eLIFE
elifesciences.org

Figures and figure supplements

Viral dark matter and virus–host interactions resolved from publicly available microbial genomes

Simon Roux, et al.

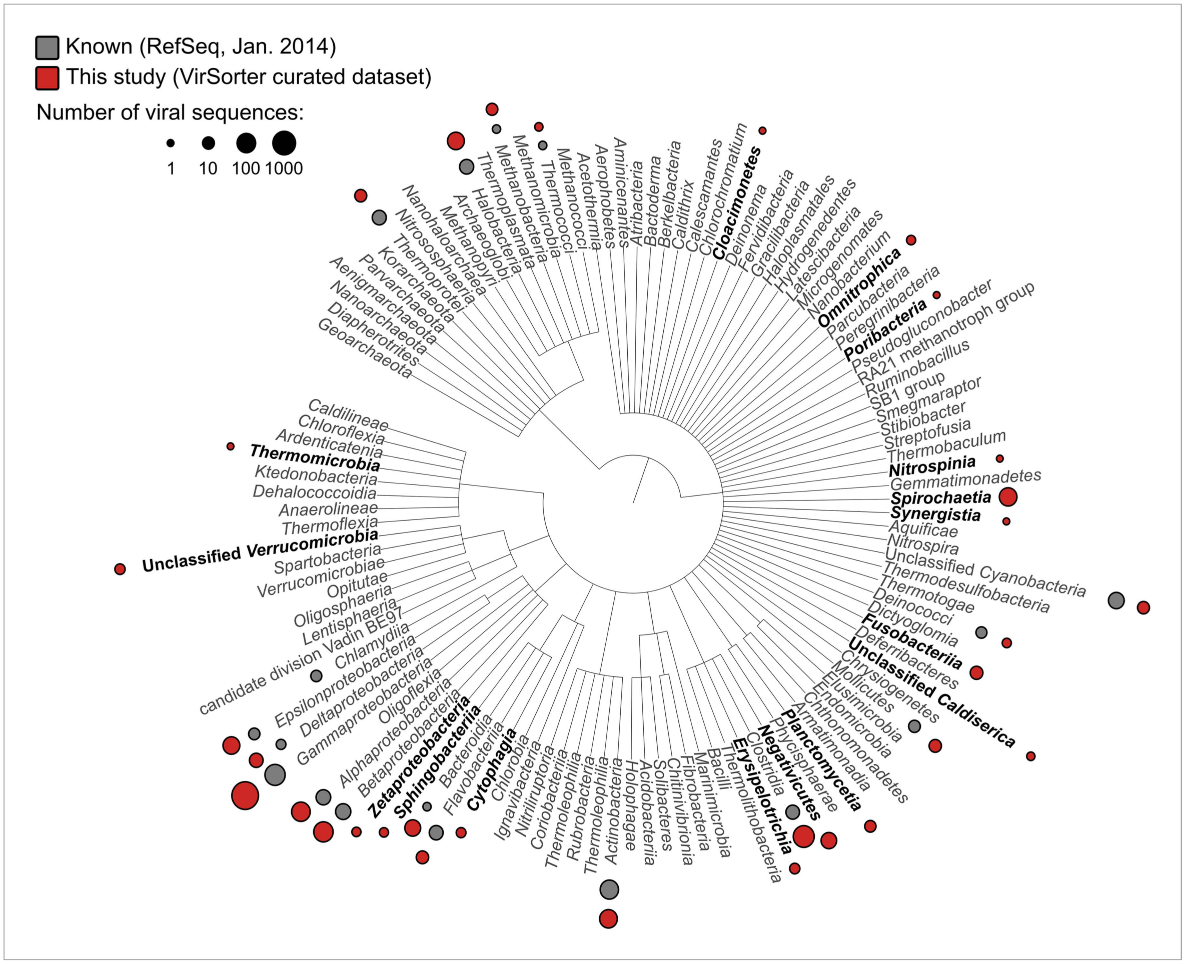


Figure 1. Distribution of viral sequences from the VirSorter curated data set across the bacterial and archaeal phylogeny. For each bacteria or archaea phylum (or phylum-level group), corresponding viruses in RefSeq (gray) and VirSorter curated data set (red) are indicated with circles proportional to the number of sequences available. Groups for which no viruses were available in RefSeq are highlighted in black.

DOI: 10.7554/eLife.08490.003

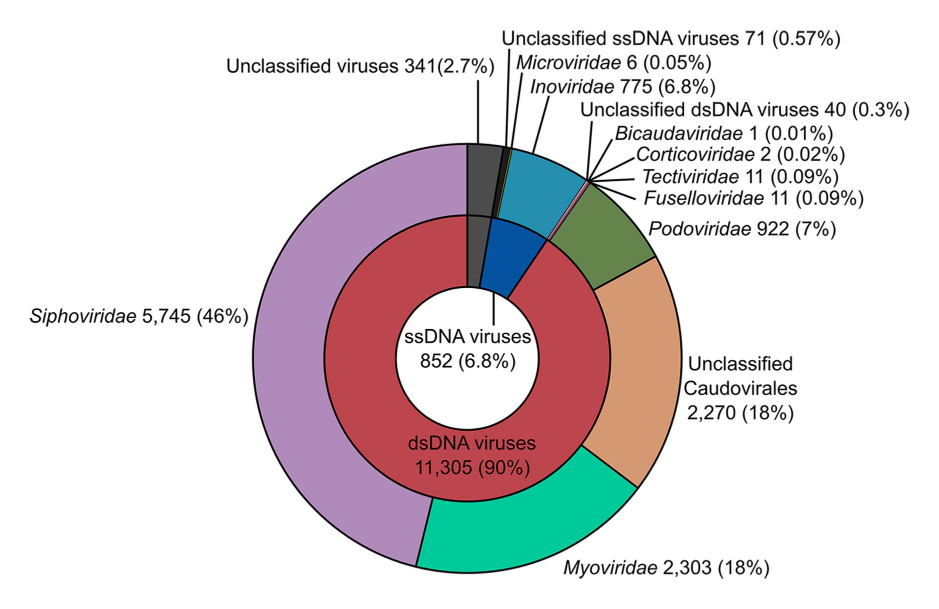


Figure 1—figure supplement 1. Viral diversity in the VirSorter data set. The best BLAST hits of predicted proteins along each sequence (i.e., within 75% of the best BLAST hit for this sequence) were used in a Lowest Common Ancestor affiliation (here displayed at the family level). 'Unclassified *Caudovirales*' gathers viruses only affiliated to the *Caudovirales* level without confident affiliation to the *Myo*-, *Sipho*-, or *Podoviridae*. The number and percentage of sequences affiliated is indicated next to each family.

DOI: 10.7554/eLife.08490.007

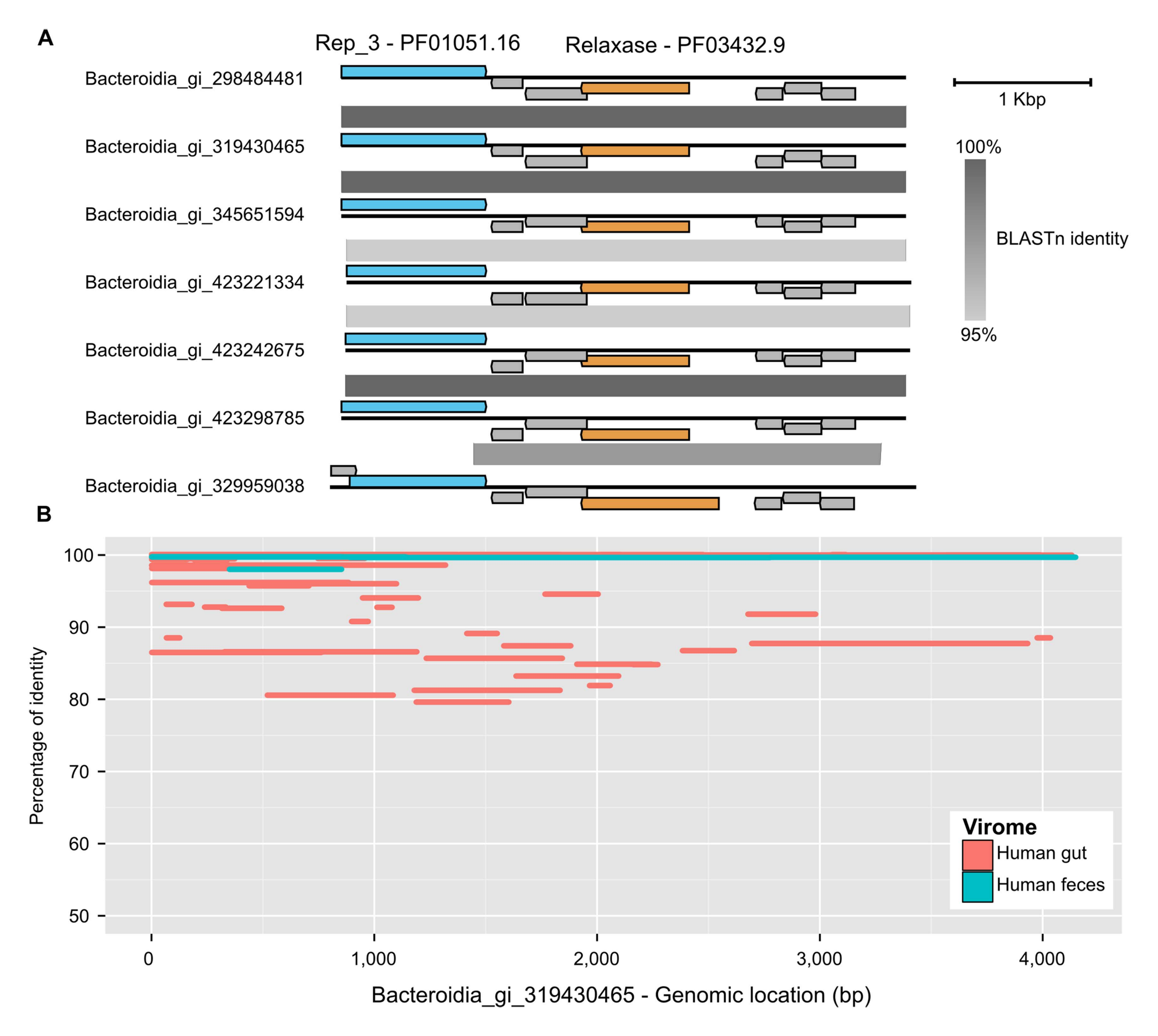


Figure 1—figure supplement 2. Genome map comparison (**A**) and recruitment plot (**B**) of *Bacteroidia* virus sequences from a putative new order. Replication-associated, Relaxase, and hypothetical proteins are depicted in blue, orange, and gray respectively. The recruitment plot includes two viromes from human feces samples from two different studies (Human gut assembly, ***Minot et al., 2012***, and Human feces, ***Kim et al., 2011***). Identity percentage is based on a blastn between virome contigs and the reference genome.
DOI: 10.7554/eLife.08490.008

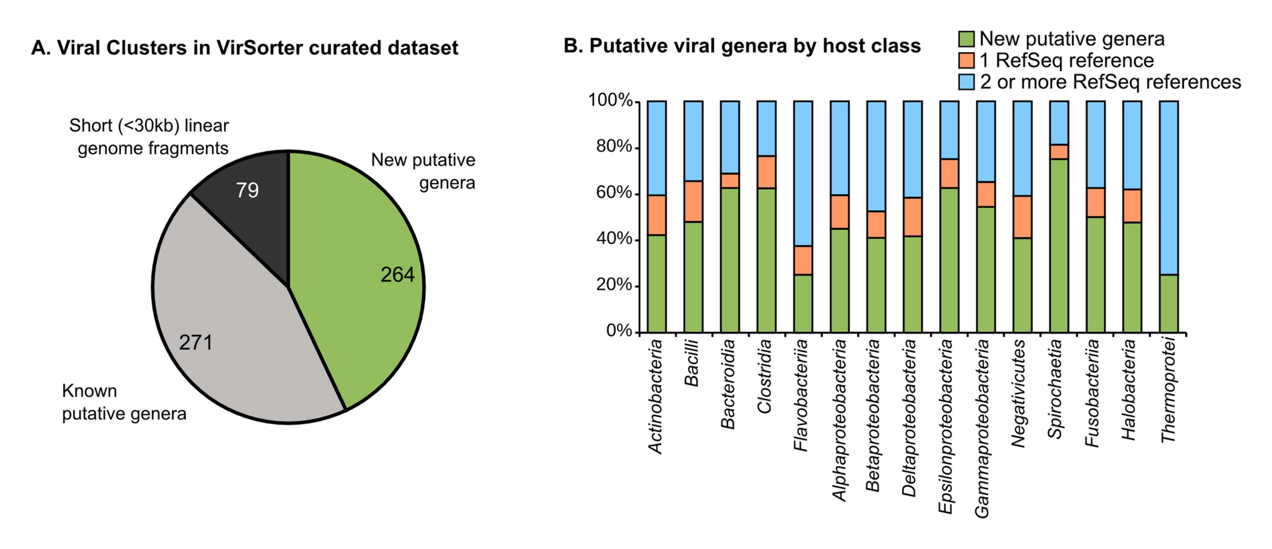


Figure 2. Degree of novelty of viruses detected in VirSorter curated data set. (**A**) Viral clusters (VCs) are considered as putative new genera when including at least one sequence larger than 30 kb, circular, or known to be a complete genome (from RefSeq). These putative genera were considered as 'new' when the VC did not include any RefSeq sequence, and 'known' otherwise. (**B**) The proportion of new VCs (containing no RefSeqABVir), VCs with only one RefSeqABVir sequence, and VCs with more than one RefSeqABVir sequence is displayed for host classes associated with more than 10 virl sequences. Only 'putative genera' VCs were considered (i.e., clusters containing a RefSeqABVir genome, a circular sequence, or a sequence with more than 30 predicted genes).

DOI: 10.7554/eLife.08490.009

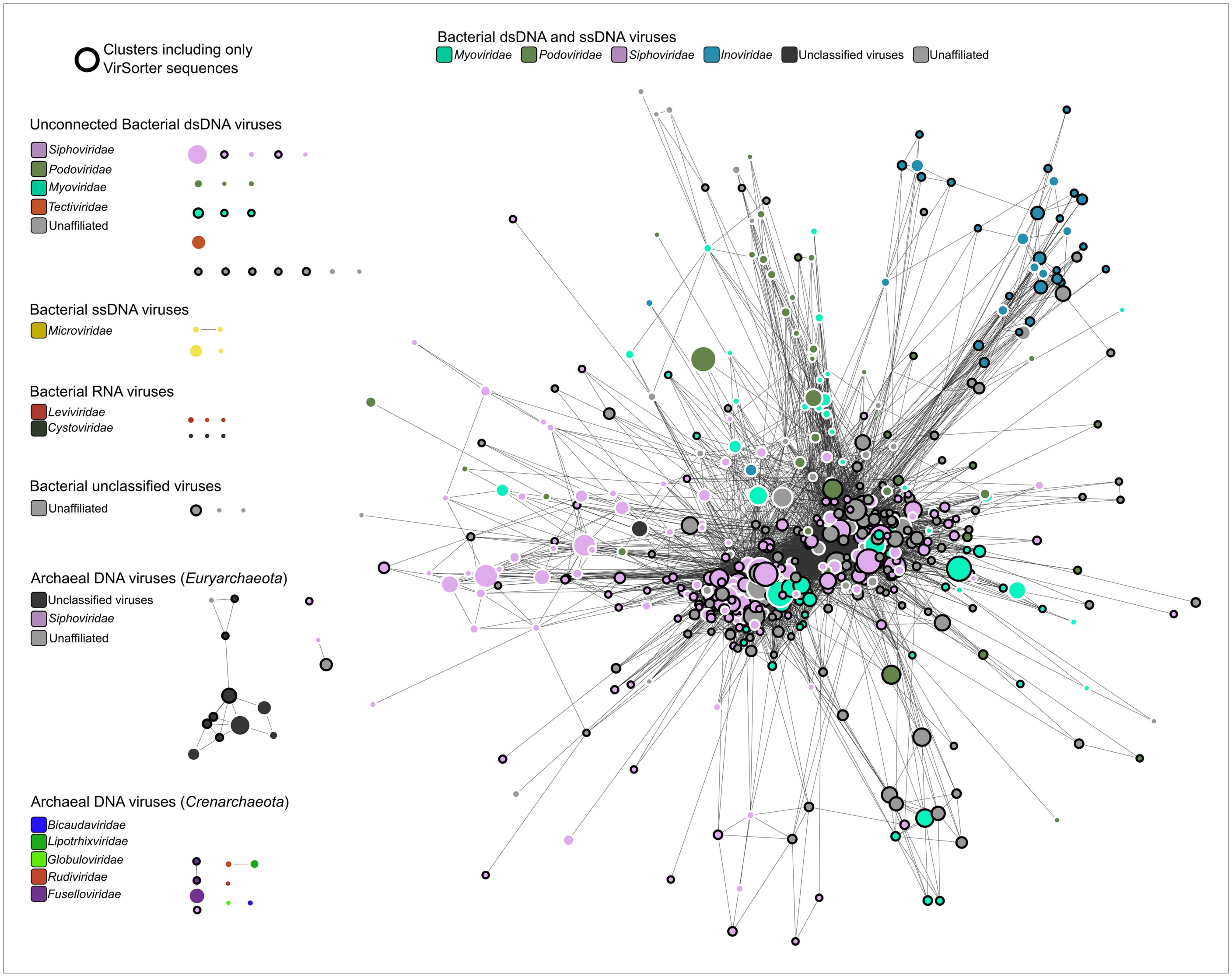
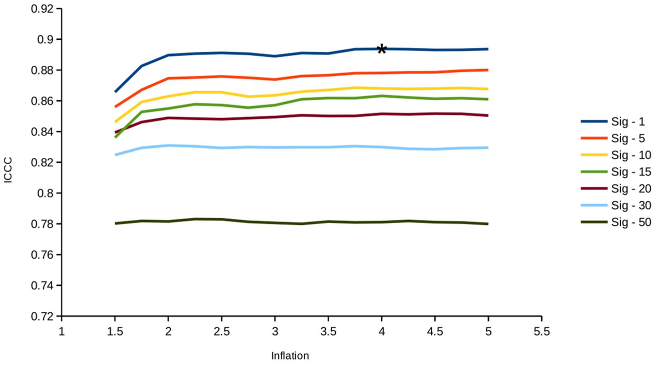


Figure 2—figure supplement 1. Structure of viral sequence space sampled in VirSorter data set. Network of virus clusters (VCs) based on gene content comparison between viral genome sequences from RefSeqABVir and VirSorter data set. VCs including only VirSorter sequences are highlighted with a black outline. The size of nodes is proportional to the number of sequences in the cluster and the color of the node corresponds to the BLAST-based affiliation (at the family level) of its members when consistent (i.e., agreement between >75% of the cluster members, otherwise clusters are indicated as 'unaffiliated').

DOI: 10.7554/eLife.08490.011

ICCC

Inflation	Sig - 1	Sig - 5	Sig - 10	Sig - 15	Sig - 20	Sig - 30	Sig - 50
1.5	0.8656	0.8560	0.8461	0.8361	0.8394	0.8248	0.7802
1.75	0.8826	0.8671	0.8592	0.8528	0.8462	0.8294	0.7818
2	0.8896	0.8746	0.8630	0.8550	0.8489	0.8309	0.7815
2.25	0.8907	0.8751	0.8655	0.8578	0.8484	0.8304	0.7831
2.5	0.8911	0.8758	0.8655	0.8572	0.8480	0.8293	0.7829
2.75	0.8906	0.8749	0.8627	0.8555	0.8487	0.8299	0.7813
3	0.8889	0.8738	0.8635	0.8571	0.8494	0.8296	0.7806
3.25	0.8910	0.8761	0.8659	0.8610	0.8506	0.8298	0.7800
3.5	0.8907	0.8766	0.8670	0.8618	0.8501	0.8298	0.7814
3.75	0.8935	0.8779	0.8685	0.8617	0.8502	0.8305	0.7809
4	0.8937	0.8780	0.8680	0.8631	0.8515	0.8299	0.7810
4.25	0.8935	0.8784	0.8677	0.8621	0.8512	0.8288	0.7818
4.5	0.8930	0.8785	0.8679	0.8613	0.8516	0.8285	0.7811
4.75	0.8931	0.8795	0.8683	0.8617	0.8515	0.8292	0.7808
5	0.8935	0.8800	0.8676	0.8610	0.8504	0.8295	0.7799



Number of clusters

Inflation	Sig - 1	Sig - 5	Sig - 10	Sig - 15	Sig - 20	Sig - 30	Sig - 50
1.5	326	340	379	433	491	564	666
1.75	394	410	450	503	556	623	708
2	432	459	496	551	601	665	743
2.25	476	499	535	593	630	698	777
2.5	510	533	564	622	670	733	806
2.75	535	552	591	651	690	755	826
3	556	574	617	672	711	783	852
3.25	576	599	644	694	734	801	877
3.5	593	610	663	715	752	820	897
3.75	603	628	681	734	771	833	911
4	617	643	692	752	783	849	928
4.25	625	654	705	765	801	859	940
4.5	636	662	717	778	814	868	960
4.75	643	668	724	785	820	877	971
5	645	675	733	796	833	887	985

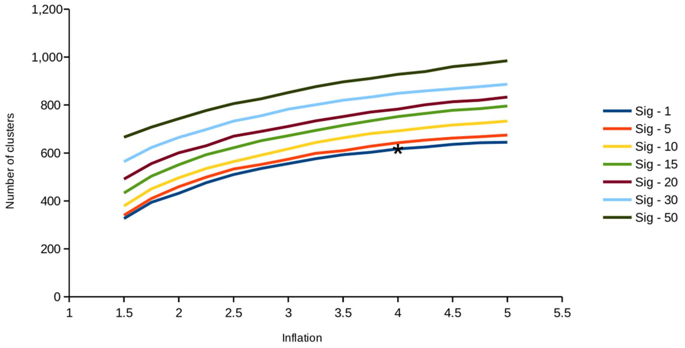


Figure 2—figure supplement 2. Benchmarks used to determine the best value for inflation and significance thresholds for virus clustering. For each pair of values (inflation and significance threshold), the genome network was computed and its overall shape evaluated with ICCC (intra-cluster clustering coefficient). The chosen values are highlighted in green in the table and with a star on the associated plot.

DOI: 10.7554/eLife.08490.012

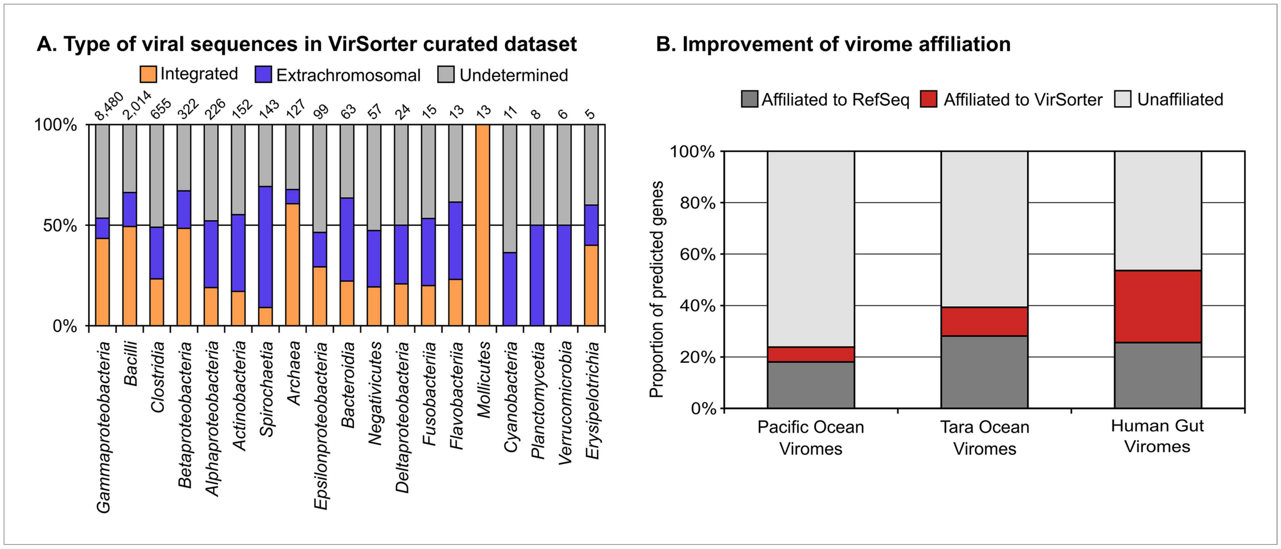


Figure 3. Extrachromosomal prophages in VirSorter curated data set and improvement in virome affiliation. (**A**) The distribution of VirSorter curated data set as 'integrated' (i.e., prophages integrated in the host chromosome), 'extrachromosomal' (i.e., >30 kb or circular sequences with no microbial genes), or 'undetermined' (<30 kb linear with no microbial genes) is indicated for each host class with at least five VirSorter curated data set sequences. The number of sequences associated with each host class in indicated above the histogram. (**B**) Improvement in the proportion of affiliated genes from viromes with VirSorter data set. Predicted genes from the Pacific Ocean Viromes (***Hurwitz and Sullivan, 2013***), Tara Ocean Viromes (***Brum et al., 2015***), and Human Gut Viromes (***Minot et al., 2012***) were compared to RefSeqVirus (May 2015) and the VirSorter data set (BLASTp, threshold of 50 on bit score and 0.001 on e-value). Predicted proteins affiliated to VirSorter (in blue) did not display any significant similarity to a RefSeq sequence.

DOI: 10.7554/eLife.08490.013

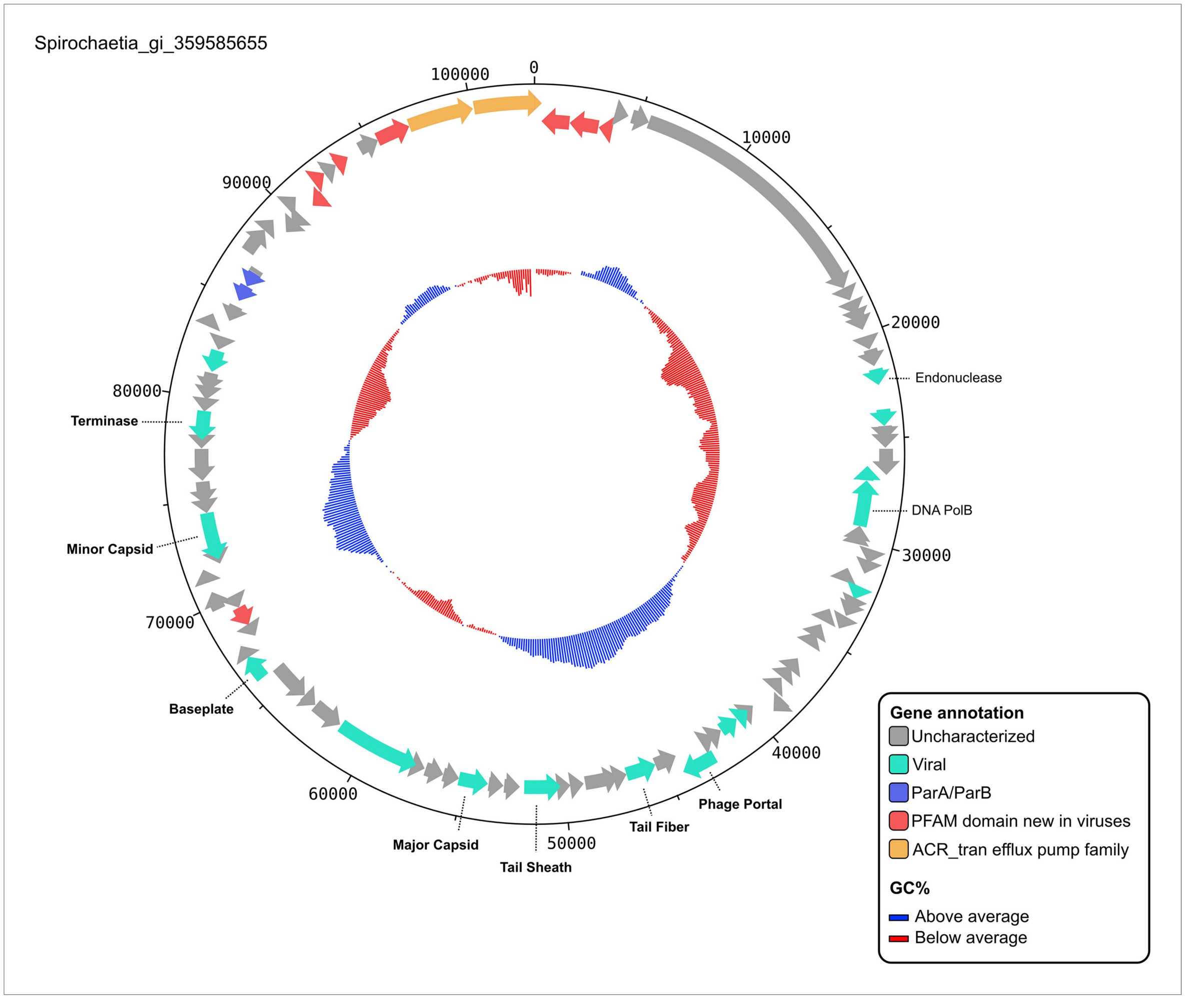


Figure 3—figure supplement 1. Contig map of a putative new extrachromosomal prophage. Contig Spirochaetia_gi_359585655 represent a complete genome (the contig was detected as circular) from a new genus (affiliated to a VC with no RefSeqABVir sequence). Functional affiliation of predicted genes is indicated on the map, with notably two genes (ParA/ParB) indicative of extrachromosomal prophages, as well as two genes (in orange) affiliated to the ACR_tran efflux pump family, of which some members are involved in antibiotic resistance phenotypes. This contig belongs to the virus cluster VC_61, composed of 35 new putative extrachromosomal prophages from different Spirochetes genomes.

DOI: 10.7554/eLife.08490.014

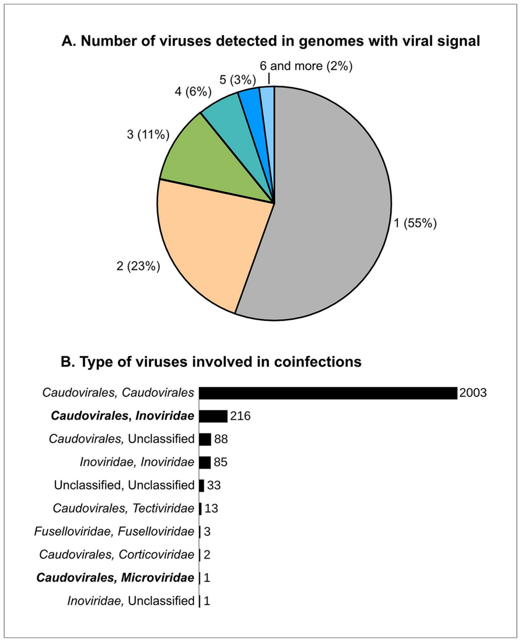


Figure 4. Scale and range of co-infection. (**A**) Number of different viral sequences detected by host genome. Numbers are based on the set of microbial genomes with at least one viral sequence detected (5492 genomes). (**B**) Affiliation of viruses involved in multiple infections of the same host. Affiliations are deduced from best BLAST hits alongside the viral sequences, as in ***Figure 1***. Co-infections involving dsDNA and ssDNA viruses are highlighted in bold.
DOI: 10.7554/eLife.08490.015

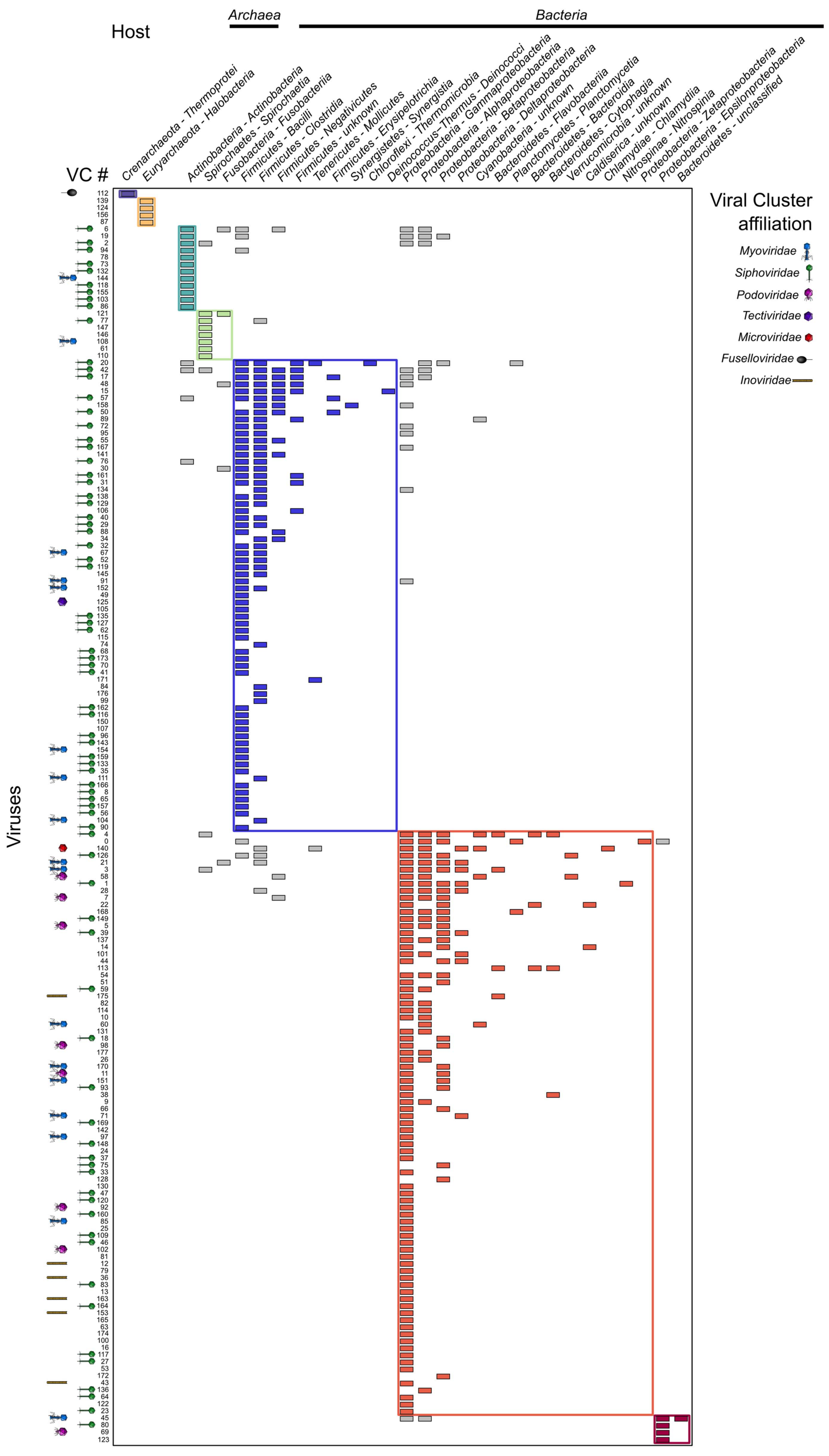


Figure 5. Virus–host network between virus clusters and host classes (matrix visualization). A cell in the matrix is colored when at least one virus from a virus cluster (VC, rows) was retrieved in a genome from a host class (columns). This virus–host network is detected as significantly modular by lp-Brim (modularity Q = 0.45; the same index computed from 99 randomly permuted matrices ranged from 0.02 to 0.17, with an average of 0.08). The different

Figure 5. continued on next page

Figure 5. Continued

modules are highlighted in color, with inter-module links in gray. Virus clusters are identified by their number and their family-level affiliation (based on BLAST-based affiliation of the cluster members) is indicated next to each cluster when available (virus clusters with inconsistent members affiliation are considered as 'unclassified', affiliations are spread along the x-axis for spacing purpose). Host phylum and class are indicated for each host column, with domains indicated above the corresponding hosts.

DOI: 10.7554/eLife.08490.016

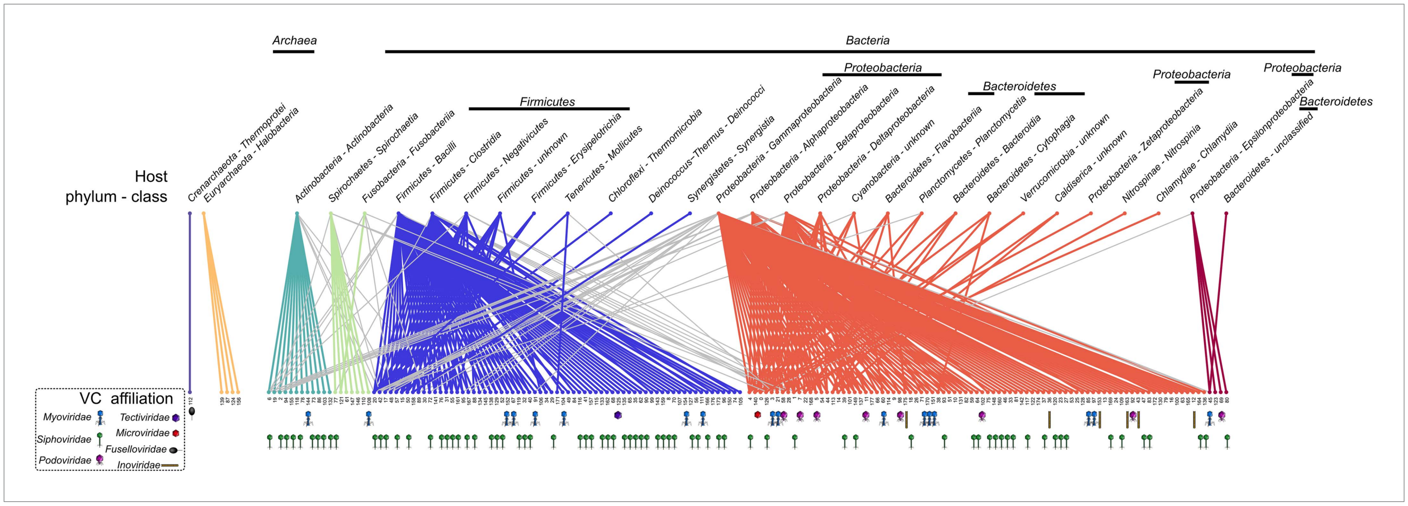

Archaea
Bacteria
Proteobacteria
Firmicutes
Bacteroidetes
Proteobacteria
Proteobacteria
Bacteroidetes
Host
phylum - class
Crenarchaeota - Thermoprotei
Euryarchaeota - Halobacteria
Actinobacteria - Actinobacteria
Spirochaetes - Spirochaetia
Fusobacteria - Fusobacteriia
Firmicutes - Bacilli
Firmicutes - Clostridia
Firmicutes - Negativicutes
Firmicutes - unknown
Firmicutes - Erysipelotrichia
Tenericutes - Mollicutes
Chloroflexi - Thermomicrobia
Deinococcus–Thermus - Deinococci
Synergistetes - Synergistia
Proteobacteria - Gammaproteobacteria
Proteobacteria - Alphaproteobacteria
Proteobacteria - Betaproteobacteria
Proteobacteria - Deltaproteobacteria
Cyanobacteria - unknown
Bacteroidetes - Flavobacteriia
Planctomycetes - Planctomycetia
Bacteroidetes - Bacteroidia
Bacteroidetes - Cytophagia
Verrucomicrobia - unknown
Caldiserica - unknown
Proteobacteria - Zetaproteobacteria
Nitrospinae - Nitrospinia
Chlamydiae - Chlamydiia
Proteobacteria - Epsilonproteobacteria
Bacteroidetes - unclassified
VC affiliation
Myoviridae
Tectiviridae
Siphoviridae
Microviridae
Fuselloviridae
Podoviridae
Inoviridae

Figure 5—figure supplement 1. Virus–host network between virus clusters and host classes (network visualization). An edge is displayed between a virus cluster (VC) and a host class when at least one virus from this cluster was retrieved in a genome from the host class. This network is detected as significantly modular by lp-Brim (modularity Q = 0.45; the same index computed from 99 randomly permuted matrices ranged from 0.02 to 0.17, with an average of 0.08). The different modules are highlighted in color, with inter-module links in gray. VCs are identified by their number and their family-level affiliation (based on BLAST-based affiliation of the cluster members) is indicated below each cluster when available (VCs with inconsistent members affiliation are considered as 'unclassified'). Host phylum and class are indicated for each host node, with phyla (when multiple class from the same phylum are included in the network) and domains indicated above the corresponding host nodes.

DOI: 10.7554/eLife.08490.017

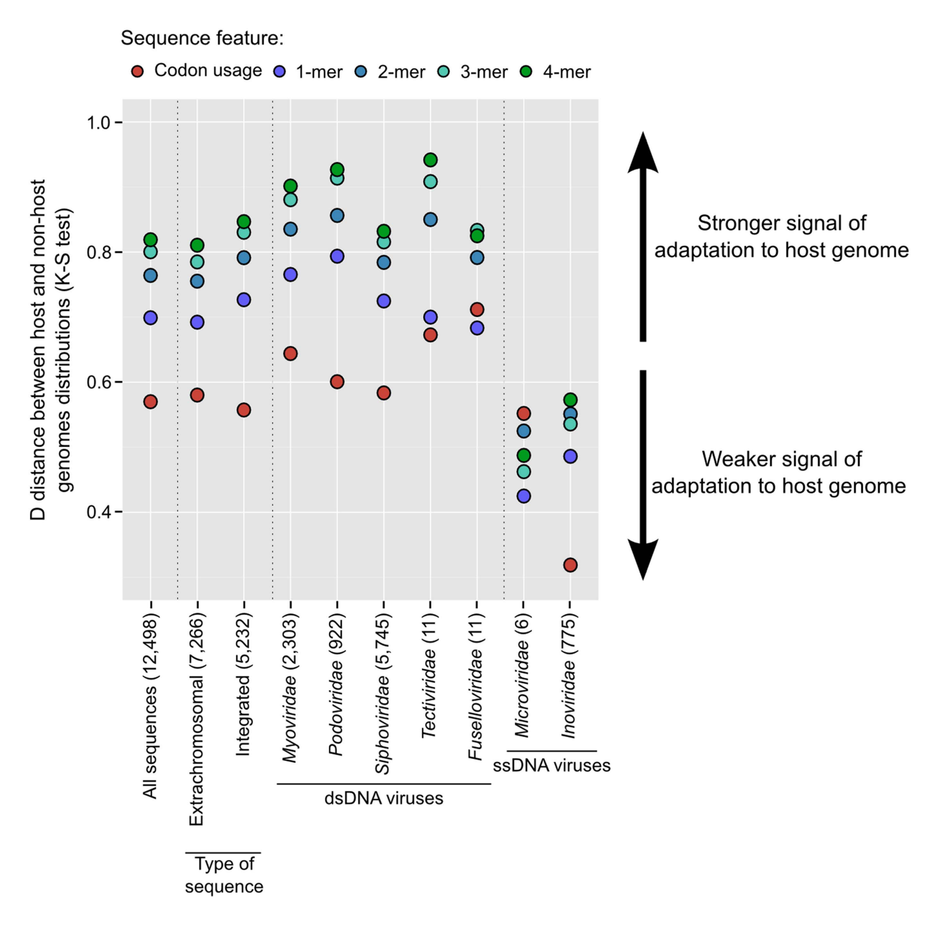


Figure 6. Adaptation of viral genome composition and codon usage to the host genome. K–S distances between distributions of virus–host distances and virus–non-host distances for each metrics (in color) and different subsets of the viral sequences (all sequences, by type, and by taxonomy). Only families with more than 5 genomes are displayed (although it should be noted that the VirSorter data set includes only 6 *Microviridae* sequences).
The number of sequences in each category is indicated in brackets. Distributions used to compute distances are displayed in ***Figure 6—figure supplement 1***.

DOI: 10.7554/eLife.08490.018

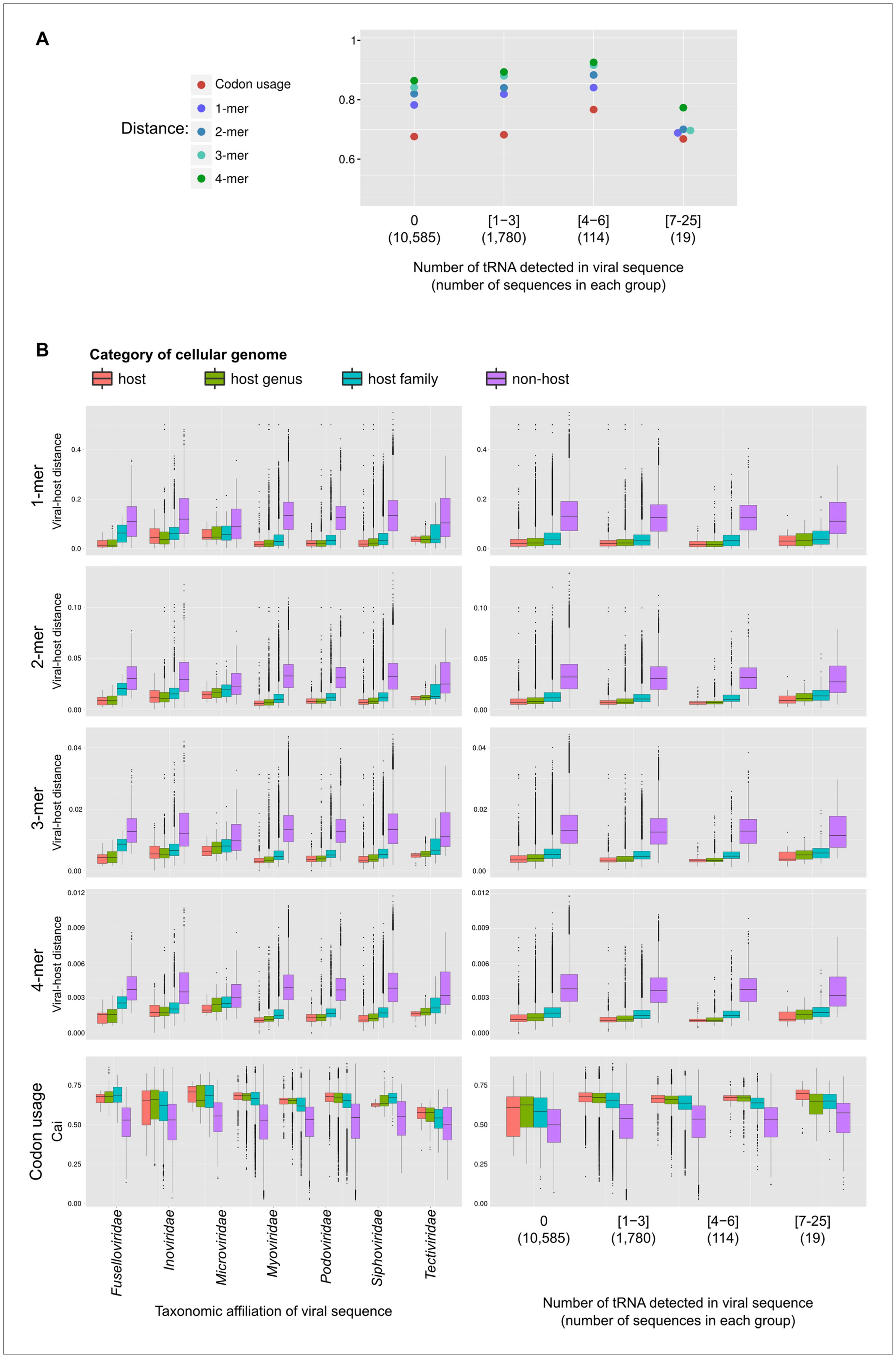


Figure 6—figure supplement 1. (**A**) K–S distances between distributions of virus–host distances and virus–non-host distances for each metrics (in color) and different subsets of the viral sequences (based on the number of tRNA

Figure 6—figure supplement 1. continued on next page

Figure 6—figure supplement 1. Continued

genes detected). The number of sequences in each category is indicated below the number of tRNA. (**B**) Distribution of k-mer distances between viral and cellular genomes and codon usage adaptation index for host, host genus, host family, and non-host (different order) genomes. For each viral genome, the distance to the host is displayed, as well as 10 randomly taken distances to genomes from each category and different subsets of the viral sequences (by taxonomy on the left column, and by number of tRNA genes on the rigth column).

DOI: 10.7554/eLife.08490.019

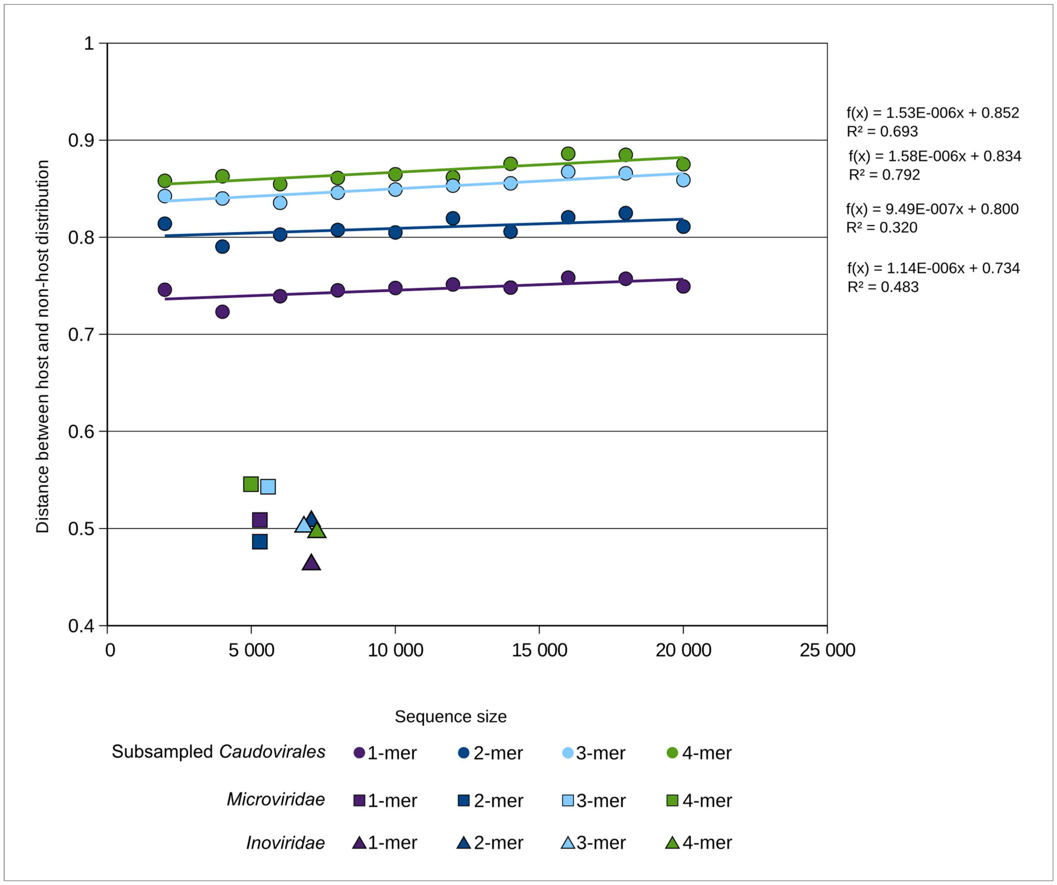


Figure 6—figure supplement 2. Distance between k-mer frequency vectors of virus genome subsamples and host genomes for *Caudovirales*. Viral genomes (1000) were randomly sub-sampled at different sizes (from 2000 to 20,000 bp). Only *Caudovirales* genomes were selected for this subsample analysis. For each size of k-mer, the result of a linear regression of distance between host or non-host and viral subsample size is indicated. The same distances for the *Microviridae* and *Inoviridae* (taken from ***Figure 6A***) are indicated for comparison, and associated with the size of the reference genome of each group (*Enterobacteria* phage phiX174 and *Enterobacteria* phage M13). For clarity's sake, the almost-identical values for 2-mer, 3-mer, and 4-mer for *Microviridae* are slightly horizontally shifted.

DOI: 10.7554/eLife.08490.020