virMine: automated detection of viral sequences from complex metagenomic samples

Andrea Garretto[1], Thomas Hatzopoulos[2] and Catherine Putonti[1,2,3,4]

[1] Bioinformatics Program, Loyola University of Chicago, Chicago, IL, United States of America
[2] Department of Computer Science, Loyola University of Chicago, Chicago, IL, United States of America
[3] Department of Biology, Loyola University of Chicago, Chicago, IL, United States of America
[4] Department of Microbiology and Immunology, Loyola University of Chicago, Maywood, IL, United States of America

ABSTRACT

Metagenomics has enabled sequencing of viral communities from a myriad of different environments. Viral metagenomic studies routinely uncover sequences with no recognizable homology to known coding regions or genomes. Nevertheless, complete viral genomes have been constructed directly from complex community metagenomes, often through tedious manual curation. To address this, we developed the software tool virMine to identify viral genomes from raw reads representative of viral or mixed (viral and bacterial) communities. virMine automates sequence read quality control, assembly, and annotation. Researchers can easily refine their search for a specific study system and/or feature(s) of interest. In contrast to other viral genome detection tools that often rely on the recognition of viral signature sequences, virMine is not restricted by the insufficient representation of viral diversity in public data repositories. Rather, viral genomes are identified through an iterative approach, first omitting non-viral sequences. Thus, both relatives of previously characterized viruses and novel species can be detected, including both eukaryotic viruses and bacteriophages. Here we present virMine and its analysis of synthetic communities as well as metagenomic data sets from three distinctly different environments: the gut microbiota, the urinary microbiota, and freshwater viromes. Several new viral genomes were identified and annotated, thus contributing to our understanding of viral genetic diversity in these three environments.

Submitted 30 November 2018
Accepted 26 February 2019
Published 10 April 2019

Corresponding author
Catherine Putonti, cputonti@luc.edu

Academic editor
Gerard Lazo

Additional Information and Declarations can be found on page 13

DOI 10.7717/peerj.6695

Subjects Bioinformatics, Computational Biology, Genomics, Microbiology, Virology
Keywords Virome, Metagenomics, Bacteriophage, Human microbiome, Freshwater virome

INTRODUCTION

In contrast to eukaryotic and prokaryotic organisms, only a small fraction of viral genomes has been sequenced and characterized. Viral metagenomic studies have been pivotal in increasing our understanding of viral diversity on Earth. Numerous habitats have been explored, such as: marine waters (*Breitbart et al., 2002*; *Yooseph et al., 2007*; *Hurwitz & Sullivan, 2013*; *Brum et al., 2015*; *Coutinho et al., 2017*; *Zeigler Allen et al., 2017*; see review *Brum & Sullivan, 2015*), soil (*Fierer et al., 2007*; *Zablocki et al., 2014*; *Adriaenssens et al., 2017*; see review *Pratama & Van Elsas, 2018*), freshwaters (*López-Bueno et al., 2009*; *López-Bueno et al., 2015*; *Roux et al., 2012*; see review *Bruder et al., 2016*), and the human

How to cite this article Garretto A, Hatzopoulos T, Putonti C. 2019. virMine: automated detection of viral sequences from complex metagenomic samples. *PeerJ* 7:e6695 http://doi.org/10.7717/peerj.6695

microbiota (e.g., *Reyes et al., 2010*; *Minot et al., 2011*; *Minot et al., 2013*; *Pride et al., 2012*; *Hannigan et al., 2015*; *Santiago-Rodriguez et al., 2015*; *Miller-Ensminger et al., 2018*; see review *Abeles & Pride, 2014*). Recent evidence has uncovered that viral members of the human microbiota (see reviews *Barr, 2017*; *Keen & Dantas, 2018*) and marine environment (see reviews *Breitbart et al., 2018*) play a more pivotal role than once thought. Regardless of the environment explored, the overwhelming majority of viral sequences produced exhibit no sequence homology to characterized viral species. Even for the well-studied marine viral communities, over 60% of the coding regions predicted are completely novel (*Coutinho et al., 2017*).

While metagenomics has been fruitful in identifying gene markers (e.g., 16S rRNA gene) and genomes of uncultivated eukaryotic and prokaryotic species (*Hug et al., 2016*), surveys of viromes face unique challenges (*Bruder et al., 2016*; *Rose et al., 2016*). First, unlike cellular organisms, there is no universally conserved gene in viruses. Viruses span a high degree of genetic diversity and are inherently mosaic (*Hatfull, 2008*). Second, even when sequencing purified virions, sequencing data often includes non-viral (host) DNA. This is further complicated by the fact that viral genomic DNA is often orders of magnitude less abundant than host cells or other organisms in the sample. In addition to the development of experimental procedures for viral metagenomics (e.g., *Conceição Neto et al., 2015*; *Hayes et al., 2017*; *Lewandowska et al., 2017*), several bioinformatic solutions have been created to aid in detecting viral sequences within mixed communities (e.g., *Roux et al., 2015*; *Hatzopoulos, Watkins & Putonti, 2016*; *Yamashita, Sekizuka & Kuroda, 2016*; *Ren et al., 2017*; *Amgarten et al., 2018*; see reviews *Hurwitz et al., 2018*; *Nooij et al., 2018*). Third, extant viral data repositories do not include sufficient representation of viral species. Thus, tools reliant upon identifying sequence homology, such as those for bacterial metagenome analysis (see review *Nayfach & Pollard, 2016*), have limited application in virome studies.

The identification of viral genomes from samples containing a single or a few viral species is relatively straight-forward, even in the presence of a large background of non-viral sequences. An example of such an inquiry would be the search for potential viral pathogens from clinical samples. Software tools including VIP (*Li et al., 2016b*), VirAmp (*Wan et al., 2015*), and VirFind (*Ho & Tzanetakis, 2014*) were designed specifically for such cases. They are, however, limited to the isolation of known viral taxa; complex viral communities pose significantly greater challenges. Typically, one of two approaches is taken. The first approach identifies contigs from metagenomic data sets based upon sequence attributes, e.g., their nucleotide usage profiles (*Ren et al., 2017*), and/or contig coverage (see reviews *Sharon & Banfield, 2013*; *Garza & Dutilh, 2015*; *Sangwan, Xia & Gilbert, 2016*). The second, more frequently pursued method, relies largely on recognizable homologies to known viral sequences, e.g., Phage Eco-Locator (*Aziz et al., 2011*), VIROME (*Wommack et al., 2012*), MetaVir (*Roux et al., 2014*), VirSorter (*Roux et al., 2015*), MetaPhinder (*Jurtz et al., 2016*), VirusSeeker (*Zhao et al., 2017*), and FastViromeExplorer (*Tithi et al., 2018*). The tool MARVEL integrates the two approaches, predicting tailed phage sequences based upon genomic features (gene density and strand shifts) and sequence homologies (*Amgarten et al., 2018*). Regardless of the approach taken, manual curation and inspection is often a critical step in the process. Several complete viral genomes have been mined from metagenomic data through inspection of sequences based upon their size, coverage, circularity, or sequence homology to annotated viral genes or genes of interest (e.g., *Inskeep et al., 2013*; *Labonté & Suttle, 2013*; *Dutilh et al., 2014*; *Smits et al., 2014*; *Smits et al., 2015*; *Bellas, Anesio & Barker, 2015*; *Rosario et al., 2015*; *Zhang et al., 2015*; *Paez-Espino et al., 2016*; *Voorhies et al., 2016*; *Coutinho et al., 2017*; *Ghai et al., 2017*; *Watkins, Sible & Putonti, 2018*). These efforts have uncovered novel viral species, furthering our understanding of genetic diversity in nature.

Here we present virMine for the identification of viral genomes within metagenomic data sets. virMine automates the process of discovery; from raw sequence read quality control through assembly and annotation. virMine incorporates a variety of publicly available tools and user-defined criteria. In contrast to previous bioinformatic tools which search for viral "signatures" based on our limited knowledge of viral diversity on Earth, virMine takes advantage of the wealth of sequence data available for cellular organisms. Thus, viral (bacteriophage and eukaryotic virus) discovery is conducted through the process of excluding what we know not to be viral. Those sequences which are not "non-viral" (i.e., putative viral sequences) are then compared to a database of viral sequences. This comparison distinguishes putative viral sequences similar to known viral sequences and those which may represent novel viruses for downstream analyses. A beta version of this tool was used to isolate viral sequences from urinary metagenome data sets (*Garretto et al., 2018*). Here we illustrate the utility of this tool using four case-studies: synthetic data sets, gut microbiomes, urinary viromes, and freshwater viromes, resulting in the identification of new strains of known viruses as well as novel viral genomes.

MATERIALS & METHODS

Pipeline development

The pipeline integrates existing tools as well as new algorithms using Python and the BioPython library (*Cock et al., 2009*). Figure 1 depicts the process employed by virMine. A key aspect of the tool is its flexibility; it was designed to be modular, allowing users to access functionality individually or execute the full pipeline. While several methods have been incorporated in this release (Table 1), new tools can be added easily. Furthermore, to facilitate targeted analyses, filtration options and customization is available for users without any programming expertise.

Users can supply either raw Illumina sequencing reads (single-end or paired-end) or assembled contigs/scaffolds. In the case in which reads are supplied, the raw sequencing data is evaluated using the quality control tool Sickle (https://github.com/najoshi/sickle). Reads are trimmed, generating high quality data for assembly. Presently, the pipeline performs assembly by one of three methods: SPAdes (*Bankevich et al., 2012*), metaSPAdes (*Nurk et al., 2017*), and MEGAHIT (*Li et al., 2016a*). These assemblers were selected as they include tools better equipped for assembly of low complexity samples (SPAdes) and those developed for complex metagenomes (metaSPAdes and MEGAHIT). In a prior study comparing tools for assembly of phage genomes from single or low complexity samples

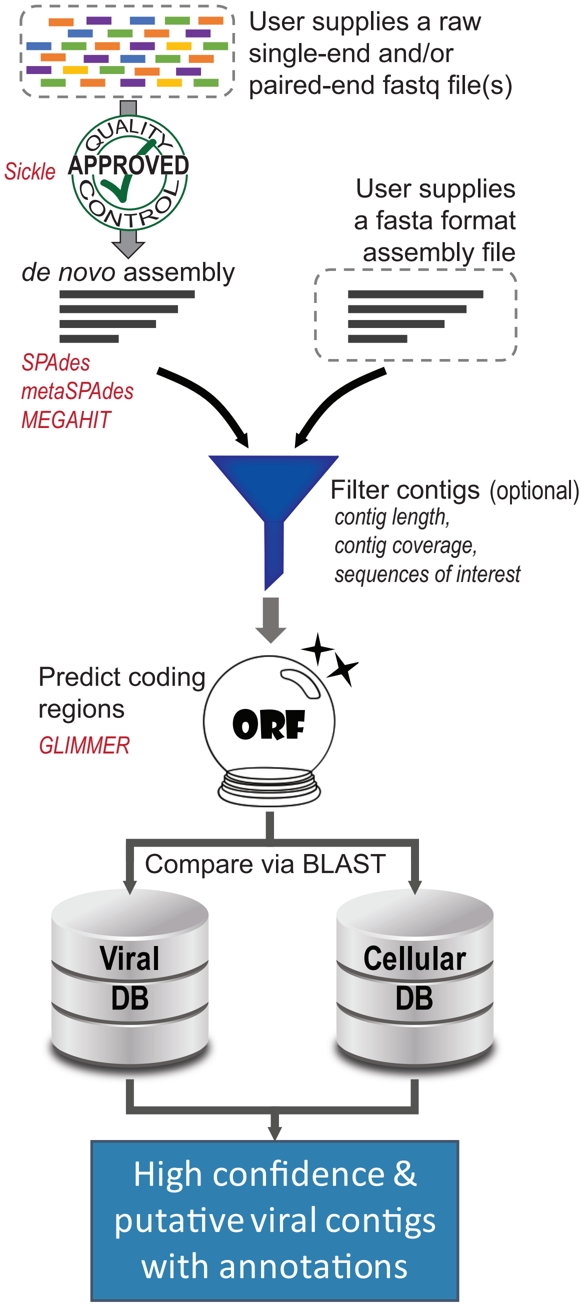

Figure 1 **Overview of virMine pipeline.** Tools integrated into the pipeline are listed in red. The sequences for viral contigs predicted with high confidence ("viral_contigs") and putative viral contigs ("unkn_contigs") are written to file.

Full-size DOI: 10.7717/peerj.6695/fig-1

(*Rihtman et al., 2016*), the SPAdes assembler (*Bankevich et al., 2012*) outperformed other tools tested. virMine also includes the assembly option "all3". This option assembles the reads using SPAdes, metaSPAdes, and MEGAHIT and selects the assembly with the highest N_{50} score for downstream analysis. The virMine command line includes a flag for the

Table 1 Software integrated into the virMine pipeline.

Tool	Version	Task	Citation
Sickle	1.33	Read trimming	https://github.com/najoshi/sickle
SPAdes	3.10.1	Assembly	*Bankevich et al. (2012)*
metaSPAdes	3.10.1	Assembly	*Nurk et al. (2017)*
MEGAHIT	1.1.4	Assembly	*Li et al. (2016a)*
BBMap	37.36	Coverage	https://sourceforge.net/projects/bbmap/
GLIMMER	3.02	Gene prediction	*Delcher et al. (1999)*
BLAST+	2.6.0	Sequence Analysis	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/

user to specify the number of threads to be used during assembly to best utilize multi-core resources.

Next, virMine includes several options for the user to filter the assembled contigs. This can include minimum and/or maximum contig length, minimum contig coverage, and presence of genes or sequences (such as CRISPR spacer sequences) of interest. Coverage is calculated by remapping the original reads to the contigs, and the per contig coverage is calculated via BBMap (https://sourceforge.net/projects/bbmap/). Coverage is not reported if this option is not selected. Alternatively, when SPAdes (*Bankevich et al., 2012*) or metaSPAdes (*Nurk et al., 2017*) is used for assembly, users can select to use the SPAdes "cov" value as a filter. Users can also provide FASTA format sequences of interest (e.g., gene sequences encoding for a specific functionality); contigs are then queried against this data set using blastx. Results with a bitscore >50 are considered real hits and only contigs containing these hits will be considered further. Any or all of these filters can be selected by the user. Furthermore, the order in which they are specified by the user determines the order in which the filters are applied.

In Step 3, coding regions are predicted for each contig. Open reading frame (ORF) prediction is conducted using the tool GLIMMER (*Delcher et al., 1999*). Coding regions are predicted using a modified GLIMMER script (available through our GitHub repository), trained to accommodate characteristics of viral genes, e.g., overlapping genes (*Chirico, Vianelli & Belshaw, 2010*) and short coding regions.

In the final step, each predicted ORF is compared to two databases—a collection of non-viral sequences and a collection of known viral sequences. These two databases can be manually curated data collections or obtained from public repositories. While the GitHub repository for virMine includes a script to generate databases from NCBI's RefSeq collection, any multi-fasta file of amino acid sequences can be used to create these databases; the user need only supply the multi-fasta files. Comparisons against these two databases are facilitated via the BLAST+ application (*Camacho et al., 2009*). Users can select to use either a blastp (protein query) or blastx (translated nucleotide) query. While blastx is more exhaustive, blastp is more expedient. Again, the threads flag is used here to expedite these comparisons. All hits are reported from both databases; the bitscores for each ORF's hits to the two databases are compared, and the ORF is called "viral" or "non-viral" based upon the hit with the greater bitscore. Contigs with more "viral" calls are predicted as viral and are written to file ("viral_contigs.fasta"), as are their ORF predictions and BLAST (either blastx or blastp) results. Contigs containing ORFs with no recognizable sequence homology to the viral database or non-viral database are classified as "unknown". These putative viral contigs ("unkn_contigs.fasta") and their ORF predictions are also written to file, as these sequences may represent truly novel species.

Tool availability

virMine is available through a Docker image at https://github.com/thatzopoulos/virMine; Docker builds the necessary environment. This repository also includes scripts for generating curated viral and bacterial databases from GenBank. The user can save the contents of their run locally, as well as supply their own input files prior to the building of the environment, by following the steps listed in the GitHub repository. This pipeline can be run on any system supporting Docker (https://www.docker.com/). Development and testing were conducted on the Ubuntu and MacOSX operating systems.

Data sets

The pipeline includes two databases for distinguishing between non-viral and viral sequences. Two data sets were created for our proof-of-concept work. The viral database includes amino acid sequences from all RefSeq (*O'Leary et al., 2016*) viral genomes and can be retrieved directly online at ftp://ftp.ncbi.nlm.nih.gov/genomes/Viruses/all.faa.tar.gz. This data set includes both eukaryotic viruses and phages. The non-viral data set used for our proof-of-concept work was created using the bacterial COGs collection (*Galperin et al., 2015*), excluding coding sequences in the category X of phage-derived proteins, transposases, and other mobilome components. The GitHub repository for virMine includes a script to create these two databases.

For the proof-of-concept studies presented in the results, four data sets were used. The first is a synthetic data set for benchmarking purposes. Sequencing read sets for a single "non-viral" sequence (*Pseudomonas aeruginosa* UW4 (NC_019670.1)) and a single viral sequence (*Pseudomonas phage* PB1 (NC_011810.1)) were created at various "concentrations" using the tool MetaSim (*Richter et al., 2008*). These synthetic data sets were made both with and without mutations introduced. (Mutations were introduced using the evolve function in which the parameters "number of leaves (Yule-Harding Tree)" and "Jukes-Cantor Model Alpha" were set to the defaults 100 and 0.0010, respectively.) Raw sequencing reads were also obtained from five different studies including the gut microbiota (*Qin et al., 2010*; *Reyes et al., 2010*), the urinary microbiota (*Santiago-Rodriguez et al., 2015*), and freshwater viromes (*Sible et al., 2015*; *Skvortsov et al., 2016*). Table 2 summarizes these data sets; details regarding the URLs for these data sets can be found in File S1.

Local BLAST searches of contigs were conducted using the complete nr/nt database (downloaded 6/24/2017). Remote BLAST queries were conducted through the NCBI website. Genome annotations were generated using RAST (*Aziz et al., 2008*), previously used for phage genome annotations (*McNair et al., 2018*). Contig mapping to complete genome sequences was performed using Bowtie2 (*Langmead & Salzberg, 2012*).

Table 2 Complex community microbiomes examined for virMine proof-of-concept study.

Sample	Study details	Sequencing technology	# samples	# reads (millions)
Synthetic	*P. aeruginosa* and *Pseudomonas phage* PB1 genomes	N/A	22	4.4
Gut Microbiomes	A subset of faecal microbiota of monozygot twins and their mothers (*Reyes et al., 2010*)	454 FLX	3	0.66
	A subset of faecal samples from 124 European individuals (*Qin et al., 2010*)	Illumina Genome Analyzer	55	1141.33
Urinary Viromes	UTI positive urine samples (*Santiago-Rodriguez et al., 2015*)	Ion Torrent PGM	10	6.22
Freshwater Viromes	A subset of samples from Lake Michigan nearshore waters (*Sible et al., 2015*)	Illumina MiSeq	4	13.46
	Viral community of Lough Neagh (*Skvortsov et al., 2016*)	Illumina MiSeq	1	4.60

RESULTS & DISCUSSION

virMine is a single tool to perform raw read quality control, assembly, annotation, and analysis (Fig. 1). The virMine tool, as described in the Methods, identifies viral sequences and putative viral sequences in metagenomic data sets by harnessing the wealth of non-viral sequence data available; contigs are scored based upon their similarity to non-viral and viral sequences. Four case studies were derived to test the efficacy of the virMine software tool, including one synthetic data set and three different environmental samples from the gut, urine, and freshwaters.

Case study 1: synthetic data sets

Sequencing reads were generated using the tool MetaSim (*Richter et al., 2008*) using a sample "non-viral" genome sequence, *Pseudomonas aeruginosa* UW4 genome (GenBank: NC_019670), and a viral genome sequence, *Pseudomonas* phage PB1 (GenBank: NC_011810). Eleven synthetic data sets were created in which 0% through 100% (increments of 10%) of the data set comprised of "reads" from the phage genome sequence. Each synthetic data set was processed independently; assemblies were generated using SPAdes (*Bankevich et al., 2012*) with the requirement that the coverage (-cov flag) be greater than or equal to three.

Figure 2 summarizes the results of the analyses. When 50% or more of the reads were from the PB1 genome, the complete PB1 genome could be reconstructed. As the N_{50} scores for each of the runs show, the length of the virMine assembled viral genome exceeds that of the PB1 genome (65,764 bps); this is a residual of the direct terminal repeats (DTRs) in the PB1 sequence. The presence of DTRs frequently leads to "wrap-around" reads contained within the genome assembly (*Merrill et al., 2016*). Each contig that did not correspond with the PB1 genome, including those identified within the 0% PB1 genome data set, was further examined via local blastn against the nr/nt database (File S2). As Fig. 2 shows, even for the synthetic data set with no reads from the PB1 genome, two contigs were predicted by virMine to be viral. We further investigated these contigs, 1021 and 1007 bp in length; the first contig is homologous to an IS3 family transposase (GenBank: AFY17357) and an IS110 family transposase (GenBank: AFY17680), respectively. As these transposases are

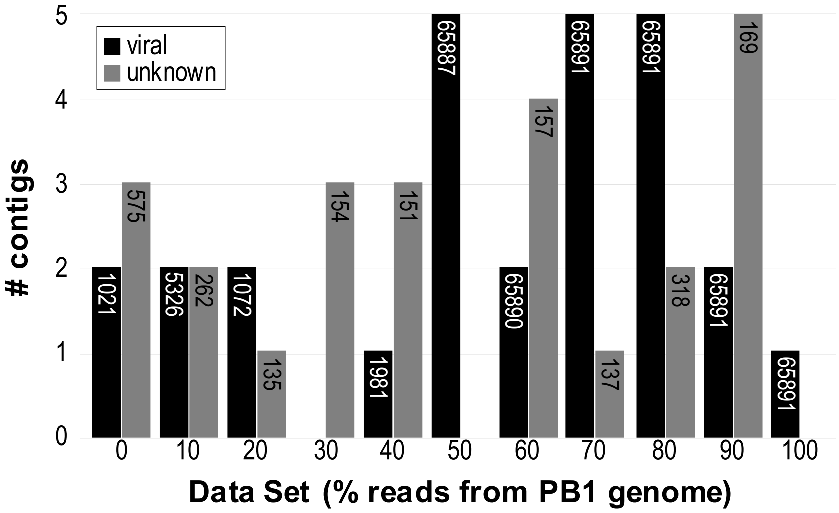

Figure 2 Number of contigs assembled for each of the synthetic data sets predicted as viral (black bars) or of unknown origin (gray bars). The N_{50} score of the assembled contigs in each group is indicated within the corresponding bars.

Full-size DOI: 10.7717/peerj.6695/fig-2

assigned COG id numbers within the category X, they were excluded from the non-viral database and thus not recognized as non-viral. Transposases are abundant in nature and have been found within phage genomes (*Aziz, Breitbart & Edwards, 2010*).

MetaSim (*Richter et al., 2008*) was used again to produce synthetic data sets for the *P. aeruginosa* and *Pseudomonas* phage PB1, this time introducing mutation (population-based random mutator; see Methods). As shown in File S2, the assemblies produced were significantly more fragmented (lower N_{50} scores); even when all reads were derived from the PB1 genome sequence, the N_{50} score was only 762 bp (in contrast to the single, full genome contig produced with the read sets generated without mutation). It is interesting to note that while the assemblers could not reconstruct the full genome or longer contigs, virMine still classified contigs as viral and subsequent blastn analyses were able to resolve the origin of the sequence.

Case study 2: gut microbiomes

Two separate gut microbiome data sets were examined (Table 2). The first includes the sequence data sets that were examined leading to the discovery of the crAssphage genome sequence (97,065 bp) (*Dutilh et al., 2014*): the data set of *Reyes et al. (2010)*. The crAssphage has since been detected in raw sewage and sewage impacted water samples (*Stachler et al., 2017*). Similar to the methods employed in the discovery of the crAssphage, both the sequence data sets of the individual samples and an aggregate of all reads were assembled by virMine using SPAdes (*Bankevich et al., 2012*). Numerous sequences predicted to be viral were identified within the individual samples (727 total) and the aggregate data set (927 total) (File S2). Local blastn analyses identified many of these contigs as representative of transposases and integrases. The abundance of transposase sequences within metagenomic sequences has previously been noted for a variety of environments (*Brazelton & Baross,*

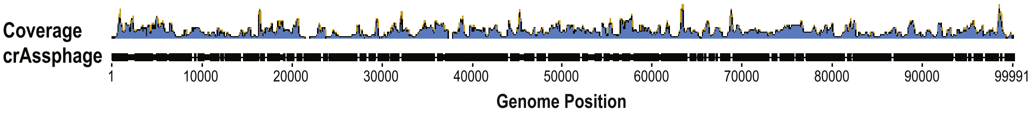

Figure 3 Coverage of crAssphage by contigs predicted by virMine as viral or unknown.
Full-size DOI: 10.7717/peerj.6695/fig-3

2009; *Aziz, Breitbart & Edwards, 2010*; *Vigil-Stenman et al., 2017*). We compared the contigs identified as viral to the crAssphage genome sequence (GenBank: JQ995537). 94.88% of the crAssphage genome was represented in 372 contigs identified as viral sequences. Coverage of the crAssphage increases when contigs classified as unknown are considered: 98.32% of the genome is represented in 613 contigs (Fig. 3). Several other complete viral genomes were also identified by virMine including a Gokushovirus and Microvirus exhibiting homology to the sequenced genomes of Gokushovirus WZ-2015a (GenBank: KT264754) and the newly discovered Microviridae sp. isolate ctci6 (GenBank: MH617627). It is worth noting that this Microviridae genome was not included in our viral database and exhibits no significant homology to other records in the current BLAST Nucleotide collection.
The second gut microbiota data set was a subset of the fecal samples from 124 European individuals (*Qin et al., 2010*). Most of this data set is bacterial in origin, with only 0.1% predicted by the authors of the study to be of eukaryotic and viral origin. Using virMine we also found that most of the sequences were likely bacterial (File S2). However, we found that the prediction of the study's authors underestimated the viral population; 1.31–38.43% of the assembled contigs were predicted by virMine to be viral in origin. We hypothesize that this discrepancy may be due to prophage sequences. As our previous analysis with the beta version of the software showed, virMine can identify prophage sequences within bacterial genome contigs as well as extrachromosomal viruses (*Garretto et al., 2018*). This underestimate may also be a result of our increased knowledge of viral diversity; the number of viral sequences in GenBank has tripled since the study of *Qin et al. (2010)* was published. The summary of our analysis of the 55 samples from this study are listed in File S2. In total 28,673 and 311,457 contigs were categorized as viral and unknown, respectively.

Case study 3: urinary viromes

Ten data sets, collected from individuals with urinary tract infections (*Santiago-Rodriguez et al., 2015*), were selected for analysis. In contrast to the gut microbiomes examined in Case Study 2, these samples were prepared such that the majority (if not all) of the sequenced DNA was representative of the viral fraction (*Santiago-Rodriguez et al., 2015*). Exploration of the urinary virome has only recently begun. Of the few viral metagenomic studies of the urinary microbiota (*Santiago-Rodriguez et al., 2015*; *Rani et al., 2016*; *Thannesberger et al., 2017*; *Garretto et al., 2018*; *Miller-Ensminger et al., 2018*; *Moustafa et al., 2018*), most of the identifiable sequences are similar to characterized phage sequences. Nevertheless, the vast majority of contigs exhibit no identifiable homology to sequence databases. As summarized in File S2, each sample consisted of more contigs in the "unknown" category than the "viral" category. We selected the larger contigs (>5,000 bp) that were predicted as viral and queried them via megablast against the nr/nt database online. Table 3 presents

Table 3 **BLAST homology for longer (>5,000 bp) contigs predicted as viral.**

SRA Run #	BLAST hit	Accession #	Contig length	% ID	% QC
MGM4568637	*Cyanothece* sp. PCC 7822	CP002198	14,157	73	0
	Choristoneura rosaceana entomopoxvirus 'L'	HF679133	11,424	66	15
MGM4568639	*Erlichia canis* strain YZ-1[a]	CP02479	12,310	73	8
	Burkholderia sp. MSMB0856	CP013427	5,156	71	5
MGM4568640	*Clostridium taeniosporum* strain 1/k	CP017253	7,987	69	2
	Escherichia phage YDC107_2	CP025713	5,479	96	88
MGM4568641	*Enterococcus faecalis* V583[a]	AE016830	16,416	95	95
	Uncultured Mediterranean phage uvMED	AP013535	13,087	79	1
	Turicibacter sp. H121	CP013476	7,825	83	0
	Enterococcus faecalis strain L9[a]	CP018004	5,086	99	100
MGM4568642	*Choristoneura rosaceana* entomopoxvirus 'L'	HF679133	9,301	66	27
	Protochlamydia naegleriophila PNK1	LN879502	5,312	83	1
MGM4568645	*Rickettsiales* bacterium Ac37b[a]	CP009217	8,302	66	11
	Rickettsiales bacterium Ac37b[a]	CP009217	8,215	68	19

Notes.

[a] Indicates BLAST homologies to annotated prophage regions.

the results of this search. virMine identified similarities to annotated prophage sequences (indicated by asterisks), extrachromosomal phages, and eukaryotic viral sequences.

Case study 4: freshwater viromes

Two freshwater viromes were considered. The first includes four samples from the Lake Michigan nearshore waters, collected by our group (*Sible et al., 2015*; *Watkins et al., 2016*). The second includes samples taken from Lough Neagh, the largest freshwater lake in Ireland (*Skvortsov et al., 2016*). The summary statistics for our analysis are included in File S2. Sequences predicted to be viral within the four Lake Michigan data sets were inspected. Hits to known viral sequences varied between samples; in total, sequence homologies were detected to 834 different phage ($n = 425$) and eukaryotic viruses ($n = 409$). Figure 4 illustrates the species, predominantly phages, with the most hits. From the Lough Neagh data set, nine contigs were identified by virMine as viral and had a length greater than 40 Kbp. In the study introducing this data set (*Skvortsov et al., 2016*), only five contigs were produced meeting this length threshold. (The IDBA-UD assembler (*Peng et al., 2012*) was used in the original analysis of this data set (*Skvortsov et al., 2016*).) Each contig was submitted to RAST (*Aziz et al., 2008*) for annotation and each was found to contain phage-related genes (File S3), suggesting that the contigs represented complete or partial phage genomes. We next queried each contig against the nr/nt database via blastn identifying only modest sequence homology to bacterial, phage, and uncultured viral isolate sequences (Table 4). These contigs thus represent likely novel viral sequences.

virMine performance

To assess the performance of virMine, the freshwater data sets were also examined using the viral sequence identification tool VirSorter (v. 1.0.3) (*Roux et al., 2015*). For all five

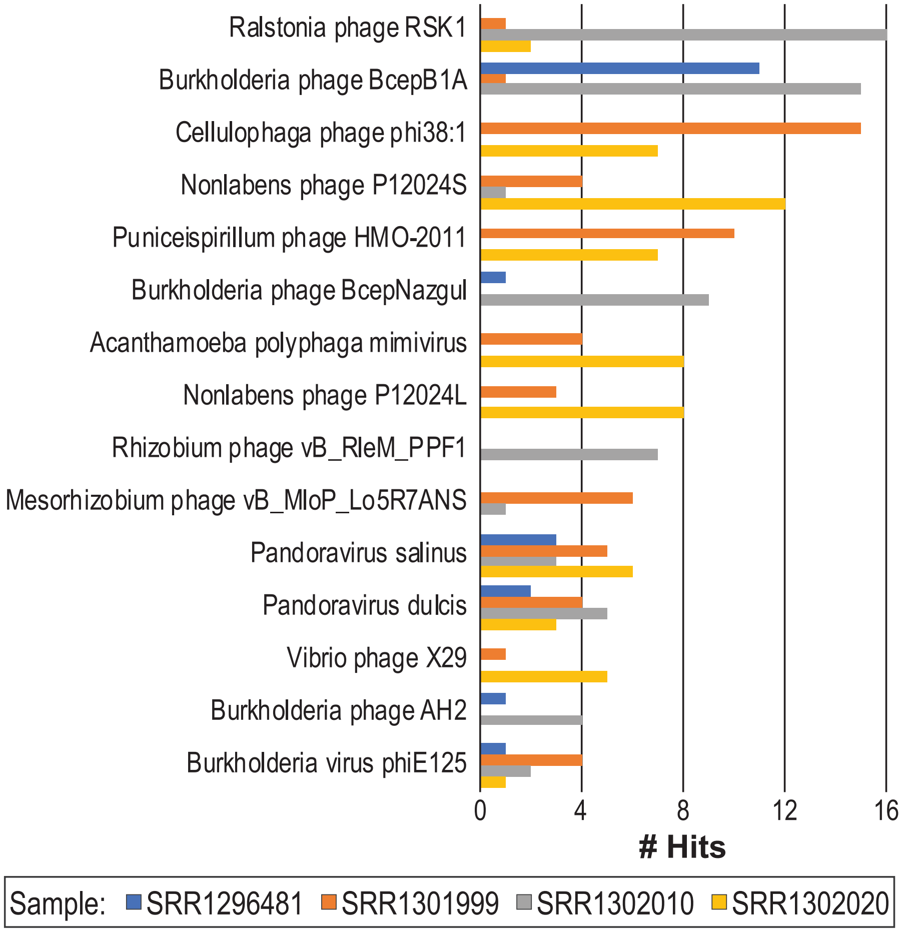

Figure 4 **Viral species most frequently detected within the Lake Michigan data sets.**

Full-size DOI: 10.7717/peerj.6695/fig-4

data sets, we found that very few contigs were predicted as viral by both tools. For instance, in the Lough Neagh data set, VirSorter only identified (a category 2 prediction) one of the nine virMine contigs (length > 40 Kbp). This prompted our manual inspection of these results. Herein we present the results for one of the samples from Lake Michigan (SRA accession number SRR1296481), representative of what we found in all sets. virMine predicted 60 of the 1,518 assembled contigs as viral. VirSorter predicted only 20 viral sequences (two category 1; five category 2; six category 3; no category 4; four category 5; and three category 6). Only two sequences were predicted by both tools. As virMine was designed for identifying viral contigs and VirSorter was designed to identify both viral contigs (categories 1–3) and prophages (categories 4–6), it is not surprising that both contigs detected by the two tools were VirSorter category 2 sequences. (While virMine can identify prophages, as was shown previously (*Garretto et al., 2018*), it will not identify prophages within large bacterial contigs.) BLAST queries to the nr/nt database of the sequences uniquely identified by virMine and VirSorter are listed in File S3; many of these predicted sequences exhibited homology to bacterial RNAs (rRNA and tRNA). Only four

Table 4 **Viral genome sequences identified by virMine from the Lough Neagh virome (*Skvortsov et al., 2016*).**

Contig	Length	# CDS	BLAST hit	Accession #	% QC	% ID	Isolation source
contig_11	46,867	71	*Chromobacterium rhizoyzae* strain JP2-74	CP031968.1	1	80	Rhizosphere
contig_12	46,702	74	Uncultured marine virus isolate CBSM-242	FJ640348.1	0	83	Chesapeake Bay sediment
contig_13	46,245	60	Bacteriophage 11b	AJ842011.2	1	68	Arctic sea ice
contig_17	40,578	56	*Methylobacterium brachiatum* strain TX0642	CP033231.1	6	67	Automobile air-conditioning evaporator
contig_18	40,568	61	*Blastochloris* sp. GI	AP018907.1	0	72	Soda dam hot springs
contig_2[a]	70,520	92	Uncultured virus YBW_Contig_50752	KU756933.1	1	72	North Sea Surface Water Virome
contig_5	56,143	55	Uncultured virus SERC 372681	KU595468.1	2	73	Rhode River surface water
contig_6	55,961	75	*Polynucleobacter asymbioticus* strain MWH-RechtKol4	CP015017.1	1	71	freshwater
contig_7	55,939	77	Uncultured virus SERC Contig 695464	KU971113.1	0	76	Rhode River surface water

Notes.

[a] Contig also predicted as viral by VirSorter (*Roux et al., 2015*).

additional sequences (two predicted by virMine and two predicted by VirSorter) exhibited homology to genes/sequences annotated as phage.

Our comparison of virMine to VirSorter highlights the importance of manual inspection of results. In contrast to VirSorter and, e.g., VirFinder, virMine not only predicts viral sequences but also reports the blast results of these sequences. This aids in the manual inspection of the virMine predictions. It is important to note that our comparison here, however, is not entirely an equivalent assessment: VirSorter relies on a different sequence database than virMine. As described in *Roux et al. (2015)*, two reference databases are used by VirSorter. These databases have been updated to version 2 since the time of its publication, and details regarding this update are not readily available. In fact, the viral databases used by existing tools varies greatly. VirSorter and MARVEL restrict their viral database to phages, all phages and dsDNA phages from the *Caudovirales* order, respectively. However, virMine includes all viral sequences—phages as well as eukaryotic viruses. As shown in Fig. 4, a number of hits to eukaryotic viruses were identified within the Lake Michigan data sets. While VirusSeeker's database is not restricted to phage sequences, as it too contains eukaryotic viral sequences, it is a curated database (last updated August 2016). Currently, MetaPhinder's and MetaVir's databases are also out of date; both were last updated in 2017. virMine's database is entirely controlled by the user and can include all data currently available. Just as virMine allows the user to create their own custom databases, so too does FastViromeExplorer. FastViromeExplorer requires the user to format files for use. In contrast, virMine only necessitates a multi-fasta file which can easily be retrieved from publicly available databases like NCBI and IMG/VR or via user-specific queries of public sequence repositories.

CONCLUSIONS

As highlighted in the recent report of the International Committee on Taxonomy of Viruses (ICTV) Executive Committee, genomes identified from metagenomic data will vastly expand our catalog of viral diversity (*Simmonds et al., 2017*). Within just the past two years, there has been an explosive growth of the number of uncultivated viral genomes identified within metagenomic data (*Roux et al., 2018*). Our analysis of complex communities has uncovered numerous novel viral genomes. virMine is capable of identifying both prophages in contigs and viral sequences. In contrast to other tools that rely solely on viral sequence availability, virMine makes use of a far larger, more comprehensive data set—non-viral sequences. Furthermore, the entire process from raw sequence quality control through analysis is packaged into a single tool providing a "consensus" solution for viral genome discovery (*Dutilh et al., 2017*). Manual inspection of virMine results can thus lead to the identification of viral sequences resembling known viruses as well as novel viral strains. As exemplified here, virMine can be used to identify viruses in any niche and thus further our understanding of this vast reservoir of genetic diversity.

ACKNOWLEDGEMENTS

The authors would like to thank Ms. Ally Miley for conversations during the development of this tool and Dr. Jason Shapiro for his feedback on earlier versions of the manuscript.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the National Science Foundation (grant number 1149387) to Catherine Putonti. Andrea Garretto was supported by Loyola University Chicago's Carbon Research Fellowship and the CRA-W's CREU program. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Science Foundation: 1149387.
Loyola University Chicago's Carbon Research Fellowship.
CRA-W's CREU program.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Andrea Garretto performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft.
- Thomas Hatzopoulos performed the experiments, analyzed the data, approved the final draft.
- Catherine Putonti conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft.

Data Availability

The following information was supplied regarding data availability:

Data is available at GitHub: https://github.com/thatzopoulos/virMine.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.6695#supplemental-information.

REFERENCES

Abeles SR, Pride DT. 2014. Molecular bases and role of viruses in the human microbiome. *Journal of Molecular Biology* **426(23)**:3892–3906 DOI 10.1016/j.jmb.2014.07.002.

Adriaenssens EM, Kramer R, Van Goethem MW, Makhalanyane TP, Hogg I, Cowan DA. 2017. Environmental drivers of viral community composition in Antarctic soils identified by viromics. *Microbiome* **5**:Article 83 DOI 10.1186/s40168-017-0301-7.

Amgarten D, Braga LPP, Da Silva AM, Setubal JC. 2018. MARVEL, a tool for prediction of bacteriophage sequences in metagenomic bins. *Frontiers in Genetics* **9**:Article 304 DOI 10.3389/fgene.2018.00304.

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. 2008. The RAST server: rapid annotations using subsystems technology. *BMC Genomics* **9**:75 DOI 10.1186/1471-2164-9-75.

Aziz RK, Breitbart M, Edwards RA. 2010. Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Research* **38**:4207–4217 DOI 10.1093/nar/gkq140.

Aziz RK, Dwivedi B, Breitbart M, Edwards RA. 2011. Phage Eco-Locator: a web tool for visualization and analysis of phage genomes in metagenomic data sets. *BMC Bioinformatics* **12**:A9 DOI 10.1186/1471-2105-12-S7-A9.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* **19**:455–477 DOI 10.1089/cmb.2012.0021.

Barr JJ. 2017. A bacteriophages journey through the human body. *Immunological Reviews* **279**:106–122 DOI 10.1111/imr.12565.

Bellas CM, Anesio AM, Barker G. 2015. Analysis of virus genomes from glacial environments reveals novel virus groups with unusual host interactions. *Frontiers in Microbiology* **6**:656 DOI 10.3389/fmicb.2015.00656.

Brazelton WJ, Baross JA. 2009. Abundant transposases encoded by the metagenome of a hydrothermal chimney biofilm. *The ISME Journal* **3**:1420–1424 DOI 10.1038/ismej.2009.79.

Breitbart M, Bonnain C, Malki K, Sawaya NA. 2018. Phage puppet masters of the marine microbial realm. *Nature Microbiology* **3**:754–766 DOI 10.1038/s41564-018-0166-y.

Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F. 2002. Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences of the United States of America* **99**:14250–14255 DOI 10.1073/pnas.202488399.

Bruder K, Malki K, Cooper A, Sible E, Shapiro JW, Watkins SC, Putonti C. 2016. Freshwater metaviromics and bacteriophages: a current assessment of the state of the art in relation to bioinformatic challenges. *Evolutionary Bioinformatics Online* **12**:25–33 DOI 10.4137/EBO.S38549.

Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, Chaffron S, Cruaud C, De Vargas C, Gasol JM, Gorsky G, Gregory AC, Guidi L, Hingamp P, Iudicone D, Not F, Ogata H, Pesant S, Poulos BT, Schwenck SM, Speich S, Dimier C, Kandels-Lewis S, Picheral M, Searson S, Tara Oceans Coordinators, Bork P, Bowler C, Sunagawa S, Wincker P, Karsenti E, Sullivan MB. 2015. Ocean plankton. Patterns and ecological drivers of ocean viral communities. *Science* **348(6237)**:Article 1261498 DOI 10.1126/science.1261498.

Brum JR, Sullivan MB. 2015. Rising to the challenge: accelerated pace of discovery transforms marine virology. *Nature Reviews Microbiology* **13**:147–159 DOI 10.1038/nrmicro3404.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**:421 DOI 10.1186/1471-2105-10-421.

Chirico N, Vianelli A, Belshaw R. 2010. Why genes overlap in viruses. *Proceedings of the Royal Society B: Biological Sciences* **277**:3809–3817 DOI 10.1098/rspb.2010.1052.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, De Hoon MJL. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**:1422–1423 DOI 10.1093/bioinformatics/btp163.

Conceição Neto N, Zeller M, Lefrère H, De Bruyn P, Beller L, Deboutte W, Yinda CK, Lavigne R, Maes P, Van Ranst M, Heylen E, Matthijnssens J. 2015. Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. *Scientific Reports* **5**:16532 DOI 10.1038/srep16532.

Coutinho FH, Silveira CB, Gregoracci GB, Thompson CC, Edwards RA, Brussaard CPD, Dutilh BE, Thompson FL. 2017. Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. *Nature Communications* **8**:Article 15955 DOI 10.1038/ncomms15955.

Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. 1999. Improved microbial gene identification with GLIMMER. *Nucleic Acids Research* **27**:4636–4641 DOI 10.1093/nar/27.23.4636.

Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GGZ, Boling L, Barr JJ, Speth DR, Seguritan V, Aziz RK, Felts B, Dinsdale EA, Mokili JL, Edwards RA. 2014. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nature Communications* **5**:Article 4498 DOI 10.1038/ncomms5498.

Dutilh BE, Reyes A, Hall RJ, Whiteson KL. 2017. Editorial: virus discovery by metagenomics: the (Im)possibilities. *Frontiers in Microbiology* **8**:Article 1710 DOI 10.3389/fmicb.2017.01710.

Fierer N, Breitbart M, Nulton J, Salamon P, Lozupone C, Jones R, Robeson M, Edwards RA, Felts B, Rayhawk S, Knight R, Rohwer F, Jackson RB. 2007. Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Applied and Environmental Microbiology* **73**:7059–7066 DOI 10.1128/AEM.00358-07.

Galperin MY, Makarova KS, Wolf YI, Koonin EV. 2015. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Research* **43**:D261–D269 DOI 10.1093/nar/gku1223.

Garretto A, Thomas-White K, Wolfe AJ, Putonti C. 2018. Detecting viral genomes in the female urinary microbiome. *The Journal of General Virology* **99**:1141–1146 DOI 10.1099/jgv.0.001097.

Garza DR, Dutilh BE. 2015. From cultured to uncultured genome sequences: metagenomics and modeling microbial ecosystems. *Cellular and Molecular Life Sciences* **72**:4287–4308 DOI 10.1007/s00018-015-2004-1.

Ghai R, Mehrshad M, Mizuno CM, Rodriguez-Valera F. 2017. Metagenomic recovery of phage genomes of uncultured freshwater actinobacteria. *The ISME Journal* **11**:304–308 DOI 10.1038/ismej.2016.110.

Hannigan GD, Meisel JS, Tyldsley AS, Zheng Q, Hodkinson BP, SanMiguel AJ, Minot S, Bushman FD, Grice EA. 2015. The human skin double-stranded DNA virome: topographical and temporal diversity, genetic enrichment, and dynamic associations with the host microbiome. *mBio* **6**:e01578–01515 DOI 10.1128/mBio.01578-15.

Hatfull GF. 2008. Bacteriophage genomics. *Current Opinion in Microbiology* **11**:447–453 DOI 10.1016/j.mib.2008.09.004.

Hatzopoulos T, Watkins SC, Putonti C. 2016. PhagePhisher: a pipeline for the discovery of covert viral sequences in complex genomic datasets. *Microbial Genomics* **2**:e000053 DOI 10.1099/mgen.0.000053.

Hayes S, Mahony J, Nauta A, Van Sinderen D. 2017. Metagenomic approaches to assess bacteriophages in various environmental niches. *Viruses* **9**:127 DOI 10.3390/v9060127.

Ho T, Tzanetakis IE. 2014. Development of a virus detection and discovery pipeline using next generation sequencing. *Virology* **471–473**:54–60 DOI 10.1016/j.virol.2014.09.019.

Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM, Amundson R, Thomas BC, Banfield JF. 2016. A new view of the tree of life. *Nature Microbiology* **1**:Article 16048 DOI 10.1038/nmicrobiol.2016.48.

Hurwitz BL, Ponsero A, Thornton J, U'Ren JM. 2018. Phage hunters: computational strategies for finding phages in large-scale 'omics datasets. *Virus Research* **244**:110–115 DOI 10.1016/j.virusres.2017.10.019.

Hurwitz BL, Sullivan MB. 2013. The Pacific Ocean virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLOS ONE* **8(2)**:e57355 DOI 10.1371/journal.pone.0057355.

Inskeep WP, Jay ZJ, Herrgard MJ, Kozubal MA, Rusch DB, Tringe SG, Macur RE, De Jennings MR, Boyd ES, Spear JR, Roberto FF. 2013. Phylogenetic and functional analysis of metagenome sequence from high-temperature archaeal habitats demonstrate linkages between metabolic potential and geochemistry. *Frontiers in Microbiology* **4**:Article 95 DOI 10.3389/fmicb.2013.00095.

Jurtz VI, Villarroel J, Lund O, Voldby Larsen M, Nielsen M. 2016. MetaPhinder-Identifying bacteriophage sequences in metagenomic data sets. *PLOS ONE* **11(9)**:e0163111 DOI 10.1371/journal.pone.0163111.

Keen EC, Dantas G. 2018. Close encounters of three kinds: bacteriophages, commensal bacteria, and host immunity. *Trends in Microbiology* **26**:943–954 DOI 10.1016/j.tim.2018.05.009.

Labonté JM, Suttle CA. 2013. Previously unknown and highly divergent ssDNA viruses populate the oceans. *The ISME Journal* **7**:2169–2177 DOI 10.1038/ismej.2013.110.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**:357–359 DOI 10.1038/nmeth.1923.

Lewandowska DW, Zagordi O, Geissberger F-D, Kufner V, Schmutz S, Böni J, Metzner KJ, Trkola A, Huber M. 2017. Optimization and validation of sample preparation for metagenomic sequencing of viruses in clinical samples. *Microbiome* **5**:Article 94 DOI 10.1186/s40168-017-0317-z.

Li D, Luo R, Liu C-M, Leung C-M, Ting H-F, Sadakane K, Yamashita H, Lam T-W. 2016a. MEGAHIT v.10: a fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**:3–11 DOI 10.1016/j.ymeth.2016.02.020.

Li Y, Wang H, Nie K, Zhang C, Zhang Y, Wang J, Niu P, Ma X. 2016b. VIP: an integrated pipeline for metagenomics of virus identification and discovery. *Scientific Reports* **6**:23774 DOI 10.1038/srep23774.

López-Bueno A, Rastrojo A, Peiró R, Arenas M, Alcamí A. 2015. Ecological connectivity shapes quasispecies structure of RNA viruses in an Antarctic lake. *Molecular Ecology* **24**:4812–4825 DOI 10.1111/mec.13321.

López-Bueno A, Tamames J, Velázquez D, Moya A, Quesada A, Alcamí A. 2009. High diversity of the viral community from an Antarctic lake. *Science* **326(5954)**:858–861 DOI 10.1126/science.1179287.

McNair K, Aziz RK, Pusch GD, Overbeek R, Dutilh BE, Edwards R. 2018. Phage genome annotation using the RAST pipeline. *Methods in Molecular Biology* **1681**:231–238 DOI 10.1007/978-1-4939-7343-9_17.

Merrill BD, Ward AT, Grose JH, Hope S. 2016. Software-based analysis of bacteriophage genomes, physical ends, and packaging strategies. *BMC Genomics* **17**:679 DOI 10.1186/s12864-016-3018-2.

Miller-Ensminger T, Garretto A, Brenner J, Thomas-White K, Zambom A, Wolfe AJ, Putonti C. 2018. Bacteriophages of the urinary microbiome. *Journal of Bacteriology* **200**:e00738-17 DOI 10.1128/JB.00738-17.

Minot S, Bryson A, Chehoud C, Wu GD, Lewis JD, Bushman FD. 2013. Rapid evolution of the human gut virome. *Proceedings of the National Academy of Sciences of the United States of America* **110**:12450–12455 DOI 10.1073/pnas.1300833110.

Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, Wu GD, Lewis JD, Bushman FD. 2011. The human gut virome: inter-individual variation and dynamic response to diet. *Genome Research* **21**:1616–1625 DOI 10.1101/gr.122705.111.

Moustafa A, Li W, Singh H, Moncera KJ, Torralba MG, Yu Y, Manuel O, Biggs W, Venter JC, Nelson KE, Pieper R, Telenti A. 2018. Microbial metagenome of urinary tract infection. *Scientific Reports* **8**:4333 DOI 10.1038/s41598-018-22660-8.

Nayfach S, Pollard KS. 2016. Toward accurate and quantitative comparative metagenomics. *Cell* **166**:1103–1116 DOI 10.1016/j.cell.2016.08.007.

Nooij S, Schmitz D, Vennema H, Kroneman A, Koopmans MPG. 2018. Overview of virus metagenomic classification methods and their biological applications. *Frontiers in Microbiology* **9**:749 DOI 10.3389/fmicb.2018.00749.

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. *Genome Research* **27**:824–834 DOI 10.1101/gr.213959.116.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* **44**:D733–D745 DOI 10.1093/nar/gkv1189.

Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, Rubin E, Ivanova NN, Kyrpides NC. 2016. Uncovering earth's virome. *Nature* **536**:425–430 DOI 10.1038/nature19094.

Peng Y, Leung HCM, Yiu SM, Chin FYL. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**:1420–1428 DOI 10.1093/bioinformatics/bts174.

Pratama AA, Van Elsas JD. 2018. The neglected soil virome—potential role and impact. *Trends in Microbiology* **26**:649–662 DOI 10.1016/j.tim.2017.12.004.

Pride DT, Salzman J, Haynes M, Rohwer F, Davis-Long C, White RA, Loomer P, Armitage GC, Relman DA. 2012. Evidence of a robust resident bacteriophage population revealed through analysis of the human salivary virome. *The ISME Journal* **6**:915–926 DOI 10.1038/ismej.2011.169.

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto J-M, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Li Y, Zhang X, Li S, Qin N, Yang H, Wang J,

Brunak S, Doré J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, MetaHIT Consortium, Bork P, Ehrlich SD, Wang J. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**:59–65 DOI 10.1038/nature08821.

Rani A, Ranjan R, McGee HS, Metwally A, Hajjiri Z, Brennan DC, Finn PW, Perkins DL. 2016. A diverse virome in kidney transplant patients contains multiple viral subtypes with distinct polymorphisms. *Scientific Reports* **6**:33327 DOI 10.1038/srep33327.

Ren J, Ahlgren NA, Lu YY, Fuhrman JA, Sun F. 2017. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**:Article 69 DOI 10.1186/s40168-017-0283-5.

Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, Gordon JI. 2010. Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466**:334–338 DOI 10.1038/nature09199.

Richter DC, Ott F, Auch AF, Schmid R, Huson DH. 2008. MetaSim—a sequencing simulator for genomics and metagenomics. *PLOS ONE* **3(10)**:e3373 DOI 10.1371/journal.pone.0003373.

Rihtman B, Meaden S, Clokie MRJ, Koskella B, Millard AD. 2016. Assessing illumina technology for the high-throughput sequencing of bacteriophage genomes. *PeerJ* **4**:e2055 DOI 10.7717/peerj.2055.

Rosario K, Schenck RO, Harbeitner RC, Lawler SN, Breitbart M. 2015. Novel circular single-stranded DNA viruses identified in marine invertebrates reveal high sequence diversity and consistent predicted intrinsic disorder patterns within putative structural proteins. *Frontiers in Microbiology* **6**:Article 696 DOI 10.3389/fmicb.2015.00696.

Rose R, Constantinides B, Tapinos A, Robertson DL, Prosperi M. 2016. Challenges in the analysis of viral metagenomes. *Virus Evolution* **2**:Article vew022 DOI 10.1093/ve/vew022.

Roux S, Adriaenssens EM, Dutilh BE, Koonin EV, Kropinski AM, Krupovic M, Kuhn JH, Lavigne R, Brister JR, Varsani A, Amid C, Aziz RK, Bordenstein SR, Bork P, Breitbart M, Cochrane GR, Daly RA, Desnues C, Duhaime MB, Emerson JB, Enault F, Fuhrman JA, Hingamp P, Hugenholtz P, Hurwitz BL, Ivanova NN, Labonté JM, Lee K-B, Malmstrom RR, Martinez-Garcia M, Mizrachi IK, Ogata H, Páez-Espino D, Petit M-A, Putonti C, Rattei T, Reyes A, Rodriguez-Valera F, Rosario K, Schriml L, Schulz F, Steward GF, Sullivan MB, Sunagawa S, Suttle CA, Temperton B, Tringe SG, Thurber RV, Webster NS, Whiteson KL, Wilhelm SW, Wommack KE, Woyke T, Wrighton KC, Yilmaz P, Yoshida T, Young MJ, Yutin N, Allen LZ, Kyrpides NC, Eloe-Fadrosh EA. 2018. Minimum information about an uncultivated virus genome (MIUViG). *Nature Biotechnology* **37**:29–37 DOI 10.1038/nbt.4306.

Roux S, Enault F, Hurwitz BL, Sullivan MB. 2015. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**:e985 DOI 10.7717/peerj.985.

Roux S, Enault F, Robin A, Ravet V, Personnic S, Theil S, Colombet J, Sime-Ngando T, Debroas D. 2012. Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLOS ONE* **7(3)**:e33641 DOI 10.1371/journal.pone.0033641.

Roux S, Tournayre J, Mahul A, Debroas D, Enault F. 2014. Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinformatics* **15**:76 DOI 10.1186/1471-2105-15-76.

Sangwan N, Xia F, Gilbert JA. 2016. Recovering complete and draft population genomes from metagenome datasets. *Microbiome* **4**:Article 8 DOI 10.1186/s40168-016-0154-5.

Santiago-Rodriguez TM, Ly M, Bonilla N, Pride DT. 2015. The human urine virome in association with urinary tract infections. *Frontiers in Microbiology* **6**:Article 14 DOI 10.3389/fmicb.2015.00014.

Sharon I, Banfield JF. 2013. Microbiology. Genomes from metagenomics. *Science* **342(6162)**:1057–1058 DOI 10.1126/science.1247023.

Sible E, Cooper A, Malki K, Bruder K, Watkins SC, Fofanov Y, Putonti C. 2015. Survey of viral populations within Lake Michigan nearshore waters at four Chicago area beaches. *Data in Brief* **5**:9–12 DOI 10.1016/j.dib.2015.08.001.

Simmonds P, Adams MJ, Benkő M, Breitbart M, Brister JR, Carstens EB, Davison AJ, Delwart E, Gorbalenya AE, Harrach B, Hull R, King AMQ, Koonin EV, Krupovic M, Kuhn JH, Lefkowitz EJ, Nibert ML, Orton R, Roossinck MJ, Sabanadzovic S, Sullivan MB, Suttle CA, Tesh RB, Van der Vlugt RA, Varsani A, Zerbini FM. 2017. Virus taxonomy in the age of metagenomics: consensus statement. *Nature Reviews Microbiology* **15**:161–168 DOI 10.1038/nrmicro.2016.177.

Skvortsov T, De Leeuwe C, Quinn JP, McGrath JW, Allen CCR, McElarney Y, Watson C, Arkhipova K, Lavigne R, Kulakov LA. 2016. Metagenomic characterisation of the viral community of Lough Neagh, the largest freshwater lake in Ireland. *PLOS ONE* **11(2)**:e0150361 DOI 10.1371/journal.pone.0150361.

Smits SL, Bodewes R, Ruiz-Gonzalez A, Baumgärtner W, Koopmans MP, Osterhaus ADME, Schürch AC. 2014. Assembly of viral genomes from metagenomes. *Frontiers in Microbiology* **5**:Article 714 DOI 10.3389/fmicb.2014.00714.

Smits SL, Bodewes R, Ruiz-González A, Baumgärtner W, Koopmans MP, Osterhaus ADME, Schürch AC. 2015. Recovering full-length viral genomes from metagenomes. *Frontiers in Microbiology* **6**:Article 1069 DOI 10.3389/fmicb.2015.01069.

Stachler E, Kelty C, Sivaganesan M, Li X, Bibby K, Shanks OC. 2017. Quantitative CrAssphage PCR assays for human fecal pollution measurement. *Environmental Science & Technology* **51**:9146–9154 DOI 10.1021/acs.est.7b02703.

Thannesberger J, Hellinger H-J, Klymiuk I, Kastner M-T, Rieder FJJ, Schneider M, Fister S, Lion T, Kosulin K, Laengle J, Bergmann M, Rattei T, Steininger C. 2017. Viruses comprise an extensive pool of mobile genetic elements in eukaryote cell cultures and human clinical samples. *The FASEB Journal* **31**:1987–2000 DOI 10.1096/fj.201601168R.

Tithi SS, Aylward FO, Jensen RV, Zhang L. 2018. FastViromeExplorer: a pipeline for virus and phage identification and abundance profiling in metagenomics data. *PeerJ* **6**:e4227 DOI 10.7717/peerj.4227.

Vigil-Stenman T, Ininbergs K, Bergman B, Ekman M. 2017. High abundance and expression of transposases in bacteria from the Baltic Sea. *The ISME Journal* **11**:2611–2623 DOI 10.1038/ismej.2017.114.

Voorhies AA, Eisenlord SD, Marcus DN, Duhaime MB, Biddanda BA, Cavalcoli JD, Dick GJ. 2016. Ecological and genetic interactions between cyanobacteria and viruses in a low-oxygen mat community inferred through metagenomics and metatranscriptomics. *Environmental Microbiology* **18**:358–371 DOI 10.1111/1462-2920.12756.

Wan Y, Renner DW, Albert I, Szpara ML. 2015. VirAmp: a galaxy-based viral genome assembly pipeline. *GigaScience* **4**:Article 19 DOI 10.1186/s13742-015-0060-y.

Watkins SC, Kuehnle N, Ruggeri CA, Malki K, Bruder K, Elayyan J, Damisch K, Vahora N, O'Malley P, Ruggles-Sage B, Romer Z, Putonti C. 2016. Assessment of a metaviromic dataset generated from nearshore Lake Michigan. *Marine and Freshwater Research* **67**:Article 1700 DOI 10.1071/MF15172.

Watkins SC, Sible E, Putonti C. 2018. Pseudomonas PB1-like phages: whole genomes from metagenomes offer insight into an abundant group of bacteriophages. *Viruses* **10**:331 DOI 10.3390/v10060331.

Wommack KE, Bhavsar J, Polson SW, Chen J, Dumas M, Srinivasiah S, Furman M, Jamindar S, Nasko DJ. 2012. VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Standards in Genomic Sciences* **6**:427–439 DOI 10.4056/sigs.2945050.

Yamashita A, Sekizuka T, Kuroda M. 2016. VirusTAP: viral genome-targeted assembly pipeline. *Frontiers in Microbiology* **7**:Article 32 DOI 10.3389/fmicb.2016.00032.

Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W, Jaroszewski L, Cieplak P, Miller CS, Li H, Mashiyama ST, Joachimiak MP, Van Belle C, Chandonia J-M, Soergel DA, Zhai Y, Natarajan K, Lee S, Raphael BJ, Bafna V, Friedman R, Brenner SE, Godzik A, Eisenberg D, Dixon JE, Taylor SS, Strausberg RL, Frazier M, Venter JC. 2007. The Sorcerer II global ocean sampling expedition: expanding the universe of protein families. *PLOS Biology* **5**(**3**):e16 DOI 10.1371/journal.pbio.0050016.

Zablocki O, Van Zyl L, Adriaenssens EM, Rubagotti E, Tuffin M, Cary SC, Cowan D. 2014. High-level diversity of tailed phages, eukaryote-associated viruses, and virophage-like elements in the metaviromes of antarctic soils. *Applied and Environmental Microbiology* **80**:6888–6897 DOI 10.1128/AEM.01525-14.

Zeigler Allen L, McCrow JP, Ininbergs K, Dupont CL, Badger JH, Hoffman JM, Ekman M, Allen AE, Bergman B, Venter JC. 2017. The Baltic Sea virome: diversity and transcriptional activity of DNA and RNA viruses. *Systems* **2**:e00125–16 DOI 10.1128/mSystems.00125-16.

Zhang W, Zhou J, Liu T, Yu Y, Pan Y, Yan S, Wang Y. 2015. Four novel algal virus genomes discovered from Yellowstone Lake metagenomes. *Scientific Reports* **5**:15131 DOI 10.1038/srep15131.

Zhao G, Wu G, Lim ES, Droit L, Krishnamurthy S, Barouch DH, Virgin HW, Wang D. 2017. VirusSeeker, a computational pipeline for virus discovery and virome composition analysis. *Virology* **503**:21–30 DOI 10.1016/j.virol.2017.01.005.