doi:10.1038/nature19094

Summary of putative viral sequence detection

We identified 125,842 putative viral sequences in 1,882 metagenomic samples. These samples were distributed as follows according with the distinct habitat types manually curated according to sample metadata: 98 metagenomic viral contigs (mVCs) from 13 air samples; 7,970 mVCs from 154 engineered samples; 393 mVCs from 49 thermal spring samples; 4,538 mVCs from 41 non-marine saline and alkaline samples; 5,596 mVCs from 99 host-associated others samples; 20,587 mVCs from 178 freshwater samples; 50,857 mVCs from 397 marine samples; 2,840 mVCs from 64 terrestrial soil samples; 1,233 mGVs from 91 host-associated plants samples; 343 mVCs from 32 terrestrial others samples; and 30,849 mVCs from 664 host-associated human samples. Note that we did not report in this manuscript the Air samples since they appear to be highly contaminated with human DNA.

Putative closed metagenomic viruses

We identified 999 closed mVCs based on overlapping 3' and 5' regions **Supplementary Table 10**) with an average size of 53,644 bp ± 45,677 bp s.d., which is consistent with the calculated average length of known isolate viruses (44,296 bp ± 83,777 bp s.d.). We manually reviewed a large number of mVCs including the largest phage genome reported thus far as well as six other incomplete viral metagenomic contigs (ranging from 350 kb to almost 600 kb) from a diversity of ecological niches (**Supplementary Table 12**).

Similarly as for all mVCs, we calculated the coverage by viral protein families of only closed mVCs to that of isolate viral genomes. Our results revealed that only 8.9% of the closed mVCs are placed within the high similarity category, with 58.9% placed within the low similarity one.

Longest phage genome

The longest phage genome detected thus far is predicted to be viral due to several lines of evidence such as: (a) the contig has been identified to be circular (by using 2 distinct and independent assemblers); (b) no recognizable plasmid replication or transmission elements have been detected and therefore we exclude the possibility that it is a plasmid; (c) tetranucleotide binning of all the contigs in the sample where this contig was identified shows that there is no relation of the putative viral contig to any bacterial or archaeal groups; (d) 11% of the genes on this contig are identified as viral genes; (e) the viral genes are distributed across the entire length of the contig; (f) the identification of two phage tail-like genes, a phage T4-like capsid assembly gene, a phage-type DNA polymerase, a phage shock protein, a phage baseplate assembly protein W, and a terminase-like gene (**Extended Data Fig. 4b** and **Supplementary Table 11**). The majority of the genes (80%) were not assigned to any known viral protein family (VPF) and their functions remain hypothetical, supporting the novel nature of this unusually large phage.

Benchmarking and comparison of vHMM pipeline and VirSorter for virus detection in metagenome datasets

Taking into account all sequence fragments from the synthetic metagenome generated to assess the accuracy of our vHMM virus detection pipeline (previously described in the **Methods** section), the precision of our pipeline was 99.6% with the recall rate of 37.5%, as compared to the precision of VirSorter of 94.5% and recall of 65%. Precision and recall of both pipelines for fragments of different lengths are shown in **Extended Data Fig. 3**. We have manually analyzed the false positive cases for both pipelines, and found that the two main sources of VirSorter's false positive predictions were plasmid replication genes with similarity to phage replication systems and ribonucleotide reductase enzymes found on shorter fragments. Additional chromosomal context, such as bacterial housekeeping genes or plasmid conjugation proteins, eliminated many of these false positives in the case of longer fragments. On the other hand, vHMM pipeline identified plasmid CP002842 in *Halopiger xanaduensis* SH-6 as potential phage, based on 5 hits to viral HMMs and very low number of genes with Pfam and KO hits, suggesting that short fragments of highly divergent plasmids may be misidentified as viruses by vHMM pipeline.

Since many fragments with prophage sequence also contained pieces of host chromosomes or plasmids, sometimes longer than the prophage sequence itself, we have calculated the number of fragments that had at least 10 kb of host sequence, taking into account VirSorter predictions of prophage coordinates for the categories 4 and 5. As shown in **Extended Data Fig. 3**, the percent of the fragments in 30-50 kb range with significant host "contamination" identified by VirSorter is twice as high as that of vHMM pipeline. This is not entirely unexpected given the difficulty of identifying prophage boundaries in fragmented sequence data. Furthermore, it is the stringent filtering of the contigs detected by viral HMMs using cutoffs on the percent of genes with Pfam and KO assignments, that removes fragments with significant fraction of host sequence, but also negatively affects the recall rate of vHMM pipeline (data not shown).

Both higher false positive rate of VirSorter and higher amount of host sequence detected as viral, represent a problem for a pipeline designed for viral sequence detection in untargeted metagenomic data. Even enriched viromes are known to be contaminated with cellular sequences, which hampers functional characterization of viral communities and can confound analysis of viral distribution[1]. Furthermore, when the sensitivity is computed on per-viral-sequence basis rather than per-fragment basis, vHMM pipeline detects at least one fragment for 116 out of 137 viral sequences (132 prophages and 5 phages), with an average of 9 fragments per viral sequence, which corresponds to an overall recall rate of 84.5%. In comparison, VirSorter detects 119 out of 137 viral sequences, with an average of 13.5 fragments per viral sequence, which corresponds to an overall recall rate of 86.9%. Interestingly, VirSorter classified more prophage sequences from the fragmented data into categories 1, 2, 4 and 5 than from the finished genome sequences. Additionally, in 3 cases adjacent prophages detected by VirSorter as a single prophage in finished genomes, were identified as separate prophages in fragmented sequences. Overall, our vHMM pipeline has higher specificity and much lower per-contig sensitivity, but comparable per-virus sensitivity to VirSorter. Given that the goal of our study was to study the overall viral diversity in the datasets dominated by host sequences, we opted for an approach with higher specificity, and supplemented it with BLASTn search using detected viral sequences as a query to identify shorter viral sequences nearly identical to those in our trusted set (see below).

We supplemented the benchmarking of our viral identification pipeline by analysing the rate of recovery of real metagenome contigs containing at least one of seven Pfams models found only in viruses, but not in cellular organisms. These Pfams (**Supplementary Table 5**) include models specific for eukaryotic and prokaryotic viruses, and they have the highest number of hits in all metagenomic datasets used in our study. Our vHMM pipeline detected 59.62% of all metagenomic sequences longer than 5 kb that had at least one gene with a hit to one of the seven viral-specific Pfams (a total of 8,575 out of 14,381 mVCs). Of the remaining 40.38%, 60% were identified as candidate viral sequences by the vHMM pipeline, but were removed by the filtering step as having too many genes associated with KO terms and/or Pfams, suggesting that they may have a significant fraction of host sequence (**Supplementary Table 6**). The remaining contigs with hits to viral-specific Pfams were not detected as candidate viral sequences by vHMM pipeline (**Extended Data Fig. 1c**) because their gene content was too divergent from the viruses in our training set and indicating that there is significant potential for improvement of the sensitivity of viral detection by performing additional iterations of clustering of viral proteins and building of vHMMs.

In addition, we have run CyVerse implementation of VirSorter on the entire set of 125,842 metagenomic viral contigs (mVCs) detected by our vHMM pipeline. Only 27% of the contigs were identified as viral by VirSorter, including 93% classified as virus (categories 1 and 2) and 7% classified as putative prophages (categories 4 and 5). We have manually inspected the longest contigs rejected by VirSorter and identified hallmark viral genes, such as capsid and tail proteins, on all of them. We conclude that our vHMM pipeline and VirSorter partially overlap in their predictions, with our pipeline specifically targeting highly divergent viral sequences and VirSorter having higher sensitivity in detection of short fragments of viruses similar to those in reference databases. Both tools will likely benefit from an expanded set of viral protein, which can be used to improve the sensitivity of vHMM pipeline and the specificity of VirSorter by more precise identification of prophage boundaries.

Comparison of AAI clustering approach to MCL clustering approach by Roux et al.

We reanalyzed the data from Roux et al., 2015[2] using our clustering approach (at least 90% average amino acid identity and 50% alignment fraction between encoded proteins). Only 83% of sequences could be clustered into 562 clusters using our method, as compared to 99.3% of sequences clustered into 614 clusters reported by Roux et al. In addition, the 10 largest clusters generated by our method included only 4,359 sequences, as compared to 5,809 sequences in the 10 largest clusters described by Roux et al. This difference in cluster granularity can be explained by the differences in clustering algorithms (90% amino acid identity cutoff versus percent of shared members in protein clusters based on reciprocal BLAST hits with e-value of 0.001 and bit score of 50). It is also consistent with the observation of Roux et al. that their sequence clusters correspond to genus level and are larger than our viral groups, which mostly clustered at species level.

Among the viral groups generated from the data of Roux et al. using our approach, we found only 5 clusters that included sequences with hosts from different classes or different phyla (based on the host taxonomy provided by Roux et al.). Unfortunately, we could not verify the taxonomic origin of the following 10 sequences from these clusters: Negativicutes_gi_546349318_12163_37890, Clostridia_gi_547873663_11997_37723, Bacteroidia_gi_557506781, Negativicutes_gi_545588341, Clostridia_gi_545037736, Betaproteobacteria_gi_483224433, Gammaproteobacteria_gi_483729537, Zetaproteobacteria_gi_483166662, Zetaproteobacteria_gi_483190635 and Clostridia_gi_545037542_3677_36763. It appears that these contigs or the genomes, from which they were extracted, have been removed from the NCBI's nucleotide non-redundant database, possibly due to contamination or quality problems, since these sequences cannot be found either by their gi numbers or by BLASTn against NCBI NR. After exclusion of these sequences of unclear origin, the remaining clusters had no inconsistencies in the host taxonomy at class level. Therefore we conclude that our observations of broad host specificity are not directly comparable to those of Roux et al., since they are obtained at different levels of taxonomy. In addition, both examples of broad specificity phages provided in our manuscript are of likely lytic phages rather than mostly temperate phages studied by Roux et al.

Determination of Host-virus association

CRISPR-Cas spacer approach to identify host-virus connections

Clustered regularly interspaced short palindromic repeats (CRISPRs), that contain viral sequences in prokaryotic genomes (spacers), have been used as a method to link viruses with their respective hosts due to their specificity, across distinct environments such as marine[3], human-associated[4], acid mine drainage[5], thermal vents[6], Arctic and Antarctic samples[7] as well as in modeled isolate genomes[8,9]. A BLAST E-value of a CRISPR spacer perfect match of 30 nt (average spacer length) in a database as large as 5 Tb is as low as 1e-03 and the expected number of occurrences of just one 30 bp spacer in any of the two strands of a 30 kb-long phage is 5.2e-14. Therefore, since CRISPR-Cas spacers have been widely used to identify host-virus connections, we implemented a large-scale search using a custom spacer database of 3.5 million sequences (developed for the purposed of this study) to link putative microbial hosts to the mVCs. Additional details can be found in the **Methods** section.

viral-tRNAs approach to identify host-virus connections

Transfer RNAs (tRNAs) are non-coding ribonucleic acids responsible for decoding a messenger RNA (mRNA) sequence in order to synthesize protein and thus ensures the precise translation of genetic information that is imprinted in DNA[10]. These 76-90 bp long sequences constitute one of the most abundant and diverse (>12 k) groups of nucleic acids[11]. After validating the hypothesis that viral tRNAs could be recruited from their host using iVGs, we queried all tRNAs found in mVCs (using ARAGORN v1.2 software; details in **Methods** section) against all available isolate genomes. We discarded the twenty most abundant viral tRNA sequences (listed as "Promiscuous" in **Supplementary Table 22**) since they were sequences highly conserved across many members (from 184 to 1227 hits) of the Gammaproteobacteria. In order to evaluate the prevalence of tRNAs in mVCs across various habitats we calculated the proportion of mVCs that contained at least one viral tRNA per habitat we observed they ranged from a maximum of 16.6% in host-associated (plants) to a minimum of 5% in the marine samples. We also calculated the average number of tRNAs per mVCs -that contained at least one of them- per habitat. All values ranged from 1.6 to 5.6 tRNAs as average from terrestrial (others) habitat and non-marine saline and alkaline habitat, respectively.

Phage clusters of Orthologous Genes (POGs) to assign viral taxonomy

Traditional classification of viruses uses phenotypic traits such as morphology (size and shape), chemical composition and structure of the genome, and mode of replication[12]. This classification has been established in two different systems: the International Committee on Taxonomy of Viruses (ICTV; http://www.ictvonline.org/) and the Baltimore classification[13]. According to the ICTV classification system, 21 POGs[14] (phage orthologous gene clusters) have been recognized as taxon-specific signatures for viral classification. We first validated this classification method using 206 iVGs for which the taxonomy was also detected via POGs. Next, we used these 21 POGs to assign taxonomic lineage to 1,165 mVCs (from 201 viral groups and 509 singletons) (**Supplementary Table 15**).

Detection of putative ssDNA viruses

The average size of all single-stranded DNA viruses deposited in NCBI corresponds to ~3.7 kb (stdev +/- 1.9 kb; median of ~2.8 kb). This, besides with the requirement in our viral detection pipeline of having exclusively viral contigs longer than 5 kb, may have limited the detection of ssDNA viruses in our analysis. However, the fact that we recruited more than 93% of all the available ssDNA viral proteins in our final set of VPFs (4,082 out of 4,371) allowed us to here hint the presence and distribution of putative ssDNA viruses. We selected all 59 POGs assigned to ssDNA viruses (not any of them recognized as taxon-specific signature) and searched for protein sequence similarity (e-value threshold of 1e-04) within our predicted list of 125,842 k viral sequences. A total of 8,252 mVCs from 1,303 samples (across all habitats) contained at least one hit (range 1-25 hits) (**Supplementary Table 7**).

Newly identified bacterial and archaeal lineages putatively infected by viruses

We identified a total of 24 microbial phyla putatively infected by mVCs, of which 16 have not been previously reported to be infected by viruses. These unreported hosts include a total of 590 mVCs putatively infecting members of microbial phyla without cultured representatives (Aigarchaeota, Nanohaloarchaeota, SR-1, Cloacimonetes, and Marinimicrobia) as well as phyla with cultured representatives (Chloroflexi, Caldiserica, Synergistes, Thermatogae, Fusobacteria, Aquificae, Acidobacteria, Spirochaetes, Verrucomicrobia, Fibrobacteres, and Chlorobi) (**Fig 2**). For the remaining 8 host phyla where viral sequences were previously assigned to be infecting, we expanded the number of known viral sequences from 742 to 9,479. In total we linked mVCs to 284 unique microbial genera, from which ~80% were not previously identified (**Supplementary Table 23**). Interestingly, these novel genera include opportunistic pathogens responsible for a number of diseases in plants and humans. These included 50 viral groups and 33 singleton sequences infecting hosts that cause multiple infections. Among these opportunistic pathogenic hosts, we found phages of Fusobacteria as Fusobacterium (important because it is associated with pharyngotonsillitis and Lemierre diseases, it has a zoonotic host range from humans to cattle and pigs and it is resistant to penicillin[15]) and Leptotrichia (involved in septicemia and endocarditis, resistant to many antibiotics and without any vaccine available[16]). We also found as a viral host species of Gardnerella (most important *G. vaginalis* that is involved in bacterial vaginosis as a result of a disruption in the normal vaginal microflora[17]), Neisseria (*N. meningitidis*, *N. sicca*, and *N. subflava* responsible for meningitis, septicemia and other opportunistic infections[18]) and others (details in **Supplementary Table 24**).

Viruses identified with broad-host range

The generally accepted view is that most viruses are specialists, and even viral generalists are known to infect a narrow range of hosts (i.e. different species within the same genus[19,20]). Here, we further explored the possibility of identifying putative broad host-range phages via CRISPR-Cas spacer complementarity. We observed a clear trend (91% of the viral groups) towards host specificity since such viral groups were exclusively targeted from spacers of genomes from single species or from the same genus or species (**Supplementary Table 19**). Interestingly, in the remaining 9% of the viral groups we found spacer matches derived from genomes of higher taxonomic distances (from family to phylum levels; **Fig 3a**). In this way, we identified 6 exact spacer matches in two nearly identical viral sequences (clustered together in the viral group 5594, AAI >97%) from two different human stool metagenomic samples. The spacers derived from *Roseburia inulinivorans* (2 spacer sequences), *Eubacterium rectale* (1 spacer sequence), and *Ruminococcus* sp. SR1/5 (3 spacer sequences), all belonging to three distinct families within the Clostridiales order and were isolated independently from human fecal samples (**Fig. 3c**). The two spacer sequences from *R. inulinivorans* were clustered in the same CRISPR array identified as a subtype II-B based on the CRISPR associated (Cas) gene content. Examples from subtype II-B CRISPR-Cas system in bacteria require the presence of a short (GG) adjacent motif (PAM) close to the 3' end of the proto-spacer (spacer sequence within the phage genome) to be incorporated into the microbial array. We observed the exact motif in the two viral members of the group 5594 for both proto-spacers derived from *R. inulinivorans*. Similarly, we investigated the three spacers detected in *Ruminococcus* sp. SR1/5 with a match to the members of the viral group 5594 and found that two of them belong to one CRISPR array (subtype I-C). The corresponding identical putative PAM (TTC sequence at the 5' end of the protospacer) was detected for those two sequences.

More striking, we detected 3 spacer matches in 3 out of the 32 viral sequences within the viral group 3098 (lowest AAI value of 52%; from 24 different human oral samples in saliva, buccal mucosa, tongue and palate). In this case, the spacers derived from *Actinomyces* sp. oral taxon 180 and *Streptococcus plurextorum* that belong to different phyla (Actinobacteria and Firmicutes, respectively) and were isolated from human and pig microbiomes respectively (**Fig. 3b**). The two spacers incorporated in one of the CRISPR arrays in *Actinomyces* sp. oral taxon 180 appeared to belong to the unknown CRISPR-Cas subtype III-U from which no PAM has been yet identified[21].

To verify our results we searched the entire sequence space of the IMG database (assembled contigs and unassembled reads, with total size of approximately 5 Tb) for matches to these spacers, allowing for with no more than 2 nucleotide changes to spacer sequence (greater than 95% sequence identity over 95% of the length of the spacer). We recovered 49 contigs containing matches to at least 1 out of 3 spacer sequences for the first phage (**Fig. 3b**), all of them from human oral metagenomes (**Extended Data Fig. 6a**). For the second phage (**Fig. 3c**) we recovered 130 contigs containing matches to at least 1 of the 7 spacers (**Extended Data Fig. 6b**). All of the matching contigs longer than 5 kb were identified as phages by our pipeline. We additionally aligned the shorter contigs to the longest sequence and found that all of them have at least 83% nucleotide identity to the longest contig over the entire length of the shorter contig. We further reassembled the samples in which the longest contigs were identified using a different assembly pipeline and found that they likely represent entire genomes of the respective phages, because the contigs could be circularized.

To assess the extent of sequence conservation of these spacer sequences, we relaxed our search criteria and found several matches with 90-95% sequence identity over 95% spacer length, which allows for at most 4 nucleotide changes. These matches were found exclusively in the viral sequences from the same viral cluster (vc_3098) and are consistent with an observation that phages quickly evolve to evade the CRISPR immunity and mutate the sequences susceptible of acquisition by CRISPR system (defined in most cases by a PAM sequence[9]). It should be noted that all proto-spacer sequences are located in protein-coding genes including structural genes and a holin, none of which is broadly conserved among phages. Thus, we conclude that they are unlikely to be under selective pressure to strictly maintain their DNA sequence.

The first of the two predicted broad specificity phages (**Fig. 3b**) was found in human oral metagenomes and two of the three spacers hitting it were found in an oral isolate sequenced as part of Human Microbiome Project (*Actinomyces* sp. oral taxon 180 F0310). The third spacer was found in *Streptococcus plurextorum* DSM 22810 isolated from a pig with pneumonia. All of the contigs with matches to the spacers in an expanded BLAST search were also found in human oral metagenomes.

The second of the predicted broad specificity phages (**Fig. 3c**) was found in human fecal metagenomes and the spacers were from fecal isolates *Eubacterium rectale* ATCC 33656, *Roseburia inulinivorans* DSM 16841, and *Ruminococcus* sp. SR1-5. The contigs with matches to the spacers in an expanded search were found mostly in human fecal metagenomes, as well as in porcine fecal metagenomes and 2 biogas bioreactor samples.

In both cases we can exclude the possibility of acquisition of the spacer(s) via lateral gene transfer, since the structure of the CRISPR arrays (both repeat sequences and contiguous spacers) is completely different. The combination of our findings suggests the existence of very wide host-range phages (phages able to infect from different families up to different phyla levels) within a habitat as well as the ability of their hosts to counterattack them via CRISPR-Cas system (**Supplementary Table 19**).

Cosmopolitan viral diversity

Our data indicate habitat type specificity for the vast majority of the viral groups and singletons (**Fig. 5** and **Extended Data Fig. 7a**). However, there were notable exceptions where we found the same viral sequence across different habitats (**Fig. 5b, c** and **Supplementary Table 27**). After excluding ambiguously classified cases, most mVCs detected in more than 1 habitat type were found in the samples from the same environmental category (e.g. 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 as either likely laboratory contaminants, or prophages with broad host specificity. However, in some cases we found examples of nearly identical members from the same viral group spanning distinct habitat types. For example, the viral groups 6422 (99% intra group AAI) and 4116 (91% intra group AAI) were independently sequenced, and were in both aquatic (wastewater; IMG sample identifier 3300001196) and host-associated (human oral and stool metagenomes; IMG sample identifier 7000000048) habitats, respectively. Similarly, the viral group 16819 was detected in aquatic (epilimnion zone from Lake Mendota, USA; IMG sample identifier 3300002835) and terrestrial (deep subsurface sample from Australia; IMG sample identifier 3300002733) habitats spanning very distinct habitats as well as extensive geographic distances. Another example was the viral group 1417 whose members were found in samples from geographically distinct freshwater lakes (IMG sample identifiers 3300003277, 3300003393, 3300003411, 3300003388, 3300003404, 3300000756), marine sediments (IMG sample identifiers 3300000126, 3300000792, 3300000124) and an Antarctic hypersaline lake (IMG sample identifier 3300001130), environments displaying large differences in salinity. In addition, members of the viral group 12572 were found in wastewater samples (petrochemical from Alberta, CA; IMG sample identifier 3300001567) and marine samples (from minimal oxygen zones; IMG sample identifiers 3300002599, 3300000137) as well as members from the viral group viral group 12236 were detected in wastewater (IMG sample identifier 3300001197) and hydrothermal vents (IMG sample identifier 3300001680) (details in **Supplementary Table 19**).

Identification of putative prophages among mVCs

A total of 7,321 mVCs had hits to one or more isolate genomes with nucleotide percent identity greater than 80% and cumulative alignment length on mVC greater than 75% of its length, suggesting that these are likely prophages (whole list in **Supplementary Table 3)**. Analysis of host distribution for these 7,321 mVCs revealed several groups with hits in genomes from multiple orders, classes and even phyla. 4 of these broad-specificity mVC groups were also ubiquitous in metagenomes from different habitats and environmental categories. One of the mVC groups had hits with >90% nucleotide identity to prophages in 19 deltaproteobacterial genomes from 3 different orders and one Chloroflexi genome (**Supplementary Table 4**). Between these 20 genomes, prophage proteins were the only ones with >90% amino acid identity, to the exclusion of normally more conserved housekeeping genes, such as those encoding ribosomal proteins. Analysis of these prophages showed that they would be grouped into a quasi-species based on AAI and alignment fraction. Only 1 sequence from this viral group has been found by the previous study (Deltaproteobacteria_gi_544815434)[19] despite the availability of at least 7 host genomes at the time of the study. The remarkable feature of this broad-host specificity prophage group was the presence of multiple and diverse sets of cargo genes (**Extended Data Fig. 8**), which included 3 different types of ABC transporters, restriction systems, methionine synthase, potential cassettes for biosynthesis of an unknown secondary metabolite and antibiotic resistance. We have analyzed the presence and distribution of this prophage group and its known isolate hosts (**Supplementary Table 4**) in metagenome samples using BLASTn approach for detection of low-abundance viruses (**Methods** section). Nucleotide sequences of 3 core subunits of DNA-dependent RNA polymerase and 30 ribosomal proteins were used as markers of the host genome with the same evalue 1.0e-50 and 90% nucleotide identity cutoffs, and filtering at cumulative alignment length of at least 10% of the length of concatenated marker genes. Hits of the marker genes or prophages were found in assembled and/or unassembled data of 183 metagenome samples, however, both of them were found in only 24 samples. Hits of the marker genes alone were found in 85 samples and hits of the prophage only were found in 74 samples (**Extended Data Fig. 9**). Previously the discordance between the presence of hosts and prophages in different samples was attributed with prophage induction[22], however, a more parsimonious explanation in the case of this prophage group is that it has an even broader host specificity than that observed in isolate genomes (mostly Deltaproteobacteria and Chloroflexi). Among the hits of the prophage sequences in metagenomic data there were two contigs over 100 kb long, containing significant portions of bacterial chromosomes in addition to prophage fragments. Examination of bacterial housekeeping genes found on these contigs indicates that one of them (IMG taxon 3300003817, contig Ga0056122_10000001) belongs to a member of Gammaproteobacteria with high similarity to *Sedimenticola* spp., while another (IMG taxon 3300005254, contig Ga0068714_10000126) likely originates from a member of the Nitrospirae phylum. These findings support the hypothesis that many "cosmopolitan" viruses are prophages with very broad host specificity, infecting hosts with disparate lifestyles from many divergent lineages including different phyla, which in turn explains the surprising global ubiquity of these viruses.

Estimation of viral abundance in fecal and oral metagenomes from Human Microbiome Project.

The heterogeneity of metagenome datasets available to us made it impossible to estimate viral fraction in metagenomics data globally, since some of the samples have assembled data, but no coverage information, other samples have no unassembled data, which could not be mapped to assembled sequences, while some samples are missing both coverage information and unassembled data. In addition, the samples varied greatly by the total amount of high-quality sequence generated, as well as by the sequencing technology (Sanger, 454, Illumina, or a combination of any of them). To obtain a ballpark estimate of the abundance of viral sequences in a large set of samples generated by untargeted metagenomics, we analyzed 550 fecal and oral samples from the Human Microbiome Project. These represent the largest collection of samples available to us generated using consistent processing protocols, comparable sequencing efforts and the same sequencing technology (http://hmpdacc.org/). Despite great efforts to achieve consistency in sample processing, we found very large variation in the amount of high-quality sequence per sample, ranging from less than 8Mb for biosample SAMN00045859 (buccal mucosa) to 16.7Gb for biosample SAMN00039874 (stool sample). Despite this variation, we detected some viral sequences in all biosamples except SAMN00045859. The fraction of viral sequences ranged from 0.2% (stool sample SAMN00037803 with unusual taxonomic composition dominated by Proteobacteria) to 54% (stool sample SAMN00081451). Per-sample amount of high quality sequence and viral fraction is provided in **Supplementary Table 25**. The average amount of viral sequence in fecal samples was 7.4% and in oral samples it was 3.4%. The average viral fraction for saliva samples was 9%; however, there are only 5 saliva samples in the dataset. 28 metagenomes including 7 oral and 21 stool samples had more than 10% of viral sequences.

crAssphage[15] was found in 64 out of 140 fecal samples, comprising from 0.00004% to 27% of the high quality sequences. Although it was the single most abundant virus in multiple samples, we identified other highly abundant viruses accounting for at least 5% of total sequence. As an example, 2 stool samples, SAMN00035505 and SAMN00081451, collected from the same individual had the highest fraction of viral sequences (47 and 54%, respectively). These samples had high abundance of viral groups 3701 and 10167 (represented by SRS058770_LANL_scaffold_30328 from IMG taxon 7000000308 and SRS017247_LANL_scaffold_14307 from IMG taxon 7000000178, respectively), which together accounted for more than 30% of the total high quality sequence in these datasets. These two groups were present in 31 and 49 samples, respectively. Remarkably, the third sample from the same individual (SAMN00070431 from the second visit) had only 5% viral fraction, and neither dominant viral group from the first and third visit was detected. Likewise, the most abundant viral group from the second visit, 3126 (represented by SRS024435_LANL_scaffold_23226 from IMG taxon 7000000624) was completely absent from the samples from the first and third visits. While this drastic change may be reflective of the highly dynamic nature of human gut virome, we cannot exclude the possibility of metadata errors. Phages with similarly high abundance were found in oral samples, including viral group 3149 (represented by SRS064423_LANL_scaffold_66139 from IMG taxon 7000000156) and viral group 5442 (represented by SRS011090_Baylor_scaffold_4389 from IMG taxon 7000000105) accounting for about 10% of all sequence in the samples from buccal mucosa and supragingival plaque, respectively. While the former group was found in 57 oral samples from 31 individuals, the second was present in only 6 oral samples from 2 individuals.

The number of viral groups and singletons per sample ranged from 22 (sample SAMN00084455 from buccal mucosa) to 806 (sample SAMN00070253 from tongue dorsum) (**Supplementary Table 25**). The average number of viral groups and singletons found in stool samples was 330, while in oral samples it was 283. For stool samples we found no correlation between the percent of viral sequences and the total amount of high quality sequence data or between the number of viral groups and singletons detected and the total amount of sequence, or between the viral fraction and the number of viral groups. Although for oral samples there was no correlation between the percent of viral sequences and the total amount of sequence, there was a correlation of 0.52 between the number of viral groups and singletons and the total amount of sequence. This may be due to the fact that there is much greater variation in the total amount of sequence between oral samples (from 42.6 Mb to 13.2 Gb) than between stool samples (from 1.7 to 16.7 Gb). Detailed information about the presence and abundance of singletons and viral groups is provided in **Supplementary Table 26**.

Supplementary References

1 Roux, S., Krupovic, M., Debroas, D., Forterre, P. & Enault, F. Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. *Open Biol* **3**, 130160, doi:10.1098/rsob.130160 (2013).

2 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**, doi:10.7554/eLife.08490 (2015).

3 Heidelberg, J. F., Nelson, W. C., Schoenfeld, T. & Bhaya, D. Germ warfare in a microbial mat community: CRISPRs provide insights into the co-evolution of host and viral genomes. *PloS one* **4**, e4169, doi:10.1371/journal.pone.0004169 (2009).

4 Pride, D. T. *et al.* Analysis of streptococcal CRISPRs from human saliva reveals substantial sequence diversity within and between subjects over time. *Genome Res* **21**, 126-136, doi:10.1101/gr.111732.110 (2011).

5 Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37-43, doi:10.1038/nature02340 (2004).

6 Oulas, A. *et al.* Metagenomic investigation of the geologically unique Hellenic Volcanic Arc reveals a distinctive ecosystem with unexpected physiology. *Environ Microbiol*, doi:10.1111/1462-2920.13095 (2015).

7 Tschitschko, B. *et al.* Antarctic archaea-virus interactions: metaproteome-led analysis of invasion, evasion and adaptation. *The ISME journal* **9**, 2094-2107, doi:10.1038/ismej.2015.110 (2015).

8 Paez-Espino, D. *et al.* Strong bias in the bacterial CRISPR elements that confer immunity to phage. *Nature communications* **4**, 1430, doi:10.1038/ncomms2440 (2013).

9 Paez-Espino, D. *et al.* CRISPR immunity drives rapid phage genome evolution in Streptococcus thermophilus. *MBio* **6**, doi:10.1128/mBio.00262-15 (2015).

10 Barciszewska, M. Z., Perrigue, P. M. & Barciszewski, J. tRNA - the golden standard in molecular biology. *Mol Biosyst*, doi:10.1039/c5mb00557d (2015).

11 Juhling, F. *et al.* tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res* **37**, D159-162, doi:10.1093/nar/gkn772 (2009).

12 Gelderblom, H. R. in *Medical Microbiology* (ed S. Baron) (1996).

13 Baltimore, D. Expression of animal virus genomes. *Bacteriol Rev* **35**, 235-241 (1971).

14 Kristensen, D. M. *et al.* Orthologous gene clusters and taxon signature genes for viruses of prokaryotes. *J Bacteriol* **195**, 941-950, doi:10.1128/JB.01801-12 (2013).

15 Citron, D. M. Update on the taxonomy and clinical aspects of the genus fusobacterium. *Clin Infect Dis* **35**, S22-27, doi:10.1086/341916 (2002).

16 Eribe, E. R. & Olsen, I. Leptotrichia species in human infections. *Anaerobe* **14**, 131-137, doi:10.1016/j.anaerobe.2008.04.004 (2008).

17 Machado, A. & Cerca, N. Influence of Biofilm Formation by Gardnerella vaginalis and Other Anaerobes on Bacterial Vaginosis. *J Infect Dis*, doi:10.1093/infdis/jiv338 (2015).

18 Ambur, O. H. *et al.* Genome dynamics in major bacterial pathogens. *FEMS Microbiol Rev* **33**, 453-470 (2009).

19 Koskella, B. & Meaden, S. Understanding bacteriophage specificity in natural microbial communities. *Viruses* **5**, 806-823, doi:10.3390/v5030806 (2013).

20 Salmond, G. P. & Fineran, P. C. A century of the phage: past, present and future. *Nat Rev Microbiol* **13**, 777-786, doi:10.1038/nrmicro3564 (2015).

21 Makarova, K. S. *et al.* Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol* **9**, 467-477, doi:10.1038/nrmicro2577 (2011).

22 Minot, S. *et al.* The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res* **21**, 1616-1625, doi:10.1101/gr.122705.111 (2011).

Supplementary Table Legends

Supplementary Table 1. List of metagenomic viral genomes derived from the RNAp and ESOM approach. List of 66 and 188 sequences derived from diverse metagenomic contigs longer than 20 kb that contained viral bins or contained a viral RNA polymerase gene, respectively, and were not captured using the previous filter of bearing 5 or more viral protein families.

Supplementary Table 2. List of genomes used for benchmarking of viral sequence detection pipeline.

Supplementary Table 3. Coordinates of prophages identified in genomes included in synthetic metagenome.

Supplementary Table 4. Broad host specificity prophage in isolate genomes identified by BLASTn of metagenomic viral contigs.

Supplementary Table 5. Sensitivity of virus identification computational approach. Description and identification of the most abundant metagenomic viral specific protein families (Pfams) from all metagenomic sequences larger than 5 kb. We detailed the number of scaffolds that we found (accessed in November 2015 -where we have a larger metagenomic dataset), and at the time of the study (using 3,042 metagenomic samples). We used only the data from those 3,042 samples to compare the number of viruses recovered by our pipeline as an approach to calculate the sensitivity of our pipeline.

Supplementary Table 6. Missed viral contigs according with viral specific Pfams or KO Terms. Taxonomic identifier of the viral sequences not recovered from our pipeline mainly because of the stringent filtering criteria.

Supplementary Table 7. Genes from metagenomic viral contigs with hits to ssDNA "phage clusters of orthologous genes" (POGs). A total of 8,252 mVCs from 1,303 samples (across all habitats) containing at least one hit (range 1-25 hits) to any of the 59 POGs assigned to ssDNA viruses (not any of them is recognized as taxon-specific signature).

Supplementary Table 8. Samples database. Global collection of 3,042 geographically and ecologically diverse metagenomic samples used in this study from the Integrated Microbial Genomes with Microbiome Samples (IMG/M) system.

Supplementary Table 9. Environmental viral protein clusters without hits to proteins from isolate viruses. Distribution of the 360,043 protein clusters with no hits to proteins in isolate viral genomes by the total number of proteins in the cluster. Clusters of proteins encoded on metagenome viral contigs were generated using the cutoffs of 30% amino acid identity and the length of the alignment between two proteins of at least 80% of the length of the shorter protein.

Supplementary Table 10. Closed metagenomic viral genomes. Identified 999 closed metagenomic viral genomes with sequence length, habitat, viral group or singleton identifier, and percent of genes present in viral protein families.

Supplementary Table 11. Genetic information of the largest phage. Detailed gene information of the scaffold “D1draft_1000006” from the metagenomic sample 3300003310 in the IMG system. Specific information is provided for the 1,148 genes contained in this closed phage, as well as the gene location within the genome.

Supplementary Table 12. The 7 largest phage genomes identified. Genomic length and habitat information of the largest phage contigs identified in this work including the largest closed phage genome to date, along with their IMG scaffold and metagenome identifiers.

Supplementary Table 13. List of all reference isolate viruses.

Supplementary Table 14. Viral Genus assignment of isolate viral genomes according to the International Committee on Taxonomy of Viruses (ICTV). Viral group or singleton identifier, the number of members per viral group, as well as the lowest intra group AAI value are reported.

Supplementary Table 15. Metagenomic viral groups and singletons assigned to ICTV viral taxonomy. Assignment of isolate viral genomes and metagenomic viral contigs to ICTV classification based on similarity of phage clusters of orthologous groups (POGs).

Supplementary Table 16. Spacer sequences and host CRISPR sources. List of all the spacer hits of the 3.5 million spacers database against all metagenomic viral contigs.

Supplementary Table 17. Detail of host spacer sequence matches against metagenomic viral contigs. Detailed taxonomy of the metagenomic viral contigs hosts, as well as the spacer sequence location within its CRISPR locus for all the BLAST matches (according with specific cutoff values explained in the **Methods** section).

Supplementary Table 18. CRISPR-Cas spacer matches from our database to the isolate viral genomes. Database of 3.5 million spacers from all bacterial and archaeal isolate genomes and metagenomes that was used to confirmed 98.5% of the matches against isolate viral genomes agreed at the genus or species level.

Supplementary Table 19. Details of habitat and host connections of all the metagenomic viral contigs. Description of all 125,842 metagenomic viral contigs reported in this study with details including length of the scaffold, percent of genes containing matches to viral protein families, viral group / singleton identifier, habitat type detection, and host connection. For the host assignment, we report the method or combination of methods that identified the connection (spacer, tRNA, or reference). When the host for one of the members of a viral group was identified based on any of the approaches, the remaining non-targeted members of the group were also assigned to the same host (using “ext”, for extensive -connected to host by extending it to the totality of members of a viral group).

Supplementary Table 20. Host assignment of isolate viral genomes based on viral tRNAs. Comparison of host taxonomy of isolate viral genomes using the information derived from phage taxonomy (according to the ICTV system) and the viral tRNA matches approach.

Supplementary Table 21. Detail of host assignment of isolate viral genomes using viral tRNA matches. Comparison at the genus level of the host of all isolate viral genomes that contain viral tRNAs and the genus taxonomy of the assigned host according to BLAST matches (perfect matches for the whole tRNA sequence). All the cases where the host agreed at the genus level (92.5% of the cases) are shown in blue, while disagreement is shown in cream.

Supplementary Table 22. Sequence of all metagenomic viral contigs tRNAs. Details of 32,449 tRNA sequences found across 9,555 metagenomic viral contigs. The amino acids for which they code for and the anticodon triplet are shown as well as the location of the tRNA within the metagenomic viral contig. All the tRNA sequences marked with a "yes" in the Promiscuous column (20 sequences) were removed from the study due to its high conservation across Gammaproteobacteria.

Supplementary Table 23. List of host genera connected with isolate viral genomes and metagenomic viral contigs. All the connections shown thus far (from NCBI and ICTV) plus the novel linkages derived from metagenomic viral contigs are shown.

Supplementary Table 24. Microbial pathogenic hosts connected to metagenomic viral contigs. Detailed taxonomy of pathogenic bacterial species for which a viral connection was not previously known, and diseases that they cause.

Supplementary Table 25. Per-sample amount of high quality sequence, viral fraction and number of viral groups and singletons detected in 549 fecal and oral metagenomes from Human Microbiome Project. Body site and body subsite information is according to GOLD.

Supplementary Table 26. Per-sample presence and abundance of viral groups and singletons detected in 549 fecal and oral metagenomes from Human Microbiome Project. Body site and body subsite information is according to GOLD, contig assignment to viral groups and singletons is according to **Supplementary Table 19**.

Supplementary Table 27. Representation of the presence of the same viral group or singletons found across different habitat types. The number of viral groups is shown in the diagonal of the figure, while cross-habitat presence is a red-colored heat map ranging from 0 (in white) to 33,309 (in brightest red). Details of each viral group and singleton habitat(s) can be found in **Supplementary Table 19**.

Supplementary Table 28. Putative prophage from all metagenomic viral contigs.