doi:10.1038/nature13459

Viral tagging reveals discrete populations in *Synechococcus* viral genome sequence space

Li Deng[1]†*, J. Cesar Ignacio-Espinoza[2]*, Ann C. Gregory[1], Bonnie T. Poulos[1], Joshua S. Weitz[3,4], Philip Hugenholtz[5]
& Matthew B. Sullivan[1,2]

Microbes and their viruses drive myriad processes across ecosystems ranging from oceans and soils to bioreactors and humans[1–4]. Despite this importance, microbial diversity is only now being mapped at scales relevant to nature[5], while the viral diversity associated with any particular host remains little researched. Here we quantify host-associated viral diversity using viral-tagged metagenomics, which links viruses to specific host cells for high-throughput screening and sequencing. In a single experiment, we screened 10^7 Pacific Ocean viruses against a single strain of *Synechococcus* and found that naturally occurring cyanophage genome sequence space is statistically clustered into discrete populations. These population-based, host-linked viral ecological data suggest that, for this single host and seawater sample alone, there are at least 26 double-stranded DNA viral populations with estimated relative abundances ranging from 0.06 to 18.2%. These populations include previously cultivated cyanophage and new viral types missed by decades of isolate-based studies. Nucleotide identities of homologous genes mostly varied by less than 1% within populations, even in hypervariable genome regions, and by 42–71% between populations, which provides benchmarks for viral metagenomics and genome-based viral species definitions. Together these findings showcase a new approach to viral ecology that quantitatively links objectively defined environmental viral populations, and their genomes, to their hosts.

Decades-old microscopic observations revealed that viruses typically outnumber microbial cells approximately tenfold in marine systems[1], recasting them from environmentally insignificant to the most abundant biological entities on Earth. Viruses are now considered important in microbial mortality, horizontal gene transfer and global biogeochemistry[2,3], with recent recognition of vast cellular metabolic reprogramming[4] during infection. However, the enormous microbial and viral diversity in nature makes it challenging to clarify and quantify these roles, particularly as viral taxonomy remains largely based on morphology and properties of isolates. Although large-scale isolate-based sequencing studies are clarifying genomic parameters for viral taxonomy—for example, defining phage 'genus' boundaries[6–8]—they remain limited to cultivated viral groups that represent only a fraction of viruses in nature.

Here we explore genetic variation in an environmentally relevant cyanobacterial model system[5,9,10]—seawater cyanophages within a Pacific Ocean viral assemblage that infect a cultured cyanobacterial host. We do so by adapting viral tagging, a high-throughput means of linking viruses to a target host[11], for use in the field. In this method, DNA in environmental viruses is labelled non-specifically with a fluorescent dye, viruses are mixed with a 'bait host' pre-labelled with isotopically heavy DNA, and infected cells are collected by fluorescence-activated flow cytometry. Isotopically light viral DNA is then separated from heavy host DNA using a density gradient, and the infecting viral DNA is quantitatively amplified[12] to produce viral-tagged metagenomes. Beyond the identification of viral populations interacting with a particular host[13], the data shed light on lineage-specific viral ecology at scales not previously possible, enabling the development of population-based measurements and models of viral ecology and evolution.

To explore Pacific Ocean cyanophage diversity linked to the cyanobacterial host *Synechococcus* sp. WH7803, we applied a traditional culture-based approach complemented by metagenomic analysis of the double-stranded (ds)DNA viral community and viral-tagged community. Of 97 new isolates, 90 were myoviruses as inferred using a marker gene (Extended Data Fig. 1), which is consistent with previous isolates on this host (88% are myoviruses; Methods). Similarly, metagenomic analysis showed that viral tagging simplified the total viral community towards one dominated by myoviruses (Extended Data Fig. 2 and Supplementary Data 1). Viral tagging an artificial viral assemblage did not enrich for myoviruses (Methods), indicating that the *Synechococcus* WH7803–myovirus interactions are specific. Furthermore, these viruses are likely to infect, rather than just adsorb to, their host given previous and current experiments in which all tested cyanophage–host interactions led to infection when positively viral tagged (5 of 5 isolates tested previously[11], and 18 of 18 isolates tested here; Extended Data Table 1) .

Beyond the expected myoviruses, viral tagging also provided evidence (genomic data) for 42 new uncultured viruses specific to *Synechococcus* WH7803 (Extended Data Fig. 3 and Supplementary Data 1), including eight podoviruses (T7-like, phiKMV-like) and one siphovirus, as well as 33 partial genomes (contigs) that were ambiguous or lacked similarity to any known viral or bacterial genes, which may represent novel viruses (Methods). The screening of ~10^7 virus particles against *Synechococcus* WH7803 probably explains why such an unprecedented diversity of specific viruses were recovered for this single host despite two decades of isolation studies.

Viral-tagging-based screening of the bulk viral community improved assembly (average contig size increased from 1.2 kilobases (kb) to 5.5 kb) to produce three nearly complete genomes (*Candidatus* genomes; CG-01, CG-03 and CG-05; 197 kb, 185 kb and 108 kb, respectively, contigs containing 94–97% of 65 T4-like core genes; Table 1) and 164 viral contigs (Supplementary Data 1) that offer genomic context and enable host-specific discoveries. Auxiliary metabolic genes[14] previously observed in viral metagenomes can now be assigned to a discrete viral entity with an experimentally defined host. For example, membrane protein 'T17', antioxidant protein 'T768' and glycosyltransferase 'T1338' were assigned to T4-like cyanophages (Supplementary File 1). Conversely, the deep sampling of *Synechococcus* cyanophages did not identify any photosystem I (PSI) genes reported in putative cyanophage metagenomic fragments[15] but lacking in cyanophage genomes[9,10,16,17], suggesting that viral-encoded PSI genes are restricted to particular locations and/or hosts.

Insights from host-linked viral-tagging data also address a fundamental and persistent challenge in microbial ecology and evolution: how to define populations and thus quantify natural community diversity. Previous marine work suggests that genomes of co-existing viruses, infecting a

[1]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85719, USA. [2]Department of Molecular and Cellular Biology, University of Arizona, Tucson, Arizona 85719, USA. [3]School of Biology, Georgia Institute of Technology, Atlanta, Georgia 30332, USA. [4]School of Physics, Georgia Institute of Technology, Atlanta, Georgia 30332, USA. [5]Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences & Institute for Molecular Bioscience, The University of Queensland, St Lucia QLB 4072, Australia. †Present address: Helmholtz Zentrum München-German Research Center for Environmental Health, Institute of Groundwater Ecology, Neuherberg 85764, Germany.
*These authors contributed equally to this work.

Table 1 | Relative abundance of T4-like myovirus populations in Pacific Ocean sea water

Rank	Name	Size (kb)	Percentage finished	No. of reads mapped (× 1,000)	Percentage of viral-tagged metagenome	Mean coverage	Percentage of isolates (PCR)
1	CG-05	108	59	2,313	18.20	1,019.0	22
	S-MbCM6/25*	176	NA				
2	CG-24	43	8	329	2.58	964.4	ND
3	CG-07	67	40	488	3.84	916.0	ND
4	CG-03	185	95	1,016	7.99	687.6	ND
5	CG-01	197	94	1,038	8.16	658.4	ND
6	CG-02	180	97	656	5.16	456.6	7.3
7	CG-09	83	57	201	1.58	304.3	ND
8	CG-06	117	37	149	1.18	159.1	ND
9	S-MbC100*	170	NA	127	1.00	93.4	17.1
10	CG-11	37	39	133	1.05	72.1	7.3
	S-MbCM7*	189	NA				
11	CG-04	114	36	65	0.51	71.4	ND
12	CG-19	44	6	25	0.20	69.8	ND
13	CG-12	33	5	18	0.14	69.6	ND
14	CG-13	42	6	21	0.17	63.9	ND
15	CG-10	68	11	29	0.23	53.1	12.2
16	CG-25	40	25	8	0.06	24.4	ND

Percentage finished refers to the estimated percentage of the complete genome captured, calculated as the fraction of the 65-gene T4 core genome observed in the resulting contig. Percentage of viral-tagging metagenome refers to the fraction of the viral-tagging metagenome reads present in the *Candidatus* genome or isolate genomes. Mean coverage refers to the average depth of coverage per *Candidatus* genome. Percentage of isolates (PCR) refers to the percentage of 41 isolates for which g23 sequences can be mapped to the *Candidatus* genomes or isolate genomes (identity >95%, only 41 of 97 g23 products of isolates were sequenced). CG, *Candidatus* genome (phylogenetically informative contig larger than 30 kb derived from the viral-tagging experimental data); NA, not applicable; ND, not detected; PCR, polymerase chain reaction.
* Isolates.

single host, range from relatively dissimilar (two co-isolated cyanophages[9] shared four-fifths of their genes at $\sim$83% average amino acid identity (AAI)) to nearly identical (five roseophages[18] with $\sim$97% average nucleotide identity (ANI)[19]). Genome sequences of hundreds of mycobacteriophages isolated using a single host have revealed 'rampant mosaicism' such that individual viral genomes are composed of assemblages of modules that challenge notions of demarcated populations and hierarchical, genome-based taxonomy (for example, see refs 20–23). Nonetheless, the mycobacteriophage sequences can be clustered into groups by nucleotide similarity, with within-group ANIs ranging from 63–99% (refs 6, 22). As in the marine case, whether these groups denote viral 'species' (that is, discrete ecological and evolutionary units) cannot be discerned given such broad ANI ranges and with only one or two phages sampled per site. Isolate-based genomics could be informative if scaled up, but a single viral-tagging experiment provides the opportunity, now, to explore this question and make four key inferences from its first field application.

First, dsDNA cyanophage genome sequence space is not a genetic continuum in nature, at least not for this particular phage type, host and site. Here, genome-wide genetic relatedness proxies[19] from conserved regions of the dominant T4-like cyanophages (Extended Data Fig. 4) generated a 'population genome landscape', revealing statistically significant discrete clustering of the viral-tagging sequences (Fig. 1) that are robust to variations in recruitment parameters (Extended Data Fig. 5). Such clustering is consistent with globally sampled mycobacteriophage groups[6,22] (see earlier), as well as population structure inferred in cyanophages using marker genes at single sites[24] and globally sampled genomes[25] (Extended Data Fig. 6), and in single-stranded (ss)DNA phages using genomes assembled from pooled natural samples interrogated by feature frequent profile analysis[26]. Yet, the viral-tagging data expand these findings by large-scale analysis at a single site to reveal discrete dsDNA viral clusters—that is, non-overlapping 'clouds' of metagenome-derived cyanophage sequence space—herein termed 'populations'. Whether these

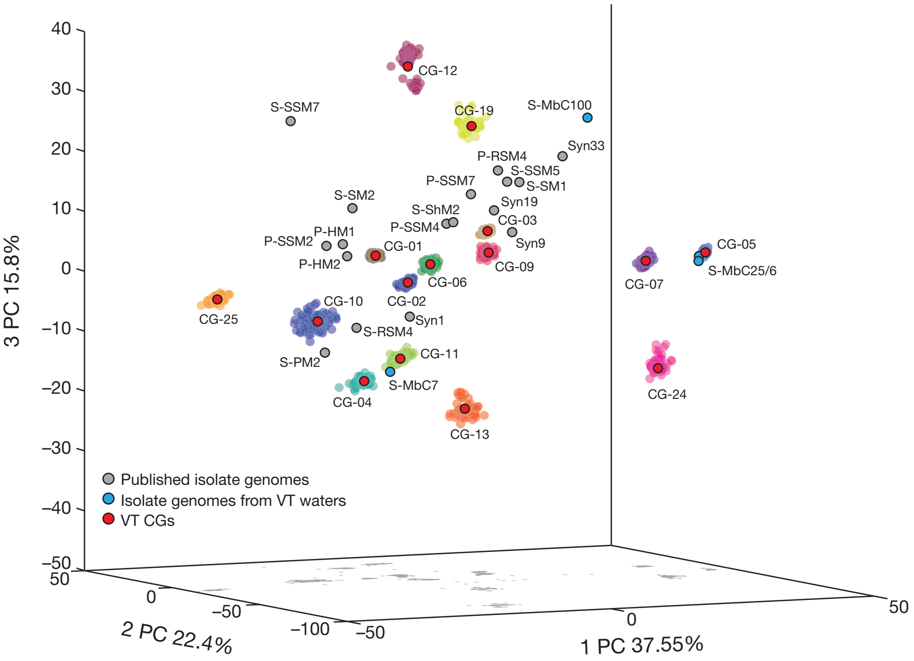


Figure 1 | Population genome landscape plot showing the genetic relationship of cultivated and viral-tagged T4-like phages of *Synechococcus* WH7803 from a single seawater sample and all available marine cyanophage genomes. Principal component (PC) projection of population-level variation within randomly resampled metagenomic data for each *Candidatus* genome (CG) recovered from the viral-tagged (VT) metagenome (the coloured 'clouds'; cloud colours are arbitrary to aid in discriminating populations). *Candidatus* genomes (phylogenetically informative contig, larger than 30 kb) were derived from the viral-tagging metagenome; accuracy of read assignation $Q = 0.9926$, Z-score = 142.2; Dunn index = 0.26, Z-score = 1,829.

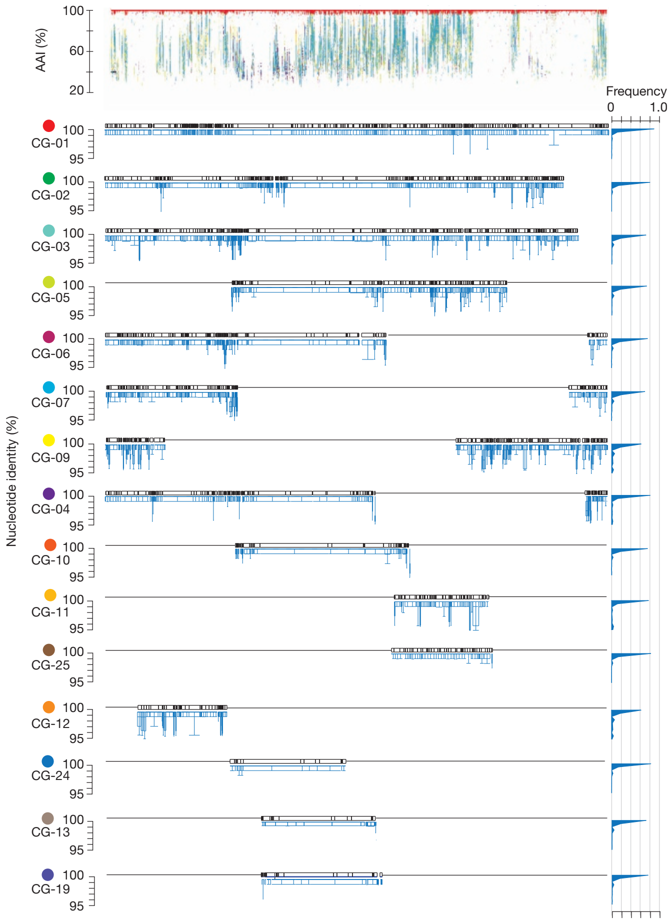


Figure 2 | The 15 dominant T4-like *Candidatus* genomes assembled from the viral-tagged metagenome. Top, reads that map to the *Candidatus* genome CG-01 (genome size 197 kb) by commonly used fragment recruitment metrics (BLASTx *e*-value < 0.001); dots represent reads assigned to the *Candidatus* genomes shown at the bottom, the match is indicated by the colour. Bottom, alignments of all T4-like *Candidatus* genomes against CG-01 are shown. The locus-by-locus nucleotide divergence of each open reading frame (blue) are plotted underneath each genome (0.09, 0.91, second and third quartile and median are shown). The histograms on the right show the summed genome-wide locus-to-locus variation. Note that most variation is concentrated in the top 1%.

populations formally represent species or not will require whole genome information and consideration of neutral and adaptive processes shaping observed variation[27].

Second, such discrete populations enable host-linked, population-based viral ecology. In this single seawater sample and for this host, there are at least 26 viral populations (a 27th is added through isolations, see later). These include 15 T4-like phage populations (Table 1), three of which include co-isolated genomes (Extended Data Fig. 7a), as well as 11 non-T4-like phage populations (Extended Data Fig. 3c). This estimate of 15 T4-like phage populations is consistent with maximal coverage depth in the larger data set of T4 contigs collected here (Extended Data Fig. 7b). Together, these 26 populations represented 0.05–18.2% of the viral-tagging metagenome reads (Table 1 and Extended Data Fig. 3c), with ~53% of the reads assignable to these populations and up to 60% if all small contigs are considered. The remaining 40% of the viral-tagging metagenome reads probably represent reads from the 'rare virosphere' (Methods). The per population, metagenome-derived coverage serves as a proxy for abundance, which enables an estimate of the first host-associated wild viral rank–abundance distribution (Extended Data Fig. 7c), analogous to long-standing ecological efforts to characterize species abundances in natural communities[28].

Third, viral tagging allows quantitative examination of cyanophage culture bias. Here, four *Myoviridae* isolates from the same waters included the first, ninth and tenth most abundant T4-like phage populations observed in the viral-tagging metagenomes on this host (Table 1). In addition, all amplicons derived from isolates using PCR with primers that target the major capsid protein (gp23) from T4-like myoviruses can be mapped to the viral-tagging populations (first, sixth, ninth, tenth, fifteenth in Table 1, and the rest to the small contigs, see Methods). This overlap between isolates and viral-tagging populations partially validates the viral-tagging procedure and suggests that, at least for T4-like cyanophages, culture bias might be relatively minimal. Notably, however, no isolates showed similarity to the 42 new viruses revealed by viral tagging, and the projected variation in sequence space recovered by a single viral-tagging experiment is larger than that associated with published global isolates (Fig. 1). Together, this suggests that culture-based studies may miss major routes of horizontal gene transfer and/or ecological interactions.

Last, we were able to document intra-population variation for wild uncultured viruses (Fig. 2), which is critical for interpreting metagenomic fragment recruitment analyses and establishing a genome-based viral species definition. Here, each population's locus-to-locus, pairwise percent nucleotide identity between the reference sequence and its 'assigned' viral-tagging metagenomic reads ranged from 95–100% ANI (mean 99.53%; for example, see insets in Fig. 2), with some populations varying more than others (see spread of clouds in Fig. 1 and box plots in Fig. 2). This is similar to >99% ANI observed across eight loci used to group 60 isolates into five clusters in *Synechococcus* cyanophage isolates[24], and >95% ANI commonly associated with microbial species definitions[19]. However, it is more conservative than most of the range (83–97%, average 90%) of ANIs observed in ten isolates in the phiKMV species complex[29]. Interestingly, intra-population ANIs from conserved and hypervariable regions are statistically indistinguishable (Fig. 2). By contrast, pairwise inter-population variation observed in the viral-tagging metagenomes suggests that nucleotide identities range from 42 to 71% between populations (Extended Data Table 2). The finding of high intra-population ANI from hypervariable regions of the captured cyanophage stands in contrast to models of rampant phage mosaicism[22], in which assemblages of modules within viral genomes suggest a horizontal, rather than vertical, evolutionary signal. It remains to be determined if this observation is exceptional or the rule for phage population structure. Similar intra- and inter-population sequence divergence levels are maintained by differences in relative recombination rates in bacteria and archaea[27]. Formal testing of whole genome data in a population genetic framework (for example, see ref. 27) is needed to assess the validity of these empirically derived populations as species. Nonetheless, these viral-tagging data already provide a much-needed benchmark for refining metagenomic analyses, albeit from a single host and sample, by suggesting an empirical cut-off (<95% ANI) for reads that probably derive from different populations.

In conclusion, viral-tagging-enabled experimental linkage of wild cyanophages to their host at a single site provides evidence that phage genome sequence space is structured in nature, just as recently posited for bacteria and archaea[27,30]. Moving forward, viral tagging has the potential to enable researchers to broadly map how viruses change over space and time. Given that such comparative viral-tagging data are genome- and host-linked, as well as population-based, these data should better elucidate the processes that drive viral population structure in nature.

METHODS SUMMARY

Viral-tagging metagenomes. Surface (10 m) water was collected from Station H3 (36° 73.5 N, 237° 98.1 E) in Monterey Bay, California, United States, 0.22 μm filtered and stored (4 °C, dark). *Synechococcus* WH7803, labelled with ^{15}N, was mixed with fluorescently labelled viral particles[11], then viral-tagged cells (increased fluorescence) were flow-cytometrically sorted, DNA was extracted, and ^{15}N-labelled heavy host DNA was separated from non-labelled light viral DNA by CsCl density ultracentrifugation. In total, 3×10^8 virions were co-incubated with 3×10^7 host cells for

60 min then 1.2×10^{7} viral-tagged cells were sorted and used for DNA extraction. Light DNA was linker amplified[12] for sequencing.

Community metagenomes. Viral concentrates were prepared from 20 l of 0.22 μm filtrate by chemical flocculation and purified using ultracentrifugation. DNA was extracted, linker amplified[12] and sequenced. However, our metagenomic DNA preparation method would strongly select against ssDNA phages, and not capture RNA phages at all.

Phage isolation and characterization. Ninety-seven cyanophages able to infect *Synechococcus* WH7803 were isolated and purified as previously described[16]. Ninety of 97 isolates were assigned to T4-like myoviruses using a specific gene marker (gp20). Four isolate genomes were assembled completely: S-MbCM6, S-MbCM7, S-MbCM25 and S-MbCM100.

Bioinformatic analyses. Quality control, filtering, assembly, protein clustering, annotation, taxonomy analyses, whole genome comparison, statistics, locus-by-locus variation and associated bioinformatics analyses were done using a set of custom scripts detailed in Methods and Extended Data Fig. 8. Scripts and associated documentation are available at http://www.eebweb.arizona.edu/faculty/mbsulli/informatics.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 5 December 2013; accepted 7 May 2014.

Published online 13 July 2014.

1. Bergh, O., Børsheim, K. Y., Bratbak, G. & Heldal, M. High abundance of viruses found in aquatic environments. *Nature* **340,** 467–468 (1989).
2. Breitbart, M. Marine viruses: truth or dare. *Annu. Rev. Mar. Sci.* **4,** 425–448 (2012).
3. Suttle, C. A. Marine viruses—major players in the global ecosystem. *Nature Rev. Microbiol.* **5,** 801–812 (2007).
4. Hurwitz, B. L., Hallam, S. J. & Sullivan, M. B. Metabolic reprogramming by viruses in the sunlit and dark ocean. *Genome Biol.* **14,** R123 (2013).
5. Flombaum, P. *et al.* Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. *Proc. Natl Acad. Sci. USA* **110,** 9824–9829 (2013).
6. Holmfeldt, K., Odić, D., Sullivan, M. B., Middelboe, M. & Riemann, L. Cultivated single-stranded DNA phages that infect marine Bacteroidetes prove difficult to detect with DNA-binding stains. *Appl. Environ. Microbiol.* **78,** 892–894 (2012).
7. Hatfull, G. F., Jacobs-Sera, D. & Lawrence, J. G. Comparative genomic analysis of 60 mycobacteriophage genomes: genome clustering, gene acquisition, and gene size. *J. Mol. Biol.* **397,** 119–143 (2010).
8. Lavigne, R., Seto, D., Mahadevan, P., Ackermann, H.-W. & Kropinski, A. M. Unifying classical and molecular taxonomic classification: analysis of the *Podoviridae* using BLASTP-based tools. *Res. Microbiol.* **159,** 406–414 (2008).
9. Sullivan, M. B. *et al.* Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. *Environ. Microbiol.* **12,** 3035–3056 (2010).
10. Mann, N. H. *et al.* The genome of S-PM2, a 'photosynthetic' T4-type bacteriophage that infects marine *Synechococcus* strains. *J. Bacteriol.* **187,** 3188–3200 (2005).
11. Deng, L., Gregory, A., Yilmaz, S., Poulos, B. T., Hugenholtz, P. & Sullivan, M. B. Contrasting life strategies of viruses that infect photo- and heterotrophic bacteria, as revealed by viral tagging. *MBio* **3,** e00373–12 (2012).
12. Duhaime, M. B., Deng, L., Poulos, B. T. & Sullivan, M. B. Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: a rigorous assessment and optimization of the linker amplification method. *Environ. Microbiol.* **14,** 2526–2537 (2012).
13. Weitz, J. S. *et al.* Phage–bacteria infection networks. *Trends Microbiol.* **21,** 82–91 (2013).
14. Sharon, I. *et al.* Comparative metagenomics of microbial traits within oceanic viral communities. *ISME J.* **5,** 1178–1190 (2011).
15. Sharon, I. *et al.* Photosystem I gene cassettes are present in marine virus genomes. *Nature* **461,** 258–262 (2009).
16. Sullivan, M. B., Coleman, M. L., Weigele, P., Rohwer, F. & Chisholm, S. W. Three *Prochlorococcus* cyanophage genomes: signature features and ecological interpretations. *PLoS Biol.* **3,** e144 (2005).
17. Millard, A. D., Zwirglmaier, K., Downey, M. J., Mann, N. H. & Scanlan, D. J. 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.* **11,** 2370–2387 (2009).
18. Angly, F. *et al.* Genomic analysis of multiple Roseophage SIO1 strains. *Environ. Microbiol.* **11,** 2863–2873 (2009).
19. Konstantinidis, K. T. Genomic insights that advance the species definition for prokaryotes. *Proc. Natl Acad. Sci. USA* **102,** 2567–2572 (2005).
20. Lawrence, J. G., Hatfull, G. F. & Hendrix, R. W. Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches. *J. Bacteriol.* **184,** 4891–4905 (2002).
21. Hendrix, R. W., Lawrence, J. G., Hatfull, G. F. & Casjens, S. The origins and ongoing evolution of viruses. *Trends Microbiol.* **8,** 504–508 (2000).
22. Hatfull, G. F. The secret lives of mycobacteriophages. *Adv. Virus Res.* **82,** 179–288 (2012).
23. Hendrix, R. W., Smith, M. C., Burns, R. N., Ford, M. E. & Hatfull, G. F. Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc. Natl Acad. Sci. USA* **96,** 2192–2197 (1999).
24. Marston, M. F. & Amrich, C. G. Recombination and microdiversity in coastal marine cyanophages. *Environ. Microbiol.* **11,** 2893–2903 (2009).
25. Ignacio-Espinoza, J. C. & Sullivan, M. B. Phylogenomics of T4 cyanophages: lateral gene transfer in the 'core' and origins of host genes. *Environ. Microbiol.* **14,** 2113–2126 (2012).
26. Labonté, J. M. & Suttle, C. A. Previously unknown and highly divergent ssDNA viruses populate the oceans. *ISME J.* **7,** 2169–2177 (2013).
27. Polz, M. F., Alm, E. J. & Hanage, W. P. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends Genet.* **29,** 170–175 (2013).
28. Whittaker, R. H. Dominance and diversity in land plant communities: numerical relations of species express the importance of competition in community function and evolution. *Science* **147,** 250–260 (1965).
29. Ceyssens, P. J. *et al.* Phenotypic and genotypic variations within a single bacteriophage species. *Virol. J.* **8,** 134 (2011).
30. Shapiro, B. J. *et al.* Population genomics of early events in the ecological differentiation of bacteria. *Science* **336,** 48–51 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements Funding was provided by the US Department of Energy (DOE) Joint Genome Institute (JGI) Community Sequencing Program, Biosphere 2, BIO5, US National Science Foundation (NSF) OCE0940390, and Gordon and Betty Moore Foundation grants to M.B.S., as well as NSF OCE1233760 and Burroughs Wellcome Fund grants to J.S.W. We thank J. Fuhrman for suggesting stable-isotope-labelled host DNA; A. Z. Worden and the CANON Initiative for the cruise opportunity; Worden laboratory members; the captain and crew of the R/V Western Flyer for operational/sampling support; J. B. Waterbury, S. W. Chisholm and A. Wichels for strains; the Tucson Marine Phage laboratory; Institute of Groundwater Ecology of Helmholtz Munich; and N. Pace, M. Young, S. W. Chisholm and S. Yilmaz for technical/analytical support and manuscript comments. We acknowledge the University of Arizona Genetics Core for viral-tagging metagenomic sequencing; iCyt and the Arizona Cancer Center and Arizona Research Laboratories (ARL) Division of Biotechnology Cytometry Core Facility for cytometry support; the University Information Technology Services Research Computing Group and the ARL Biotechnology Computing for high-performance computing clusters (HPCC) access and support. Community metagenomic sequencing was provided by the DOE JGI Community Sequencing Program under the Office of Science of the US DOE contract no. DE-AC02-05CH11231.

Author Contributions L.D., P.H. and M.B.S. designed the experiments. L.D. collected samples. L.D., A.C.G. and B.T.P. performed the experiments. L.D., J.C.I.-E., J.S.W., P.H. and M.B.S. analysed data, interpreted results and wrote the paper.

Author Information Data for viral genomes have been deposited in GenBank under accession numbers JN371768 and KF156338-40; metagenomic data have been deposited in CAMERA under accession numbers CAM_P_0001068 and CAM_P_0000915; raw data including gp23 sequences and informatic pipelines, assemblies and data for figures are available at http://datadryad.org/resource/doi:10.5061/dryad.gr3ks. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.B.S. (mbsulli@email.arizona.edu).

METHODS

Experimental methods. Detailed protocols for chemical flocculation, viral purification, transmission electron microscopy (TEM), viral tagging, and linker amplification are available at http://eebweb.arizona.edu/Faculty/mbsulli/protocols.

Strains and culturing conditions. *Synechococcus* WH7803 was grown at 20–22 °C under a 14 h:10 h light–dark cycle at 15–18 μE m^{-2} s^{-1} in SNAX medium[31], made from filtered (100 kDa membrane, nominal molecular weight limit) and autoclaved water collected from surface Pacific Ocean (10 m depth, near Scripps Pier, San Diego, California, United States; 32° 52.0 N, 117° 15.4 W) in April 2009. Growth of cultures in liquid medium was followed by *in vivo* phycobiliprotein autofluorescence as a proxy for biomass in arbitrary fluorescence units (AU) using an Appliskan plate reader (Thermo Electron) with excitation wavelengths of 485 ± 20 nm and measured emission wavelengths of 590 ± 40 nm.

Source waters for viral tagging and community viral metagenomics. Water samples were collected from the surface (10 m depth) at Station H3 (36° 73.5 N, 237° 98.1 E) in Monterey Bay, California, United States on 1 October, 2009. Samples were immediately 0.22 µm filtered (Millipore Express Plus, Millipore) and stored at 4 °C in the dark in acid-washed polycarbonate bottles until further analysis. Viral concentrates for community metagenomic sequencing were prepared from 20 l of 0.22 µm filtrate by chemical flocculation as described previously[32], followed by purification using ultracentrifugation as described previously[33].

Phage isolation. Cyanophages able to infect *Synechococcus* WH7803 were isolated using plaque assays as previously described[34] using the same source water for the viral-tagging experiments described earlier (herein termed 'VT water'). Individual plaques, which appeared after 7–35 days, were collected in agarose plugs using sterile glass Pasteur pipettes and three more rounds of plating were used to ensure the cyanophage isolates were clonal. DNA of clonal isolates was extracted, linker amplified as described later and sequenced using 454 Roche Titanium chemistry. Genomes were assembled as described later for the metagenomic samples. Four genomes were assembled completely: S-MbCM6, S-MbCM7, S-MbCM25 and S-MbCM6100; although S-MbCM6 and 25 were nearly identical (>99% nucleotide identity, but ~100-fold greater than sequencing error). Isolations were classified and named according to the 2012 release of ICTV: the first letter indicates the initial host strain, where S = *Synechococcus*; the second to forth letter indicates the place from which the sample was collected, where MbC = Monterey Bay coastal site; the last letter indicates morphotype, where M = *Myoviridae*; a number is used to distinguish between otherwise similar isolates of the same type.

T4-like myovirus-specific marker genes (gp23 and gp20) were employed as previously described[9,34]. PCRs using gp20 primers were positive for 90 of the 97 isolates, suggesting that the bulk are T4-like phages (Extended Data Fig. 1), similar to the taxonomy pattern of previous isolations on the same host (92 of 105, 88%)[35–44]. Using non-optimized PCR conditions, only 41 of the 97 isolates yielded gp23 products, and these were further sequenced (Table 1).

TEM. Viral lysates were positively stained by 2% uranyl acetate (Ted Pella) for 30 s followed by three 10 s washes in ultra-pure water (Milli-Q, Millipore) on formvar-covered copper support grids. Deposited material was then wicked away by filter paper. Grids were then dried at ambient conditions overnight and stored in a desiccator until analysis. Prepared grids were examined at 65,000–100,000 magnification using a transmission electron microscope (Philips CM12, FEI) with 100 kV accelerating voltage. Micrographs were collected using a Macrofire Monochrome CCD camera (Optronics).

Viral-tagging experiments. Viral-tagging experiments were performed between December 2010 to January 2011 as previously described[11]. SYBR Gold commercial stock (with a concentration of 10,000×) was diluted to 50× in TE buffer (10 mM Tris, 1 mM EDTA; pH 8.0) for storage at −20 °C in the dark. The ultracentrifugal devices (10 kDa cut-off; Nanosep, PALL, catalogue no. 29300-608) were pretreated by incubating 0.5 ml of 0.02 µm filter-sterilized 1% bovine serum albumin (BSA) (Bioexpress, catalogue no. E531-1.5ML) in phosphate buffered saline (PBS) for 60 min at room temperature. Filtered water samples (0.22 µm) were stained with SYBR Gold (final concentration of 5×; Invitrogen, catalogue no. S11494) at 80 °C for 10 min and washed six times by TE buffer in the pretreated ultracentrifugal devices. Fifty-microlitres TE buffer was added back for every 500 µl water samples and sonicated (VWR Signature Ultrasonic cleaner B1500A-DTH, VWR) for 3 min using the settings of 50 W at 42 kHz, resulting in a tenfold concentration of viruses from the original water sample. Stained and washed viruses were co-incubated with cells of *Synechococcus* WH7803 in the late log phase at concentrations and ratios suitable for flow cytometer analysis, typically 10^5 cells per ml and virus-to-bacterium ratio (VBR) of ten. *Synechococcus* WH7803 cells were acclimatized through five inoculations before mixing with viruses. Various co-incubation times from 10 to 120 min (for rationale see ref. 34) were tested, the percentage of viral-tagged cells plateaued at 46% after 60 min, so 60 min was chosen for the co-incubation time in the rest of the viral-tagging experiment. In total, 3×10^7 cells and 3×10^8 virus particles were mixed, and 1.2×10^7 viral-tagged cells (fluorescently labelled cells) were separated from unlabelled cells and were collected together with the virus particles tagged to them using maximum purity sorting settings. Hence, at least 1.2×10^7 viral particles were screened given the association of at least one virus per viral-tagged cell. Viral-tagging experiments were done with a negative control, which was prepared identically to the stained and washed virus samples except without viruses; this controlled for free dye creating the appearance of false positive viral-tagged cells.

Flow cytometer analyses. Samples were examined using an iCyt Reflection flow cytometer (Sony Biotechnology) equipped with a 200 mW 488 nm air-cooled solid-state laser or a MoFlo XDP cytometer (Beckman Coulter). Fluorescence was detected using a 520/40 band pass filter with an amplified photomultiplier tube. Events were detected using a Forward Scatter trigger, and data collected in logarithmic mode then analysed with WinList 6.0 software (Verity Software House). Fluorescent polystyrene FLOW Check microspheres (1 µm yellow–green beads; Polysciences, catalogue no. 23517-10) were used as an internal standard. Samples were typically run with a concentration of 10^5 cells ml^{-1} and 10^6 viruses ml^{-1}.

Isotope labelling of host cell DNA for viral-tagging metagenomes. To minimize the bacterial DNA in the viral-tagging metagenome, axenic *Synechococcus* WH7803 was grown in medium amended with 800 µM ^{15}N-ammonium chloride (Cambridge Isotope Laboratories, catalogue no. NLM-467-1) for 4 weeks, that is, 4 transfers. Cells were harvested each week by centrifugation at 8,000*g* for 15 min at 20 °C and re-suspended in the labelled medium. After 4 weeks, 72% of the *Synechococcus* WH7803 DNA was isotopically heavy, and so ready for use as hosts for the wild viral-tagging experiments. The 5th week cells were centrifuged, and then washed twice and re-suspended in non-labelled medium. DNA was extracted from the flow-cytometry-sorted viral-tagged cells and ^{15}N-ammonium-chloride-labelled, isotopically heavy DNA (bacterial) was separated from non-labelled, isotopically light DNA (viral with few non-labelled bacterial) by CsCl (density ρ 1.7) in a TV865 vertical rotor (Sorvall) at 44,000 r.p.m. for 48 h at 18 °C.

Linker amplification and sequencing. DNA from the viral-tagged community was extracted immediately after sorting. DNA extracted from both viral-tagged and community viral concentrates was linker amplified in March 2011, as previously described[12] for metagenomic sequencing. Please note that our linker ligation step of the library preparation strongly selects against ssDNA and, furthermore, RNA viruses would not be sequenced in a DNA metagenome[45]. DNA extracted from community viral concentrates was sequenced using Roche 454 Titanium chemistry at the University of Arizona Genetics Core (http://uagc.arl.arizona.edu/), to yield 254,642 sequence reads (herein termed '454 reads'). DNA extracted from viral-tagging viral concentrates was sequenced initially by 454 in parallel with the community sample (132,052 reads), and additionally by the Department of Energy Joint Genome Institute using Illumina HiSeq 2000 chemistry (herein Illumina reads) to obtain deeper coverage (22,589,436 bp).

Phage infectivity. Phage isolates (12 infecting *Prochlorococcus*: six myoviruses, four podoviruses and two siphoviruses; 11 infecting *Synechococcus*: ten myoviruses and one podovirus; and three phages[46] of *Cellulophaga baltica* (phylum *Bacteroidetes*); Extended Data Table 1) were screened for their infectivity on *Synechococcus* WH7803 using the plaque assays described earlier. Each interaction between a phage and *Synechococcus* WH7803 was performed in duplicate on at least two different occasions.

Viral-tagging signals of all phages correspond to plaque assays. Among 23 cyanophages, two co-isolated myoviruses (S-SSM1 and S-SSM2), where one can infect *Synechococcus* WH7803 (S-SSM2) and the other cannot (S-SSM1), revealed corresponding positive and negative viral-tagging signals. All cyanophage–host mixtures resulted in only 45% of host cells being viral-tagged (Extended Data Table 1; T = 20 min, VBR = 10) as previously reported, except for three that were lower (P-SSM4, 5.72 ± 0.71%; P-SSM6, 3.54 ± 0.52%; P-RSM2, 4.40 ± 0.75%; n = 4). These low percentages of viral-tagged cells are probably due to low efficiencies of infection. In particular, two phages (P-SSM4 and P-SSM6) were previously characterized as non-infective on the given host by plaque assays. However, we found here that at a higher viral concentration (~10^9) and VBR (10), these phages did indeed lyse the host. Phage P-RSM2 lysed its host in a previous[42] and the present study; while the viral concentration and VBR employed previously was not available in the decades-old experiments, we assume that high viral titres were used.

Viral-tagging adsorption preference validation. In order to check whether the viral-tagging conditions favour a certain morphology of phages, we performed viral-tagging experiments using phages of *Myoviridae* (P-MM105), *Podoviridae* (P-MP121) and *Siphoviridae* (P-MS209), together with their host strain *Pseudomonas putida* IsoF, which is also the strain they were originally isolated on. Host cells were acclimatized through three inoculations prior to mixing with phages. Viral-tagged percentages of cells were 100% for all three phages using the same viral-tagging condition described earlier, a VBR of ten, and a co-incubation time of 60 min of phages and host cells. Then, we mixed three phages at a 1:1:1 ratio (VBR = 30), co-incubated with host cells at the same condition (60 min), and 100% of cells were viral-tagged. We collected those viral-tagged cells in two replicating experiments, and performed TEM as described earlier to determine the morphologies of tagged phages. The recovered ratio of myo-, podo- and siphoviruses ($52 \pm 3 : 48 \pm 5 : 56 \pm 3$; $n = 4$) was statistically indistinguishable (replicated G-test; $P = 0.336$) from the original 1:1:1 input ratio, which indicates that viral tagging did not introduce a preferential selection for viral morphotype.

Bioinformatic analyses. Quality control, filtering, assembly and bioinformatics analyses were done using a set of custom scripts. Scripts (referenced below by name) and associated documentation are available at http://www.eebweb.arizona.edu/faculty/mbsulli/informatics. Informatic pipelines, assemblies, data for figures and raw data are available at http://datadryad.org/resource/doi:10.5061/dryad.gr3ks. For the current project two kinds of data were used: 454 data that came from the DNA prepared from both the community and viral-tagging metagenomes, and Illumina data for the viral-tagging metagenome.

Quality control. Quality control for 454 data was done as previously described[33]. Illumina data quality control consisted of trimming ends with a quality score lower than 25 as well as sequences containing ambiguous bases; only reads longer than 100 bp were kept. Additionally, because the Illumina sequencing, derived from linker-amplified DNA, was mixed 1:1 with phiX174 DNA to minimize base-calling issues in Illumina software, full-length reads matching (>98% identity) to the phiX174 genome were removed. From the remaining reads, the linkers were removed and run through the quality control process described earlier.

Assembly. Contigs were assembled from post-quality-control reads using Velvet (version 1.2.01) with a conservative k-mer size of 57 and the -cov_cutoff option set to 10 as done previously[47]. Iterative assembly was used whereby reads incorporated into the largest contigs were removed to compensate for highly variable coverage (30–500×) found across the genomes in these natural samples. After 15 rounds of assembly, 26 large contigs were obtained (>30 kb that were 'representative' regions of the genome, see later) and referred to as '*Candidatus* genomes' (CGs) in the manuscript. These 26 CGs constitute a total of ~40% of the available reads; the remaining 60% of the data presumably belong to rare members of this coastal phage community. The community metagenome was assembled using Newbler[48], as only 454 data were available, requiring >40 base pairs of overlap at >95% identity.

Protein clustering. Open reading frames (ORFs) were predicted using Prodigal[49] from all contigs >1.5 kb, including the CGs, as well as on all 454 reads that were not used in assembly. ORFs were clustered using cd-hit[50] with a cut-off of 75% identity. Clusters with two or more members were considered bona fide, which is notably permissive (as compared to ref. 51 'high confidence clusters') to maximize the data mappable to protein clusters. Individual reads then were mapped to protein clusters using BLASTx with an *e*-value cut-off of 0.001, only top hits were used. Rarefaction curves were calculated using a custom perl script (Rarefaction.pl). The Chao-1 index was calculated from the protein cluster data as described previously[52]. The Simpson diversity index[53] (D) was calculated as $D = \sum n(n-1)/N(N-1)$ and the Shannon–Wiener diversity index (H') was calculated as $H' = -\sum p \ln p$, where $p = n/N$, $n =$ number of reads in each protein cluster and N represents the total number of reads assigned to protein clusters.

Contig annotation. Assembled contigs >1.5 kb were annotated as follows. ORFs were predicted using Prodigal (as described earlier) and functionally annotated using manually curated data resulting from BLASTp analyses against the non-redundant protein database of GenBank, and custom databases of T4 phage gene clusters (T4-GCs[9]) and Microbial Metabolic Genes[14]. We were able to assign gene function and taxonomy to the majority of the contigs (87%); the lack of identities in a small fraction of the viral-tagging contigs (18% or 13%) raised the possibility that they were of bacterial origin. While sequence-similarity-dependent methods for identifying contigs in viral metagenomes are a well-known problem, they are much less of a problem for microbes as those reference databases are relatively well populated. The identification of phage genes within bacterial genomes represents an equivalent problem; they rely on the search for certain characteristics that may suggest a region is of viral origin. Strand bias, gene density and sequence homology are among the most widely used characteristics but usually require at least 10 kb (ref. 54). The size of contigs in our experiment was not sufficient (1.5–2.5 kb) for applying similar methods, instead we developed the following approach to test for identification of microbial genes in similarly sized bacterial contigs. We obtained the latest release of RefSeq (60; $N = 197{,}527$ contigs) and artificially generated 10,000 contig fragments of similar size (1.5, 2.0 and 2.5 kb). This was done by randomly selecting a contig from the RefSeq database and then randomly extracting a 1.5 to 2.5 kb fragment from the contig. These 10,000 contig fragments for each size (total = 30,000 contigs) were used as queries for a BLASTp (using translated ORFs already predicted in the genomes) against the complete RefSeq database (v.60). Self matches were ignored, that is, we counted the number of times each contig had only proteins that only hit themselves in databases. We found self-hits only in 16 (1.5 kb contigs), 5 (2.0 kb contigs) and 2 (2.5 kb contigs) of the 30,000 contigs. These results suggest that for 1.5–2.5 kb contig sizes that are derived from microbes, we would have a > 99% success rate at identifying them. Thus, we speculate that with >99% certainty the 33 ambiguous contigs represent non-microbial taxa.

Gp23 sequences matches. Forty-one of 97 isolates yielded product using a primer set specific for g23 (ref. 34); all g23 fragments were sequenced and can be mapped to the viral-tagging contigs (identity >95%). In total, 66% of the isolates can be mapped to CGs, listed in Table 1, the rest can be mapped to small contigs (17% to Contig00058 and 17% to Contig00161).

Taxonomy analyses. To estimate the relative proportion of reads associated with particular viral types, we used a BLASTx search against the phage genomes available in NCBI ($n = 1{,}218$, October 2013), and assigned taxonomy to metagenomic reads by the taxon lineage associated with their top hit (requiring an *e*-value $<1 \times 10^{-3}$); we used read2family.pl available with the rest of the scripts.

Whole genome comparisons. To estimate the relatedness of the new whole genomes and CGs generated in this study, we adopted commonly used metrics for microbial genome comparisons—ANI/AAI[19]. For the broader comparisons, we used AAI rather than ANI due to the low nucleotide conservation across viral genomes.

First, we performed *in silico* 'sizing' and 'positioning' evaluations to empirically determine how to interpret fragmented genomes resulting from viral-tagging metagenomic assemblies using a custom perl script (SizeAndLocation.pl). Specifically, fragments (20, 25, 30, 35, 40, 45, 50 and 55 kb) were generated from each complete genome on a sliding window of 5 kb. We then calculated the ANI between the fragment and our database of 21 full genomes (17 available cyanophage genomes plus 4 isolates genomes in this study). A custom script (Pearsons.pl) was used to compare the resulting vector (similarity profile) of ANI values (fragment versus genomes) to that of the originating complete genome (genome versus genomes). The result was converted to a correlation-based distance ($= 1 - r$, where r corresponds to Pearson's correlation coefficient; only positive values of r were obtained) to assess how well any given fragment represents a full genome. We found that fragments as short as 30 kb could be assigned to their original source, so long as they originated from areas of the genome with high Pearson's r values (Extended Data Fig. 4). Then, fragments assembled from the metagenome fulfilling the previous characteristics were used in the downstream analysis and are referred to as *Candidatus* genomes (CGs).

The genetic relatedness of all T4-like 15 CGs from the viral-tagging metagenome was compared using AAI, against a fixed database of 21 genomes (17 available cyanophage genomes plus 4 isolate genomes identified in this study). AAI was calculated only from conventionally defined pairs of homologous genes[9,55]. Homology was defined as sequence similarity over 40% covering at least 60% of the length of the shortest gene.

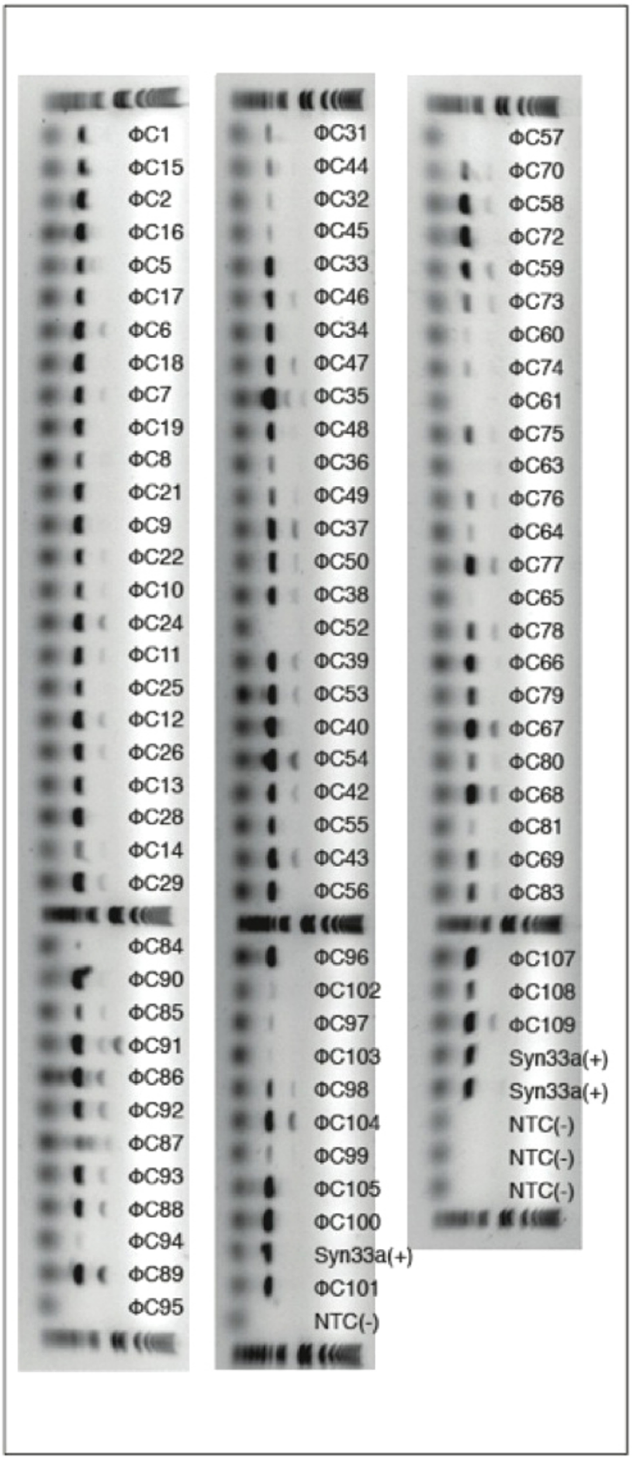
To estimate the variability within a population from the available metagenomic data, random CGs were generated as follows using a series of custom perl scripts. First (Extended Data Fig. 8), we recruited reads to each CG requiring at least 95% identity and a coverage of 95% of the entire length of the read (Recruit2CloudV1.pl). Each read was non-redundantly assigned and aligned to a CG using default parameters in MUSCLE[56]. For each CG population, we generated 100 random CG sequences using the metagenomic data that were recruited to consensus sequences, with each base having a probability of being assigned from its relative abundance in the underlying metagenomic sequence data (sequences deposited in http://datadryad.org/resource/doi:10.5061/dryad.gr3ks). The matrix of pairwise AAI genome comparisons (size: $1{,}500 \times 21$) was used in principal component analysis. The first three components account for 75% of the variation. The Euclidean distances of the reference genomes in this three-dimensional coordinate system are a good proxy for their phylogenetic relationships (Extended Data Fig. 6).

Statistical evaluation. We evaluated the clusterdness of the viral-tagging data using the following approaches. First, we defined the accuracy of the assignation, Q. We calculated the distances between each random sequence generated from the CGs and each of its consensus sequences. Each randomly generated sequence is assigned to the consensus that is closest to it, independent of its origin. Only the first three coordinates were used as these three PCs account for 75% of the variation and serve as a good proxy for phylogenetic distances (see earlier). We compiled this information in an assignation matrix, A, where rows are the actual consensus sequence sources and the columns are the assigned (closest) sequences. If the random sequences are highly similar to the source, then the assignment matrix should be nearly diagonal. The accuracy of the assignation is defined as $Q = \mathrm{Tr}(A)/N$, where N is the total number of randomizations and $\mathrm{Tr}(A)$ denotes the trace of the matrix A. Alternatively, Q is equivalent to the fraction of true positive assignations (that is, the number of times in which a genome was correctly assigned to its true source divided by the total number of generated genomes). To statistically evaluate the significance of the observed value of Q we used a randomization scheme as follows (Acc.m and AccRdm.m). Labels were randomly assigned to fragments, then Q was calculated as described earlier, this was done 1,000,000 times, in no case did we obtain a higher value of Q in a randomization trial than in the observed data. We measured the effect size in terms of a Z-score: $Z = (Q_e - Q_r)/\sigma$, where $Q_e = 0.9906$ is the observed Q value, $Q_r = 0.0665$ is the average value of our randomization scheme and σ is the standard deviation of Q values in the randomization scheme, $\sigma = 0.0065$, $Z = 142.17$. This Z score implies that the observed Q is very far from any observed value obtained by random chance.

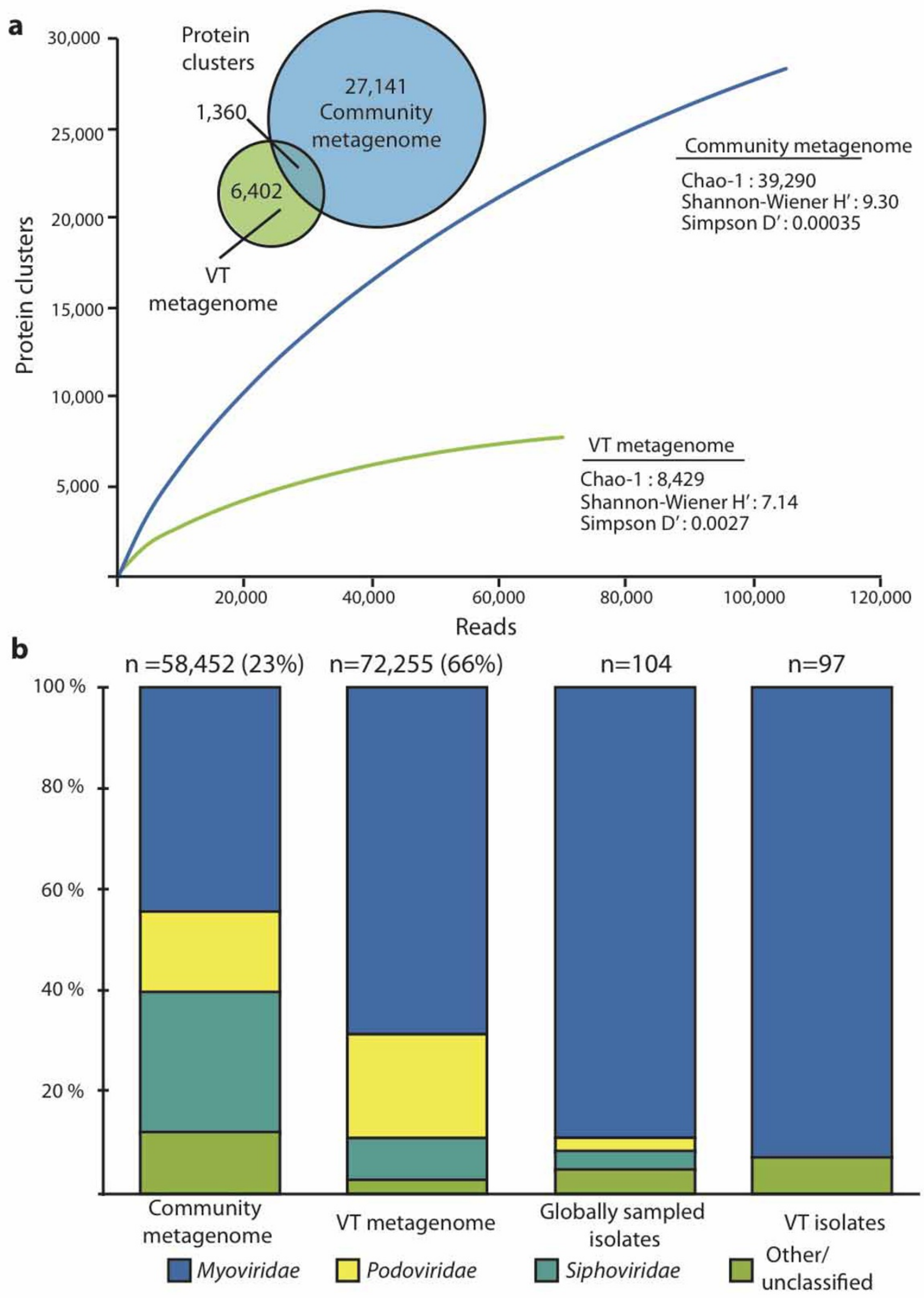
Since a value of *Q* close to 1 can result from loose clusters that are well separated in space we also wanted to quantify the compactness of the cluster. To do this, we used the Dunn index[57] (dunns.m and DunnRdm.m). Briefly, this index corresponds to the ratio of the smallest distance between all pairs of clusters divided by the maximum distance within a cluster. We ran a similar randomization scheme as stated earlier; out of 1,000,000 repetitions the measured Dunn index of the CGs data was larger than that observed in any of the randomization trials. The *Z*-score for the Dunn index was 1,829, again suggesting that the observed clustering is highly unlikely to have occurred by chance.

Locus-by-locus variation. To get beyond genome-wide averaged genetic diversity metrics, we examined the underlying sequence data for each population to estimate variation at the level of a predicted ORF. Those reads that mapped to reference genomes (95% identity over 95% read length) were further examined to determine the locus-by-locus genetic diversity (average pairwise per cent nucleotide identity per ORF) using a custom perl script (LocusbyLocus.pl). While most loci in these populations are nearly 100% identical, box plots (0.09, 0.91, second and third quartile and median) show the range of variability in the identity of reads assigned to each locus (Fig. 2 and Extended Data Fig. 3). A lower stringency recruitment (90% identity at 90% coverage of read length as well as 80% identity at 90% coverage of read length) resulted in nearly identical results (Extended Data Fig. 5) indicating that the genetic diversity shown in the population genome landscape reflects biology rather than bioinformatic artefact. Fragment recruitment plots were generated using a custom script (read2genome.pl, modified from ref. 58) that plots the BLASTx results of metagenomic reads against a reference genome, where alignment location information is only considered for hits with >20% identity and longer than 45 amino acids.

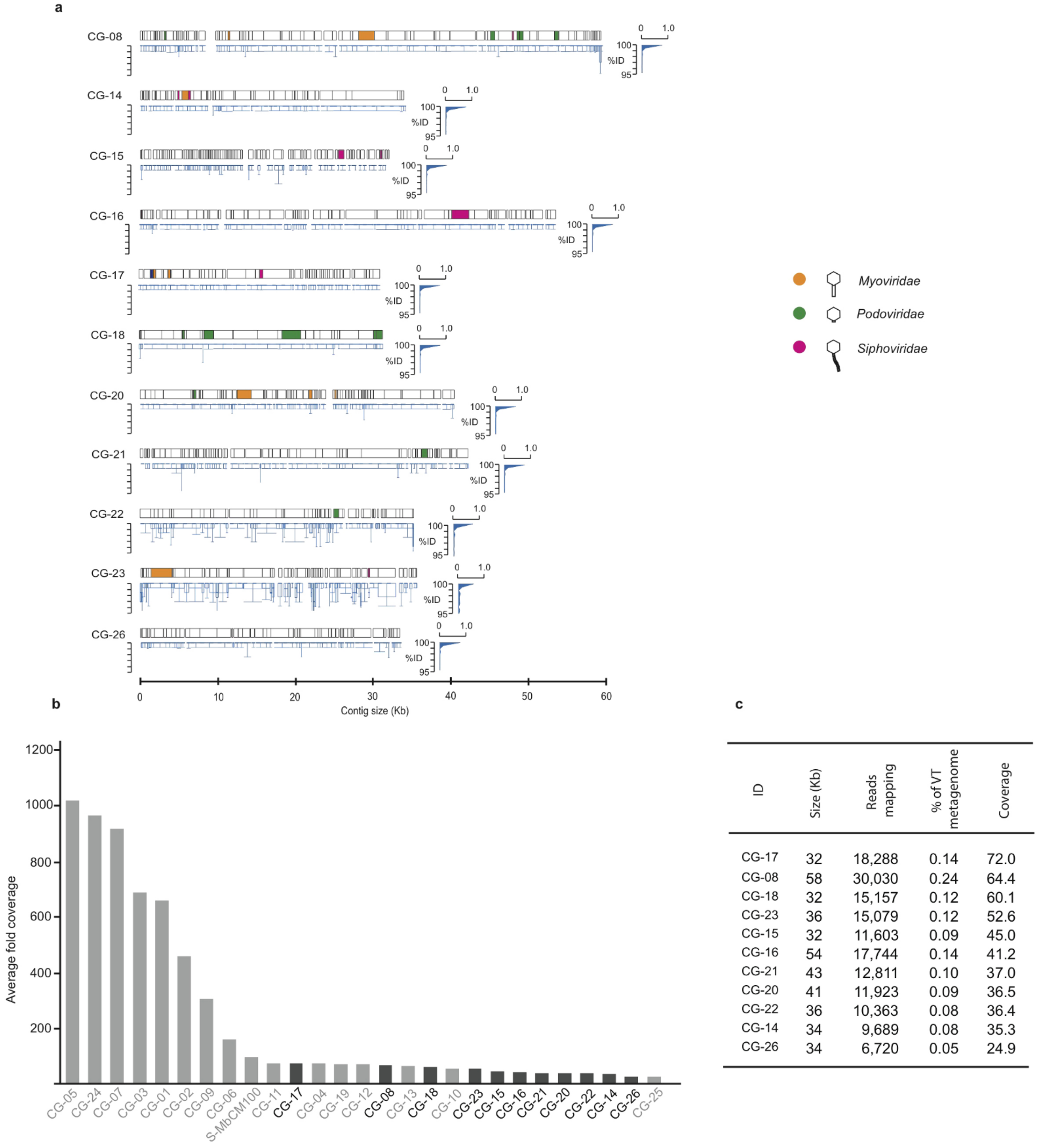
31. Waterbury, J. B., Watson, S. W., Valois, F. W. & Franks, D. G. Biological and ecological characterization of the marine unicellular cyanobacterium *Synechococcus*. *Can. Bull. Fish. Aquat. Sci.* **214,** 71–120 (1986).
32. John, S. G. *et al.* A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environ. Microbiol. Rep.* **3,** 195–202 (2011).
33. Hurwitz, B. L., Deng, L., Poulos, B. P. & Sullivan, M. B. Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. *Environ. Microbiol.* **15,** 1428–1440 (2013).
34. Deng, L. & Hayes, P. K. Evidence for cyanophages active against bloom-forming freshwater cyanobacteria. *Freshw. Biol.* **53,** 1240–1252 (2008).
35. Suttle, C. A. & Chan, A. M. Marine cyanophages infecting oceanic and coastal strains of *Synechococcus*: abundance, morphology, cross-infectivity and growth characteristics. *Mar. Ecol. Prog. Ser.* **92,** 99–109 (1993).
36. Waterbury, J. B. & Valois, F. W. Resistance to co-occurring phages enables marine *Synechococcus* communities to coexist with cyanophage abundant in seawater. *Appl. Environ. Microbiol.* **59,** 3393–3399 (1993).
37. Wilson, W. H., Joint, I. R., Carr, N. G. & Mann, N. H. Isolation and molecular characterization of five marine cyanophages propogated on *Synechococcus* sp. strain WH 7803. *Appl. Environ. Microbiol.* **59,** 3736–3743 (1993).
38. Fuller, N. J., Wilson, W. H., Joint, I. R. & Mann, N. H. Occurrence of a sequence in marine cyanophages similar to that of T4 g20 and its application to PCR-based detection and quantification techniques. *Appl. Environ. Microbiol.* **64,** 2051–2060 (1998).
39. Lu, J., Chen, F. & Hodson, R. E. Distribution, isolation, host specificity, and diversity of cyanophages infecting marine *Synechococcus* spp. in river estuaries. *Appl. Environ. Microbiol.* **67,** 3285–3290 (2001).
40. Chen, F. & Lu, J. Genomic sequence and evolution of marine cyanophage P60: a new insight on lytic and lysogenic phages. *Appl. Environ. Microbiol.* **68,** 2589–2594 (2002).
41. Marston, M. F. & Sallee, J. L. Genetic diversity and temporal variation in the cyanophage community infecting marine *Synechococcus* species in Rhode Island's coastal waters. *Appl. Environ. Microbiol.* **69,** 4639–4647 (2003).
42. Sullivan, M. B., Waterbury, J. B. & Chisholm, S. W. Cyanophages infecting the oceanic cyanobacterium *Prochlorococcus*. *Nature* **424,** 1047–1051 (2003).
43. Wang, K. & Chen, F. Prevalence of highly host-specific cyanophages in the estuarine environment. *Environ. Microbiol.* **10,** 300–312 (2008).
44. Kuznetsov, Y. G., Chang, S.-C., Credaroli, A., Martiny, J. & McPherson, A. An atomic force microscopy investigation of cyanophage structure. *Micron* **43,** 1336–1342 (2012).
45. Solonenko, S. A. & Sullivan, M. B. Preparation of metagenomic libraries from naturally occurring marine viruses. *Methods Enzymol.* **531,** 143–165 (2013).
46. Holmfeldt, K. *et al.* Twelve previously unknown phage genera are ubiquitous in the global oceans. *Proc. Natl Acad. Sci. USA* **110,** 12798–12803 (2013).
47. Hess, M. *et al.* Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* **331,** 463–467 (2011).
48. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376–380 (2005).
49. Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W. & Hauser, L. J. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11,** 119 (2010).
50. Niu, B., Fu, L., Sun, S. & Li, W. Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinformatics* **11,** 187 (2010).
51. Yooseph, S. *et al.* The Sorcerer II global ocean sampling expedition: expanding the universe of protein families. *PLoS Biol.* **5,** e16 (2007).
52. Chao, A. & Lee, S. M. Estimating the number of classes via sample coverage. *J. Am. Stat. Assoc.* **87,** 210–217 (1992).
53. Simpson, E. H. Measurement of diversity. *Nature* **163,** 688 (1949).
54. Akhter, S., Aziz, R. K. & Edwards, R. A. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combine similarity- and composition-based strategies. *Nucleic Acids Res.* **40,** e126 (2012).
55. Kettler, G. C. *et al.* Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet.* **3,** e231 (2007).
56. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5,** 113 (2004).
57. Dunn, J. C. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Cybern. Syst.* **3,** 32–57 (1973).
58. Coleman, M. L. *et al.* Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* **311,** 1768–1770 (2006).



Extended Data Figure 1 | Agarose gel of PCR products used for screening the 97 cyanophage isolates derived from this study. Primers used have a well-understood and strong history in the literature and amplify a ~400 bp region of the portal protein encoded gene (gp20) of T4-like phages. ΦC refers to phages S-MbC.

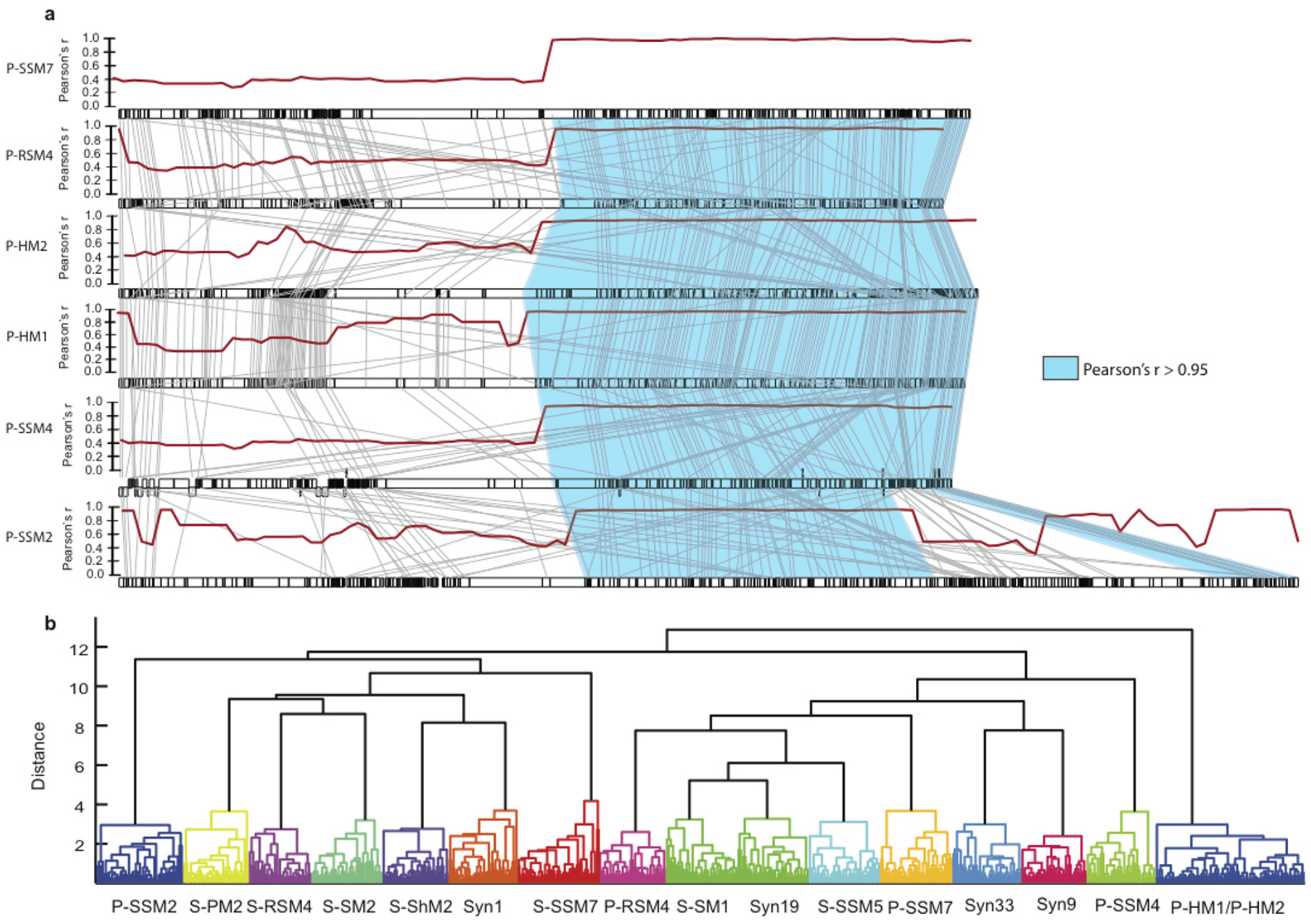


Extended Data Figure 2 | The viral-tagging metagenome is less complex than the whole viral community metagenome. **a**, Diversity of the viral-tagging (VT) metagenome shows a five-to-tenfold reduction when compared to the community metagenome by different metrics applied to protein clusters with only 17.5% (1,360 of 7,762) of the viral-tagging protein clusters occurring in the community metagenome (Venn diagram). **b**, The viral taxonomic profiles from each metagenome assigned by BLASTx search (*e*-value <0.001) against all phage genomes present in NCBI (1,218 genomes, December, 2013), and compared against the designations from cultured isolates (*n* indicates number of reads on top of metagenome bars and number of phage isolates on top of isolates bars; percentage of metagenome bars represent the percentage of reads used).

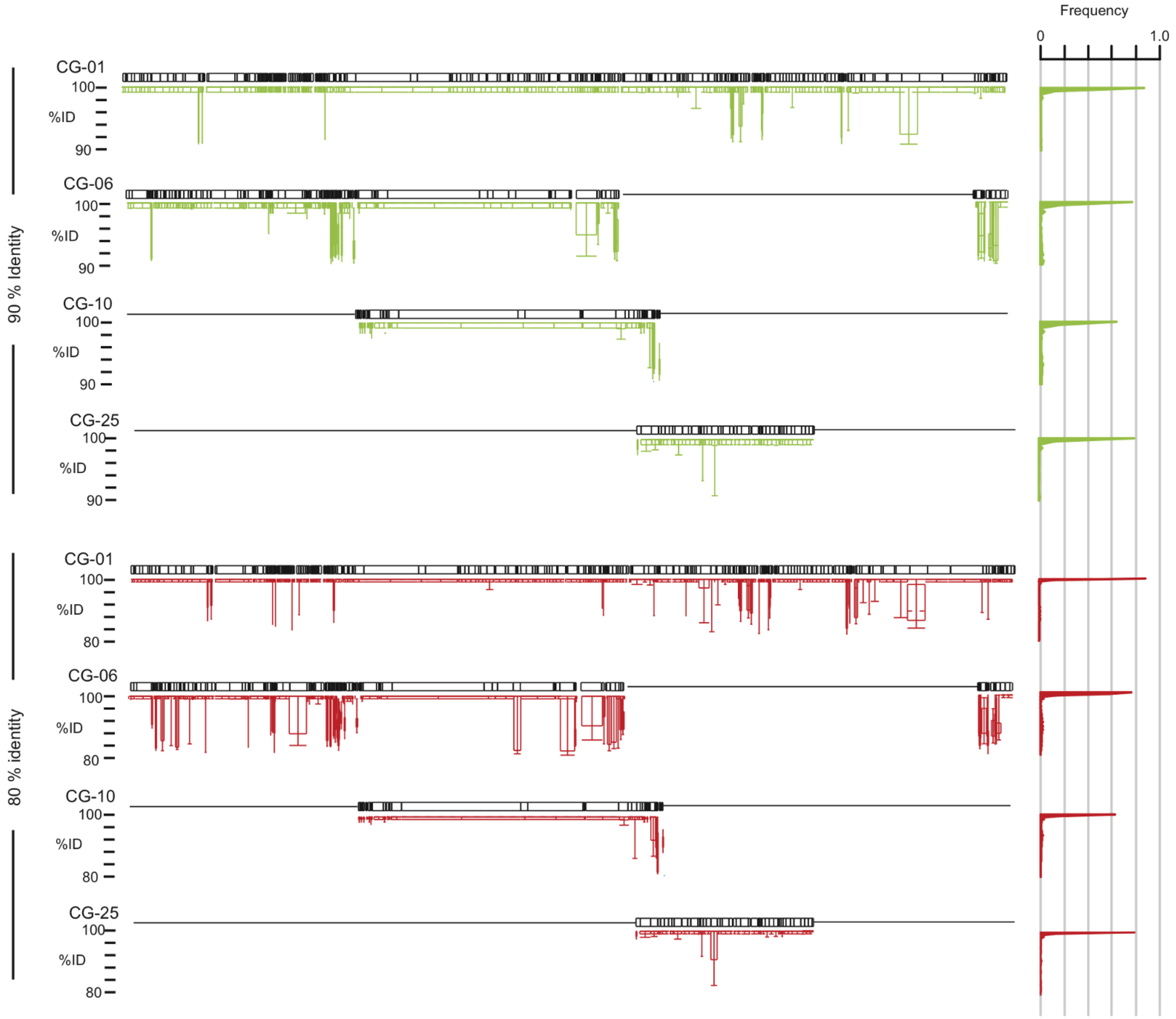


ID	Size (Kb)	Reads mapping	% of VT metagenome	Coverage
CG-17	32	18,288	0.14	72.0
CG-08	58	30,030	0.24	64.4
CG-18	32	15,157	0.12	60.1
CG-23	36	15,079	0.12	52.6
CG-15	32	11,603	0.09	45.0
CG-16	54	17,744	0.14	41.2
CG-21	43	12,811	0.10	37.0
CG-20	41	11,923	0.09	36.5
CG-22	36	10,363	0.08	36.4
CG-14	34	9,689	0.08	35.3
CG-26	34	6,720	0.05	24.9

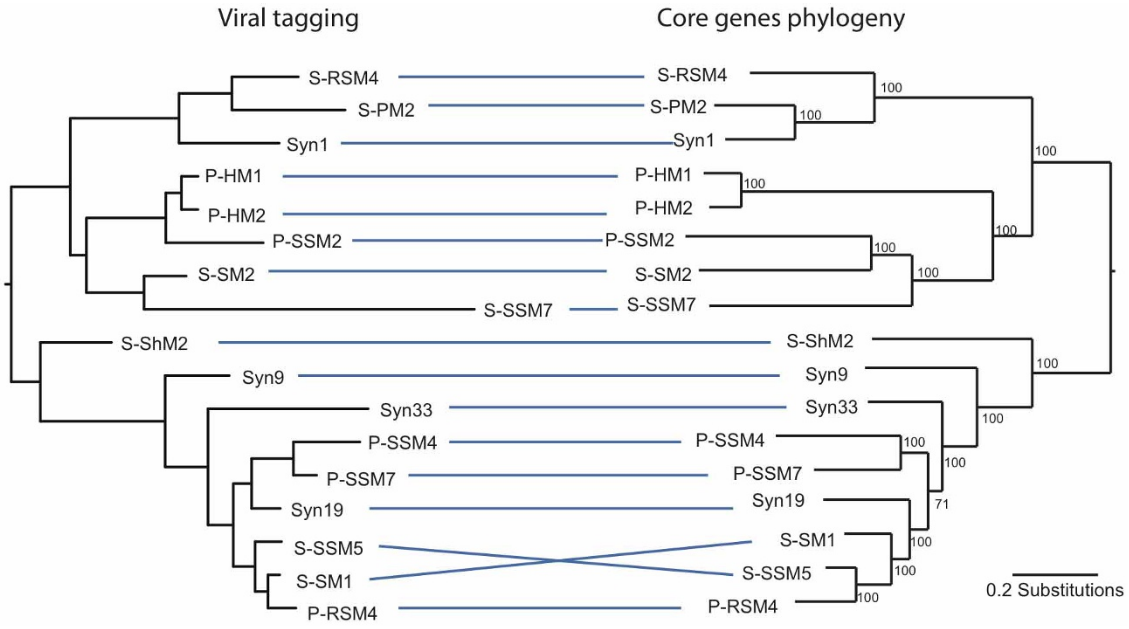
Extended Data Figure 3 | *Candidatus* genomes assembled from the dominant viral-tagging metagenome populations that were not T4-like myoviruses. a, Black boxes represent the predicted ORFs, blue box-plots reflect the intrapopulation locus-to-locus variation as described in Fig. 2. A cumulative distribution plot of the genome-wide locus-by-locus percentage nucleotide identity is represented to the right of each genome. Colours denote the taxonomic assignment for each gene, based on blastp best hits to nr, for detailed annotation refer to Supplementary Data 1. CGs, *Candidatus* genomes. **b**, Normalized (corrected for contig length) coverage of all *Candidatus* genomes including the rare non-T4-like ones (in dark grey). **c**, Quantification of the relative abundances of the non-T4-like *Candidatus* genomes.



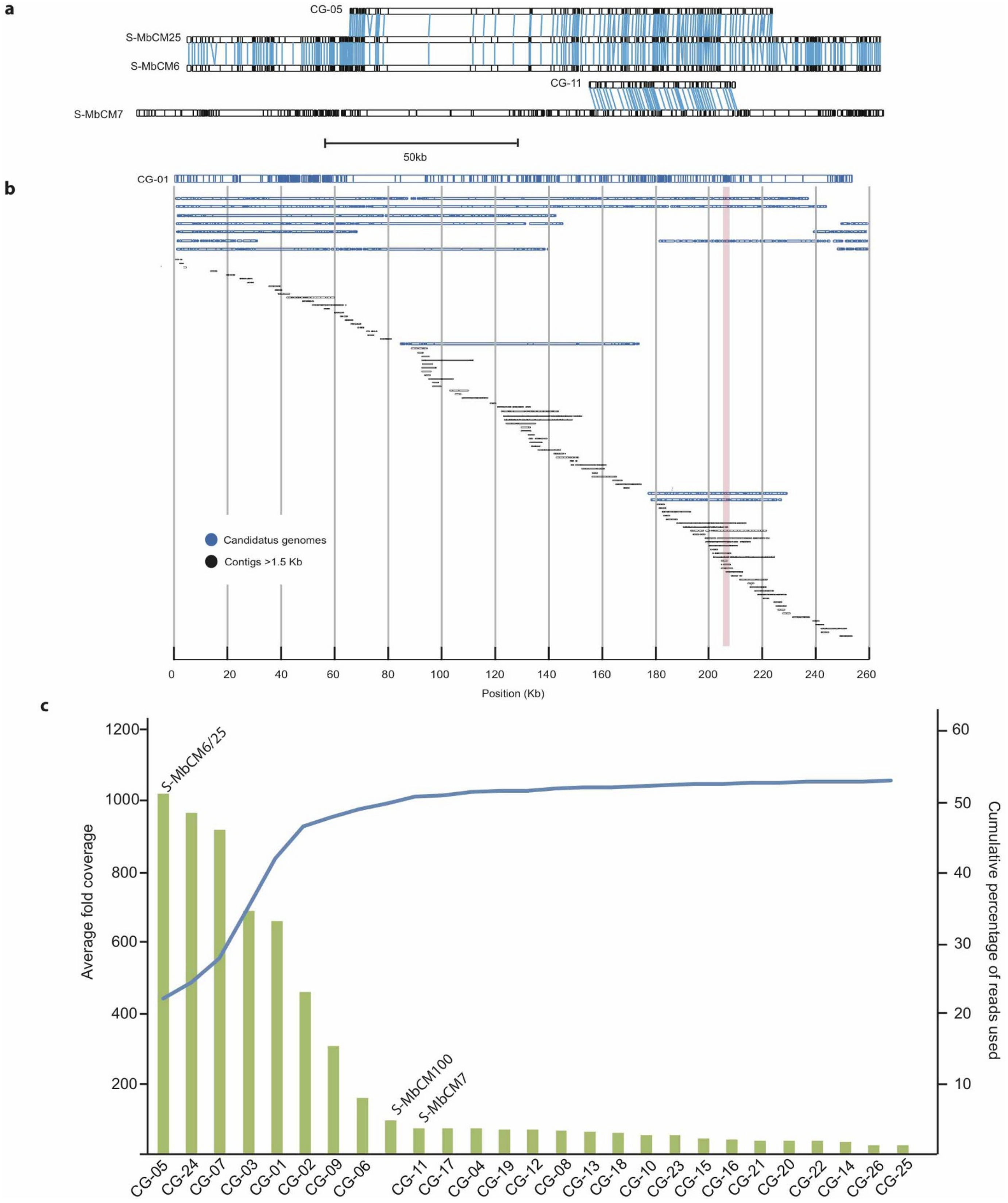
Extended Data Figure 4 | Genome sizing and location *in silico* experiments. **a**, Variation along the genome was investigated by *in silico* breaking the genome into 30 kb fragments using a 5 kb sliding window to compare the similarity profile (ANI of the genome or fragment versus the reference genomes) of the fragment to that derived from the whole genome, just a subset shown. Where Pearson's *r* is high (>0.95, the right side of the genomes in blue) the fragment profile parallels that of a genome-wide profile. **b**, These similarity profiles were converted to a correlation distance (1 − Pearson's *r*) and then clustered using hierarchical clustering (linkage = 'complete' or furthest distance). Comparison of the clustering patterns showed that 30 kb fragments from within the 'blue' region of the genome are more closely related to those derived from their own genomes than other genomes, except for the co-isolated phages P-HM1 and P-HM2, which were the most similar genomes in the data set.



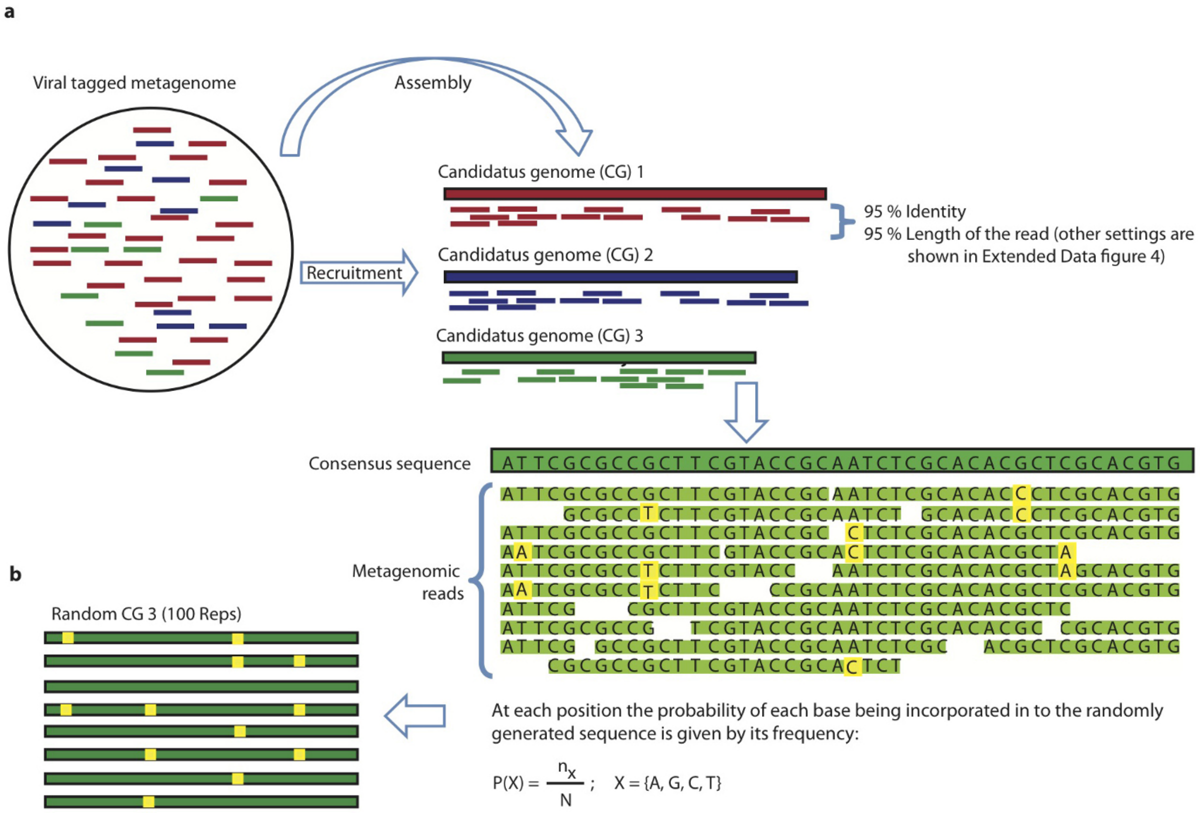
Extended Data Figure 5 | Sensitivity analyses for recruitment parameters. Recruitment, as described in the main text, required 95% nucleotide identity over 95% of the length of the reads, whereas here we examine lower stringency recruitment including 90% nucleotide identity over 90% of the length of the read, and 80% nucleotide identity over 90% of the length of the read. These results were consistent with those described in the main text (Fig. 2)—that is, most of the recruited reads are in the top 2% (see histogram on the right). Only four representative *Candidatus* genomes (CGs) are shown here.



Extended Data Figure 6 | Comparison of neighbour-joining trees derived from viral-tagging and phylogenomic analyses. The left panel represents Euclidean distances of the three-dimensional space reconstructed with the first three principal components in Fig. 1. The right panel represents the currently accepted cyano-T4 phage core phylogeny derived from analysis of 57 concatenated proteins totalling 20,638 amino acids.



Extended Data Figure 7 | Exploring viral-tagging metagenomic population sequence space. **a**, Whole genome comparisons of isolates S-MbCM25 and S-MbCM6 show that they are part of the same population as CG-05, while isolate S-MbCM7 appears to be part of the CG-11 population (lines connecting reciprocal blast hits >95% identity). Note in Fig. 1 how the variation ('cloud') associated with CG-05 and CG-11 overlap with their representative isolates. **b**, Alignment, based on homologues sequences within each contig, of all assembled T4-like viral-tagging contigs (>1.5 kb) against CG-01 as a reference genome. At the deepest point (around the 205 kb mark, orange bar) there are a total of 14 to 17 overlapping contigs. **c**, Rank abundance curve for the 26 most abundant *Candidatus* genomes (CGs) in the viral-tagging source waters. Values are derived from mean contig coverage values. The blue line quantifies the cumulative use of reads as more genomes are added.


a
Viral tagged metagenome
Assembly
Recruitment
Candidatus genome (CG) 1
Candidatus genome (CG) 2
Candidatus genome (CG) 3
95 % Identity
95 % Length of the read (other settings are shown in Extended Data figure 4)
Consensus sequence
ATTCGCGCCGCTTCGTACCGCAATCTCGCACACGCTCGCACGTG
Metagenomic reads
ATTCGCGCCGCTTCGTACCGCAATCTCGCACACCCTCGCACGTG
GCGCCTCTTCGTACCGCAATCT GCACACCCTCGCACGTG
ATTCGCGCCGCTTCGTACCGC CTCTCGCACACGCTCGCACGTG
AATCGCGCCGCTTCGTACCGCACTCTCGCACACGCTA
ATTCGCGCCTCTTCGTACC AATCTCGCACACGCTAGCACGTG
AATCGCGCCTCTTC CCGCAATCTCGCACACGCTCGCACGTG
ATTCG CGCTTCGTACCGCAATCTCGCACACGCTC
ATTCGCGCCG TCGTACCGCAATCTCGCACACGC CGCACGTG
ATTCG GCCGCTTCGTACCGCAATCTCGC ACGCTCGCACGTG
CGCGCCGCTTCGTACCGCACTCT
b
Random CG 3 (100 Reps)

At each position the probability of each base being incorporated in to the randomly generated sequence is given by its frequency:

$$P(X) = \frac{n_x}{N}; \quad X = \{A, G, C, T\}$$

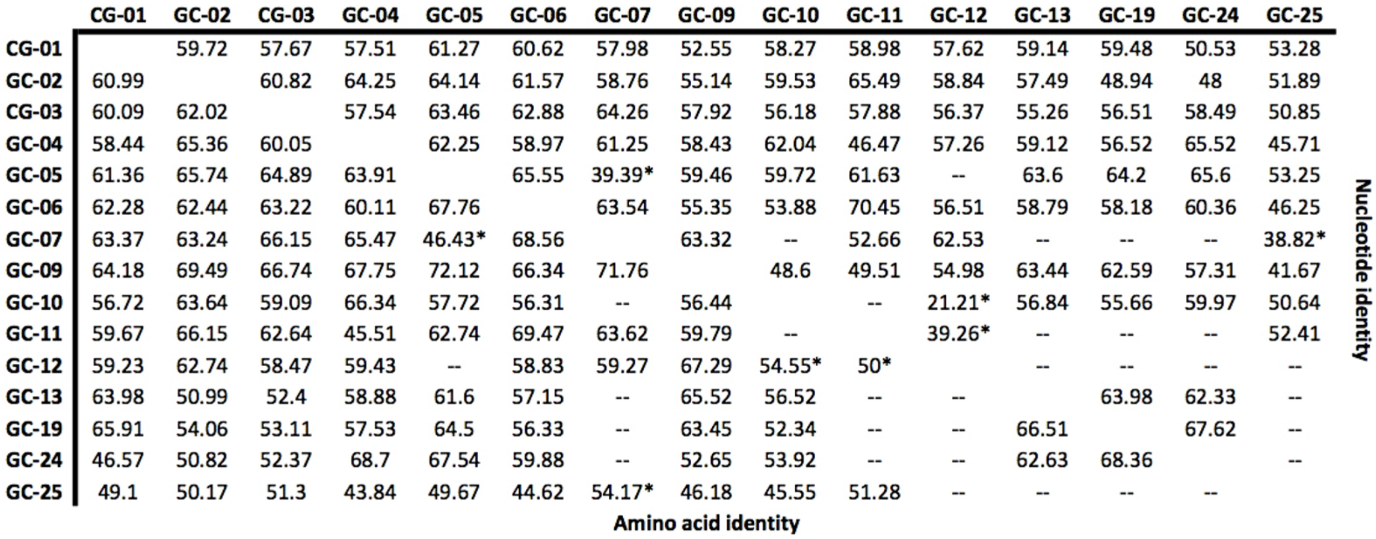
Extended Data Figure 8 | Flow diagram describing the bioinformatics processing steps. **a**, We recruited reads to each *Candidatus* genome requiring at least 95% identity and a coverage of 95% of the entire length of the read. Each read was non-redundantly assigned and aligned to a *Candidatus* genome using default parameters in MUSCLE. **b**, For each *Candidatus* genome population, we generated 100 random *Candidatus* genome sequences by probabilistically resampling (using the observed occurrences) the metagenomic data that went into generating their consensus sequences.

Extended Data Table 1 | Phage isolates used in the study of phage infectivity indication by viral tagging and plaques assay

Phage Name	Locale	Latitude and Longitude	Date Collected	TEM Morphology	Isolation host	Published	This study	Percent of viral tagged cells (%; VBR=10, T=20min; n=4)	Reference
P-RSM2	Red Sea	29°28'N, 34°55'E	Sep-00	M	*P* NATL1A	●	●	4.40±0.75	39
P-RSM3	Red Sea	29°28'N, 34°55'E	Sep-00	M	*P* NATL2A	●	●	41.18±2.21	39
P-SSM1	BATS	31°48'N, 64°16'W	Jun-00	M	*P* MIT9303	○	○	0	39
P-SSM2	BATS	31°48'N, 64°16'W	Jun-00	M	*P* NATL1A	○	○	0	39
P-SSM4	BATS	31°48'N, 64°16'W	Jun-00	M	*P* NATL2A	○	●	5.72±0.71	39
P-SSM6	BATS	31°48'N, 64°16'W	Sep-99	M	*P* NATL2A	○	●	3.54±0.52	39
P-RSP2	Red Sea	29°28'N, 34°53'E	Jul-00	P	P MIT9302	○	○	0	39
P-SSP5	BATS	31°48'N, 64°16'W	Sep-99	P	*P* MIT9515	○	○	0	39
P-SSP6	BATS	31°48'N, 64°16'W	Sep-99	P	*P* MIT9515	○	○	0	39
P-SSP7	BATS	31°48'N, 64°16'W	Sep-99	P	*P* MED4	○	○	0	39
P-SS1	Slope	37°40'N, 73°30'W	Sep-01	S	*P* MIT9313	○	○	0	39
P-SS2	Slope	37°40'N, 73°30'W	Sep-01	S	*P* MIT9313	○	○	0	39
S-PM2	English Channel	50°18'N, 4°12'W	Sep-92	M	*S* WH7803	●	●	42.46±1.10	56
S-SM1	Slope	37°40'N, 73°30'W	Sep-01	M	*S* WH6501	○	○	0	39
S-SSM1	Sargasso Sea	34°24'N, 72°03'W	Sep-01	M	*S* WH6501	○	○	0	39
S-SSM2	Sargasso	34°24'N, 72°03'W	Sep-01	M	*S* WH8102	●	●	45.26±4.73	39
S-WHM1	Woods Hole	41°31'N, 71°40'W	Aug-92	M	*S* WH7803	●	●	49.14±4.84	55
Syn-1	Woods Hole	41°31'N, 71°40'W	Aug-90	M	*S* WH8101	●	●	47.91±1.03	55
Syn-19	Sargasso Sea	34°06'N, 61°01'W	Jul-90	M	*S* WH8012	●	●	43.25±1.76	55
Syn-2	Sargasso Sea	34°06'N, 61°01'W	Jul-90	M	*S* WH8109	●	●	44.28±2.44	55
Syn-33	Gulf Stream	25°51'N, 79°26'W	Jan-95	M	*S* WH7803	●	●	45.16±1.25	55
Syn-9	Woods Hole	41°31'N, 71°40'W	Oct-90	M	*S* WH8109	●	●	41.48±2.21	55
Syn-5	Sargasso Sea	34°06'N, 61°01'W	Jul-90	P	*S* WH8012	○	○	0	55
Φ17:2	Sweden	56°2'N, 12°37'E	2005	P	#17	○	○	0	63
Φ4:1	Sweden	56°2'N, 12°37'E	2005	P	#4	○	○	0	63
Φ46:4	Sweden	56°2'N, 12°37'E	2005	P	NN016046	○	○	0	63

See also Methods. The first letter of phage name refers to: P, *Prochlorococcus* phages; S, *Synechococcus* phages; Φ, *Bacteriodetes* phages. Phages were grouped according to their isolation locations. Transmission electron microscopy morphology designations were as follows: P, *Podoviridae*; M, *Myoviridae*; S, *Siphoviridae*. Open circles indicate negative infection and filled circles indicate positive infection by plaque assay on *Synechococcus* WH7803 either previously published or from this study.

Extended Data Table 2 | Pairwise nucleotide and amino acid identity calculated between all shared genes for each pair of *Candidatus* genomes



	CG-01	GC-02	CG-03	GC-04	GC-05	GC-06	GC-07	GC-09	GC-10	GC-11	GC-12	GC-13	GC-19	GC-24	GC-25
CG-01		59.72	57.67	57.51	61.27	60.62	57.98	52.55	58.27	58.98	57.62	59.14	59.48	50.53	53.28
GC-02	60.99		60.82	64.25	64.14	61.57	58.76	55.14	59.53	65.49	58.84	57.49	48.94	48	51.89
CG-03	60.09	62.02		57.54	63.46	62.88	64.26	57.92	56.18	57.88	56.37	55.26	56.51	58.49	50.85
GC-04	58.44	65.36	60.05		62.25	58.97	61.25	58.43	62.04	46.47	57.26	59.12	56.52	65.52	45.71
GC-05	61.36	65.74	64.89	63.91		65.55	39.39*	59.46	59.72	61.63	--	63.6	64.2	65.6	53.25
GC-06	62.28	62.44	63.22	60.11	67.76		63.54	55.35	53.88	70.45	56.51	58.79	58.18	60.36	46.25
GC-07	63.37	63.24	66.15	65.47	46.43*	68.56		63.32	--	52.66	62.53	--	--	--	38.82*
GC-09	64.18	69.49	66.74	67.75	72.12	66.34	71.76		48.6	49.51	54.98	63.44	62.59	57.31	41.67
GC-10	56.72	63.64	59.09	66.34	57.72	56.31	--	56.44		--	21.21*	56.84	55.66	59.97	50.64
GC-11	59.67	66.15	62.64	45.51	62.74	69.47	63.62	59.79	--		39.26*	--	--	--	52.41
GC-12	59.23	62.74	58.47	59.43	--	58.83	59.27	67.29	54.55*	50*		--	--	--	--
GC-13	63.98	50.99	52.4	58.88	61.6	57.15	--	65.52	56.52	--	--		63.98	62.33	--
GC-19	65.91	54.06	53.11	57.53	64.5	56.33	--	63.45	52.34	--	--	66.51		67.62	--
GC-24	46.57	50.82	52.37	68.7	67.54	59.88	--	52.65	53.92	--	--	62.63	68.36		--
GC-25	49.1	50.17	51.3	43.84	49.67	44.62	54.17*	46.18	45.55	51.28	--	--	--	--	

Amino acid identity

Nucleotide identity

The absence of values indicates that no overlap in gene content (see Fig. 2) was observed between the two *Candidatus* genomes (CGs).
* Only two genes shared.