debroas D: **Assessing the diversity and specificity of two freshwater viral communities through metagenomics**. *PLoS ONE* 2012, **7**:e33641.

40. Hurwitz BL, Deng L, Poulos BT, Sullivan MB: **Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics**. *Environ Microbiol* 2012 http://dx.doi.org/10.1111/j.1462-2920.2012.02836.x.

41. Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W *et al.*: **The Sorcerer II global ocean sampling expedition: expanding the universe of protein families**. *PLoS Biol* 2007, **5**:e16.

42. • Rohwer F: **Global phage diversity**. *Cell* 2003, **113**:141.
This author was bold enough to attempt to provide a rational starting place for estimating global phage diversity.

43. Kristensen DM, Kannan L, Coleman MK, Wolf YI, Sorokin A, Koonin EV, Mushegian A: **A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches**. *Bioinformatics* 2010, **26**:1481-1487.

44. Nehrt NL, Clark WT, Radivojac P, Hahn MW: **Testing the ortholog conjecture with comparative functional genomic data from mammals**. *PLoS Comput Biol* 2011, **7**:e1002073.

45. Kristensen DM, Cai X, Mushegian A: **Evolutionarily conserved orthologous families in phages are relatively rare in their prokaryotic hosts**. *J Bacteriol* 2011, **193**:1806-1814.

46. Abrescia NGA, Bamford DH, Grimes JM, Stuart DI: **Structure unifies the viral universe**. *Annu Rev Biochem* 2012, **81**:795-822.

47. Huang S, Wang K, Jiao N, Chen F: **Genome sequences of siphoviruses infecting marine Synechococcus unveil a diverse cyanophage group and extensive phage–host genetic exchanges**. *Environ Microbiol* 2012, **14**:540-558.

48. Hatfull GF, Jacobs-Sera D, Lawrence JG, Pope WH, Russell DA, Ko CC, Weber RJ, Patel MC, Germane KL, Edgar RH *et al.*: **Comparative genomic analysis of 60 Mycobacteriophage genomes: genome clustering, gene acquisition, and gene size**. *J Mol Biol* 2010, **397**:119-143.

49. Kwan T, Liu J, DuBow M, Gros P, Pelletier J: **The complete genomes and proteomes of 27 *Staphylococcus aureus* bacteriophages**. *Proc Natl Acad Sci U. S A* 2005, **102**:5174-5179.

50. Kwan T, Liu J, DuBow M, Gros P, Pelletier J: **Comparative genomic analysis of 18 *Pseudomonas aeruginosa* bacteriophages**. *J Bacteriol* 2006, **188**:1184-1187.

51. Castro-Nallar E, Chen H, Gladman S, Moore SC, Seemann T, Powell IB, Hillier A, Crandall KA, Chandry PS: **Population genomics and phylogeography of an Australian dairy factory derived lytic bacteriophage**. *Genome Biol Evol* 2012, **4**:382-393.

52. Curtis TP, Sloan WT, Scannell JW: **Estimating prokaryotic diversity and its limits**. *Proc Natl Acad Sci U S A* 2002, **99**:10494-10499.

53. John SG, Mendez CB, Deng L, Poulos B, Kauffman AKM, Kern S, Brum J, Polz MF, Boyle EA, Sullivan MB: **A simple and efficient method for concentration of ocean viruses by chemical flocculation**. *Environ Microbiol Rep* 2011, **3**:195-202.

54. Duhaime MB, Deng L, Poulos BT, Sullivan MB: **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* 2012, **14**:2526-2537.

55. Zhao Y, Temperton B, Thrash JC, Schwalbach MS, Vergin KL, Landry ZC, Ellisman M, Deerinck T, Sullivan MB, Giovannoni SJ: **Abundant SAR11 viruses in the ocean**. *Nature* 2013, **494**:357-360.

56. Van Etten JL, Lane LC, Dunigan DD: **DNA viruses: the really big ones (giruses)**. *Annu Rev Microbiol* 2010, **64**:83-99.

57. Haegeman B, Hamelin J, Moriarty J, Neal P, Dushoff J, Weitz JS: **Robust estimation of microbial diversity in theory and in practice**. *ISME J* 2013, **7**:1092-1101.

58. Mora C, Tittensor DP, Myers RA: **The completeness of taxonomic inventories for describing the global diversity and distribution of marine fishes**. *Proc Biol Sci* 2008, **275**:149-155.