High-throughput identification of viral termini and packaging mechanisms in virome datasets using PhageTermVirome

Julian R. Garneau, Véronique Legrand, Martial Marbouty, Maximilian O. Press, Dean R. Vik, Louis-Charles Fortier, Matthew B. Sullivan, David Bikard, Marc Monot

scientific reports

Check for updates

OPEN

High-throughput identification of viral termini and packaging mechanisms in virome datasets using PhageTermVirome

Julian R. Garneau[1,8]✉, Véronique Legrand[2,8], Martial Marbouty[3], Maximilian O. Press[4], Dean R. Vik[5], Louis-Charles Fortier[6], Matthew B. Sullivan[5], David Bikard[7] & Marc Monot[1]✉

Viruses that infect bacteria (phages) are increasingly recognized for their importance in diverse ecosystems but identifying and annotating them in large-scale sequence datasets is still challenging. Although efficient scalable virus identification tools are emerging, defining the exact ends (termini) of phage genomes is still particularly difficult. The proper identification of termini is crucial, as it helps in characterizing the packaging mechanism of bacteriophages and provides information on various aspects of phage biology. Here, we introduce PhageTermVirome (PTV) as a tool for the easy and rapid high-throughput determination of phage termini and packaging mechanisms using modern large-scale metagenomics datasets. We successfully tested the PTV algorithm on a mock virome dataset and then used it on two real virome datasets to achieve the rapid identification of more than 100 phage termini and packaging mechanisms, with just a few hours of computing time. Because PTV allows the identification of free fully formed viral particles (by recognition of termini present only in encapsidated DNA), it can also complement other virus identification softwares to predict the true viral origin of contigs in viral metagenomics datasets. PTV is a novel and unique tool for high-throughput characterization of phage genomes, including phage termini identification and characterization of genome packaging mechanisms. This software should help researchers better visualize, map and study the virosphere. PTV is freely available for downloading and installation at https://gitlab.pasteur.fr/vlegrand/ptv.

Viruses play key roles in diverse microbial ecosystems. They are recognized to influence biogeochemical cycling, modulate microbial populations and metabolism, and act as a driving force in gene flow in soil and the oceans[1–6]. In humans, viruses can also provide a first line of non-host derived immunity[7] and are associated with health status[8–12]. Such discoveries have led to an increasing interest in characterizing viruses and their impact on ecosystems. However, although metagenomic surveys have revealed hundreds of thousands of viral genomes and large genome fragments[13–18], these new genomes likely only scratch the surface of the viruses existing in these environments[15,18].

A considerable proportion of the reads obtained when sequencing the human gut virome typically has no significant homology to known viral genomes[19]. For example, Manrique et al. reported that only a small subset of active phages detected in the gut microbiome could be taxonomically classified and that more than half (≈ 60%) could represent entirely unknown novel bacteriophages. Similarly, datasets from the deeply sequenced Global Ocean Viromes show that only a minimal fraction of reads (≈ 10–20%) can be mapped to large de novo assembled viral reference genome databases, such as GOV1 and GOV2 databases[13,14]. The term "viral dark matter" has been used to describe this vast unknown sequence space[19–21]. Recent efforts to establish viral clusters in gene-sharing networks have revealed structures that will greatly help in virus identification and classification, but a large

[1]Biomics Platform, C2RT, Institut Pasteur, 75015 Paris, France. [2]Infrastructure et Ingénierie Scientifique, Institut Pasteur, 75015 Paris, France. [3]Institut Pasteur, Unité Régulation Spatiale des Génomes, UMR 3525, CNRS, 75015 Paris, France. [4]Phase Genomics Inc, Seattle, WA 98109, USA. [5]Department of Microbiology, Ohio State University, Columbus, OH 43210, USA. [6]Faculty of Medicine and Health Sciences, Department of Microbiology and Infectious Diseases, Université de Sherbrooke, Sherbrooke, QC J1E 4K8, Canada. [7]Département de Microbiologie, Institut Pasteur, Groupe Biologie de Synthèse, 75015 Paris, France. [8]These authors contributed equally: Julian R. Garneau and Véronique Legrand. ✉email: julian.garneau@pasteur.fr; marc.monot@pasteur.fr

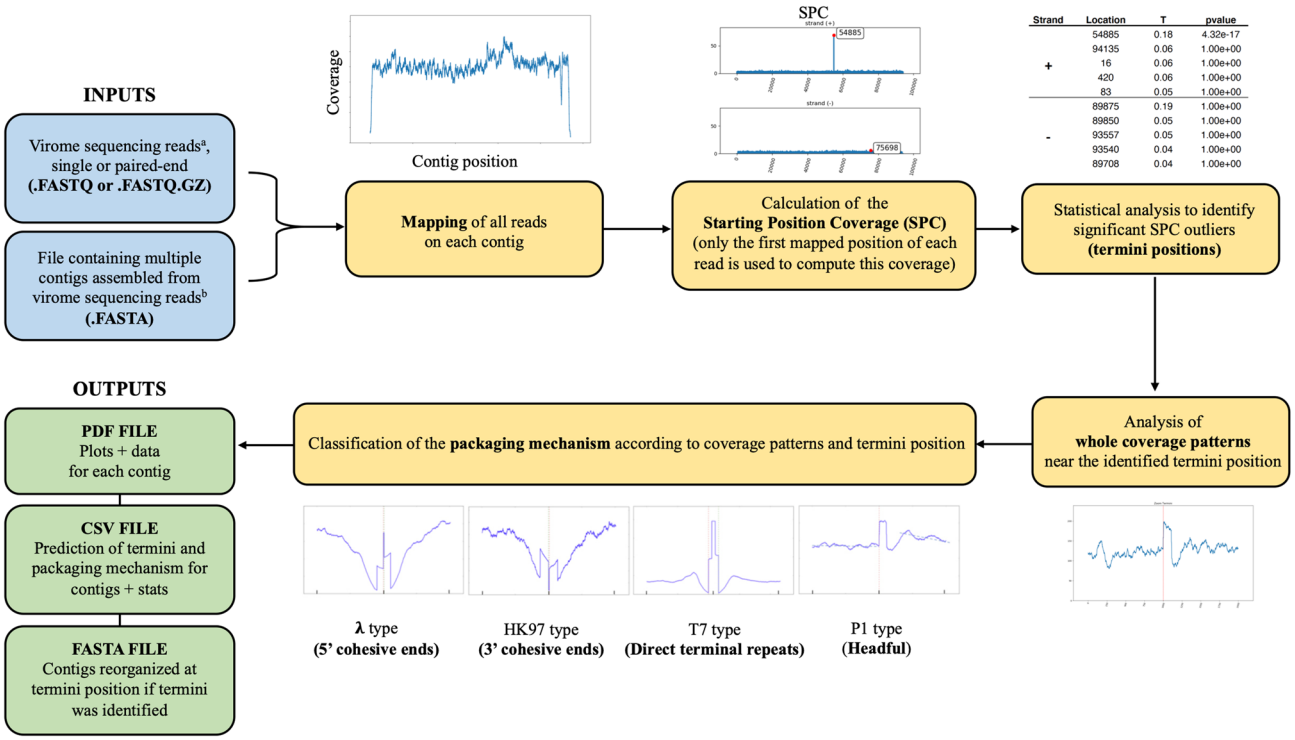


Figure 1. Overview of the different steps involved in the PhageTermVirome workflow. Virome sequencing reads are simultaneously mapped to numerous contigs assembled from the virome dataset. After alignments for each contig, PTV performs peak detection and statistical analyses to determine the position of significant termini. On all contigs for which termini were detected, a coverage pattern analysis is achieved to predict the corresponding genome packaging mechanism. [a]The sequencing method and library preparation require the preservation of genome ends (ex: all methods involving random DNA fragmentation prior to library preparation and long-read sequencing of intact DNA). [b]PhageTermVirome can process massive amount of contigs as input (multifasta), but still allow users to analyze a single phage contig (fasta).

number of unknown clusters likely remain to be discovered[22,23]. Despite the continuous rise in the number of sequences compiled in these databases, researchers will undoubtedly keep finding unknown sequences in abundance for the foreseeable future. It is thus still important to ascertain the viral origin of sequences discovered in metagenome analysis with higher confidence. Phage and virus biologists have therefore focused their efforts on developing orthogonal complementary prediction approaches, such as the analysis of k-mer frequencies and viral genome features, also recently implemented through machine-learning algorithms[24,25].

It has been previously shown that the ends of linear DNA, a hallmark of genomic DNA (gDNA) packaged in the capsid of free phage particles, can be easily detected from shotgun sequencing data if the sequencing library was prepared in a way that preserves such DNA ends (i.e., random fragmentation of DNA prior to library preparation)[26,27]. Shotgun sequencing approaches typically rely on the ligation of adapters to randomly fragmented DNA. During this process, DNA fragments with one end that corresponds to natural genome termini will be strongly enriched relative to other DNA fragments that end at random positions due to the random fragmentation of DNA during library preparation. This enables the recognition of termini by analyzing the number of reads that start at each position along a given phage sequence, normalized against the whole sequence coverage. The number of read starts mapping precisely at the natural termini will be over-represented, because it directly correlates with the number of genomes (i.e., viral particles) present in a sample. We previously used this concept and approach to identify the genome termini and characterize packaging mechanisms of individual pure phages[27].

Here, we introduce PhageTermVirome (PTV), an easy-to-use bioinformatics software that efficiently identifies critical information such as phage genome termini and packaging mechanisms in modern large-scale viral metagenomics datasets. Due to its ability to detect termini present in capsid-packaged linear DNA, PTV also offers an entirely orthogonal and complementary approach to predict the true viral origin of metagenomic contigs, along with other existing phage prediction tools[28–36].

Results and discussion

Overview of the PhageTermVirome workflow. The key steps of the PTV analysis workflow are depicted in Fig. 1. They consist of: (i) mapping all viral metagenomic reads on contigs assembled from the same read dataset, (ii) calculation of the starting position coverage (SPC) and identification of positions on the contig for which the SPC is significantly higher, which may represent genome ends (termini) and secondary terminase cutting sites, (iii) if applicable, characterization of the corresponding packaging mechanism by analysis of the coverage patterns near the identified termini, and (iv) aggregation of the prediction results and coverage plots for

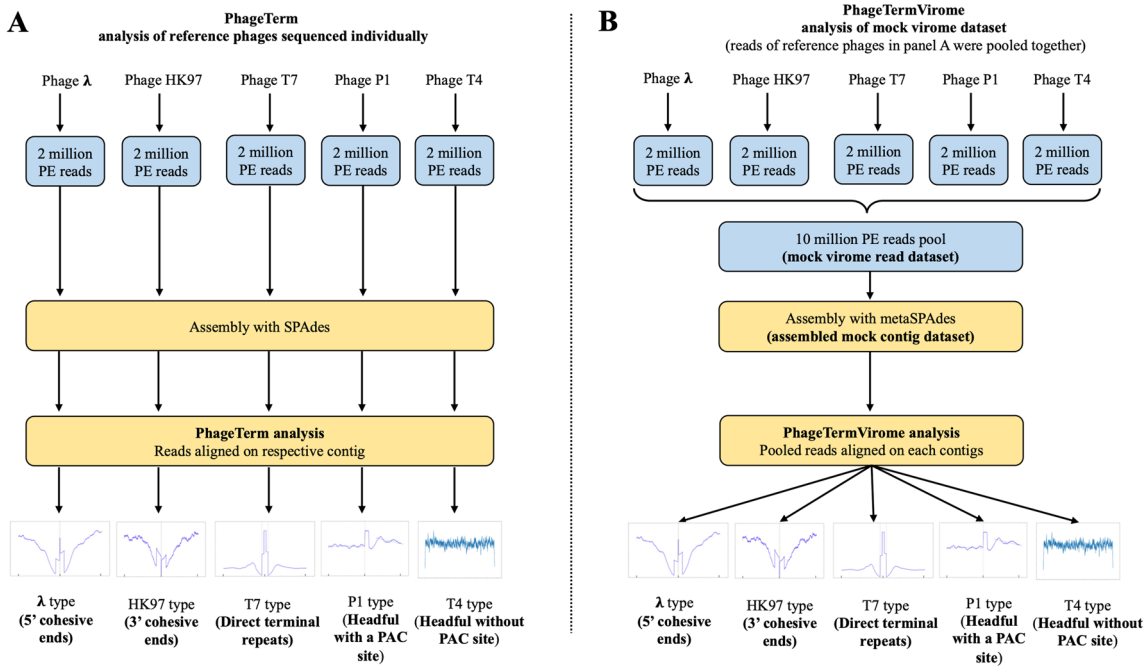


Figure 2. Comparison of results obtained with PhageTerm versus PhageTermVirome when using a mock virome dataset. (**A**) Individually sequenced phages, separately analyzed using PhageTerm previously confirmed the well-established packaging mechanism for each reference phage. (**B**) Reads of reference phages used in panel (**A**) were pooled to form the mock virome read dataset and were then assembled with metaSPAdes to generate the associated mock viral contig dataset. The pool of 10 million reads were mapped to the assembled contigs to obtain coverage values, statistics, termini positions, and packaging mechanism predictions for each contig.

all contigs submitted to PTV into a single PDF file for visual inspection. A CSV file containing the predictions and statistics for each contig is also provided to users for easier consultation and manipulation of the results in a table format, which can be useful when there is a very large number of contigs. For contigs with successful predictions, PTV also outputs a new fasta file containing the contigs reorganized to start at the predicted termini.

The PTV algorithm was developed to allow the analysis of multiple contigs in a single workflow (i.e., users do not have to re-run the script for each contig they want to analyze). The mandatory inputs into PTV required for analysis are (i) a multifasta (.fasta) file containing the assembled contigs and (ii) a fastq file (.fastq) containing the reads used to assemble the contigs found in the fasta file. When available, the user can also provide the corresponding paired-end fastq read file. Although not required, the use of paired-end reads is recommended, as it has been shown to improve termini and packaging mechanism prediction[27].

The PTV algorithm is largely based on the mapping of all reads of a dataset to each contig provided, which can represent a daunting task for ultra-deeply sequenced viromes containing a vast amount of reads and a large number of contigs. Additional options and parameters have thus been added to the PTV scripts to accelerate and scale the process. For example, in addition to the multi-threading option initially implemented in the original PhageTerm software[27] (using the command-line option --core), a new multi-machine functionality has also been deployed in PTV (using the command-line option --mm). The multi-machine option can be used to go over the limit of the number of cores available on a machine when processing large datasets and thus accelerate the analysis. It is intended for advanced users who have the possibility to perform analyses on multi-machine computer clusters. As PTV can still rapidly process average-sized virome datasets using only multi-threading options, the multi-machine option becomes more attractive for extra-large datasets (e.g., $\gtrsim$100 million reads with $\gtrsim$1000 contigs). Many factors can influence the required time to complete a PTV analysis, such as the total number of reads, the total number and length of contigs to analyze, the speed of the processors, and the number of cores available for parallelization.

Detection of phage termini and packaging mode in a control mock virome dataset. We first assessed the ability of PTV to correctly identify termini and packaging modes in metagenomics datasets using a mock viral metagenomics dataset. This dataset was built by pooling 1×10^7 paired-end reads from five reference phages, using 2×10^6 paired-end reads from each phage previously sequenced individually[27]. HK97, T4, T7, P1, and Lambda phage paired-end reads were used, as these phages harbor a diversity of well-described termini and packaging mechanisms. After running the PTV workflow on the mock read dataset and the reference phage genomes, the expected termini positions were identified and corresponding packaging mechanisms successfully defined by PTV for all five reference phages (Fig. 2): Lambda (5′cos), HK97 (3′cos), P1 (headful with pac site), and T7 (direct terminal repeat, DTR). As expected, PTV returned no signal for T4 (headful without precise pac site). It is known that T4 type phages generate no precise or detectable termini, as their DNA is randomly packaged inside the capsid. This type of phage packages full genomes with redundant ends, starting and ending at

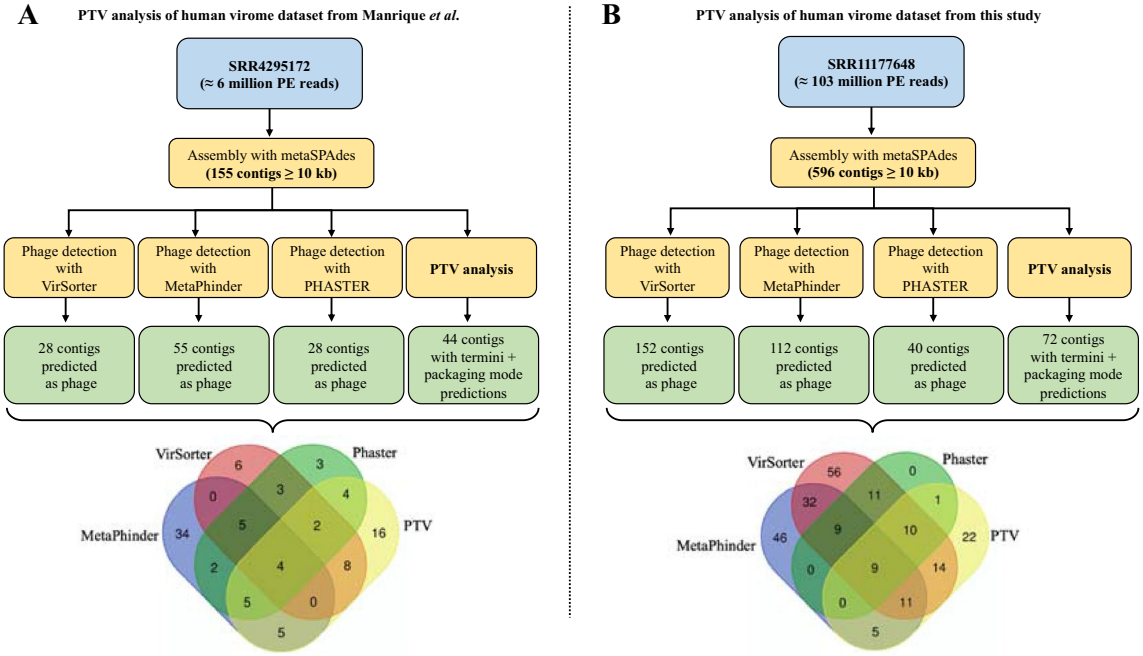


Figure 3. PhageTermVirome analysis with two human virome datasets. (**A**) PTV analysis of 155 contigs assembled from the study of Manrique et al. (**B**) PTV analysis of 596 contigs assembled from the virome dataset generated for this study.

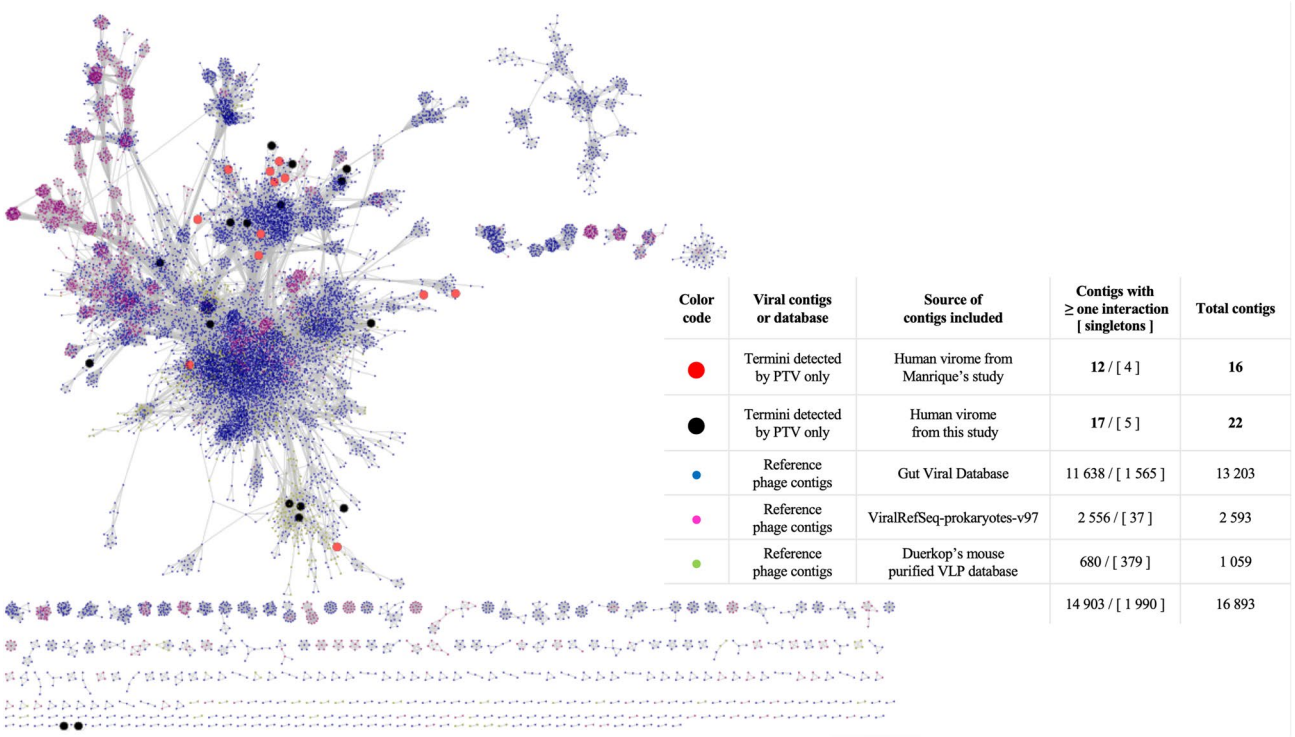
random positions, generating no observable over-represented sequencing biases after sequencing. Our results on the mock virome dataset suggest that PTV algorithms, coverage analyses, and statistical models are sufficiently robust (assuming sufficient coverage) to accurately detect termini and classify packaging mechanisms when the reads of different phages are merged, as is the case for real virome metagenomic datasets.

Characterization of phage termini and packaging mode in two real virome metagenomics datasets. Given the successful characterization of phage termini and packaging mechanisms in the mock virome dataset, we next applied PTV to two real virome datasets. The first dataset selected to test PTV was a human gut virome previously sequenced and published in the study by Manrique et al.[37]. The second dataset was generated during the course of this study and also corresponds to a human gut virome (Fig. 3). In addition to analyzing the two virome datasets with PTV, we also performed a phage prediction analysis with three widely used software programs, MetaPhinder[30], VirSorter[29] and Phaster[31]. Concomitant analysis of the assembled contigs was performed with these software programs and PTV to verify whether our tool performs better on contigs that are typically predicted to be phages. Thus, we produced Venn diagrams showing contigs individually and commonly predicted to be phage by each tool and compared the results with the successful termini PTV predictions.

For the Manrique dataset, PTV predicted termini and the packaging mode for 44 contigs from a total of 155 assembled contigs. Among the 44 contigs, 28 were predicted to be phage by the other softwares, meaning that PTV could determine termini and the packaging mode on contigs (N = 16) that were not declared to be phage by any other prediction tools. Detailed analysis of these 16 contigs showed a diversity of predicted packaging modes (4 *cos* phages, 10 *pac* phages, and 2 DTR phages).

In the second human virome dataset, generated for this study, PTV was able to predict termini and the packaging mode for 72 contigs (from a total of 596), among which 50 were also predicted to be phage by the other phage prediction tools. Thus, PTV obtained predictions for a number of contigs (N = 22) in this dataset that were not identified as phage by any of the other tools. The predicted packaging modes for these 22 contigs were also diverse, with 8 *cos* phages, 8 *pac* phages, and 6 DTR phages. One finding that arises from this comparative analysis is that each phage prediction software performs unevenly on different datasets and that it is useful, even necessary, to use a combination of different tools to obtain a more complete and precise vision of the landscape of the phages in a virome dataset.

In light of these results, we wished to understand why certain contigs with successful PTV predictions could not be established as phage by any of the three prediction tools used in this study. Thus, the 38 contigs (16 + 22) were annotated with PROKKA using the latest PHASTER phage protein database (as of December 2020). For each contig, if an ORF was not annotated with a known protein in the database, a second annotation was performed using PROKKA's default bacterial protein database. Proteins of phage origin could be found on most of the contigs but, in many cases, represented a minority of the total predicted ORFs. The high number of hypothetical proteins found on these contigs likely explains, at least partially, why they were not identified as phage by the other prediction tools used here. Another hypothesis is that these contigs may be fragments of a partially



Color code	Viral contigs or database	Source of contigs included	Contigs with ≥ one interaction [singletons]	Total contigs
●	Termini detected by PTV only	Human virome from Manrique's study	**12** / [4]	**16**
●	Termini detected by PTV only	Human virome from this study	**17** / [5]	**22**
●	Reference phage contigs	Gut Viral Database	11 638 / [1 565]	13 203
●	Reference phage contigs	ViralRefSeq-prokaryotes-v97	2 556 / [37]	2 593
●	Reference phage contigs	Duerkop's mouse purified VLP database	680 / [379]	1 059
			14 903 / [1 990]	16 893

Figure 4. Viral cluster analysis with vConTACT2 using a gene-sharing network. The large red dots represent phage contigs (Manrique dataset) for which PTV predicted a terminus and packaging mode but not predicted to be viral by MetaPhinder, Phaster, or VirSorter. The large black dots represent phage contigs (dataset from this study) for which PTV predicted a terminus and packaging mode but not predicted to be viral by MetaPhinder, Phaster, or VirSorter. Network analysis was performed using three reference contig collections: GVD[8], ViralRefSeq V.97[41], and a database of curated virus-like particles from mice[42].

assembled phage genome, as is often the case in virome metagenomic datasets[38]. In such fragmented assemblies, hallmark phage genes required by certain tools to declare the contig as phage may be lacking. However, given sufficient read coverage, PTV is still able to determine termini and packaging modes on a contig, even if a large part of the corresponding phage genome is missing from the assembly. This feature also allows PTV to discriminate between sequences of active phages and those of purely decaying or inactive prophages (which can be present in contaminating bacterial genomes, but which ultimately do not form viral particles packaging linear DNA with termini). It is difficult to confirm the completeness of a phage contig obtained after the sequencing of a virome sample, especially when using short-read sequencing. To date, the most reliable way to ensure full genome recovery is still probably the sequencing of pure cultured phage lysates. Sequencing of virome DNA using long-read technologies also shows great promise in retrieving more complete phage genomes[39].

Since only a few phage proteins could be identified after annotation of our contigs for which we obtained a termini and packaging mode prediction, we sought to perform a more in-depth analysis using vConTACT2[23] in order to obtain additional clues about the taxonomy of those contigs (Fig. 4). In this analysis, 12 of 16 contigs from the Manrique dataset clustered with at least one contig from the reference viral databases included here. It also showed that 17 of 22 contigs from the dataset generated for the study also clustered with contigs from the reference viral databases. Further in-depth analyses of the contigs found to be singletons will help determine the nature of these sequences and whether they represent unknown and distant phages or other entities capable of packaging DNA, such as phage-inducible chromosomal islands (PICIs), genomic islands, or other classes of transducing agents[40].

Limitations and perspectives for PhageTermVirome. The capacity of PTV to detect termini and identify packaging mechanisms depends on the protocol used to prepare the nucleic-acid libraries prior to sequencing. Indeed, the detection of DNA termini can only occur if they are preserved during library preparation. Methods such as tagmentation, used for example in the Nextera kits from Illumina, insert adapters within DNA fragments in a non-random fashion (creating non-natural sequencing biases for read start positions) and are therefore incompatible with the natural bias-detection procedure of PTV. As previously mentioned, to enable data analysis using PTV, the sequencing library must be prepared through the ligation of adapters to DNA that has been randomly fragmented (e.g., Covaris, sonication), or any DNA preparation in which the natural DNA ends have been preserved. (e.g., Illumina TruSeq PCR free and TruSeq Nano, NebNext, Accel-NGS 1S PluS DNA library Kit). Note that the approach is constrained by the sequencing library preparation methods but not the sequencing technology itself. Although only Illumina sequencing was assessed here, other sequencing technologies or approaches, such as SMRT PacBio sequencing, Nanopore sequencing, and the recently developed

VirION2 approach[43], are theoretically also well suited to work with PTV. These sequencing technologies exhibit higher error rates than short-read sequencing approaches, but this should not strongly affect the ability of PTV to detect termini and determine packaging mechanisms. The first step of the PTV pipeline is based on the perfect alignment of only short seed sequences, of which the length can be adjusted if needed (default is 20-mers at the start of each read).

It is important to keep in mind that the algorithm of PTV is based on read alignments and the detection of over-represented sequences. Thus, the performance of PTV may be affected by very low coverage depth. Users must keep in mind that PTV will make more confident calls for contigs with a reasonable mean coverage. PTV can work at low mean coverage (e.g., 10×) in some cases, but often requires higher mean coverage for the unequivocal detection of termini and packaging mechanism (e.g., around 30×). Coverage depth can be variable from one contig to another in virome datasets and can be very low for viral particles that are scarce in the sample. PTV can often identify the position of potential termini for contigs with very low coverage, but may have difficulty in properly determining the packaging mechanism. PTV issues a warning for contigs exhibiting very low coverage and invites users to manually inspect the results and consult the coverage-pattern plots to confirm ambiguous results.

For extra-large virome datasets (a massive number of reads and a very large quantity of contigs), PTV analysis may require long processing times, especially if the analysis is being performed with very few computing cores. In this first version of PTV, a multi-threading option is available, as well as a newly implemented multi-machine option. When parallelization is not possible, alternatives can be used to reduce preventable computing time, such as excluding undesired or irrelevant contigs (e.g., very short contigs ≤ 1 kb or other contigs representing non-viral contamination). The minimal contig limit size can be customized in PTV using the "--limit" option (default is 500 bp). However, as previously mentioned, we expect PTV to be useful for termini detection and characterization for entities known to package small non-random DNA fragments (e.g., *Dinoroseobacter shibae* gene transfer agents can package 4.2 kb dsDNA[44]. Other examples include particles packaging very small linear DNA or RNA fragments, such as satellite DNA or RNA viruses, as well as linear satellite RNA (satRNA). Group 1 large single-stranded satRNAs can package linear RNA of 0.7–1.5 kb and group 2 single-stranded satRNAs typically package linear RNA fragments < 700 bp (21,994,595). Thus, we recommend carefully choosing the metagenomics contigs excluded from the analysis. Another recommendation, also left to the discretion of the user, is to preselect contigs with a minimal coverage threshold (e.g., reject contigs ≤ 10× mean coverage), which will reduce the processing time and also help to generate more robust results.

The PhageTerm approach is naturally restricted to the detection of viruses that package linear nucleic acids. In viral metagenomic samples, dsDNA phages of the caudovirales order represent most sequences known to date, but a large number of diverse phages are yet to be discovered. Recent studies have shed light on the previously underestimated prevalence of single-stranded DNA (ssDNA) phages, including *Innoviridae* and *Microviridae* phages[45–47], but the ssDNA viruses discovered thus far appear to package circular DNA and therefore escape detection and characterization by PTV. As certain eukaryotic viruses can package linear ssDNA with inverted terminal repeats, such as Parvoviruses and Bidensoviruses[48,49], we have to consider the possibility that linear ssDNA phages may also be found in nature and that they may be detected and characterizable by PTV. Our strategy is also not restricted to bacteriophages. Thus, viruses that package linear nucleic acids that infect any kingdom, including ssDNA viruses and linear RNA viruses, should also be characterizable by PTV. The detection of the termini of linear RNA viruses requires custom sample preparation protocols which were not explored here. Protocols akin to those used to identify transcription starts in mRNA sequencing could readily be used to identify the 5′ end of RNA viruses[50], whereas methods similar to those used to identify transcription terminators could be used to identify the 3′ end[51]. In addition, emerging methods involving the direct sequencing of RNA fragments in a sample (e.g., Nanopore direct RNA sequencingg)[52,53] are expected to be compatible with PTV and should also allow the efficient detection of RNA phage termini and the characterization of their packaging mechanism.

We hypothesized that PTV should be able detect DNA ends not only in phages and viruses, but also in other biological entities, such as PICIs[54,55] and other transducing agents with conserved DNA ends. For example, a recent study has described the packaging of non-random fragments of bacterial genomes inside phage capsids in more detail[40]. The authors of this study showed that the precise packaging of bacterial DNA is made possible by the presence of sites resembling phage pac sites at multiple locations throughout the host genome. Until recently, gene transfer agents (GTA), which are phage-like particles, were thought to package bacterial DNA in a random fashion only, but recent studies have shown that certain GTAs (i.e., those in the dinoflagellate-associated bacterium *Dinoroseobacter shibae*) can package non-random DNA fragments, presumably also using a headful packaging mechanism[56]. Like phages, these entities are also of great interest, as they are increasingly alleged to be involved in horizontal gene transfer, host adaptation, and bacterial virulence[57,58]. The ability to define the exact start and end of these sequences by identifying the termini should help in more precisely defining and characterizing these particles and their role in various ecosystems.

Going forward, the detection and thorough characterization of viral contigs in metagenomic datasets will benefit from the use of a combination of tools based on various approaches[59,60]. In summary, PTV is a new and distinct tool for the high-throughput characterization of phage genomes that should substantially accelerate the comprehensive study of the virosphere.

Materials and methods

Origin of samples, library preparation, sequencing, and contig assembly. Three sequence datasets were used in this study: (i) a simulated mock virome containing reads collected from five reference phages sequenced in our previous study (Lambda, HK97, T7, P1, Mu)[27], in which the mock virome was generated by randomly subsampling 2×10^6 paired-end reads from each phage dataset and grouping them in a single

forward and reverse read file (1×10^7 paired-end reads in total), (ii) a human gut virome sequencing dataset (SRR4295172) from a previously published study[37], and (iii) a human gut virome sequencing dataset generated for this study. For this study, viral particles were separated and purified as follows: 500 mg of fecal matter was resuspended in sodium citrate solution (50 mL) and centrifuged at 300×*g* for 10 min at 4 °C. The supernatant was recovered and centrifuged at 5000×*g* for 45 min at 4 °C to pellet the bacteria. The supernatant was again recovered and diluted (1:5) with cold SM buffer (200 mM NaCl, 10 mM $MgSO_4$, 50 mM Tris pH 7.5) and filtrated using 0.45 µm and 0.2 µm polyethersulfone membranes. For viral particle precipitation, PEG 6000 was added to a final concentration of 10% and the samples maintained at 4 °C overnight. The following morning, samples were centrifuged at 6000×*g* for 1 h at 4 °C and the pellets resuspended in 2 ml cold SM buffer and treated with an equal volume of chloroform. After centrifugation at 15,000×*g* for 5 min at 4 °C, the aqueous phase containing viral particles was recovered. SDS (0.5% final concentration) and proteinase K (20 mg/mL) were added and the samples incubated for 3 h at 65 °C. DNA was extracted using phenol chloroform, precipitated with ethanol, and resuspended in 10 mM Tris–HCl. DNA was sonicated using a Covaris S220 apparatus. The sequencing library was prepared using a TruSeq Illumina kit, 2 × 75 bp paired end, and sequenced on a NextSeq550 Illumina sequencer. The run generated 102,581,131 paired-end reads (15.6 Gbases total). The sequence dataset was deposited under SRR11177648.

Read quality control and virome assembly. To generate a proper assembly of all virome datasets, raw reads were trimmed and cleaned using AlienTrimmer v.0.4.0[61] and the NextFlex PCR Free adapter list, with the following options: -l 30. Cleaned reads were then assembled using metaSPAdes v.3.10.0[62] with the following parameters: --meta -k 21,33,55,77,99,127 --threads 12. Contigs ≥ 10 kbp were retained for further analyses.

Prediction and detection of phage contigs. Assembled contigs were analyzed for viral/phage detection using three different popular user-friendly software programs: MetaPhinder v.2.1[30], VirSorter v.1.0.3[29], and PHASTER (using multiple separate contigs option)[31]. All software was used with the default parameters.

PhageTermVirome analyses. PTV v.4.0.0 was run in the paired-end mode with the default options for all datasets analyzed in this study.

Phage contig annotation. Selected contigs were annotated with PROKKA v.1.14.0[63], run locally, using the PHASTER curated database as the primary source of trusted phage proteins (as of December 22, 2020), with an E-value threshold of 10^{-3}.

Protein network analysis with vConTACT2. Network analysis with vConTACT2 v.0.9.22 was performed using three reference contig databases (GVD[8], ViralRefSeq V.97[41], and a database of curated virus-like particles from the mouse from Duerkop et al.[42]. The visualization of the protein-sharing network was done using Cytoscape software v.3.1.1; http://cytoscape.org/)[64]. The edge-weighted spring-embedded model was used to position genomes sharing most protein clusters in proximity to one another.

Nucleotide sequence accession numbers. The raw read data of the human fecal virome from this study was deposited in the sequence read archive (SRA) under the accession number SRR11177648.

Data availability

PhageTermVirome source code is available at GitHub as a standalone software written in python3 (https://gitlab.pasteur.fr/vlegrand/ptv). A conda virtual environment is distributed with the software and all installation instructions are specified in the included "README.txt" file. Other datasets generated during the current study are available from the corresponding author on reasonable request.

Received: 31 July 2021; Accepted: 27 August 2021
Published online: 15 September 2021

References

1. Suttle, C. A. Viruses in the sea. *Nature* **437**, 356–361. https://doi.org/10.1038/nature04160 (2005).
2. Suttle, C. A. Marine viruses—Major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801–812. https://doi.org/10.1038/nrmicro1750 (2007).
3. Brum, J. R. & Sullivan, M. B. Rising to the challenge: Accelerated pace of discovery transforms marine virology. *Nat. Rev. Microbiol.* **13**, 147–159. https://doi.org/10.1038/nrmicro3404 (2015).
4. Fuhrman, J. A. Marine viruses and their biogeochemical and ecological effects. *Nature* **399**, 541–548. https://doi.org/10.1038/21119 (1999).
5. Forterre, P. The virocell concept and environmental microbiology. *ISME J.* **7**, 233–236. https://doi.org/10.1038/ismej.2012.110 (2013).
6. Rosenwasser, S., Ziv, C., van Creveld, S. G. & Vardi, A. Virocell metabolism: Metabolic innovations during host–virus interactions in the ocean. *Trends Microbiol.* **24**, 821–832. https://doi.org/10.1016/j.tim.2016.06.006 (2016).
7. Barr, J. J. *et al.* Bacteriophage adhering to mucus provide a non-host-derived immunity. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 10771–10776. https://doi.org/10.1073/pnas.1305923110 (2013).
8. Gregory, A. C. *et al.* The gut virome database reveals age-dependent patterns of virome diversity in the human gut. *Cell Host Microbe* **28**, 724-740.e8. https://doi.org/10.1016/j.chom.2020.08.003 (2020).
9. Broecker, F., Klumpp, J. & Moelling, K. Long-term microbiota and virome in a Zürich patient after fecal transplantation against *Clostridium difficile* infection. *Ann. N. Y. Acad. Sci.* **1372**, 29–41. https://doi.org/10.1111/nyas.13100 (2016).

10. Ma, Y., You, X., Mai, G., Tokuyasu, T. & Liu, C. A human gut phage catalog correlates the gut phageome with type 2 diabetes. *Microbiome* **6**, 24. https://doi.org/10.1186/s40168-018-0410-y (2018).
11. Monaco, C. L. *et al.* Altered virome and bacterial microbiome in human immunodeficiency virus-associated acquired immunodeficiency syndrome. *Cell Host Microbe* **19**, 311–322. https://doi.org/10.1016/j.chom.2016.02.011 (2016).
12. Norman, J. M. *et al.* Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* **160**, 447–460. https://doi.org/10.1016/j.cell.2015.01.002 (2015).
13. Gregory, A. C. *et al.* Marine DNA viral macro- and microdiversity from pole to pole. *Cell* **177**, 1109-1123.e14. https://doi.org/10.1016/j.cell.2019.03.040 (2019).
14. Roux, S. *et al.* Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**, 689–693. https://doi.org/10.1038/nature19366 (2016).
15. Brum, J. R. *et al.* Ocean plankton. Patterns and ecological drivers of ocean viral communities. *Science* **348**, 1261498. https://doi.org/10.1126/science.1261498 (2015).
16. Paez-Espino, D. *et al.* Uncovering Earth's virome. *Nature* **536**, 425–430. https://doi.org/10.1038/nature19094 (2016).
17. Emerson, J. B. *et al.* Host-linked soil viral ecology along a permafrost thaw gradient. *Nat. Microbiol.* **3**, 870–880. https://doi.org/10.1038/s41564-018-0190-y (2018).
18. Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D. & Lawley, T. D. Massive expansion of human gut bacteriophage diversity. *Cell* **184**, 1098-1109.e9. https://doi.org/10.1016/j.cell.2021.01.029 (2021).
19. Aggarwala, V., Liang, G. & Bushman, F. D. Viral communities of the human gut: Metagenomic analysis of composition and dynamics. *Mob. DNA* **8**, 12. https://doi.org/10.1186/s13100-017-0095-y (2017).
20. Youle, M., Haynes, M. & Rohwer, F. Scratching the surface of biology's dark matter. In *Viruses Essent. Agents's Life* (ed. Witzany, G.) 61–81 (Springer Netherlands, 2012).
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**, e08490. https://doi.org/10.7554/eLife.08490 (2015).
22. Bolduc, B. *et al.* vConTACT: An iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. *PeerJ* **5**, e3243. https://doi.org/10.7717/peerj.3243 (2017).
23. Bin Jang, H. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* **37**, 632–639. https://doi.org/10.1038/s41587-019-0100-8 (2019).
24. Amgarten, D., Braga, L. P. P., da Silva, A. M. & Setubal, J. C. MARVEL, a tool for prediction of bacteriophage sequences in metagenomic bins. *Front. Genet.* **9**, 304. https://doi.org/10.3389/fgene.2018.00304 (2018).
25. Tampuu, A., Bzhalava, Z., Dillner, J. & Vicente, R. ViraMiner: Deep learning on raw DNA sequences for identifying viral genomes in human samples. *PLoS One* **14**, e0222271. https://doi.org/10.1371/journal.pone.0222271 (2019).
26. Li, S. *et al.* Scrutinizing virus genome termini by high-throughput sequencing. *PLoS One* **9**, e85806. https://doi.org/10.1371/journal.pone.0085806 (2014).
27. Garneau, J. R., Depardieu, F., Fortier, L.-C., Bikard, D. & Monot, M. PhageTerm: A tool for fast and accurate determination of phage termini and packaging mechanism using next-generation sequencing data. *Sci. Rep.* **7**, 8292. https://doi.org/10.1038/s41598-017-07910-5 (2017).
28. Guo, J. *et al.* VirSorter2: A multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, 37. https://doi.org/10.1186/s40168-020-00990-y (2021).
29. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: Mining viral signal from microbial genomic data. *PeerJ* **3**, e985. https://doi.org/10.7717/peerj.985 (2015).
30. Jurtz, V. I., Villarroel, J., Lund, O., Voldby Larsen, M. & Nielsen, M. MetaPhinder-identifying bacteriophage sequences in metagenomic data sets. *PLoS One* **11**, e0163111. https://doi.org/10.1371/journal.pone.0163111 (2016).
31. Arndt, D. *et al.* PHASTER: A better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* **44**, W16–W21. https://doi.org/10.1093/nar/gkw387 (2016).
32. Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. & Sun, F. VirFinder: A novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**, 69. https://doi.org/10.1186/s40168-017-0283-5 (2017).
33. Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: Automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* **8**, 90. https://doi.org/10.1186/s40168-020-00867-0 (2020).
34. Ajami, N. J., Wong, M. C., Ross, M. C., Lloyd, R. E. & Petrosino, J. F. Maximal viral information recovery from sequence data using VirMAP. *Nat. Commun.* **9**, 3205. https://doi.org/10.1038/s41467-018-05658-8 (2018).
35. Tithi, S. S., Aylward, F. O., Jensen, R. V. & Zhang, L. FastViromeExplorer: A pipeline for virus and phage identification and abundance profiling in metagenomics data. *PeerJ* **6**, e4227. https://doi.org/10.7717/peerj.4227 (2018).
36. Auslander, N., Gussow, A. B., Benler, S., Wolf, Y. I. & Koonin, E. V. Seeker: Alignment-free identification of bacteriophage genomes by deep learning. *Nucleic Acids Res.* **48**, e121. https://doi.org/10.1093/nar/gkaa856 (2020).
37. Manrique, P. *et al.* Healthy human gut phageome. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 10400–10405. https://doi.org/10.1073/pnas.1601060113 (2016).
38. Sutton, T. D. S., Clooney, A. G., Ryan, F. J., Ross, R. P. & Hill, C. Choice of assembly software has a critical impact on virome characterisation. *Microbiome* **7**, 12. https://doi.org/10.1186/s40168-019-0626-5 (2019).
39. Somerville, V. *et al.* Long-read based de novo assembly of low-complexity metagenome samples results in finished genomes and reveals insights into strain diversity and an active phage system. *BMC Microbiol.* **19**, 143. https://doi.org/10.1186/s12866-019-1500-0 (2019).
40. Kleiner, M., Bushnell, B., Sanderson, K. E., Hooper, L. V. & Duerkop, B. A. Transductomics: Sequencing-based detection and analysis of transduced DNA in pure cultures and microbial communities. *Microbiome* **8**, 158. https://doi.org/10.1186/s40168-020-00935-5 (2020).
41. Brister, J. R., Ako-Adjei, D., Bao, Y. & Blinkova, O. NCBI viral genomes resource. *Nucleic Acids Res.* **43**, D571–D577. https://doi.org/10.1093/nar/gku1207 (2015).
42. Duerkop, B. A. *et al.* Murine colitis reveals a disease-associated bacteriophage community. *Nat. Microbiol.* **3**, 1023–1031. https://doi.org/10.1038/s41564-018-0210-y (2018).
43. Zablocki, O. *et al.* VirION2: A short- and long-read sequencing and informatics workflow to study the genomic diversity of viruses in nature. *Microbiology* https://doi.org/10.1101/2020.10.28.359364 (2020).
44. Bárdy, P. *et al.* Structure and mechanism of DNA delivery of a gene transfer agent. *Nat. Commun.* **11**, 3034. https://doi.org/10.1038/s41467-020-16669-9 (2020).
45. Roux, S. *et al.* Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. *Nat. Microbiol.* **4**, 1895–1906. https://doi.org/10.1038/s41564-019-0510-x (2019).
46. Tisza, M. J. *et al.* Discovery of several thousand highly diverse circular DNA viruses. *Elife* **9**, e51971. https://doi.org/10.7554/eLife.51971 (2020).
47. Creasy, A., Rosario, K., Leigh, B. A., Dishaw, L. J. & Breitbart, M. Unprecedented diversity of ssDNA phages from the family microviridae detected within the gut of a protochordate model organism (*Ciona robusta*). *Viruses* https://doi.org/10.3390/v10080404 (2018).
48. Krupovic, M. & Forterre, P. Single-stranded DNA viruses employ a variety of mechanisms for integration into host genomes. *Ann. N. Y. Acad. Sci.* **1341**, 41–53. https://doi.org/10.1111/nyas.12675 (2015).

49. Krupovic, M. & Koonin, E. V. Evolution of eukaryotic single-stranded DNA viruses of the Bidnaviridae family from genes of four other groups of widely different viruses. *Sci. Rep.* **4**, 5347. https://doi.org/10.1038/srep05347 (2014).
50. Stamatoyannopoulos, J. A. Illuminating eukaryotic transcription start sites. *Nat. Methods* **7**, 501–503. https://doi.org/10.1038/nmeth0710-501 (2010).
51. Dar, D. *et al.* Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria. *Science* **352**, aad9822. https://doi.org/10.1126/science.aad9822 (2016).
52. Garalde, D. R. *et al.* Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* **15**, 201–206. https://doi.org/10.1038/nmeth.4577 (2018).
53. Lorenz, D. A., Sathe, S., Einstein, J. M. & Yeo, G. W. Direct RNA sequencing enables m6A detection in endogenous transcript isoforms at base-specific resolution. *RNA* **26**, 19–28. https://doi.org/10.1261/rna.072785.119 (2020).
54. Martínez-Rubio, R. *et al.* Phage-inducible islands in the Gram-positive cocci. *ISME J.* **11**, 1029–1042. https://doi.org/10.1038/ismej.2016.163 (2017).
55. Penadés, J. R. & Christie, G. E. The phage-inducible chromosomal islands: A family of highly evolved molecular parasites. *Annu. Rev. Virol.* **2**, 181–201. https://doi.org/10.1146/annurev-virology-031413-085446 (2015).
56. Tomasch, J. *et al.* Packaging of Dinoroseobacter shibae DNA into gene transfer agent particles is not random. *Genome Biol. Evol.* **10**, 359–369. https://doi.org/10.1093/gbe/evy005 (2018).
57. Soucy, S. M., Huang, J. & Gogarten, J. P. Horizontal gene transfer: Building the web of life. *Nat. Rev. Genet.* **16**, 472–482. https://doi.org/10.1038/nrg3962 (2015).
58. Fogg, P. C. M. Identification and characterization of a direct activator of a gene transfer agent. *Nat. Commun.* **10**, 595. https://doi.org/10.1038/s41467-019-08526-1 (2019).
59. Bolduc, B., Youens-Clark, K., Roux, S., Hurwitz, B. L. & Sullivan, M. B. iVirus: Facilitating new insights in viral ecology with software and community data sets imbedded in a cyberinfrastructure. *ISME J.* **11**, 7–14. https://doi.org/10.1038/ismej.2016.89 (2017).
60. Paez-Espino, D. *et al.* IMG/VR v.2.0: An integrated data management and analysis system for cultivated and environmental viral genomes. *Nucleic Acids Res.* **47**, D678–D686. https://doi.org/10.1093/nar/gky1127 (2019).
61. Criscuolo, A. & Brisse, S. AlienTrimmer removes adapter oligonucleotides with high sensitivity in short-insert paired-end reads. Commentary on Turner (2014) Assessment of insert sizes and adapter content in FASTQ data from NexteraXT libraries. *Front. Genet.* **5**, 130. https://doi.org/10.3389/fgene.2014.00130 (2014).
62. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: A new versatile metagenomic assembler. *Genome Res.* **27**, 824–834. https://doi.org/10.1101/gr.213959.116 (2017).
63. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics (Oxf., Engl.)* **30**, 2068–2069. https://doi.org/10.1093/bioinformatics/btu153 (2014).
64. Kohl, M., Wiese, S. & Warscheid, B. Cytoscape: Software for visualization and analysis of biological networks. In *Data Min. Proteomics Stand. Appl.* (eds Hamacher, M. *et al.*) 291–303 (Humana Press, 2011).

Author contributions

J.R.G., M.Mo. and D.B. designed the study. J.R.G., V.L., M.Mo., and D.B. wrote the PhageTermVirome code. M.Ma. prepared the data from the human virome. J.R.G., M.Mo., V.L., D.R.V. and M.O.P. performed the bioinformatics analyses. J.R.G., M.Mo., D.B. and M.B.S. wrote the manuscript. J.R.G., V.L., M.Mo., D.B., M.O.P., D.R.V., L.-C.F., M.Ma., M.B.S. contributed to the manuscript and have read and accepted the final version.

Funding

This work was supported by the French Government's program Investissement d'Avenir program; Laboratoire d'Excellence 'Integrative Biology of Emerging Infectious Diseases' [ANR-10-LABX-62-IBEID] and the Natural Sciences and Engineering Research Council of Canada (NSERC #341450-2010). Biomics Platform, C2RT, Institut Pasteur, Paris, France, is supported by France Génomique (ANR-10-INBS-09-09) and the Infrastructures de recherche en biologie, santé et agronomie (IBISA).

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.R.G. or M.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.