ARTICLE

doi:10.1038/nature19094

Uncovering Earth's virome

David Paez-Espino[1], Emiley A. Eloe-Fadrosh[1], Georgios A. Pavlopoulos[1], Alex D. Thomas[1], Marcel Huntemann[1], Natalia Mikhailova[1], Edward Rubin[1,2,3], Natalia N. Ivanova[1] & Nikos C. Kyrpides[1]

Viruses are the most abundant biological entities on Earth, but challenges in detecting, isolating, and classifying unknown viruses have prevented exhaustive surveys of the global virome. Here we analysed over 5 Tb of metagenomic sequence data from 3,042 geographically diverse samples to assess the global distribution, phylogenetic diversity, and host specificity of viruses. We discovered over 125,000 partial DNA viral genomes, including the largest phage yet identified, and increased the number of known viral genes by 16-fold. Half of the predicted partial viral genomes were clustered into genetically distinct groups, most of which included genes unrelated to those in known viruses. Using CRISPR spacers and transfer RNA matches to link viral groups to microbial host(s), we doubled the number of microbial phyla known to be infected by viruses, and identified viruses that can infect organisms from different phyla. Analysis of viral distribution across diverse ecosystems revealed strong habitat-type specificity for the vast majority of viruses, but also identified some cosmopolitan groups. Our results highlight an extensive global viral diversity and provide detailed insight into viral habitat distribution and host–virus interactions.

Viruses are the most abundant entities across all habitats, and a major reservoir of genetic diversity[1] affecting biogeochemical cycles and ecosystem dynamics[1]. Exploration of viral populations in oceans of the world and within the human microbiome has illuminated considerable genetic complexity[2,3]; however, there are significant gaps in the global virome catalogue. There are an estimated 10^{31} viral particles infecting microbial populations[4]; yet fewer than 2,200 genomes from double-stranded DNA (dsDNA) viruses and retroviruses are deposited in NCBI, compared to over 45,000 bacterial genomes[5]. Culture-independent approaches have provided a broader view of the diversity and distribution of dsDNA viruses[6]. However, their accurate detection and quantification using targeted sequencing remains challenging owing to the lack of universally conserved genomic signatures and complex experimental protocols[7].

Beyond gaps in characterized diversity, the scope of host–viral interactions is poorly understood, although it has been hypothesized that all cellular organisms are prey to viral attack[8]. Methods for studying host–viral interactions rely almost exclusively on cultured virus–host systems; however, recent *in silico* approaches have revealed that a much broader range of hosts is susceptible to viral infections[9,10]. Given the role that viruses play in host metabolism reprogramming, gene flow, and structuring of microbial communities, it is critical to capture viral linkages with their hosts.

Currently, a plethora of metagenomic data exists that present a unique opportunity for viral sequence discovery[11]. Although most of these data sets were generated by untargeted approaches without viral particle enrichment, they contain a wealth of viral sequences. Here, we developed a computational approach to explore the viral content of more than 3,000 metagenomic samples. We uncovered 2.1 Gb of viral sequence data, which increases the known viral sequence space by an order of magnitude, enables the prediction of previously unknown host–viral interactions and provides a global view of viral biogeography.

Global expansion of viral sequence space

In the absence of universally conserved markers, previous studies attempted to identify viruses using proteins present exclusively in viruses[12]. To overcome the limitations of a biased collection of isolate viruses (iVGs), we complemented the viral protein families of the iVGs (derived from dsDNA viruses and retroviruses in the NCBI database) with a set of viral protein families from 1,800 manually identified metagenomic viral contigs (mVCs). This set was used as a bait to identify putative viral sequences in a large collection of assembled metagenomic contigs longer than 5 kb (Methods; Extended Data Figs 1–3; Supplementary Tables 1–7). These contigs were obtained from 3,042 metagenomes in the Integrated Microbial Genomes with Microbiome Samples (IMG/M) system[11] (Supplementary Table 8), representing a collection of geographically and ecologically diverse samples according to metadata from the Genomes OnLine Database[5,13]. This led to the identification of 125,842 putative DNA metagenomic viral contigs, increasing the viral sequence size in base pairs by 17.3-fold and the number of viral genes by 16.6-fold (Fig. 1a; Methods). These encode more than 2.79 million proteins, 75% of which have no sequence similarity to proteins from known isolate viruses, consistent with previous studies[12,14]. Sequence similarity clustering of proteins encoded by the mVCs resulted in a total of 418,541 clusters with 2 or more members and 765,991 singletons (Methods; Supplementary Table 9). Benchmarking was performed to validate our computational pipeline, and indicated that 70% of the sequences identified in this study would have been missed by other methods (Methods; Extended Data Fig. 3; Supplementary Tables 2–6).

To evaluate the coverage of the viral protein space by the newly identified sequences, we estimated the rate of accumulation of protein clusters as a function of the number of samples (Fig. 1b). In agreement with recent reports[9,15], the curves of cluster accumulation in the two most heavily sampled habitats, human-associated and marine, appear to reach saturation. However, the rate of cluster discovery does not plateau when all samples are considered, suggesting that the global viral sequence space is largely uncharacterized.

To compare the coverage of mVCs and iVGs by viral protein families, we calculated the percentage of genes with hits to viral protein families relative to the total number of genes on each contig (Fig. 1c). On the basis of this percentage, viral contigs were classified into three categories: those with at least 70% of genes in viral families (highly covered with strong similarity to viruses in the training set); those with 35–70% of genes in viral families; and those with less than 35% of genes in viral families (low covered with low similarity to viruses

[1]Department of Energy, Joint Genome Institute, Walnut Creek, California 94598, USA. [2]Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. [3]Metabiota Inc., San Francisco, California 94104, USA.

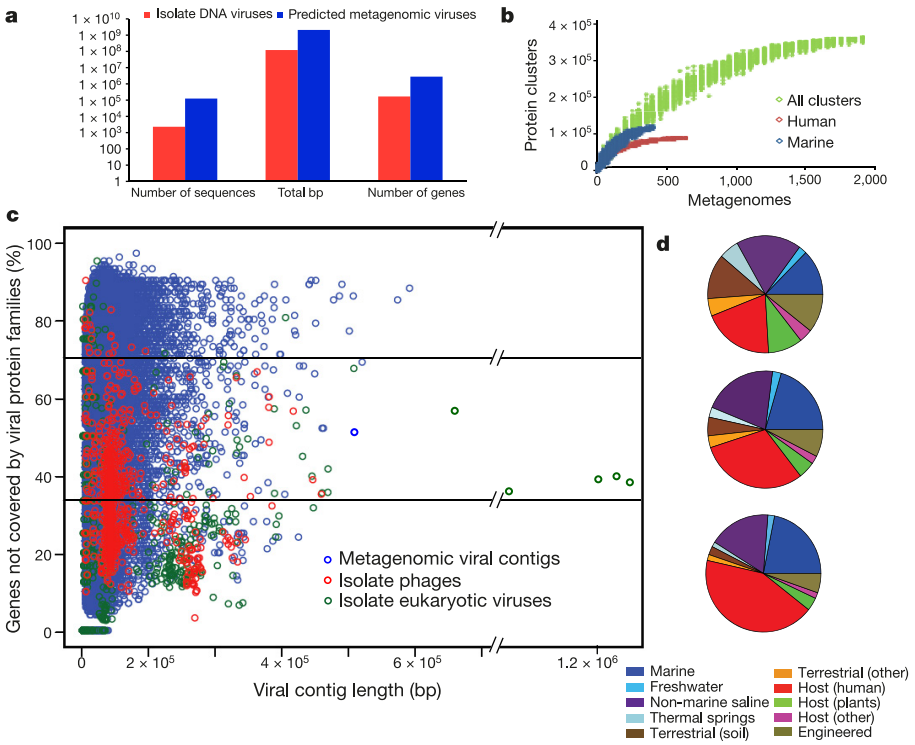


Figure 1 | Identification of metagenomic viral sequences and habitat distribution. a, Number of metagenomic viral contigs compared to isolate viral genomes. **b**, Accumulation curves showing the protein cluster growth with increased sampling. Green, blue, and red represent clusters from all, aquatic, or human metagenomes, respectively. The ranges represent 100 random replicates. **c**, Distribution of metagenomic viral contigs (blue), isolate phages (red) and isolate eukaryotic viruses (green) by length and coverage with viral protein families. Viral sequences are divided into three categories: high, medium and low coverage with protein families. **d**, Distribution of metagenome viral contigs by habitat type is shown.

in the training set) (Extended Data Fig. 4a). The highly covered category included 67% of isolate viruses, but only 24.5% of mVCs (Fig. 1c), the majority of them from marine and human-associated habitats, where more reference viruses were available (Fig. 1d). Another 24.2% of mVCs placed in the low-covered category were typically found in soil, plant-associated, and engineered samples (Fig. 1d). The differences were even more pronounced when the data were normalized by total sequence length per habitat (Methods), suggesting the need for more extensive sampling of these environments.

The length of mVCs ranges from 5 kb to nearly 600 kb (average 16,625 ± 18,057 bp) (Fig. 1c). On the basis of the end overlaps, 999 of mVCs were probably circular, representing complete viral genomes (Supplementary Table 10). The average size of the circular mVCs (53,644 ± 45,677 bp) is consistent with the calculated average length of isolate dsDNA viruses (44,296 ± 83,777 bp) (Supplementary Information). Among circular contigs, we identified the largest phage recovered to date, a 596 kb contig from a bioreactor sample[16], with many signature genes of tailed viruses, but no recognizable housekeeping genes of bacteria or plasmids (Methods; Extended Data Fig. 4b; Supplementary Table 11; Supplementary Information). We identified six more mVCs ranging from 350 to 470 kb, probably representing fragments of other large phage genomes (Supplementary Table 12). As the sizes of viral particles and viral genomes are correlated[17], these mVCs found in many ecological niches point to the hidden diversity and abundance of very large phages, probably avoiding detection by the conventional enrichment methods.

Sequence grouping to gauge viral diversity

To quantify the amount of taxonomic diversity, mVCs and isolate viral genomes were clustered into quasi-species groups on the basis of the average amino acid identity[18] (AAI) of all proteins and single-linkage clustering, using an approach analogous to the whole-genome-based classification scheme developed for prokaryotes[19] (Methods; Extended Data Fig. 5a; Supplementary Table 13). 64,160 mVCs and 2,536 isolate contigs were clustered into 18,470 viral groups, ranging from 2 to 365 members per group. Most groups (57%) had only 2 members and only 3.7% had more than 10 members (Extended Data Fig. 5b). Similar to previous studies[12,14,20], the vast majority of viral groups (95.9%) did not contain isolate viruses.

218 viral groups and 842 singletons contained at least one iVG with genus- and species-level taxonomic assignment according to the International Committee on Taxonomy of Viruses (Methods; Supplementary Tables 14, 15). Our method recapitulates current species-level groupings in 87% of the cases with the remainder grouping at genus level (Supplementary Table 14; Extended Data Fig. 5c–e). We compared our method with sequence-based classification used in previous studies, which applied protein cluster occurrence to generate mostly genus-level groups[21]. In agreement with an assessment that our groups represent quasi-species, our approach resulted in smaller clusters and more singletons (Supplementary Information). Next we proceeded to predict host specificity and determine environmental distribution of these species-level viral groups and singletons.

Host–virus connectivity revealed

We used a suite of computational methods to identify putative host–virus connections. First, we projected the isolate viral–host information onto a group, resulting in host assignments for 2.4% of viral groups (Fig. 2a). Then we used the CRISPR–Cas prokaryotic immune system, which holds a 'library' of genome fragments from phages (proto-spacers) that have previously infected the host[22]. These fragments retained by the host in the form of spacers can be matched to phage genomes linking phages with their hosts[23–25]. We amassed a database of 3.5 million spacers from prokaryotic isolate genomes and metagenomes in IMG (Supplementary Tables 16, 17). As a control, 98.5% of spacer matches against isolate viral genomes agreed with their known host specificity at genus or species level (Methods; Supplementary Information; Supplementary Table 18). Spacers from isolate microbial genomes with matches to mVCs were identified for 4.4% of the viral

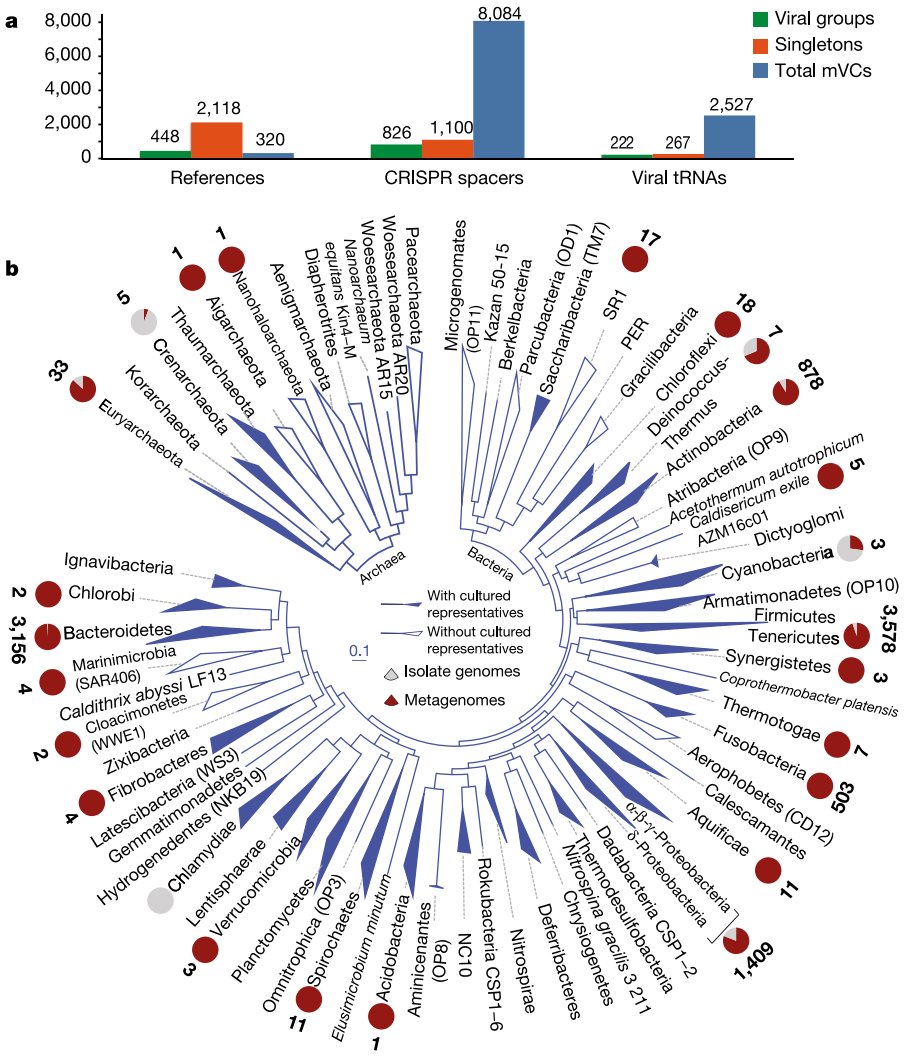


Figure 2 | Host–virus connectivity. a, Total number of host assignments to metagenomic viruses with three approaches. Total assignments for viral groups, singletons, and metagenomic viral contigs are shown above each bar. **b**, Phylogenetic distribution of bacterial and archaeal hosts. For each phylum a pie chart indicates the fraction of sequences assigned to this phylum from metagenomic viral contigs (red), and isolate viruses (grey). The number of metagenomic viral contigs assigned to each phylum is indicated by the numbers next to pie charts. Clades in blue represent phyla with cultivated representatives. Clades in white represent candidate phyla without cultured representatives.

groups and 1.7% of singletons (Supplementary Table 19). Finally, we explored the hypothesis that viral transfer RNA (tRNA) genes originate from their host[26]. Using stringent sequence identity cutoffs, viral tRNAs identified in 7.6% of the mVCs were matched to isolate genomes from a single species or genus (Methods; Supplementary Information; Supplementary Tables 20–22). The specificity of tRNA-based host-viral assignment was confirmed by CRISPR–Cas spacer matches showing a 94% agreement at the genus level (Supplementary Table 19).

Overall, these approaches identified 9,992 putative host–virus associations enabling host assignment to 7.7% of mVCs. The majority of these connections were previously unknown, and include hosts from 16 prokaryotic phyla for which no viruses have previously been identified (Supplementary Table 23), such as the first instance of viruses infecting the candidate phylum SR1 (Fig. 2b). We also connected mVCs to pathogenic species for which no viral connections were known, including *Fusobacterium* and *Leptotrichia* that cause oral and skin infections in mammals (Supplementary Information; Supplementary Table 24). The discovery of phages infecting these and other pathogens could be exploited for phage therapy applications[27,28].

It is widely assumed that most viruses specialize in infecting related hosts, as broad host range is negatively correlated with infection success[29]. However, this may be an artefact[29], and viral generalists that infect hosts across taxonomic orders do exist[30]. Our data suggested a trend for narrow host range with some notable exceptions. Whereas most CRISPR spacer matches were from viral sequences to hosts within one species or genus (Fig. 3a), some mVCs were linked to multiple hosts from higher taxa, including different phyla. A viral group comprised of mVCs from human oral samples contained three distinct proto-spacers with nearly exact matches to spacers in Actinobacteria and Firmicutes (Fig. 3b). In another case (Fig. 3c), proto-spacers from two mVCs derived from faecal samples were linked to spacers in three distinct Clostridiales families (Extended Data Fig. 6; Supplementary Information). As viruses exploit the host transcription/translation machinery, the existence of viruses with a surprisingly broad range of hosts opens opportunities for identification of novel enzymes or regulatory sequences, with biotechnological applications.

Biogeographic patterns of viral diversity

Previous studies of viral biogeography mainly focused on single habitats[3,31,32], and only a handful of small-scale studies explored viromes across environments[33–35]. As our pipeline was designed to identify longer viral contigs that probably represent more abundant populations, we explored the dispersal of the predicted viruses by aligning the contigs against all assembled and unassembled

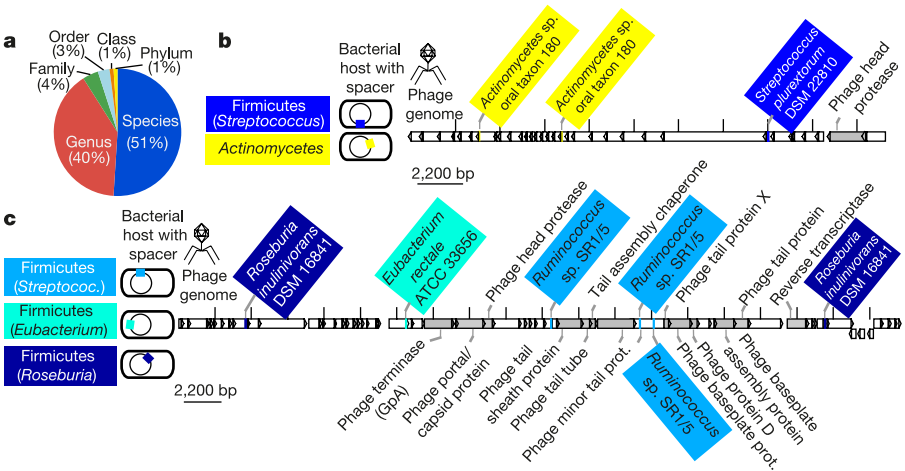


Figure 3 | Expanded host-range specificity identifies viral generalists. a, Proportion of viral groups connected with predicted hosts at various taxonomic levels. **b**, Three proto-spacers encoded on mVCs identified in human oral metagenomic samples that were linked to CRISPR spacers from hosts from distinct phyla, *Actinomycetes* sp. oral taxon 180 (Actinobacteria) and *Streptococcus plurextorum* DSM 22810 (Firmicutes). **c**, Seven proto-spacers encoded on two nearly identical mVCs deriving from independent faecal samples, and linked to CRISPR spacers from *Roseburia inulinivorans* DSM 16841 (Lachnospiraceae family), *Eubacterium rectale* ATCC 33656 (Eubacteriaceae family), and *Ruminococcus* sp. SR 1/5 (Ruminococcaceae family) (details in Supplementary Information).

metagenome sequences (Methods; Supplementary Information). This approach revealed that 86% of the viral sequences were found in more than one sample, whereas 73% were present in at least 5 samples (Extended Data Fig. 7a), mostly from relatively well-sampled marine and human-associated habitats. This enabled a detailed investigation of viral distribution patterns across these environments (Fig. 4).

The distribution of viral sequences in marine samples is characterized by distinct spatial patterns based on water column depth and distance from the shore (Fig. 4a). Although viral assemblages in coastal waters from distant biogeographical provinces are markedly different, bathypelagic samples from different oceanic basins display very similar viral profiles, in agreement with observations that deep ocean phylum-level composition of microbial communities is relatively uniform[36]. Although viral sequences are mostly partitioned into zone-specific groups, some are present in diverse samples across zones and oceanic provinces, including one viral group found in 95% of all twilight samples and in 44% of deep ocean samples (Extended Data Fig. 7b–c).

The distribution of viral sequences in human microbiome samples (Fig. 4b) also shows clear body-site specificity with only a few viral groups and singletons found in both faecal and oral samples (Supplementary Tables 25, 26). In contrast to previous studies[2,37–39], many viral sequences, mostly of phage origin, were shared between samples from the same body site of unrelated individuals. More than 30% of intestinal and 50% of oral viral sequences were shared by at least 10% of sampled subjects (Extended Data Fig. 7d, e). Approximately 0.5% of sequences in both body sites were shared by more than 80% of sampled individuals, whereas 17% and 9% of intestinal and oral viral sequences, respectively, were unique to each individual. We used raw sequencing reads to estimate the amount of viral sequences in 550 faecal and oral samples. Viral fraction varied from 0.2 to 54% of the total amount of high quality sequence in the sample, with the average of 3.4% in oral samples and 7.4% in stool samples, which is higher than previously reported 2.5 to 3.5% in stool[2] (Supplementary Information).

Although 84% of our quasi-species viral groups found in multiple samples resided within a single habitat type (Fig. 5a), 14% were found in two habitat types, typically, within the same broader environmental category (Fig. 5b, c), and a small number of groups were spanning two or more environmental categories (Fig. 5c, d; Supplementary Information; Supplementary Table 27). Most of these were due to uncertainty of habitat classification (for example, plant rhizosphere samples classified as host-associated) (Fig. 5c). A more detailed analysis of the most ubiquitous viral sequences revealed that they are probably human and laboratory contaminants, including ΦX and λ phages used as vectors, sequencing and molecular weight standards[40], and *Propionibacterium acnes* phages, common inhabitants of human skin. Several viruses recovered in a wide variety of environments were found to be prophages with broad host specificity (Supplementary Information; Supplementary Table 28) infecting hosts with different habitat preferences. Some of these prophages were found to carry a variety of cargo genes, presumably conferring competitive advantage to their hosts and explaining their broad distribution[41] (Extended Data Figs 8, 9). However, in a few cases, the presence of viral groups in diverse environments could not be attributed to metadata discrepancies, ambiguity of habitat classification, contamination or broad host specificity. A small number of viral groups was found in aquatic samples with large differences in salinity, such as freshwater and hypersaline lakes, whereas other groups were found in oil-contaminated wastewater, and in human oral and faecal samples (Fig. 5c, d; Supplementary Information). Our observations of a limited number of ubiquitous viruses expand on

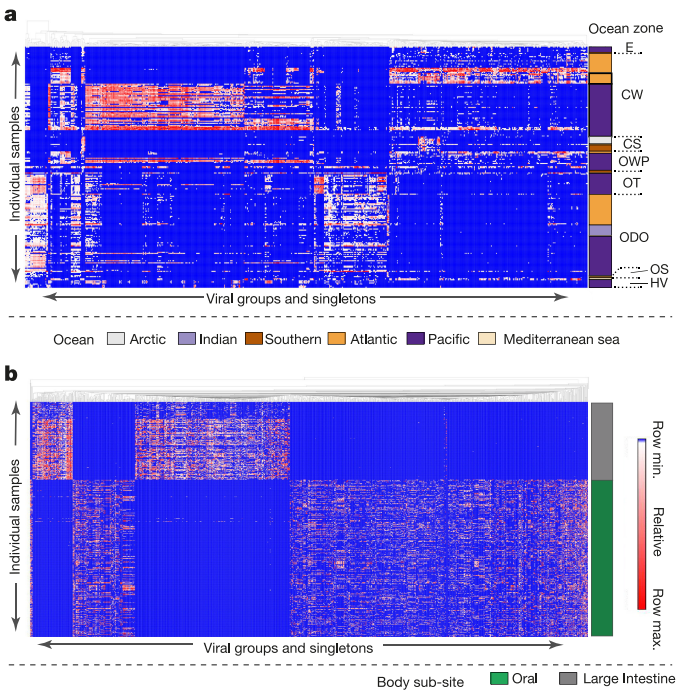


Figure 4 | Viral distribution patterns in marine and human samples. **a**, **b**, Hierarchical clustering of viral groups and singletons across marine (**a**) and human samples (**b**). Data sets were grouped according to environmental metadata or body sub-site, respectively. Oceanic zones (**a**) include Estuary (E), Coastal waters (CW), Coastal sediments (CS), Oceanic water photic (OWP; surface to 200m depth), Oceanic twilight (OT; 200m to 750m depth), Oceanic deep ocean (ODO; below 750m), Oceanic sediment (OS), and Hydrothermal vents (HV). Virus coverage is colour-coded from white (lowest coverage) to red (highest coverage) as shown in **b** inset. Blue represents absence of the corresponding viral sequence.

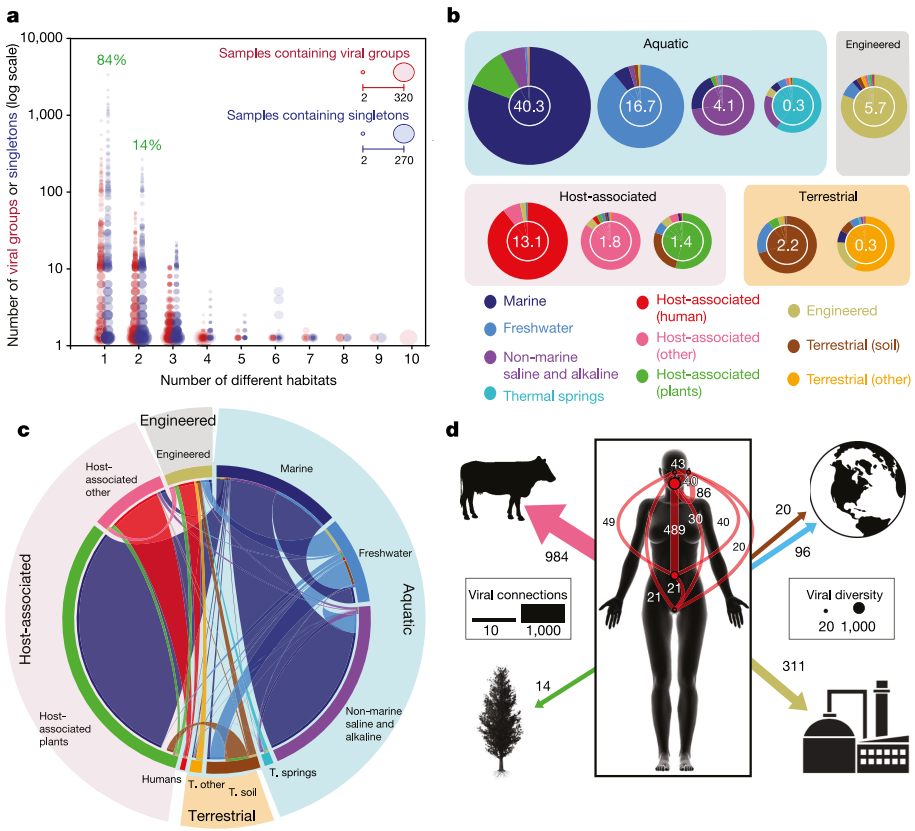


Figure 5 | Habitat distribution of metagenomic viruses. a, Distribution of viral groups and singletons (red and blue bubbles, respectively) against the number of habitats. Bubble size (inset) reflects the number of samples containing viral sequence. **b**, Distribution of viral sequences per habitat type and environmental category. Pie charts size and the number inside show viral groups and singletons per habitat type grouped into four environmental categories; values are given in units of 10^3. **c**, Pairwise connection between habitat types based on shared viral sequences; **d**, Distribution of viral sequences across human body sub-sites and across habitat types. Red dots indicate viral groups and singletons exclusive for each sub-site.

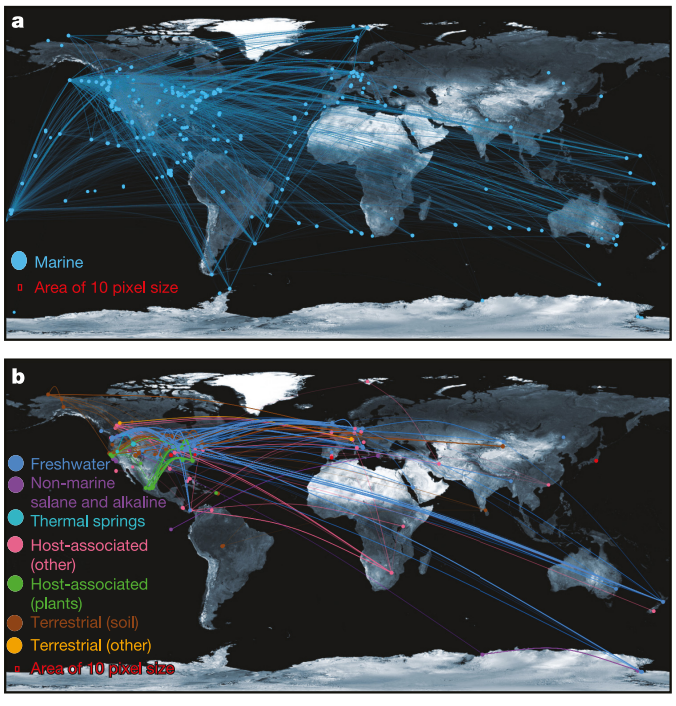


Figure 6 | Global distribution of viral diversity. The presence of the same viral groups or singletons across samples (circles) is represented by connecting lines. Only samples over 10 pixels apart and sharing 2 or more viral groups or singletons are shown. The colours of the circles indicate the habitat type. **a**, Connections between marine samples. Transparency of the lines reflects the number of shared viral groups. **b**, Connections between non-marine samples of the same habitat type. Connections between different habitat types are shown in Extended Data Fig. 10. Equirectangular projection of the world map is used as background image (http://visibleearth.nasa.gov/view.php?id=57752).

previous studies[35,42] that shed light on the mechanisms underlying their 'cosmopolitanism'.

To generate a global map of viral dispersal, we linked the viral sequences with geographic coordinates of the corresponding samples. Many viruses were found in similar ecological niches across large geographic distances, with the most prominent connectivity within extensively sampled marine biome (Fig. 6a; Supplementary Table 8), which is in agreement with previous studies, suggesting that viruses are passively transported along oceanic currents[3]. We also observed sparse but non-negligible connections between non-marine viral groups and singletons in samples of the same habitat type across biomes (Fig. 6b) and across different ecosystems (Extended Data Fig. 10).

Discussion

This study shows that in-depth exploration of ecosystems by untargeted metagenome sequencing is a powerful approach to fill knowledge gaps and address fundamental questions of viral ecology. Our analysis led to a notable increase in the number of viral sequences and putative virus–host connections, demonstrating that a much larger prokaryotic diversity than previously known is preyed upon by viruses, expanding on a recent report of microbial lineages containing prophages[21]. Consistent with previous observations, the environmental viral quasi-species were mostly found to have a narrow host range with a few notable exceptions of phages with broad taxonomic host specificity, including examples of hosts from different phyla.

The global maps of viral biogeography show that viruses are predominantly found in similar habitats, regardless their geographic proximity. This pattern was most prominent for marine viruses as previously observed[3], yet was also striking across seemingly isolated locales such as lakes, plant-associated habitats and soils, where the dispersal mode is not immediately obvious. More surprising was the significant human virome sharing between unrelated individuals and the identification of viral quasi-species distributed across markedly different habitats.

Overall, this study demonstrates the value of untargeted *de novo* metagenomic analysis as compared to reference-based and targeted

virome approaches, highlighting the importance of globally sampled metagenomic data sets[43] to vastly improve viral sequence discovery. Ultimately, large-scale computational exploration of uncharted viral sequence space will assist in addressing the remaining mysteries of viral ecology.

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 23 November 2015; accepted 8 July 2016.
Published online 17 August 2016.

1. Suttle, C. A. Marine viruses—major players in the global ecosystem. *Nat. Rev. Microbiol.* **5,** 801–812 (2007).
2. Reyes, A. *et al.* Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466,** 334–338 (2010).
3. Brum, J. R. *et al.* Ocean plankton. Patterns and ecological drivers of ocean viral communities. *Science* **348,** 1261498 (2015).
4. Whitman, W. B., Coleman, D. C. & Wiebe, W. J. Prokaryotes: the unseen majority. *Proc. Natl Acad. Sci. USA* **95,** 6578–6583 (1998).
5. Reddy, T. B. *et al.* The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. *Nucleic Acids Res.* **43,** D1099–D1106 (2015).
6. Chow, C. E. & Suttle, C. A. Biogeography of viruses in the sea. *Annu Rev Virol* **2,** 41–66 (2015).
7. Rohwer, F. & Edwards, R. The Phage Proteomic Tree: a genome-based taxonomy for phage. *J. Bacteriol.* **184,** 4529–4535 (2002).
8. Fuhrman, J. A. Marine viruses and their biogeochemical and ecological effects. *Nature* **399,** 541–548 (1999).
9. Brum, J. R. & Sullivan, M. B. Rising to the challenge: accelerated pace of discovery transforms marine virology. *Nat. Rev. Microbiol.* **13,** 147–159 (2015).
10. Edwards, R. A., McNair, K., Faust, K., Raes, J. & Dutilh, B. E. Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiol. Rev.* **40,** 258–272 (2016).
11. Markowitz, V. M. *et al.* IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Res.* **42,** D568–D573 (2014).
12. Edwards, R. A. & Rohwer, F. Viral metagenomics. *Nat. Rev. Microbiol.* **3,** 504–510 (2005).
13. Ivanova, N. *et al.* A call for standardized classification of metagenome projects. *Environ. Microbiol.* **12,** 1803–1805 (2010).
14. Hurwitz, B. L. U'Ren, J. M. & Youens-Clark, K. Computational prospecting the great viral unknown. *FEMS Microbiol. Lett.* (2016).
15. Ignacio-Espinoza, J. C., Solonenko, S. A. & Sullivan, M. B. The global virome: not as big as we thought? *Curr. Opin. Virol.* **3,** 566–571 (2013).
16. Lu, H. *et al.* Membrane biofouling in a wastewater nitrification reactor: Microbial succession from autotrophic colonization to heterotrophic domination. *Water Res.* **88,** 337–345 (2016).
17. Serwer, P., Hayes, S. J., Thomas, J. A. & Hardies, S. C. Propagating the missing bacteriophages: a large bacteriophage in a new class. *Virol. J.* **4,** 21 (2007).
18. Varghese, N. J. *et al.* Microbial species delineation using whole genome sequences. *Nucleic Acids Res.* **43,** 6761–6771 (2015).
19. Simmonds, P. Methods for virus classification and the challenge of incorporating metagenomic sequence data. *J. Gen. Virol.* **96,** 1193–1206 (2015).
20. Hurwitz, B. L., Brum, J. R. & Sullivan, M. B. Depth-stratified functional and taxonomic niche specialization in the 'core' and 'flexible' Pacific Ocean Virome. *ISME J.* **9,** 472–484 (2015).
21. Roux, S., Hallam, S. J., Woyke, T. & Sullivan, M. B. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *eLife* **4,** (2015).
22. Mojica, F. J., Díez-Villaseñor, C., García-Martínez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155,** 733–740 (2009).
23. Andersson, A. F. & Banfield, J. F. Virus population dynamics and acquired virus resistance in natural microbial communities. *Science* **320,** 1047–1050 (2008).
24. Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315,** 1709–1712 (2007).
25. Lum, A. G. *et al.* Global transcription of CRISPR loci in the human oral cavity. *BMC Genomics* **16,** 401 (2015).
26. Bailly-Bechet, M., Vergassola, M. & Rocha, E. Causes for the intriguing presence of tRNAs in phages. *Genome Res.* **17,** 1486–1495 (2007).
27. Goren, M. G., Yosef, I. & Qimron, U. Programming Bacteriophages by Swapping Their Specificity Determinants. *Trends Microbiol.* **23,** 744–746 (2015).
28. Salmond, G. P. & Fineran, P. C. A century of the phage: past, present and future. *Nat. Rev. Microbiol.* **13,** 777–786 (2015).
29. Holmfeldt, K., Middelboe, M., Nybroe, O. & Riemann, L. Large variabilities in host strain susceptibility and phage host range govern interactions between lytic marine phages and their Flavobacterium hosts. *Appl. Environ. Microbiol.* **73,** 6730–6739 (2007).
30. Peters, D. L., Lynch, K. H., Stothard, P. & Dennis, J. J. The isolation and characterization of two *Stenotrophomonas maltophilia* bacteriophages capable of cross-taxonomic order infectivity. *BMC Genomics* **16,** 664 (2015).
31. Emerson, J. B. *et al.* Virus-host and CRISPR dynamics in archaea-dominated hypersaline Lake Tyrrell, Victoria, Australia. *Archaea* **2013,** 370871 (2013).
32. Tschitschko, B. *et al.* Antarctic archaea-virus interactions: metaproteome-led analysis of invasion, evasion and adaptation. *ISME J.* **9,** 2094–2107 (2015).
33. Breitbart, M. & Rohwer, F. Here a virus, there a virus, everywhere the same virus? *Trends Microbiol.* **13,** 278–284 (2005).
34. Dinsdale, E. A. *et al.* Functional metagenomic profiling of nine biomes. *Nature* **452,** 629–632 (2008).
35. Breitbart, M., Miyake, J. H. & Rohwer, F. Global distribution of nearly identical phage-encoded DNA sequences. *FEMS Microbiol. Lett.* **236,** 249–256 (2004).
36. Salazar, G. *et al.* Global diversity and biogeography of deep-sea pelagic prokaryotes. *ISME J.* **10,** 596–608 (2016).
37. Abeles, S. R. & Pride, D. T. Molecular bases and role of viruses in the human microbiome. *J. Mol. Biol.* **426,** 3892–3906 (2014).
38. Wylie, K. M. *et al.* Metagenomic analysis of double-stranded DNA viruses in healthy adults. *BMC Biol.* **12,** 71 (2014).
39. Robles-Sikisaka, R. *et al.* Association between living environment and human oral viral ecology. *ISME J.* **7,** 1710–1724 (2013).
40. Mukherjee, S., Huntemann, M., Ivanova, N., Kyrpides, N. C. & Pati, A. Large-scale contamination of microbial isolate genomes by Illumina PhiX control. *Stand. Genomic Sci.* **10,** 18 (2015).
41. Bondy-Denomy, J. & Davidson, A. R. When a virus is not a parasite: the beneficial effects of prophages on bacterial fitness. *J. Microbiol.* **52,** 235–242 (2014).
42. Short, C. M. & Suttle, C. A. Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. *Appl. Environ. Microbiol.* **71,** 480–486 (2005).
43. Kyrpides, N. C., Eloe-Fadrosh, E. A. & Ivanova, N. N. Microbiome data science: understanding our microbial planet. *Trends Microbiol.* **24,** 425–427 (2016).

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

Acknowledgements We thank A. Visel and H. Maughan for critical reading and feedback, A. Pati for help in earlier versions, and the IMG and GOLD teams for their support. This work was conducted by the US Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, under contract number DE-AC02-05CH11231 and used resources of the National Energy Research Scientific Computing Center, supported by the Office of Science of the US Department of Energy.

Author Contributions D.P.E., N.N.I., and N.C.K. conceived and led the study. All authors participated in the analysis and interpretation of data. D.P.E., E.E.F., E.R., N.N.I., and N.C.K. wrote the paper.

Author Information 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 N.C.K. (nckyrpides@lbl.gov).

Reviewer Information *Nature* thanks C. A. Suttle and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Data availability. All the sequence data and metadata from the samples used in this work could be accessed through the Integrated Microbial Genomes with Microbiomes system IMG/M database[11] (https://img.jgi.doe.gov) using both metagenome and scaffold identifiers provided throughout the manuscript and the Supplementary Information. Thus, by using these identifiers in the Genome Search tool or Scaffold Search tool (under 'Find Genomes' tab) in the user interface, the corresponding sequences, their annotations, as well as their associated metadata can be retrieved. Moreover, Hidden Markov Models (HMMs) of viral protein families as well as the predicted DNA viral sequences in fasta format are available at the following public FTP site: (http://portal.nersc.gov/dna/microbial/prokpubs/EarthVirome_DP/).

Metagenomic samples used in this study. All publicly available metagenomic data sets from the IMG/M system (3,042 samples comprising 5 terabase pairs of sequences) were used for this analysis[11]. The sample collection included 1,729 environmental samples, 1,079 host-associated samples, and 234 engineered samples according to the Genomes OnLine sample classification[5]. We identified putative viral contigs in 1,882 out of the 3,042 metagenomic data sets. The metadata for these data sets including sample collection information, library construction and sequencing protocols, as well as assembly strategy were retrieved from GOLD database[5]. Based on GOLD metadata, the vast majority of these data sets were generated from dsDNA using an untargeted approach (that is, only 59 samples underwent viral particle enrichment, viral DNA enrichment or library construction with sequencing protocols optimized for the recovery of viral sequences). All of the data sets have been annotated by the IMG metagenome annotation pipeline[44], which performs gene prediction and functional annotation through assignment of predicted proteins to protein families, such as Pfam[45] and KEGG Orthology (KO) clusters[46]. Some of the data sets included both assembled and unassembled data, while others had only assembled sequences (Supplementary Table 8). An assembly pipeline used for each data set is described in GOLD. In addition, the contiguity of assembled sequences varied greatly from sample to sample. The ecosystem subcategories here used were manually curated according to sample metadata establishing 10 distinct habitat types: marine, freshwater, non-marine-saline and alkaline, thermal springs, terrestrial soil, terrestrial others (including mostly deep subsurface samples), host-associated human, host-associated plants, host-associated others (including host animal-associated other than human), and engineered (for example, bioreactor) (Supplementary Table 8). Only contigs longer than 5 kb (59.5 Gb from 5.1 million contigs) were primarily included in this study.

Normalization factors. We normalized the data sets by the size of the sample (measured as total number of bp from sequences larger than 5 kb) per habitat type. The normalization factor used in each habitat type was: marine, 29,602 Mb; freshwater, 96,314 Mb; non-marine saline and alkaline, 2,825 Mb; thermal springs, 1,828 Mb; terrestrial (soil), 5,794 Mb; terrestrial (other), 1,659 Mb; host-associated (human), 10,349 Mb; host-associated (plants), 3,909 Mb; host-associated (others), 23,452 Mb; engineered, 10,486 Mb.

Isolate reference viruses (iVGs). We used a combination of 2,353 iVGs composed of all isolate dsDNA viruses and retroviruses from the NCBI server (http://www.ncbi.nlm.nih.gov/genome/viruses/, data accessed on 04/2015) to extract all viral proteins and to establish, after filtering, the first round of viral protein families. Additionally, we used a list of 5,042 reference viruses (Supplementary Table 13) extracted from the IMG/M system (that included all RNA and DNA eukaryotic and prokaryotic referenced viral genomes) to generate and validate our viral genome clustering method and also to calculate the average length of all reference viruses as 44,296 bp $\pm$ 83,777 bp s.d. (Supplementary Table 13).

Generation of viral protein families. 167,042 protein coding genes were collected from 2,353 iVGs (dsDNA viruses and retroviruses combined) from the NCBI server (http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=10239#). After dereplication using 70% identity in usearch[47], 98,000 protein sequences were obtained from which 83,500 were clustered into 15,900 groups using the Markov Cluster (MCL) algorithm[48]. Proteins within clusters were aligned using MAFFT[49] and a set of 14,296 viral protein families was created using hmmbuild[50]. After manual curation of the viral families with high representation in prokaryotic genomes, viral protein families were compared against the 5.1 million metagenomic contigs longer than 5 kb. 62,000 contigs with 5 or more viral protein families were collected, and these were reduced to 9,000 putative viral contigs after removing contigs below 50 kb. An additional filtering step was performed to exclude contigs with a high number of Kegg Orthology (KO) terms and Pfams (10% and 25% respectively); this reduced the number of putative viral contigs to 1,589. These were complemented with 66 and 188 sequences derived from diverse metagenomic contigs longer than 20 kb that were binned with viruses or contained a viral RNA polymerase gene, respectively, and were not captured using the previous filter of bearing 5 or more viral protein families (detailed in section below; Extended Data Fig. 2; Supplementary Table 1). A total of 1,843 mVCs encoding 191,000 proteins were used to complement the original set of 167,042 proteins derived from iVGs. Repeating the steps described above (that is, *usearch* 70% for de-replication, MCL clustering, MAFFT alignment and *hmmbuild* with a filter for viral families abundant in prokaryotes) the final list of 25,000 viral protein families was obtained and used for further exploration.

Identification of metagenomic viral contigs for a training set via manual curation, binning and DNA-dependent RNA polymerase alignment. To expand the training set of viral sequences, metagenome contigs identified as high-confidence viral sequences in the first iteration of our pipeline (Extended Data Fig. 1) were complemented with additional metagenome contigs and scaffolds, not captured using viral protein families generated from isolate viruses. The first approach used kmer-based binning of 6 metagenome samples that contained the highest number of candidate viral sequences, which were not satisfy high-confidence threshold due to insufficient number of hits to protein models. These data sets were binned by Emergent Self Organizing Maps (ESOM; by Ultsch) as described previously[51] and contig sets outside the bins corresponding to cellular organisms were manually checked (Extended Data Fig. 2a). K-mer-based binning identified 66 putative novel mVCs from diverse habitat types (freshwater, wastewater, thermal vents, and marine with IMG sample identifiers 3300000553, 3300001592, 3300001681, 3300000116, and 3300001450, respectively).

The second approach relied on identification of contigs containing RNA polymerase with domain composition reminiscent of RNA polymerase (RNAp) found in cellular life forms, which could not be placed into one of three domains on the tree of life based on their sequence similarity. First, 2,551 representative sequences of the genes encoding the three major subunits (α, β, β') of the RNAp gene from bacteria, as well as their eukaryotic and archaeal counterparts, were collected from IMG database. Next, the domains of these genes were extracted using Pfam models and aligned with MAFFT[49]. Alignments were manually inspected and HMM models were built using hmmbuild[50]. These models were used to scan metagenomic sequences longer than 5 kb and identified 39,109 contigs with matches for at least one core RNAp subunit. After filtering short matches and a dereplication step, we obtained 7,437 metagenomic sequences that were combined with 2,551 reference isolates to build a tree with 9,309 RNAp sequences using FastTree[52] with default parameters (Extended Data Fig. 2d). The tree was visualized using Dendroscope[53] and RNAp branch corresponding to large eukaryotic DNA viruses was identified on the basis of reference sequences from isolate genomes. In addition to eukaryotic viruses, another set of metagenomic RNAp sequences branching separately from cellular references, turned out to comprise phage RNAp with domain composition similar to bacterial enzyme (detailed in Extended Data Fig. 2e). A total of 188 contigs longer than 20 kb containing viral and phage RNAp sequences were added to the training set.

Assignment of metagenomic sequences to viruses. The 25,000 viral protein families were used to identify 125,842 DNA metagenomic viral contigs (mVCs) longer than 5 kb using 3 distinct filters. First, mVCs were identified from metagenomic contigs that had at least 5 hits to viral protein families, the total number of genes covered with KO terms on the contig was $<$20%; the total number of genes covered with Pfams $\leq$40%; and the number of genes covered with viral protein families $>$10%. Second, metagenomic sequences were selected as mVCs when the number of viral protein families on the contig were equal or higher than the number of Pfams. Finally, metagenomic contigs for which the number of viral protein families was equal or higher than 60% of the total of the genes were also assigned to mVCs. Benchmarking and modelling of this DNA viral discovery computational approach are detailed below, demonstrating a specificity of 99.6% for viral detection with a 37.5% recall rate (sensitivity to identify all viral sequences).

Benchmarking of computational approaches for virus detection. In order to assess the accuracy of our DNA vHMM virus detection pipeline, we generated a synthetic metagenome, consisting of finished genomes of 32 bacteria, 3 archaea and 5 viruses (Supplementary Table 2), which included 88 replicons. Bacterial genomes include representatives of 4 phyla. A total of 132 prophage sequences were identified including 99 prophages identified by CyVerse[54] implementation of VirSorter[55] in the categories 1, 2, 4, and 5, and 33 prophages identified by manual curation based on the presence of hallmark phage genes and analysis of synteny with closely related strains. Coordinates of 35 prophages predicted by VirSorter had to be manually adjusted to eliminate bacterial genes (including ribosomal RNAs and other housekeeping genes) and to separate 2 prophage sequences called as one prophage over an intervening stretch of bacterial genes. Coordinates of the prophages are provided in Supplementary Table 3. None of the viruses or prophages used in the synthetic metagenome were included in the training set used to generate viral HMMs for our pipeline.

The genome sequences were fragmented to generate 63,222 contigs of length 5 kb to 60 kb. The distribution of fragmented contigs include 28,497 5-kb-long

fragments, 14,228 10-kb-long fragments, 7,096 20-kb-long fragments, 4,723 30-kb-long fragments, 3,525 40-kb-long fragments, 2,810 50-kb-long fragments and 2,343 60-kb-long fragments. The resulting synthetic metagenome is dominated by bacterial and archaeal chromosomal fragments with an admixture of a relatively small number of plasmid and viral sequences, which is a faithful representation of a typical metagenome data set generated by an untargeted approach rather than by targeted virome sequencing approach. The metagenome was submitted to a CyVerse implementation of VirSorter and also processed by our vHMM pipeline. Only the categories 1, 2, 4 and 5 of Virsorter predictions were considered, as manual inspection showed that categories 3 and 6 contained mostly false positives. Sequence fragments with at least 3 kb of phage or prophage sequence were considered as true positive viral sequences; those with less than 3 kb of phage or prophage sequence were considered true negative.

Calculating the rate of viral protein cluster accumulation and the number of proteins with high similarity to proteins encoded by isolate viruses. The 125,842 metagenomic viral contigs longer than 5 kb encoded a total of 2.79 million proteins. BLASTp[56] with an *e*-value of 1.0×10^{-5} was used and 1 hit per query protein with >60% sequence identity and >80% alignment on the shorter sequence. Proteins encoded by mVCs were clustered using CD-HIT[57] at 60% sequence identity and >80% alignment on the shorter sequence. For each sample count, 100 random metagenome sets were generated and the total number of protein clusters found on the contigs from this set was calculated. This analysis was repeated separately for metagenome samples classified as 'aquatic' ($n = 656$) and 'human' ($n = 673$).

Comparison of mVCs protein clusters against all iVGs. Sequence similarity of mVCs to iVGs was computed using BLASTp[56] with an *e*-value threshold of 1.0×10^{-5} and alignment length of at least 80% of the shorter protein. No percentage identity or bit-score cutoffs were applied (Supplementary Table 9).

Identification of complete metagenomic viral genomes. To assess the number of closed DNA mVCs, we searched for overlapping sequences in the 3′ and 5′ region of all the 125,842 metagenomic contigs. Extractseq[58] was used to trim the first 100 bp of each contig and BLAT[59] was used to search each 100-bp fragment against the respective contig. Only exact overlapping matches for both the 3′ and 5′ regions were considered. This resulted in the identification of 999 putatively closed mVCs, ranging from 5,037 bp to 630,638 bp in length (average, 53,644 bp ± 45,677 bp s.d.). Supplementary Table 10 lists all putatively closed mVCs.

Viral genome clustering and designation of viral groups. A sequence-based classification framework was developed for systematically linking closely related viral genomes based on their overall protein similarity. The framework relies on both AAI and total alignment fraction (AF) for pairwise comparisons of viral sequences, and enables natural grouping of related iVGs and mVCs. The 125,842 mVCs were combined with all iVGs (DNA and RNA viruses) for the generation of the viral group classification framework (Supplementary Information). To reduce the number of the AAI comparisons, only mVCs that contained at least one protein match with ≥70% identity across ≥50% of the shortest protein length were selected for pairwise computations. This filter reduced the number of total pairwise comparisons from 9.5 billion to 15.9 million. The bidirectional average amino acid identity (AAI) was performed as previously described[18] for all of the 15.9 million pairwise comparisons. This method implements usearch[47] for rapid blast, and selects the bidirectional best hit for each protein encoded on the mVC and outputs the AAI and the AF. The output was subsequently filtered to include only matches that had ≥90% AAI and ≥50% AF which were the observed parameters that best reproduced the existing taxonomy of iVGs (Supplementary Information; Extended Data Fig. 5a). The high-quality filtered AAI results were then clustered using single-linkage hierarchical clustering and visualized in Cytoscape[60] (Extended Data Fig. 5c–e).

Validation of viral groups generated. As a validation of our clustering method we observed that 87% of the iVGs (920 out of the 1,060 viral groups or singletons) with a taxonomic assignment according to the International Committee on Taxonomy of Viruses (ICTV) clustered in agreement with the ICTV-designated species. All the remaining 13% of iVGs clustered at the genus-level. From this 13% (represented by 140 viral groups that contain at least one iVG) we found that only 49 were phage groups, with high pairwise (over 90% AAI) values for the reference viruses within each group, suggesting that despite their taxonomic assignments, they were also probably members of the same species (Supplementary Table 14). These analyses show that our viral groups are taxonomically relevant and provide a useful method for organizing distinct viral types.

Viral host assignment using the CRISPR–Cas system. A CRISPR–Cas spacer database of 3.5 million sequences was created using a modified version of the CRISPR Recognition Tool[61] (CRT) detailed in ref. 44 against 40,623 isolates and 6,714 metagenomes (all data sets from the IMG system as of 9 July 2015). All identified spacers were queried for exact sequence matches against all iVGs using the BLASTn-short function from the BLAST+ package with parameters: *e*-value threshold of 1.0×10^{-10}, percentage identity of 95%, and using 1 as a maximum target sequence[56]. 98.5% of the detected 1,340 spacer hits were to a putative bacterial or archaeal host whose taxonomic assignment was in agreement at the species or genus level with the existing viral taxonomy (Supplementary Table 18). From the remaining matches, 1.2% of the hits agreed at the family level and only 0.3% of the spacers (2 cases where *Pseudomonas* spacers matched a *Rhodothermus* phage, and *Methylomicrobium* spacers that matched *Pseudomonas* and *Burkholderia* phage) were above family, validating our approach of host assignment based on CRISPR–Cas spacer matches. Subsequently, all 3.5 million spacers were compared against the 125,842 mVCs, requiring at least 95% identity over the whole spacer length, and allowing only 1–2 SNPs at the 5′ end of the sequence. A total of 12,576 proto-spacers (that is, spacer sequences within a phage genome) were identified. Based on CRISPR–Cas spacer matches exclusively from microbial isolate genomes we assigned host taxonomy to 8,084 mVCs (representing 6.42% of all the mVCs), comprising 826 viral groups (~4.47% of the total) plus 1,100 viral singletons (~1.71%) (Fig. 2a; Supplementary Table 19).

Host–virus assignment using viral tRNA matches. Identification of tRNAs from mVCs was performed with ARAGORN v1.2 (ref. 62) using the '-t' option. In order to validate this approach, 2,181 tRNA sequences were recovered from 344 referenced viruses (~7% of the total). These were compared against all genomes and metagenomes in the IMG system using BLAST, leading to 16,089 perfect hits (100% length and 100% sequence identity) after removing self-hits and duplicates. The taxonomic assignment of the tRNAs found in iVGs was compared against the taxonomic information of the isolate microbial genomes showing that 92.5% of the matches agreed at the genus or species level (Supplementary Table 18). After culling the top-20 most abundant viral-tRNA sequences (sequences conserved across members of the gammaproteobacteria class; Supplementary Table 22) and repeating the above steps with mVCs, 32,449 tRNAs within 9,555 mVCs (7.6% out of the 125,842 total) were identified, enabling the host assignment for 2,527 mVCs (Supplementary Information; Supplementary Table 19).

Low abundance virus detection. In order to detect the presence of any of the mVCs in lower abundances across different habitat types, we expanded our analysis to include not only assembled data (that probably represent the most abundant viruses) but also unassembled data from 4,169 samples currently available in IMG/M database, which comprises more than 5 Tb of sequences. We used BLASTn program in the Blast+ package[56] to find hits to our 125,842 predicted viral sequences with an *e*-value cutoff of 1×10^{-50}, at least 90% identity, and the hits from the sample covering at least 10% of the length of the viral contig. This filtering of BLAST results excluded matches to short highly conserved fragments of viral sequences, such as tRNAs, and other spurious hits. Our filtering criteria were optimized for the type of metagenome data sets available to us, and are significantly more stringent than those used in some previous studies for similar data (e.g. 95% identity over 75 nt alignment used in ref. 63) or tBLASTx with *e*-value of 1.0×10^{-5} recommended by ref. 64. However, it was less stringent than the 75% coverage used in the analysis of Tara Oceans Viromes[3], which relied on viral enrichment to increase viral sequence coverage. For the largest metagenome available to us (IMG taxon 3300002568, Grasslands soil microbial communities from Hopland, California, USA), this new analysis was able to detect 500 nt of viral sequence in 138,769,704,035 nt of total metagenome sequence, which corresponds to the abundance of 3.06×10^{-07}%.

Habitat type specificity of predicted viral sequences based on their BLASTn hits in assembled and unassembled data shows their presence even at low abundance, depending on the sequence coverage for each specific metagenome (Fig. 4a, b). The distribution of less abundant viruses supports the trend that viruses have a strong specificity for a particular habitat type since ~84% of all the viral groups are found exclusively in a single habitat type. About 14% of the viral groups were found in 2 habitat types, and most of these cases could be explained by the uncertainty of habitat type classification. For instance, algae-associated microbiomes were classified as plant host-associated and shared viral groups with marine samples, whereas loose soil samples classified as terrestrial habitat type shared viral groups with rhizosphere samples, which were classified as plant host-associated. After excluding ambiguously classified cases, most viral groups detected in more than 1 habitat type were found in the samples from the same environmental category (for example, in different aquatic habitats or in different mammalian hosts). We further report the finding of ~0.2% of the viral groups in 5 or more habitats types and discuss the main types of these 'cosmopolitan' viruses (probably laboratory contaminants, prophages with broad-host specificity, and bona fide lytic phages with unexpectedly broad habitat type distribution).

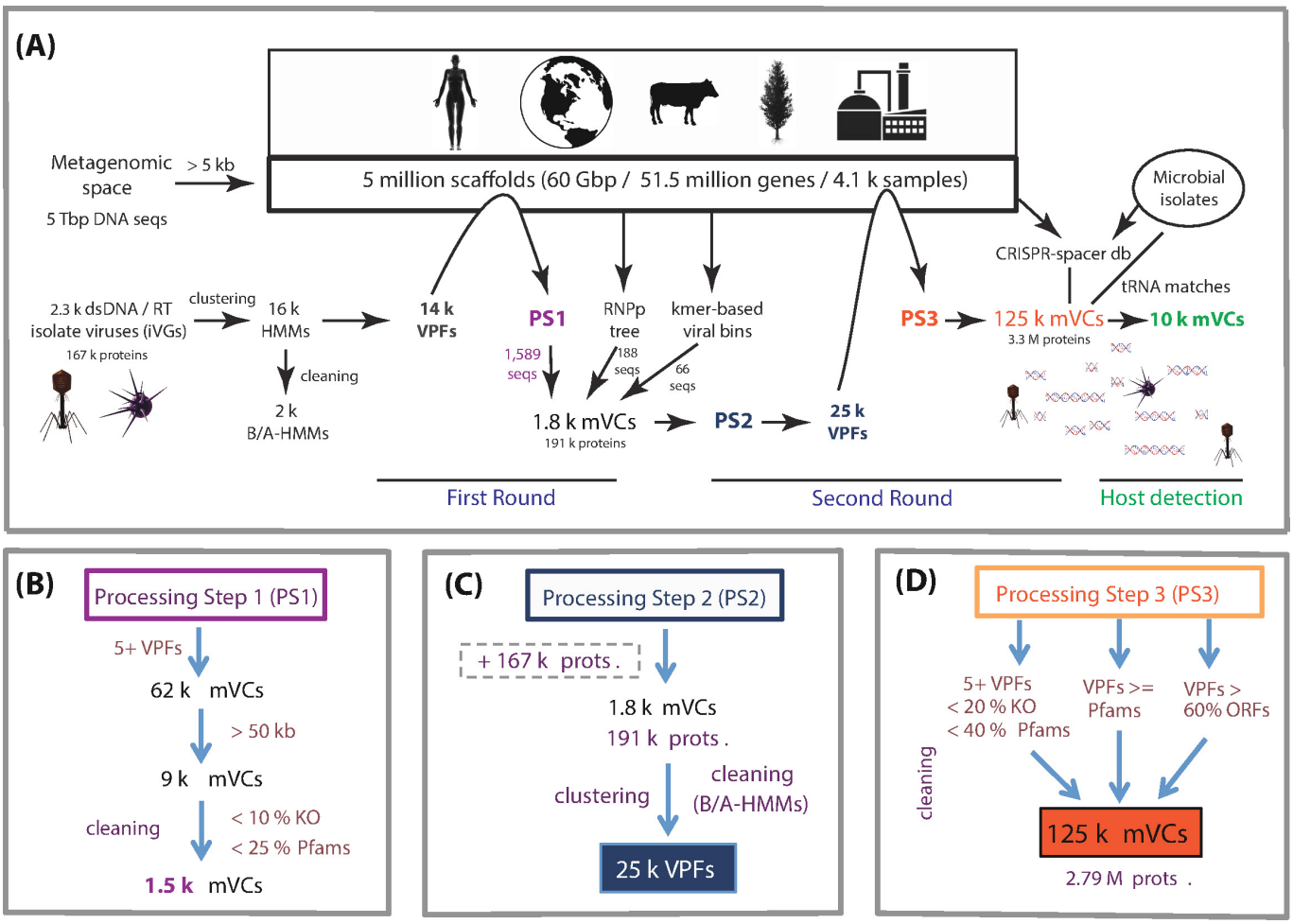
Estimation of viral abundance in faecal and oral metagenomes from Human Microbiome Project. Raw reads for faecal and oral metagenomes were retrieved from the Short Read Archive (http://www.ncbi.nlm.nih.gov/sra/) based on the metadata available in GOLD. The reads were quality-filtered and quality-trimmed

using rqcfilter tool from BBtools package (https://sourceforge.net/projects/bbtools/) with default settings: kmer length for trimming of 23, minimum average quality of 5, trim quality threshold of 10, reads shorter than 45 nt after trimming were discarded. Quality-filtered and trimmed reads were digitally normalized and error corrected using bbnorm tool from BBtools package with default settings. Normalized reads were assembled using SPADES 3.6.2 (ref. 65) and kmers of 19, 39, 59, 79, 99, selecting an optimal kmer length based on the maximal N50. Average contig and scaffold coverage of assembled data was calculated by mapping the quality-filtered and -trimmed reads to the assembly using bbmap tool from BBtools with default kmer length of 13 and minimum percentage identity cutoff of 95%. The unmapped reads were merged using bbmerge tool from BBtools package and the sequences shorter than 100 nt were discarded. mVCs were aligned against these data using BLASTn and filtered as described above. Only 1 best hit per sequence was retained. Coverage of each mVC by sample data was calculated as alignment length multiplied by the coverage of the subject sequence and summed over all sequences in the sample with hits to this mVC.

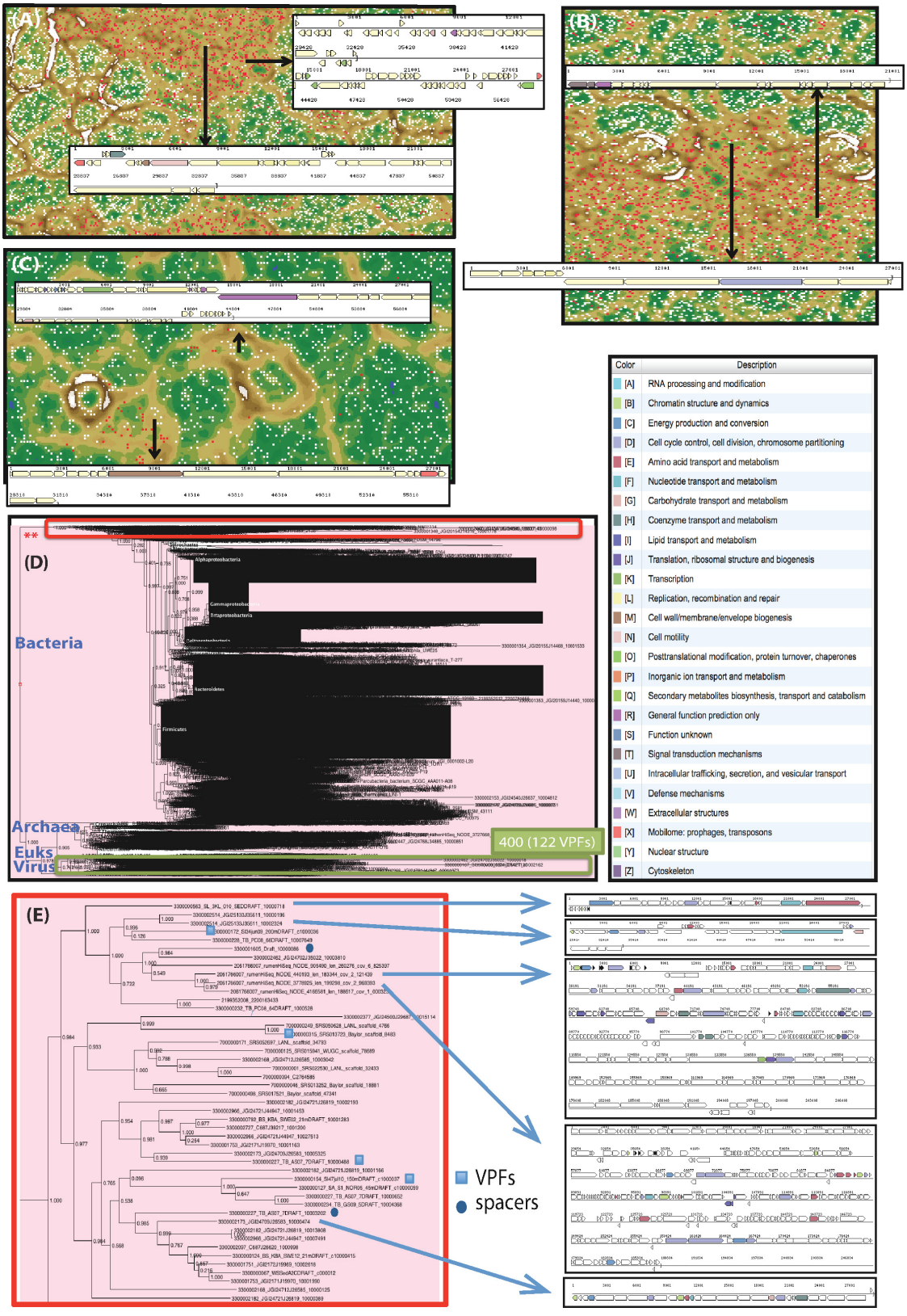
Putative prophage identification. We have identified putative prophages among 125,842 mVCs using these contigs as a query and running BLASTn[56] comparison of 'blast+' package against all isolate genomes in the IMG database. *e*-value cutoff of 1.0×10^{-50} and percentage identity of 80% were used, and mVCs with cumulative alignment of at least 75% of mVC length against an isolate genome were considered prophage candidates (Supplementary Table 4).

Global virus distribution maps. Visualization was made with the use of Processing programming language (https://processing.org/) and a freely available equirectangular projection of the world map (http://eoimages.gsfc.nasa.gov/images/imagerecords/57000/57752/land_shallow_topo_2048.jpg) was used as a background image. Sample points are positioned by latitude and longitude coordinates of Biosamples obtained from GOLD[5]. Points are coloured based on a customized reclassification of the GOLD hierarchical ecosystem classification (habitat types). Lines between points indicate samples that share at least 2 viral groups or singletons.

44. Huntemann, M. *et al.* The standard operating procedure of the DOE-JGI Microbial Genome Annotation Pipeline (MGAP v.4). *Stand. Genomic Sci.* **10,** 86 (2015).
45. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44** (D1), D279–D285 (2016).
46. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44** (D1), D457–D462 (2016).
47. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26,** 2460–2461 (2010).
48. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30,** 1575–1584 (2002).
49. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30,** 3059–3066 (2002).
50. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39,** W29–W37 (2011).
51. Dick, G. J. *et al.* Community-wide analysis of microbial genome sequence signatures. *Genome Biol.* **10,** R85 (2009).
52. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26,** 1641–1650 (2009).
53. Huson, D. H. & Scornavacca, C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Syst. Biol.* **61,** 1061–1067 (2012).
54. Merchant, N. *et al.* The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences. *PLoS Biol.* **14,** e1002342 (2016).
55. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3,** e985 (2015).
56. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 421 (2009).
57. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22,** 1658–1659 (2006).
58. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16,** 276–277 (2000).
59. Kent, W. J. BLAT–the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).
60. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–2504 (2003).
61. Bland, C. *et al.* CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8,** 209 (2007).
62. Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* **32,** 11–16 (2004).
63. Dutilh, B. E. *et al.* A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* **5,** 4498 (2014).
64. Aziz, R. K., Dwivedi, B., Akhter, S., Breitbart, M. & Edwards, R. A. Multidimensional metrics for estimating phage abundance, distribution, gene density, and sequence coverage in metagenomes. *Front. Microbiol.* **6,** 381 (2015).
65. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19,** 455–477 (2012).

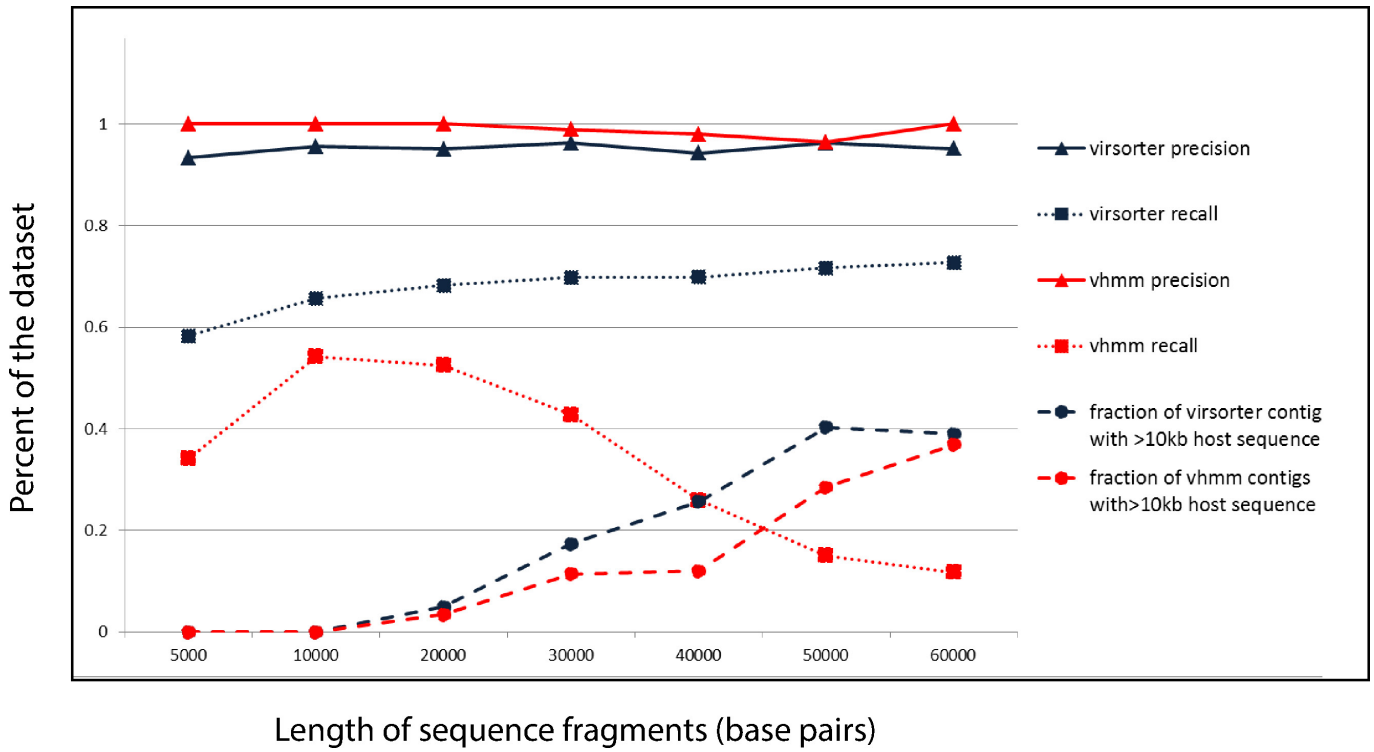


Extended Data Figure 1 | Detailed workflow for the identification of viral sequences from metagenomic data. a, Overview of the acquisition and filtering of viral protein families in two rounds and their use for the identification of metagenomic viral contigs larger than 5 kb. In the first round, proteins from 2,300 double-stranded DNA viruses were grouped into 16,000 protein families, which were aligned to generate Hidden Markov Models (HMMs). These HMMs were used in combination with analysis of k-mer composition and phylogenetic analysis of DNA-dependent RNA polymerase genes to identify 1,843 high-confidence metagenome viral contigs. **b**, **c**, These contigs were validated by manual analysis (**b**) and the proteins from this set were combined with the isolate viral proteins to generate a final set of 25,000 viral protein families (**c**). **d**, HMMs generated from alignment of these protein families were used to identify 125,842 metagenomic viral contigs. Processing steps detailed in **b**–**d** are described in the Methods. The final mVCs were then grouped and assigned to their hosts via CRISPR–Cas spacer matches and viral tRNA matches against isolate microbes (not shown in this figure).

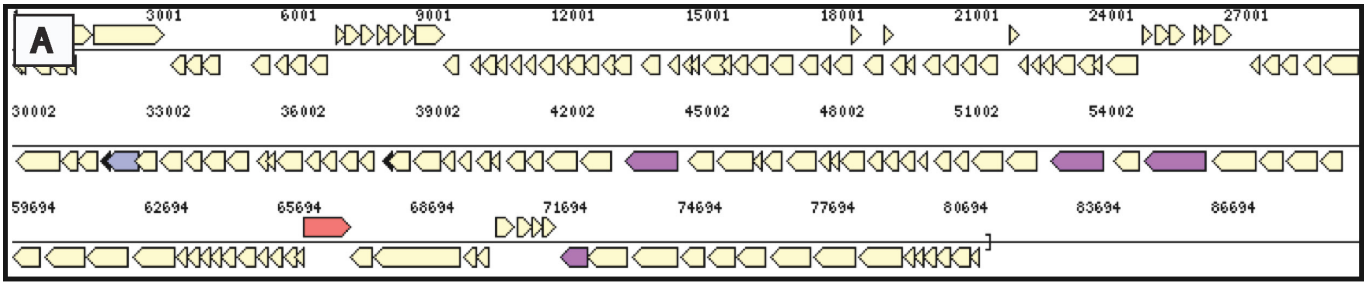


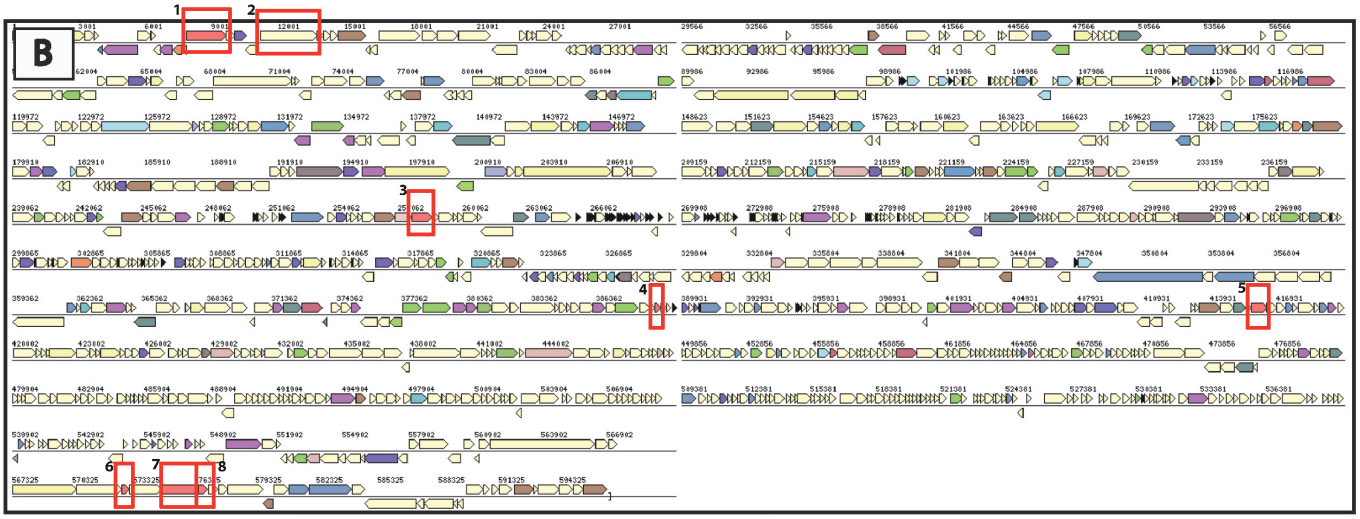
Extended Data Figure 2 | See next page for caption.

Extended Data Figure 2 | Identification of metagenomic viral contigs via binning and DNA-dependent RNA polymerase alignment. **a**–**c**, Three distinct metagenomic examples of tetranucleotide Emergent Self Organizing Maps (ESOM) as a binning method for identification of candidate viral sequences in metagenome data sets. Tetranucleotide binning of metagenomic samples (full list in Supplementary Table 1) was used to identify highly divergent viral sequences, which were left undetected using viral protein families generated from isolate viruses. Each dot on the maps represents a 10 kb fragment of a metagenomic scaffold longer than 20 kb. 'Bubbles' (ESOM structures) correspond to fragments with similar tetranucleotide composition probably originating from the same genome. Red dots represent viral sequences detected by viral protein families generated for isolate viruses; white dots represent highly divergent viral sequences with no hits to viral protein families. **a**, ESOM of freshwater sample (combined assembly of freshwater microbial communities from Lake Mendota and Trout Bog Lake, IMG identifier 3300000553) shows 2 putative viral sequences previously unidentified (IMG scaffold identifiers 10001161 and 10001271). **b**, ESOM of marine sample (marine microbial communities from Delaware Coast, sample from Delaware MO Spring March 2010, IMG identifier 3300000116) shows 2 putative viral sequences sequences (IMG scaffold identifiers c10000689 and c10000429). **c**, ESOM of hydrothermal vent sample (black smokers hydrothermal plume microbial communities from Abe, Lau Basin, Pacific Ocean, IMG identifier 3300001681) showing 2 viral sequences (IMG scaffold identifiers 10000222 and 10000095). Metagenome samples can be found in IMG using IMG identifiers and 'Quick Search' or 'Genome Search' tools; metagenome scaffolds can be using scaffold identifier and 'Scaffold Search' tool on the respective 'Microbiome Details' page. **d**, **e**, DNA-dependent RNA polymerase genes of likely viral origin from metagenomic sequences longer than 5 kb. **d**, Hidden Markov Models (HMMs) were built for sequences corresponding to α, β, and β' subunits of bacterial DNA-dependent RNA polymerase for a representative set of 2,551 cellular organisms (archaea, bacteria, and eukaryotes) and viruses. These models were used to search the proteins encoded by metagenomic contigs longer than 5 kb and the proteins with hits were aligned against the HMMs. A total of 7,437 nearly full-length metagenomic sequences were combined with 2,551 reference sequences to reconstruct the phylogenetic tree using FastTree tool. Two distinct branches on this tree were separated from the sequences from cellular organisms and included RNA polymerase genes from eukaryotic viruses (green box) and putative phage sequences with domain structure similar to that of bacterial RNA polymerase (red box, marked with double asterisk). Only 122 out of the 400 contigs in the eukaryotic viral RNA polymerase branch were captured by isolate protein families. **e**, Detailed view of the RNA polymerase tree branch with putative phage sequences. Metagenome contigs detected as viral by viral protein families and by spacer hits are marked with a square or circle next to it. Gene structure for selected contigs (IMG chromosomal neighbourhood view) is shown in the boxes. In the examples, genes are coloured based on predicted function category (using Clusters of Orthologous Genes prediction) and are specified in the figure. White-coloured genes correspond to those with hypothetical or unknown function.



Extended Data Figure 3 | Benchmarking of vHMM-based pipeline and VirSorter on synthetic metagenome data. Precision (solid lines) and recall (dotted lines) for vHMM pipeline (red) and VirSorter (blue) is plotted against the length of sequence fragments in base pairs. The percentage of contigs detected as viral, but which have at least 10 kb of host sequence is shown by dashed lines for vHMM pipeline (red) and VirSorter (blue).

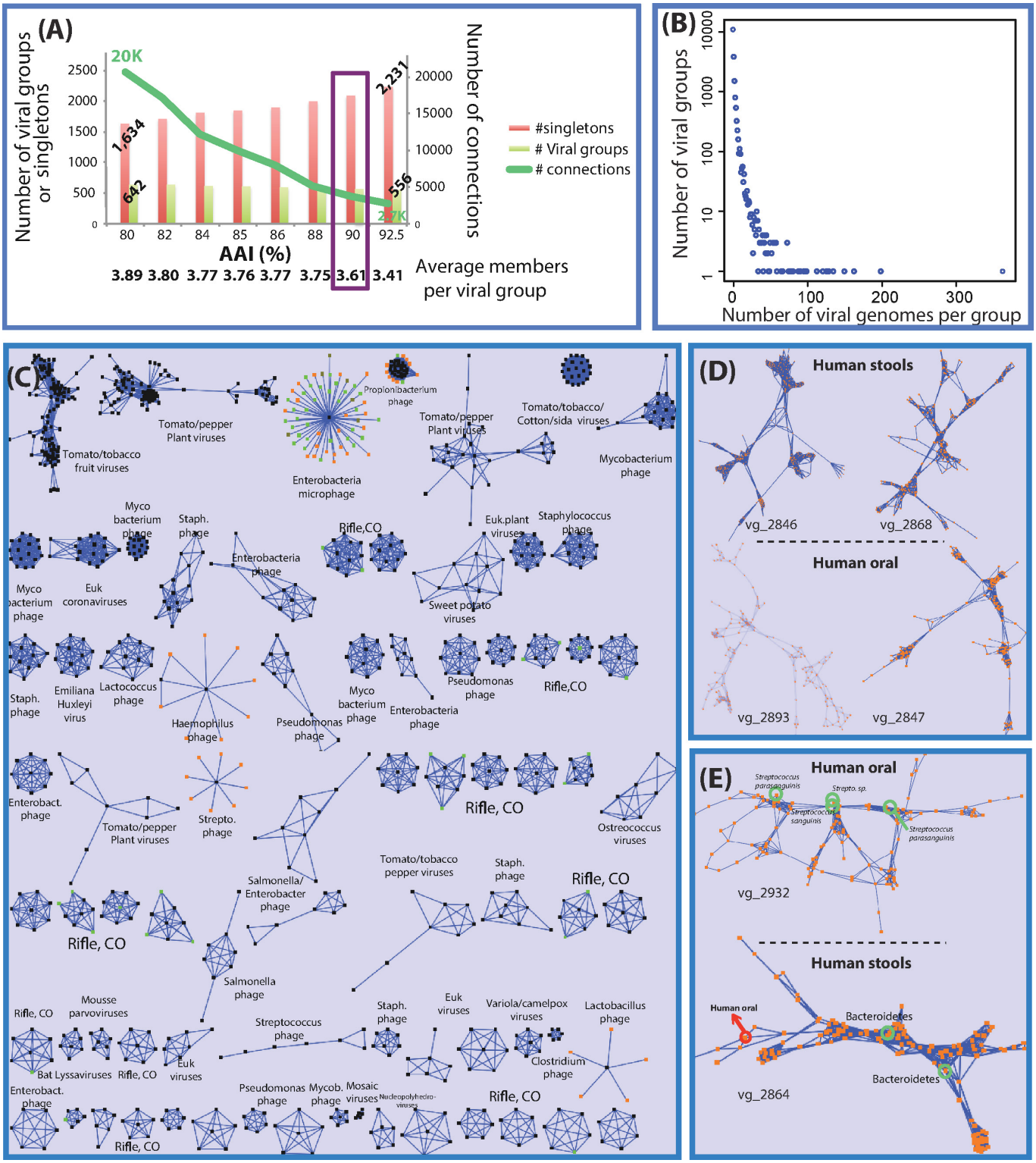




Color	Description
[A]	RNA processing and modification
[B]	Chromatin structure and dynamics
[C]	Energy production and conversion
[D]	Cell cycle control, cell division, chromosome partitioning
[E]	Amino acid transport and metabolism
[F]	Nucleotide transport and metabolism
[G]	Carbohydrate transport and metabolism
[H]	Coenzyme transport and metabolism
[I]	Lipid transport and metabolism
[J]	Translation, ribosomal structure and biogenesis
[K]	Transcription
[L]	Replication, recombination and repair
[M]	Cell wall/membrane/envelope biogenesis
[N]	Cell motility
[O]	Posttranslational modification, protein turnover, chaperones
[P]	Inorganic ion transport and metabolism
[Q]	Secondary metabolites biosynthesis, transport and catabolism
[R]	General function prediction only
[S]	Function unknown
[T]	Signal transduction mechanisms
[U]	Intracellular trafficking, secretion, and vesicular transport
[V]	Defense mechanisms
[W]	Extracellular structures
[X]	Mobilome: prophages, transposons
[Y]	Nuclear structure
[Z]	Cytoskeleton

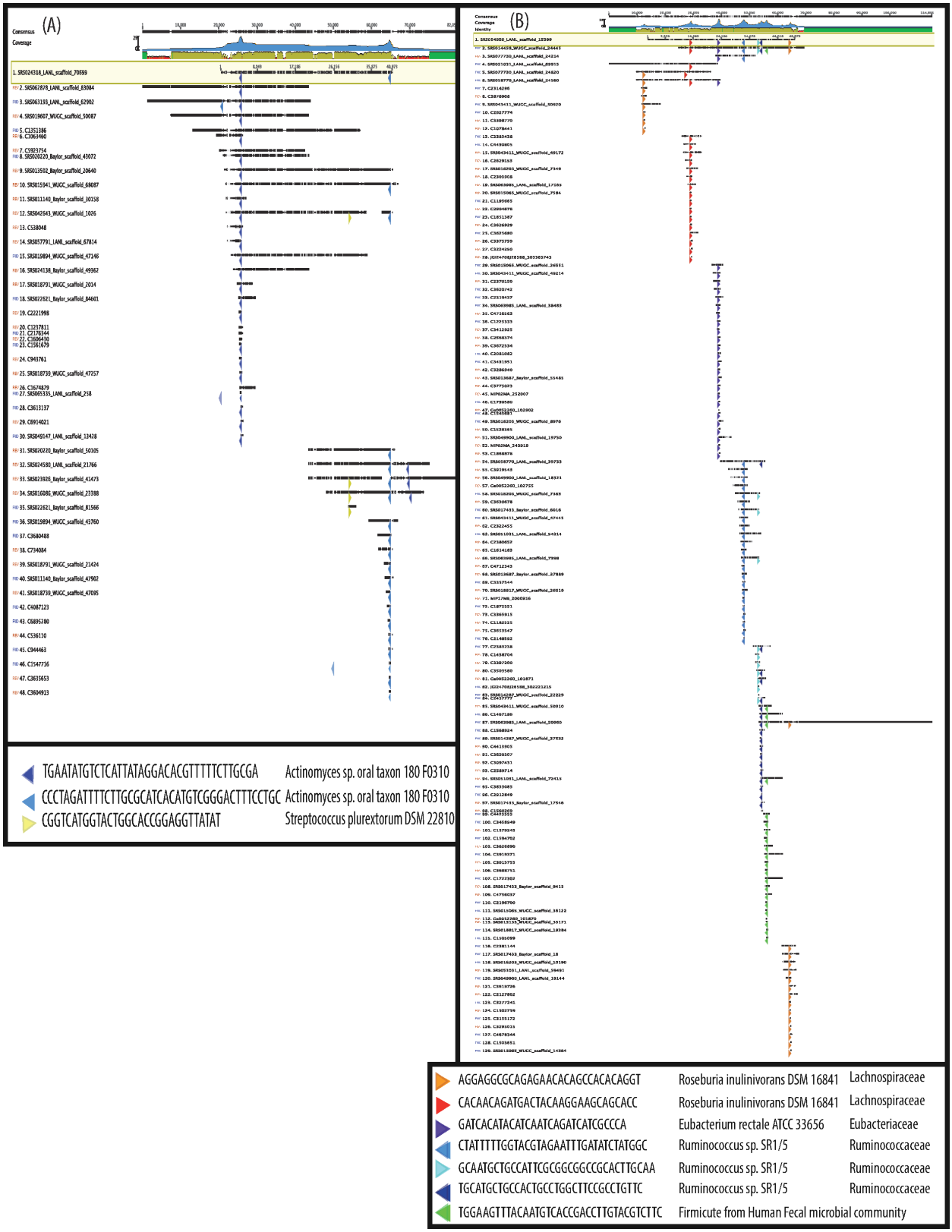
	Gene product name	Start Coord	End Coord	Strand
1	Terminase	7,846	9,678	+
2	T4-like capsid prot (gp20)	11,212	13,776	+
3	Mu-like tail prot gpP	257,685	258,626	+
4	Phage repressor prot C	389,035	389,328	+
5	Phage prot gp37	415,752	416,477	+
6	Phage baseplate assembly prot W	572,696	573,109	+
7	Phage tail prot (Tail P2 I)	574,656	576,485	+
8	Tail fiber prot	576,482	576,946	+

Extended Data Figure 4 | Detailed gene content of singular metagenomic viral contigs examples. a, Gene content of the metagenomic partial viral genome with the lowest gene coverage by viral protein families. This length of the partial viral genome is 81,542 bp (guanine and cytosine (GC) content of 43%; 163 total genes) and was identified from a bovine rumen metagenome (IMG scaffold identifier, rumenHiSeq_NODE_3763566_len_81492_cov_5_518198; IMG metagenome identifier, 2061766007). White-coloured genes correspond to those with hypothetical or unknown function. Only 3% of the genes were covered by VPFs. **b**, Gene content of the largest closed viral genome identified to date. The length of the closed (circular) viral genome is 596,617 bp (GC, 40%; 1,148 total genes) and was identified from a bioreactor metagenome (IMG scaffold id: D1draft_1000006, from Bioreactor L1-648F-DHS sludge microbial communities sample). Predicted gene function is coloured based on Clusters of Orthologous Genes. Black triangles indicate tRNAs sequences (**a**, **b**). A total of 11% of the genes were covered by VPFs. Specific viral genes distributed across the genome are boxed in red, identified with a number, and described in the legend table. The detailed information of the whole gene content of this viral genome is located in Supplementary Table 11.

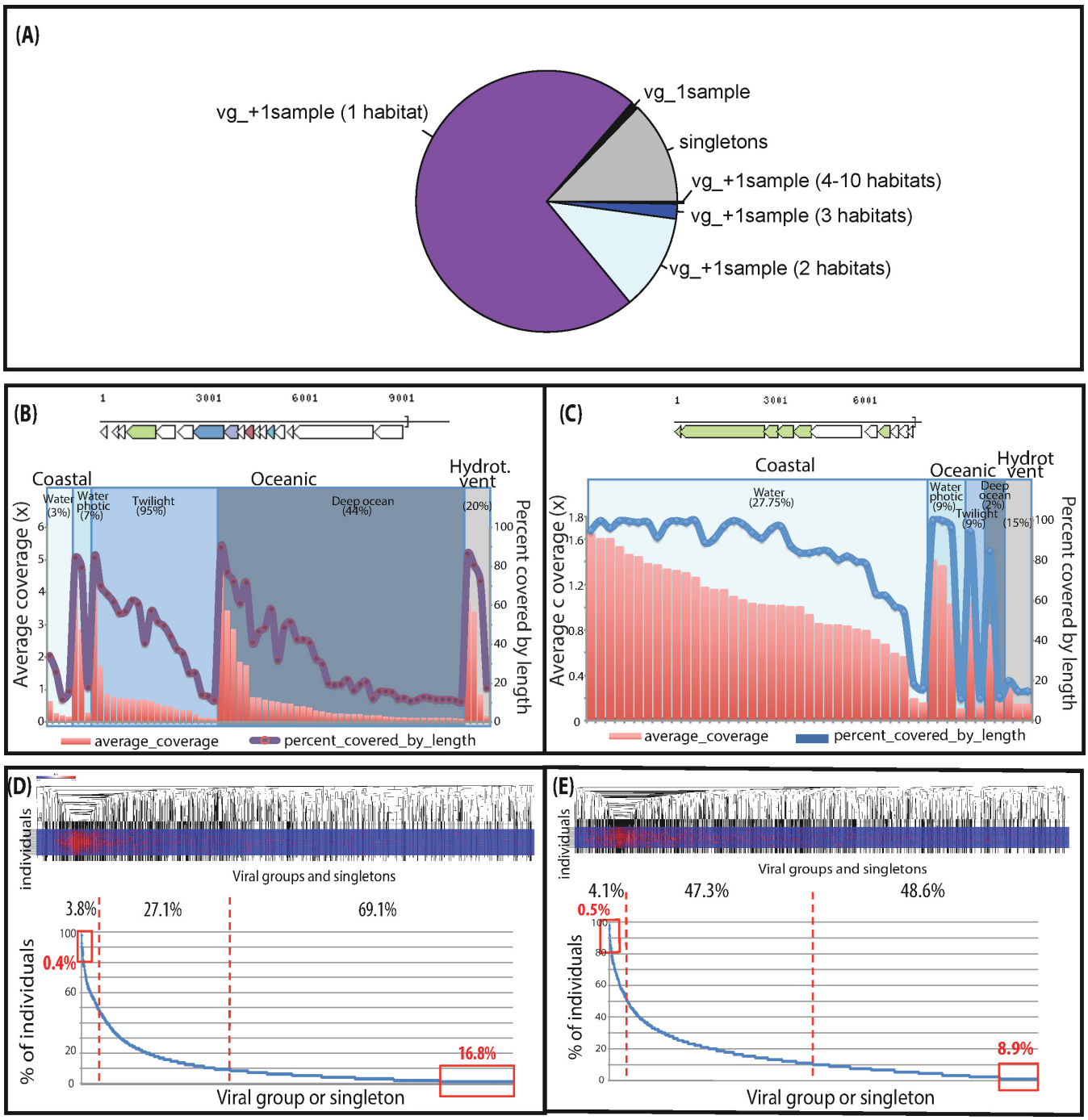


Extended Data Figure 5 | See next page for caption.

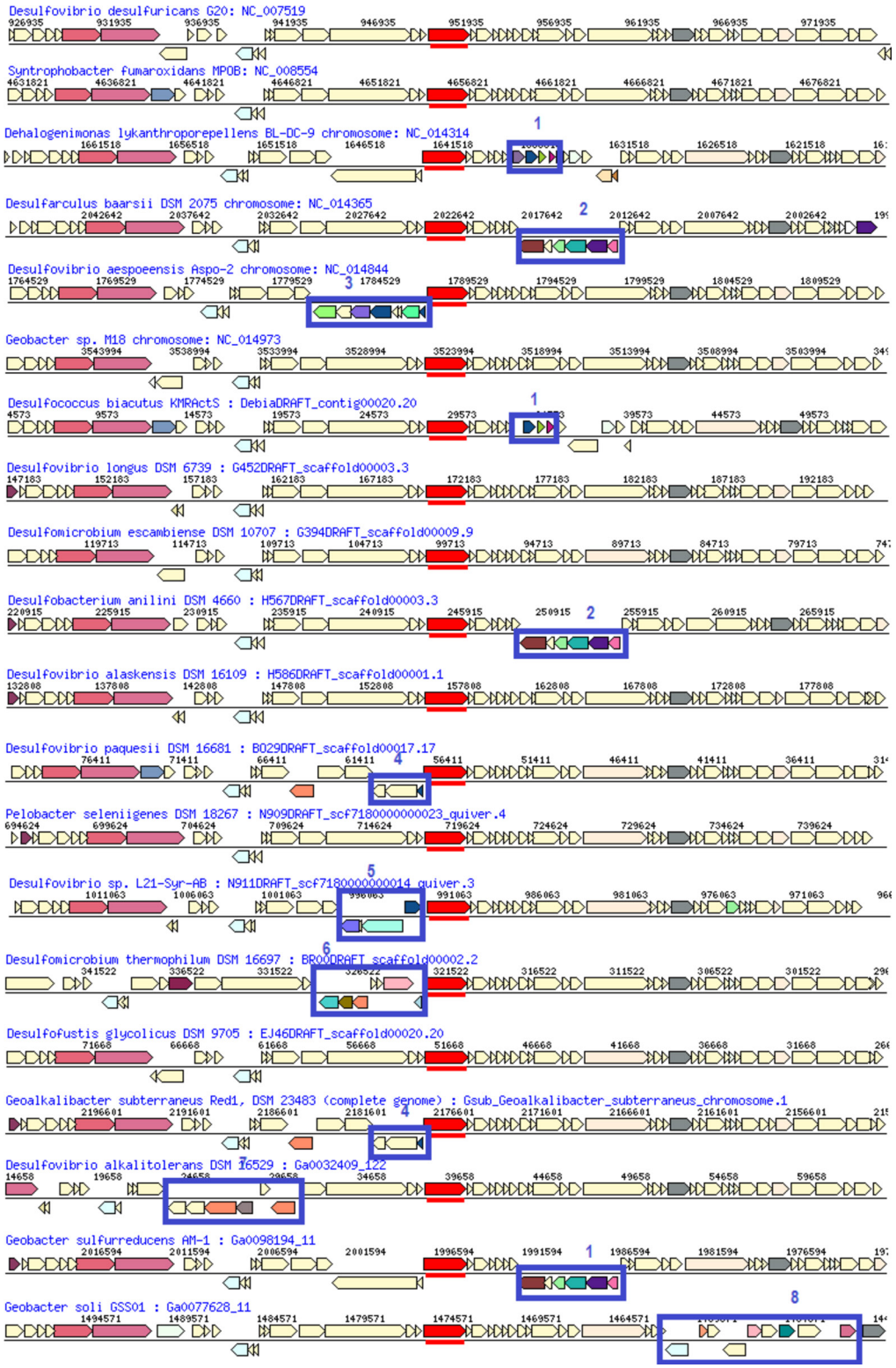
Extended Data Figure 5 | Viral group clustering method. a, Parameters used in the clustering of viral sequences. We used all 5,042 reference isolate viral genomes (iVGs) to group them using single-linkage hierarchical clustering (SLC) with different combinations of AAI and AF values to validate the clustering approach. The thresholds for AAI and AF were set at 90% and 50%, respectively, (boxed in purple) and were selected based on the accurate grouping of iVGs that was in agreement at the genus level, and the vast majority at the species level, according to the ICTV classification system (Supplementary Information). Further, these thresholds reduced the number of total connections (green line referred to secondary *y* axis) compared with lower AAI thresholds, without altering the total number of singletons and viral groups created (red and light green bars referred to primary *y* axis, respectively), as well as the average number of members per viral group (shown at the bottom of the figure). **b**, Size distribution of viral groups. Distribution of the 66,696 viral genomes clustered into 18,470 viral groups. Number of viral members (spanning from 2 to 365) per viral groups is shown. **c**–**e**, The cytoscape visualization of some viral groups. **c**, Major reference isolated viral groups created using SLC with AAI and AF values of 90% and 50%, respectively. Cytoscape force-directed (unweighted) layout option was used to visualize these groups. Black nodes represent isolated viral genomes whereas orange and green nodes represent metagenomic viral contigs clustered with isolates from host-associated and environmental samples, respectively. Group edges connect viral groups based on the above cutoffs. **d**, The four largest viral groups created from metagenomic viral contigs (containing 365, 201, 165, and 152 members, respectively). Specific habitat information of the samples as well as the viral group identifier is shown in the figure. **e**, Examples of viral groups (vg_2932 and vg_2864) containing proto-spacers (indicated by green circles) found in the CRISPR–Cas system of the indicated bacterial taxon. All the metagenomic viral contigs clustered in both viral groups were found in the same habitat subtype: human oral samples for vg_2932, and human faecal samples for vg_2864 (with a sole exception in the latter group that derived from an oral sample, indicated with a red arrow).



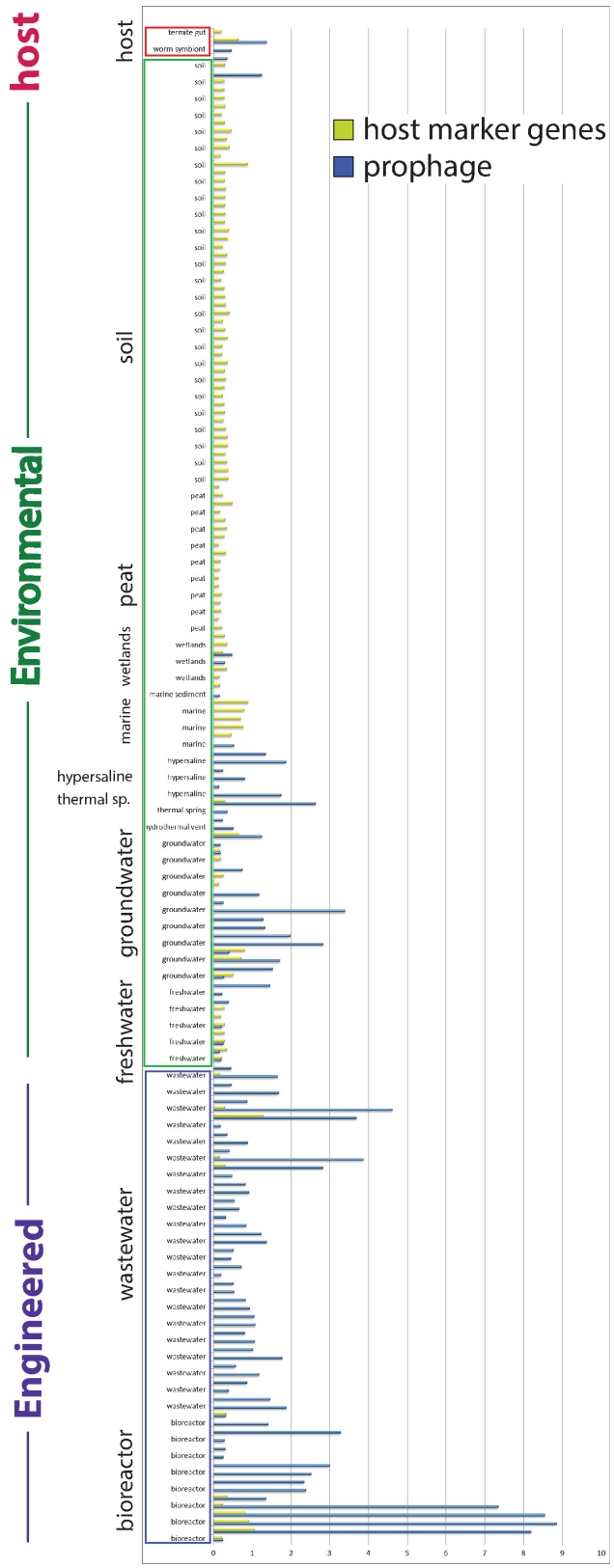
Extended Data Figure 6 | Verification of viruses identified with broad-host range. a, **b**, Alignments of all contigs found in the IMG database containing any of the 3 spacer matches present in a viral group potentially infecting 2 different phyla or any of the 7 spacer matches present in a viral group potentially infecting 3 different families are shown in **a** and **b**, respectively. Alignments were performed by mapping all the matches (48 for **a**, and 128 for **b**; named with an IMG scaffold identifier) to a viral representative using the 'map to reference' package of Geneious software (http://www.geneious.com). Black lines represent 100% sequence identity to the reference virus. The location of the 3 spacers (that derived from 2 different phyla) in **a** as well as the 7 spacers (that derived from 3 different families) in **b** is indicated with triangles with different colours. Spacer sequences, as well as the genomes that contain them in a CRISPR locus is boxed at the bottom.



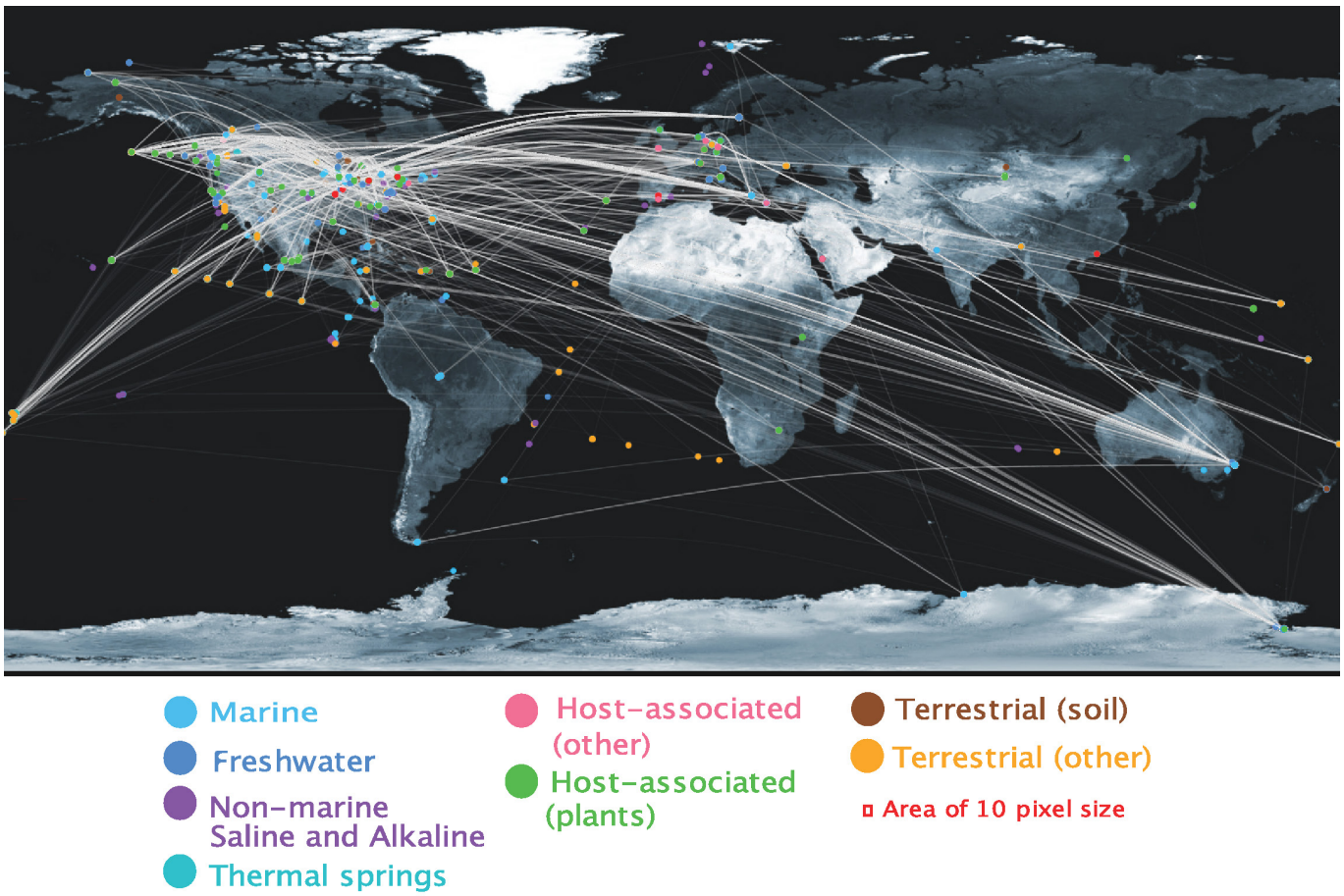
Extended Data Figure 7 | Habitat type specificity of all viral diversity and specific examples. a, Distribution of the presence of the total viral diversity of metagenomic viral contigs (viral groups and singletons) across distinct number of habitat types. A total of 85.9% of all viral diversity resided in a single habitat type (either as a singleton 19.7%, as a viral group found in a single sample 1.8%, or as a viral group found in 2 or more samples 64.4%), whereas only a small fraction (0.31% of all mVCs) were found in 4 or more different habitat types. **b**, **c**, Examples of viral groups found in diverse samples across different oceanic zones and provinces. Presence of a single viral group across distinct marine samples based on average coverage values (red bars; *y* axis on the left) and total percentage of the viral sequence length recovered per sample (purple line; *y* axis on the right). Samples were grouped by marine zones and indicate the percentage of the total samples per zone. **b**, Representative of viral group 2463 (IMG taxon id: 3300001450 and IMG scaffold id: JGI24006J15134_100002847) was found exclusively in marine biomes at depth and with reduced exposure to sunlight (across 95% of all twilight samples and in 44% of deep ocean samples). **c**, Representative of viral group 10643 (IMG taxon id: 3300000216 and IMG scaffold id: SI53jan11_150mDRAFT_c1002499) detected preferentially across coastal water samples (28% of all samples of this zone, preferentially in oxygen minimum zones), but also present in twilight, deep ocean, and hydrothermal vent samples. This viral group was identified as a SUP05-infecting phage. The genes of the viral contig representatives were coloured by the phylogenetic distribution of the best hit in the database (white, unknown; green, Proteobacteria; blue, Chlorophyta, red, unclassified virus; turquoise, Firmicutes; purple, Deinococcus). **d**, **e**, The distribution of viral sequences of distinct body sub-sites across different individuals. Hierarchical clustering (average linkage using Jaccard distance) was used for both axes (samples and individuals) across 'large intestine' (**d**) and 'oral' metagenomes (**e**), respectively (top chart in both panels). Presence or absence of viral groups or singletons per sample is colour-coded as red or blue, respectively. The line chart of both panels show the percentage of viral sharing for >50%, 50–10%, and <10% of the individuals (vertical lines) highlighting in red boxes the percentage of viral sharing for >80% as well as viral sequences only present in a single individual.



Extended Data Figure 8 | Alignment of broad-host specificity prophage in 20 isolate genomes in IMG using 'Gene Neighborhood' tool. The gene 'adenine-specific DNA methyltransferase' is used as an anchor for the alignment (in red). Genes are coloured according to COG cluster annotation, with light yellow representing genes without COG assignment. Blue boxes highlight likely cargo genes inserted in prophage genomes. These include: (1) alkyl hydroperoxide reductase system in *Dehalogenimonas lykanthroporepellens, Desulfococcus biacutus* and *Geobacter sulfurreducens*, (2) efflux ABC transporter in *Desulfoarculus baarsii* and *Desulfobacterium anilini*, (3) possible secondary metabolite biosynthesis genes in *Desulfovibrio aespoenensis*, (4) restriction system in *Desulfovibrio paquesii* and *Geoalkalibacter subterraneous*, (5) methionine synthase in *Desulfovibrio sp. L21-Syr-AB*, (6) molybdate ABC transporter in *Desulfomicrobium thermophilum*, (7) ABC transporter involved in multi-copper enzyme maturation in *Desulfovibrio alkalitolerans*; and (8) likely antibiotic resistance cassette in *Geobacter soli*. Details in Supplementary Table 24.



Extended Data Figure 9 | Distribution of hits to broad-host prophage and its potential hosts in metagenomic samples. The hits to prophage sequences and host marker genes (RNA polymerase subunits and ribosomal proteins) were identified by BLASTn with *e*-value 1.0×10^{50}; 90% nucleotide identity and cumulative alignment length of at least 10% of the length of the prophage or concatenated marker genes. Metagenome samples grouped by habitat are shown on the *y* axis; boxes correspond to broad environmental categories. Red box surrounds non-human host-associated samples (worm and termite symbionts), green box surrounds environmental samples (aquatic and terrestrial), blue box surrounds engineered samples (wastewater and bioreactors). Average coverage of the prophage and concatenated host marker genes is plotted on the *x* axis.



Extended Data Figure 10 | Global connectivity of viral diversity from different habitat types. Geographic location of metagenomic samples containing the same viral groups and singletons represented by a white connecting line across metagenomes from different habitats. Only samples sharing 2 or more viral groups or singletons that are more distant than 10 pixels (area shown as a red square in the figure) are connected. The colours of the samples (circles) indicate the habitat type according with the legend. A freely available equirectangular projection of the world map was used as a background image (http://visibleearth.nasa.gov/view.php?id=57752).