Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data

David Paez-Espino, Georgios A Pavlopoulos, Natalia N Ivanova & Nikos C Kyrpides

Joint Genome Institute, Department of Energy, Walnut Creek, California, USA. Correspondence should be addressed to D.P.-E. (adpaezespino@lbl.gov).

Published online 27 July 2017; doi:10.1038/nprot.2017.063

The analysis of large microbiome data sets holds great promise for the delineation of the biological and metabolic functioning of living organisms and their role in the environment. In the midst of this genomic puzzle, viruses, especially those that infect microbial communities, represent a major reservoir of genetic diversity with great impact on biogeochemical cycles and organismal health. Overcoming the limitations associated with virus detection directly from microbiomes can provide key insights into how ecosystem dynamics are modulated. Here, we present a computational protocol for accurate detection and grouping of viral sequences from microbiome samples. Our approach relies on an expanded and curated set of viral protein families used as bait to identify viral sequences directly from metagenomic assemblies. This protocol describes how to use the viral protein families catalog (~7 h) and recommended filters for the detection of viral contigs in metagenomic samples (~6 h), and it describes the specific parameters for a nucleotide-sequence-identity-based method of organizing the viral sequences into quasi-species taxonomic-level groups (~10 min).

INTRODUCTION

The improvement of high-throughput sequencing technologies has led to a marked increase in the number and size of microbiome studies during the past few years[1,2]. Among these studies, only a small fraction (3.8%) correspond to viral metagenome sequence libraries generated by classical concentration techniques of viral particles from various types of samples, also referred as environmental viromes[3–7]. The generation of these virome samples requires expertise in complex experimental protocols[8,9] and is usually performed by virologists. Our current understanding of the dynamics, diversity, and distribution of viruses[4,5,10–13], and their impact on biogeochemical cycles in nature, has primarily been established by the study of these targeted samples. In contrast to these, the vast majority of microbiome data sets are generated through untargeted approaches without viral particle enrichment.

Different computational methods have been developed to identify viral sequences. Of these methods, the vast majority are designed to identify viral sequences directly from microbial isolated genomes (prophages)[14–17]. These prophage predictors primarily rely on sequence similarity between microbial genome regions and isolated viral sequences, and their use to detect free-living lytic viruses in uncharacterized samples is very limited.

More recently, a tool named VirSorter was developed to detect prophage regions, as well as other viral fragments in larger-scale microbial fragmented genomic data sets[18]. This approach uses viral protein families from viruses that infect microbes (archaea and bacteria) in the RefSeq database (similar to the recent prokaryotic Virus Orthologous Groups (pVOGs)[19]) complemented by viral proteins from three specific microbiomes (marine, freshwater, and human-associated samples). However, it is known from the analysis of the viral distribution across diverse ecosystems that the vast majority of viruses show a strong habitat-type specificity[20], and therefore targeting highly divergent viral sequences from less common habitat types (or even different sub-habitats within the most sampled microbiomes) would require a broader set of specific viral protein families.

Uniqueness of this method

Here we present a protocol that is specifically aimed at the detection of lytic viral sequences directly from metagenomic samples, followed by grouping of the identified viruses[20]. The uniqueness of this method relies on the complementation of the viral isolated protein families (that include both prokaryotic and eukaryotic viruses) with a set of manually curated families from a plethora of distinct habitat types to generate a global set of 25,281 models. As a comparison, the pVOGs database contains 9,518 orthologous groups from nearly 3,000 microbial isolated viral genomes[19] and VirSorter uses 15,673 clusters from the viral RefSeq database (infecting only microbes) and known phages from marine, freshwater, and human samples[18].

Our method has been used (i) to computationally detect 264,413 viral contigs from >6,000 ecologically diverse metagenomic samples according to metadata from the Genomes OnLine Database[21,22], increasing the sensitivity of detection of short fragments of viruses from uncharacterized samples, and (ii) to group approximately half of them into ~39,701 genetically distinct quasi-species clusters. All these sequences have been deposited into the IMG/VR database[23]. This viral sequence repository will be used as a reference for this protocol for the viral grouping step and will be used to hint at the presence of viruses with low abundance in a target metagenome.

This protocol (**Fig. 1**) includes a description of (i) how to use the viral protein families for accurate detection of viral sequences directly from any metagenomic sample, (ii) the optimal parameters for grouping them into quasi-species, and (iii) the identification of viruses with low abundance from metagenomic data sets. The total running time for this protocol can vary depending on the size and number of samples that must be analyzed. With a single Intel Xeon 2.50GHz E5-2670 CPU, it takes ~13 h to analyze a sample containing 164 Mb of assembled sequences >5 kb plus 2.88 Gb of unassembled reads.

Benchmarking of computational approaches for virus detection

We assessed the accuracy of our virus detection pipeline and compared it with VirSorter[18]. This benchmarking is detailed in the supplementary information of ref. 20.

Briefly, we generated a synthetic metagenome, composed of 32 finished and taxonomically diverse bacterial genomes (that contained a total of 132 prophage sequences), 3 archaea, and 5

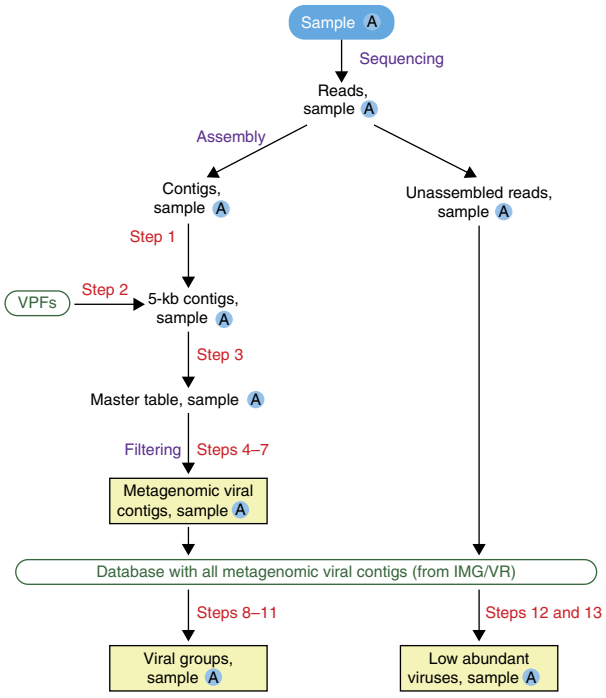

Figure 1 | Overview of the computational workflow. General pipeline of the protocol, showing the different steps required for the detection of abundant and low-abundant metagenomic viral contigs, as well as their classification into viral groups. Viral protein family models (VPFs) and all metagenomic viral contigs from IMG/VR are available through the aforementioned FTP site.

viruses, comprising a total of 88 replicons. After a genome fragmentation step, the resulting synthetic metagenome was dominated by bacterial and archaeal chromosomal fragments with an admixture of a relatively small number of plasmid and viral sequences. This composition represents a typical metagenome data set generated by an untargeted approach rather than by a targeted virome sequencing approach. The synthetic metagenome was submitted to a CyVerse[24] implementation of VirSorter and also processed by our virus discovery approach. Overall, our viral-protein-families-based pipeline has higher specificity and lower per-contig sensitivity than VirSorter (a 99.6% precision with a 37.5% recall rate, as compared with a 94.5% and a 65%, respectively), but comparable per-virus sensitivity (84.5% recall of our method, as compared with an 86.9% recall rate in VirSorter). As the goal of this protocol is the discovery of viral sequences in microbiome data sets dominated by host sequences, the approach with higher specificity is preferable.

For the final comparison, we ran a CyVerse implementation of VirSorter on the viral data set of 133,491 metagenomic viral contigs detected by our pipeline published in ref. 23 using the 'virome decontamination' option. 68.7% of the contigs were identified as viral by VirSorter, of which 98% were classified as virus (categories 1 and 2) and 2% were classified as putative prophages (categories 4 and 5). The examination of the 41,528 sequences missed by VirSorter revealed that 68% of them (28,412 sequences) had at least 50% of their genes covered by our manually curated viral protein family models, confirming their viral origin. Another 28% of the viral sequences missed by VirSorter had between 25 and 49% of their genes covered by our models. In the remaining 4% of the putative viral sequences not captured by VirSorter, the percentage of coverage by our viral models ranged between 7 and 24%, with the vast majority of the remaining genes being unknown or hypothetical. However, even for those sequences, we detected hallmark viral genes and the absence of plasmid or microbial signature gene sequences for most of them.

We conclude that our pipeline and VirSorter partially overlap in their predictions, with our pipeline specifically targeting highly divergent viral sequences and VirSorter having higher sensitivity in detection of short fragments of viruses similar to those in reference databases. Both tools would probably benefit from an expanded set of viral proteins, which can be used to improve the sensitivity of the viral protein families pipeline and the specificity of VirSorter by allowing more precise identification of prophage boundaries.

Diversity of potential applications and complementary tools

Viruses are considered to be the most abundant biological entities across all habitats, representing a great reservoir of genetic diversity[25], and ultimately affecting all biogeochemical cycles and ecosystem dynamics. Efforts to improve the detection of viral sequences from microbiomes by untargeted metagenome sequencing represent a powerful approach to filling knowledge gaps and addressing fundamental questions of viral ecology.

This protocol highlights the value of an untargeted *de novo* metagenomic analysis to improve viral sequence discovery and the grouping with known metagenomic viral sequences to obtain more information about viral distribution across habitats and hosts. Undoubtedly, this protocol can benefit from the use of complementary tools. Especially important are those computational approaches that are aimed at detecting viral–host relationships[26] or tools specifically designed for such predictions (e.g., HostPhinder[27]). Similarly, prophage detection tools (e.g., PHASTER[17], PhiSpy[14], and VirSorter[18]) could complement this protocol in the detection of the viral component from a metagenomic sample. Detection of viral sequences and their respective host(s) will open opportunities for identification of novel enzymes or regulatory sequences, with biotechnological applications[20], as well as potential phage therapy treatments[28,29].

Limitations of the method

The protocol presented here is aimed at the detection of free viruses from metagenomic samples, avoiding, to the extent possible, the presence of prophage sequences. To increase the sensitivity of this method, we exclusively consider assembled sequences of >5,000 bp (5 kb). This considerably reduces the number of ssDNA viruses that can be detected, as the majority of the isolated ssDNA viruses contained in the databases are shorter than this cutoff. Therefore, it is suited primarily to the detection of dsDNA viruses infecting both eukaryote and prokaryote organisms. In addition, as detailed in the Benchmarking section, the overall recall rate of this protocol is estimated at 37.5%.

Experimental design

Generation of viral protein families. To generate the viral protein families upon which the pipeline is based, we went through the following 12 stages:

Stage 1. Collection of protein-coding genes from isolated dsDNA viral genomes and retroviruses from the NCBI server. At the time of the study[20], we collected 167,042 genes from 2,353

isolated viruses and retroviruses. This step can be updated by adding any novel viral information from the databases.

Stage 2. De-replication of protein-coding genes using 70% identity in USEARCH[30]. We obtained 98 k protein sequences.

Stage 3. Clustering into groups using the Markov cluster algorithm, a fast and scalable unsupervised cluster algorithm for graphs that is based on simulation of stochastic flow in graphs[31]. We obtained 15.9 k groups that included 83.5 k protein sequences.

Stage 4. Alignment of proteins within clusters using MAFFT[32].

Stage 5. Creation of a set of viral protein families using hmmbuild[33]. We created a set of 14,296 viral protein families.

Stage 6. Manual curation of the viral families with high representation in prokaryotic genomes, using functional annotations of proteins and gene neighborhoods as guides for removal of protein families common in plasmids and bacterial/archaeal chromosomes.

Stage 7. Comparison of viral protein families against all metagenomic contigs longer than 5 kb in the IMG/M system[34]. At the time of the study, 5.1 million metagenomic contigs were longer than 5 kb.

Stage 8. Collection of metagenomic contigs containing five or more viral protein families and selection of those >50 kb. We collected 62 k metagenomic contigs with five or more viral protein families; this number was reduced to 9 k putative viral contigs after removing contigs of <50 kb.

Stage 9. Exclusion of contigs for which >10 and >25% of their predicted genes hit models from the Kegg Orthology (KO) database (http://www.genome.jp/kegg/ko.html) terms and protein families (Pfams), respectively. We reduced the number of putative viral contigs to 1,589.

Stage 10. Complementation of long (>50 kb) viral metagenomic contigs with either metagenomic contigs longer than 20 kb (binned with viruses) or those containing a viral RNA polymerase gene (not captured using the previous filter of bearing five or more viral protein families). We complemented the 1,589 viral contigs for these two data sets with 66 and 188 extra sequences, respectively, identified using the parameters described in **Box 1**, to generate a final set of 1,843 metagenomic viral contigs.

Stage 11. Merging of the reference isolated viral protein data set with the identified metagenomic viral protein data set. We merged the 167,042 proteins derived from isolated viruses (in stage 6) with the 191,000 viral proteins from the 1,843 identified metagenomic viral contigs from Stage 10.

Stage 12. Repetition of stages 2–5 (de-replication, clustering, alignment, and creation of viral protein families). We obtained a final set of 25,281 viral protein families to use for metagenomic viral contig detection.

Assignment of metagenomic sequences as viruses (Steps 4–7). The 25,281 viral protein families were used to screen all DNA metagenomic contigs longer than 5 kb for the identification of contigs passing any of the following three filters:

Filter 1. Metagenomic contigs that had at least the following:

5 hits to viral protein families; AND

Total number of genes covered with KO terms on the contig ≤20%; AND

Box 1 | Identification of metagenomic viral contigs for a training set. A training set is a primary set of data used to discover predictive relationships. A training set can be created via manual curation, binning, and DNA-dependent RNA polymerase alignment.

Additional sequences were identified to complement the set of the viral metagenomic contigs in Stage 10 of the process used to identify viral protein families, using two approaches:

1. Kmer-based metagenome binning of samples containing a high number of candidate viral sequences.
   - We binned six metagenomic samples with the highest number of candidate viral sequences (not satisfying the high-confidence threshold of the number of hits to protein models) using emergent self-organizing maps by Ultsch, as described previously[44,45].
   - We manually checked the contig sets outside the bins corresponding to cellular organisms and identified 66 long putative novel metagenomic viral contigs from diverse habitat types (freshwater, wastewater, thermal vents, and marine), which were added to the training set.
2. Identification of contigs containing viral RNA polymerase genes.
   - We collected 2,551 representative sequences of the genes encoding the three major subunits (α, β, and β′) of the *RNAp* gene from bacteria, as well as their eukaryotic and archaeal counterparts from the IMG system[34].
   - We extracted the domains of these genes using *Pfam* models and aligned with MAFFT[32].
   - We manually inspected alignments and built HMM models using hmmbuild[33].
   - These models were used to scan all the metagenomic contigs longer than 5 kb. We identified 39,109 contigs with matches to at least one core *RNAp* subunit.
   - We filtered out all short matches and de-replicated the sequences. This resulted in the identification of 7,437 metagenomic contigs, which were then added to the 2,551 reference isolates.
   - A phylogenetic tree with all 9,309 *RNAp* sequences was generated using FastTree[46] with default parameters. The tree was visualized using Dendroscope[47].
   - We identified (i) an *RNAp* branch corresponding to large eukaryotic DNA viruses and (ii) another set of metagenomic *RNAp* sequences branching separately from cellular references (comprising phage *RNAp* with domain composition similar to that of bacterial enzyme).
   - A total of 188 contigs longer than 20 kb containing viral and phage *RNAp* sequences from both viral *RNAp* branches described previously were added to the training set.

Total number of genes covered with Pfams ≤40%; AND
Total number of genes covered with viral protein families ≥10%.

Filter 2. Metagenomic contigs that had the following:
5 hits to viral protein families; AND
Number of viral protein families ≥number of Pfams.

Filter 3. Metagenomic contigs with the following:
5 hits to viral protein families; AND
Number of viral protein families ≥60% of the total number of genes.

Viral genome clustering, designation of viral groups, and validation of the method (Steps 8–11). We developed a sequence-based classification framework for systematically linking closely related viral genomes on the basis of their overall nucleotide similarity. This framework relies on both average nucleotide identity and total alignment fraction for pairwise comparisons of viral sequences, enabling a scalable and natural grouping of related isolated viral genomes and metagenomic viral contigs. This method has been recently applied for grouping 264,413 metagenomic viral contigs combined with 3,908 isolated reference DNA viruses, generating a total of 39,701 viral clusters (ranging from 2 to 349 members per group) and 122,665 singletons[23].

We used the BLASTn program in the Blast+ package[35] to find hits of all the viral sequences to themselves with an e-value cutoff of 1×10^{-50}, at least 90% identity across ≥75% of the shortest sequence length, and at least one hit over 1 kb in length. This filtering of BLAST results excluded matches against short highly conserved fragments of viral sequences, such as tRNAs, and other spurious hits.

As a validation of our clustering method, we observed that 97% of the isolated viral genomes (4,890 of the 5,007 viral groups or singletons) with a taxonomic assignment according to the International Committee on Taxonomy of Viruses (ICTV) clustered in agreement with the ICTV-designated species. All the remaining 3% of isolated viral genomes clustered at the genus level.

These analyses show that our viral groups are taxonomically relevant and provide a useful method for organizing distinct viral types.

Detection of viruses with low abundance (Steps 12 and 13). To detect the presence of any of the previously identified metagenomic viral contigs from ref. 20 in lower abundances within the examined sample(s), we recommend the following approach. By applying it, the viral analysis can be expanded to include the identification of known viral sequences from raw sequence reads.

We used the BLASTn program from the Blast+ package[35] to find hits to the 125,842 predicted viral sequences (file mVGs_sequences_v2.fna) that are deposited in the FTP site (http://portal.nersc.gov/dna/microbial/prokpubs/EarthVirome_DP/) with an e-value cutoff of 1×10^{-5} and at least 90% identity, as well as hits from the sample covering at least 10% of the length of the known viral contig. This filtering of BLAST results excluded matches to short highly conserved fragments of viral sequences, such as tRNAs, and other spurious hits[20]. Our filtering criteria were optimized for the type of metagenome data sets available to us, and are substantially more stringent than those used in some previous studies for similar data (e.g., 95% identity over 75-nt alignment used in ref. 36 or tBLASTx with e-value of 1×10^{-5}, recommended in ref. 37). However, our criteria were less stringent than the 75% coverage used in the analysis of Tara Oceans Viromes[10], which relied on viral enrichment to increase viral sequence coverage.

The above cutoff suggested by this protocol (90% identity over 10% of the viral sequence length) to identify viruses with low abundance could be customized by the user (as described in Step 13).

As an alternative to using BLASTn, raw reads can be aligned to the database more quickly using specific approaches (e.g., BBMap (http://www.sourceforge.net/projects/bbmap/), bwa (ref. 38), or bowtie2 (ref. 39)) to obtain comparable results.

MATERIALS

EQUIPMENT

Equipment and software requirements to run the protocol

- A computer with an average ~4 GB RAM with a Unix terminal and access to the Internet (e.g., a Macintosh with a 2.7-GHz processor, an Intel Core i7 and 16 GB of RAM). Extensive computations (e.g., very large microbiome data sets or detection of low-abundance viruses using Blast) could be optimized and parallelized using multiple CPUs ▲ **CRITICAL** Java JRE and Perl must be supported. Most scripts are written in Java and Perl. Protein family (Pfam) assignments are required for gene annotation, and licensed KO term annotations are highly recommended.
- Blast+[35] and hmmer[40] software packages to perform the corresponding query searches. These tools can be freely downloaded from the https://blast.ncbi.nlm.nih.gov/ and http://www.hmmer.org/ servers, respectively

Sequences

- Nucleotide and amino acid .fasta sequences of metagenomic assembled contigs of length >5 kb, prepared as described in the Equipment Setup section
- The 25,281 viral protein families generated in Paez-Espino *et al.*[20]. The complete list can be accessed from the FTP site (http://portal.nersc.gov/dna/microbial/prokpubs/EarthVirome_DP/final_list.hmms) or from the IMG/VR database (https://img.jgi.doe.gov/vr/)[23]

Example data sets

- To apply our protocol (**Fig. 1**) and detect 640 dominant metagenomic viral contigs plus 8,436 putative viruses with low abundance, we processed a publicly available metagenomic sample freely accessible through the IMG/M system (IMG Genome ID: 3300001348) (**Fig. 2**) and deposited in the FTP site (http://portal.nersc.gov/dna/microbial/prokpubs/EarthVirome_DP/Nature_Protocols)

EQUIPMENT SETUP

Preparation of nucleotide and amino acid .fasta sequences of metagenomic assembled contigs of length >5 kb From the nucleotide raw reads, trim, decontaminate, and assemble. We recommend the use of bbduck.sh and bbnorm.sh from the bbtools package (detailed information in http://jgi.doe.gov/data-and-tools/bbtools/) before assembly. We extensively used SPAdes[41] or MEGAHIT[42] as assemblers. For longer contigs, we recommend the use of SPAdes. For gene annotations, we highly recommend that users deposit their assembled metagenomic samples in the IMG/M system[34] to obtain, among other valuable metadata, a full gene annotation of the metagenomic contigs (including genes with Pfams and KO terms). All data can then be accessed from *IMG/M Genome Portal* through the *Download Data* button in the IMG/M landing page of a metagenome (**Fig. 2**). Alternatively, assembled contigs can be annotated using the open-source gene finder Prodigal[43] (http://prodigal.ornl.gov/) and protein families annotation using the *hmmsearch* command against the models in the latest release of the Pfam database with model-specific trusted cutoffs.

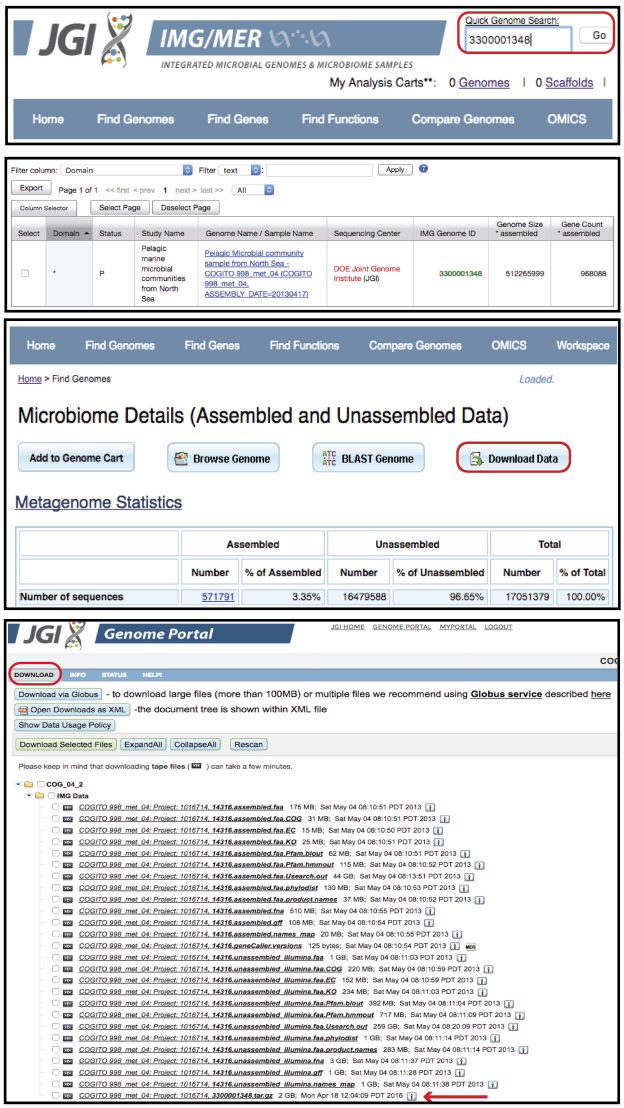

Figure 2 | Metagenome sample used as an example. Details of sample 3300001348 from the IMG/M system. (**a**) Access to the sample via *Quick Genome Search* (red box). (**b**) After clicking *Go*, data for the sample from the 'Pelagic marine microbial communities from North Sea' will appear. (**c**) Landing page of the example sample, summarizing the statistics for the number of assembled and unassembled sequences that it contains. From the *Download Data* button (red box), the Joint Genome Institute (JGI) *Genome Portal* can be accessed. (**d**) Download page (red box) of the JGI *Genome Portal*, from which all data and annotations of the given example can be retrieved. The red arrow points to the compressed *.tar.gz* file containing all files needed (protein and nucleotide assembled and unassembled data files, and annotated Pfam and KO files for the assembled contigs).

PROCEDURE

Preparation of master table ● TIMING ~8 h

1| Filter out all assembled sequences shorter than 5 kb, using option A if the data are obtained from the IMG/M system (fully annotated) or option B if the assembled data are in .fasta format:

(A) Filtering of assembled data from the IMG/M system (fully annotated):

(i) Access the metagenomic sample (e.g., '3300001348', if using the example provided) via the *Quick Genome Search* function from IMG/M (https://img.jgi.doe.gov/mer/).

(ii) From the *Scaffold Search (assembled)* option, select sequences larger than 5,000 (also provide a maximum number of base pairs for the sequences, e.g., 1,000,000).

(iii) For the given example, the query fetches 13,594 sequences. Select the sequences and add to the *Scaffold cart*, and export as a .fasta nucleic acid file.

(B) Filtering of assembled data in .fasta format

(i) Run the following command:

```
awk 'BEGIN{RS=">"{NR>1{sub("\n","\t"); gsub("\n",""); print RS$0}'
<YOUR_assembled_nt_FASTA_file.fna>| awk '{if(length($2)>=5000) print' | perl -p -e
's/\t/\n/g' | fold -w 60 >OUTPUT_assembled_nt_FASTA_file_5kb.fna
```

This will generate your assembled FASTA file in the following format: `YOUR_assembled_nt_FASTA_file.fna` (*3300001348_assembled.fna* file, for the given example).

The command `OUTPUT_assembled_nt_FASTA_file_5kb.fna` generates the results, which are stored in *3300001348_assembled_5kb.fna* in the FTP site.

▲ **CRITICAL STEP** Note that headers of the .fasta file must be 60 characters maximum. Option B provides the same output as option A, i.e., 13,594 sequences larger than 5 kb.

? TROUBLESHOOTING

2| Calculate viral protein family models against genes from contigs >5 kb using *hmmsearch* by running the following command:

```
hmmsearch --cpu 0 -E 1.0e-05 --tblout <OUTPUT_File> <VIR_HMM_File> <YOUR_Input_Pro-
tein_FASTA_File>,
```

where `OUTPUT_File` is the name of the output file. In the example in the FTP site, it is *3300001348_hits_to_vHMMs.hmmout*.

`VIR_HMM_File` contains 25,281 viral protein families (*hmm models*) generated in ref. 20. Users can access and retrieve these viral protein families from the FTP site: http://portal.nersc.gov/dna/microbial/prokpubs/EarthVirome_DP/final_list.hmms.

`YOUR_Input_Protein_FASTA_File` is the amino acid .fasta file of the genes contained in sequences larger than 5 kb (file *3300001348_genes_in_scafs_5kb.faa*). The header of this file contains the identifiers of the scaffolds and genes (*Scaffold_ID* and *Gene_ID*), separated by a pipe symbol ('|').

3| Generate a master table that includes all the required information for the metagenomic contigs larger than 5k bp with at least one hit to the viral protein families (*Master_Table_3300001348.txt* file).

This includes identifier (*Scaffold_ID*), number of hits to viral protein families (*hits_to_VPFs*), number of genes per scaffold (*#_of_genes*), and the percentage of the total genes covered by viral protein families (*%covered_VPFs*), number of genes that contain at least one Pfam (*genes_with_pfams*), along with the percentage they represent (*%genesPfams*), as well as the number of genes that contain a KO term (*genes_with_KO*) and the percentage that they represent (*%genesKO*) (**Supplementary Table 1**).
▲ **CRITICAL STEP** The information about gene assignments to Pfams (required) and KO terms (optional) is automatically generated by IMG/M (files *3300001348.a.pfam.txt* and *3300001348.a.ko.txt*) and is accessible via the FTP site. Alternatively, users can use their own annotations.

Application of filters to detect viral contigs from metagenomic data ● TIMING 10–30 min

4| Apply Filter 1 (scaffolds with at least five hits to viral protein families; the number of genes with KO term assignments is <20% (optional); Pfam-assigned genes <40%; and total number of genes with viral protein families hits >10%) by running the following command:

```
cat Master_Table_3300001348.txt | awk '$2 >= 5' | awk '$8 <= 20' | awk '$6 <= 40' |
awk '$4 >= 10' | cut -f 1 > Filter1.out
```

The file *Filter1.out* contains 569 metagenomic viral contigs.

5| Apply Filter 2 (scaffolds with at least five hits to viral protein families, and the number of genes with hits to viral protein families is > the number of genes with hits to Pfams) by running the following command:

```
cat Master_Table_3300001348.txt | awk '$2 >= 5' | awk '$2 >= $5' | cut -f
1 > Filter2.out
```

The file *Filter2.out* contains 576 metagenomic viral contigs.

6| Apply Filter 3 (scaffolds with at least five hits to viral protein families, and the number of genes with hits to viral protein families is at least 60%) by running the following command:

```
cat Master_Table_3300001348.txt | awk '$2 >= 5' | awk '$4 >= 60' | cut -f
1 > Filter3.out
```

The file *Filter3.out* contains 250 metagenomic viral contigs.

7| Obtain the total nonredundant list of metagenomic viral contigs (*Viral_contigs_3300001348.txt* file) from the sample 3300001348 by running the following command:

```
cat Filter1.out Filter2.out Filter3.out | sort | uniq > Viral_contigs_3300001348.txt
```

The *Viral_contigs_3300001348.txt* file consists of 640 viral sequences.

? TROUBLESHOOTING

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Viral genome clustering: grouping of viruses detected in public metagenomes with the viral contigs from the metagenome of interest ● TIMING 5–10 min

8| Run the following Blast command:

```
blastn -query <your_viral_contigs.fna> -db <reference_database> -outfmt '6 std
qlen slen' -out <your_viruses_vs_mVCs.blout> -evalue 1.0e-50 -perc_identity 80
-num_threads (optional)
```

`your_viral_contigs.fna` is the file *Viral_contigs_3300001348.fna* (a nucleotide .fasta format containing the 640 viral sequences larger than 5 kb) in our example.

Reference_database:
mVCs_PaezEspino_Nature.fna, located at http://portal.nersc.gov/dna/microbial/prokpubs/EarthVirome_DP/Nature_Protocols/reference_metagenomic_virus_database/ (Blast database containing 125,842 metagenomic viral contigs[20]).

`your_viruses_vs_mVCs.blout` is the file (*Viral_contigs_3300001348_vs_mVCs.blout*) containing 11,898 hits against the metagenomic viral contigs database.

9| Remove self-hits from the Blast output by running the following command:

```
cat <Viral_contigs_3300001348_vs_mVCs.blout>|awk '$1 != $2' > <your_viruses_vs_mVCs_
noSelfHits.blout>
```

`your_viruses_vs_mVCs_noSelfHits.blout` is the file containing 10,743 non-self hits (*Viral_contigs_3300001348_vs_mVCs_noSelfHits.blout*).

10| Parse the output using specific cutoffs (similarity ≥90%; covered length ≥75%; covered length requires at least one contig of >1,000-bp length) by running the following command:

```
java Parse_BLAST <your_viruses_vs_mVCs_noSelfHits.blout> > <your_viruses_vs_mVCs_no-
SelfHits_parsed.blout>
```

`your_viruses_vs_mVCs_noSelfHits_parsed.blout` is the file containing 493 pairwise hits between the metagenome of study and all metagenomic viral contigs from all other samples (*Viral_contigs_3300001348_vs_mVCs_noSelfHits_parsed.blout* file).

The script `Parse_BLAST` is located in the *script1* folder and is compiled using Java SE Runtime Environment build 1.7.0_51-b13. For simplicity, after compilation, users are advised to run the executable directly from the folder in which the *.class* files are located.

11| Carry out single linkage clustering by running the following command:

```
perl SLC.pl <your_viruses_vs_mVCs_noSelfHits_parsed.blout> <viral_groups.slc>
```

`viral_groups.slc` is the file *3300001348_viral_groups.slc*, which contains 246 viral groups composed of 268 metagenomic viral contigs from the 3300001348 marine sample, as well as 457 metagenomic viral contigs from 32 other different metagenomes according to metadata from the IMG/M system with taxon identifiers (e.g., 3300000101, 3300000115, 3300000116, 3300000128, 3300000130, 3300000188, 3300000224, 3300000265, 3300001351…). These 32 metagenomic samples all correspond to aquatic marine habitats (from neritic to oceanic marine zones), in agreement with the studied sample. The number of members per viral group ranged from 2 to 15. The remaining 372 viral sequences from the sample of study (i.e., 3300001348) remained as singletons.

The script `SLC.pl` is located in the *script2* folder.

Hinting at the presence of viruses with low abundance ● TIMING 4h 20 min

12| Run the following Blast command:

```
blastn -task megablast -query <Your_unassembled_file.fasta> -db <reference_metagenomic_virus_database>-outfmt '6 std qlen slen' -out <Your_output.blout>-evalue 1.0e-05 -perc_identity 90 -num_threads (optional)
```

The `Your_unassembled_file.fasta` file is *3300001348_u.fna*.

The `reference_metagenomic_virus_database` is *mVCs_PaezEspino_Nature.fna*.

The `Your_output.blout` file is *3300001348_u_vs_mVCs.blout*.

? TROUBLESHOOTING

13| Parse Blast output (see the *script3* folder) by running the following command:

```
java Coverage_VIRUSES_10Percent <Your_output.blout> > output_unassembled_vs_mVCs
```

Results will be automatically saved in the same folder as the original input file. The output file will have the same name as the input file, with the appended extension *.10percent.txt*. We compiled all current executables using Java SE Runtime Environment build 1.7.0_51-b13. For simplicity, we advise users to run the file from the same folder in which the *.class* files are located.

The `output_unassembled_vs_mVCs` file is *3300001348_u_vs_mVCs.blout.10percent.txt* in the example.

We detected 12,963 viral sequences (from 8,436 distinct viral species, i.e., unique viral groups) with at least 10% of their length covered by unassembled reads (with at least 90% sequence identity) from the metagenome of interest (3300001348) (**Supplementary Table 2**).

▲ **CRITICAL STEP** This recommended cutoff can be customized by the user by modifying the value '10' from 'line 166' (`vvv.metagenome_length) >= 10`) of the `Coverage_VIRUSES_10Percent.java` script.

? TROUBLESHOOTING

? TROUBLESHOOTING

Troubleshooting advice can be found in **Table 1**.

TABLE 1 | Troubleshooting table.

Step	Problem	Possible reason	Solution
General advice	Scalability can be a bottleneck as inputs (.fasta files) can vary from a few megabytes to many gigabytes in size	Loading inputs in RAM and iterating through them or directly using them for processes such as Blast can significantly slow down the whole pipeline	A common strategy is to split a very large .fasta file into smaller chunks and process them in parallel. To do that, we provide a perl script named *fasta_splitter.pl* (located in the *accessory_scripts* folder). You can run it with the following command: *./fasta_splitter.pl <input_fasta_file> > <number_of_sequences_per_file>* This will split the input .fasta file into multiple files carrying the same filename as the input file. with the added extension *.partN*, where *N* is an ascending counter
1	Header length limitation to 60 characters in .fasta files or very long, difficult-to-follow Unix commands	For simplicity and in order to be able to use basic Unix commands, we often switch between .fasta and tab-delimited formats to filter the .fasta files	We provide two Java scripts, named *Tab2Fasta.java* and *Fasta2Tab.java* (located in the *accessory_scripts* folder), which are implemented to convert a .fasta file to a tab-delimited format and vice versa. A typical run would be: *java Fasta2Tab <input_file> > <output>*. Scripts are designed to parse the files without loading them into RAM, and therefore they run regardless of the size of the file. Current executable files were compiled with Java SE Runtime Environment (build 1.7.0_51-b13)

(continued)

TABLE 1 | Troubleshooting table (continued).

Step	Problem	Possible reason	Solution
7	Very few (or no) viral sequences are detected	Although the specificity of our DNA viral discovery computational approach has been benchmarked and modeled as high as 99.6% for viral detection, the recall rate (sensitivity to identify all viral sequences) is still relatively low: 37.5%. The number of viral sequences detected in a sample is not only a reflection of the 'real' virus abundance within a sample but a positive correlation with the total number of sequences >5 kb	We might want to reassemble the metagenomic sample to increase the number of sequences >5 kb. We strongly recommend using an appropriate assembler for each type of metagenomic sample. By default, we use SPAdes[41] or MEGAHIT[42] as assemblers
12 and 13	Viruses with low-abundance detection take too long and processing is computationally expensive	Searching for viruses with low abundance through multiple metagenomes at the same time is computationally heavy, resource-consuming, and tricky	The *script3* provided in this protocol must be used to parse the corresponding Blast output one sample at a time. Use only one metagenomic sample at a time to run *script3* accurately

● TIMING

The computing time to run this protocol depends on the number of samples and sample(s) size of contigs larger than 5 kb to be analyzed (abundant viruses) and the number of unassembled reads used to hint at the presence of lower-abundance viruses. The example data set described in this protocol contains 0.51 Gb of assembled sequences—of which 164 Mb corresponds to sequences >5 kb—as well as 2.88 Gb of unassembled reads (**Fig. 2**). To run this protocol, we used a single thread; the total estimated computing time consumed to complete it was ~13 h. Much shorter computing times can be achieved by using more cores in parallel.

Step 1, filtering of assembled sequences shorter than 5 kb (seconds)
Step 2, use of viral protein family models against genes from contigs >5 kb using hmmsearch: 7 h 20 min
Step 3, generation of master table: 0.5–1 h
Steps 4–7, application of filters to detect viral contigs in metagenomic data: 10–30 min
Steps 8–11, viral genome clustering: 5–10 min
Steps 12 and 13, hinting at the presence of viruses with low abundance: 4 h 20 min

ANTICIPATED RESULTS

The protocol presented here (**Fig. 1**) creates several tab-delimited .txt output files as a result of the discovery of viral sequences directly from assembled data of a metagenomic sample, the grouping of these detected viruses with those in the databases (at a quasi-species viral taxonomic level), and the prediction of viruses with low abundance (**Supplementary Fig. 1**).

The total number of abundant viral sequences detected with this method depends largely on the length, quality, and amount of assembled sequence data. All the required files needed to run this protocol, databases, scripts, and outputs generated are deposited in the FTP site: http://portal.nersc.gov/dna/microbial/prokpubs/EarthVirome_DP/Nature_Protocols/.

In the example used in this protocol (Pelagic microbial community sample from the North Sea; IMG/M taxon identifier 3300001348), we detected 640 dominant metagenomic viral contigs (**Supplementary Table 1**) and grouped them into 246 viral clusters. We also hint at the presence of 8,436 distinct viral species (unique viral clusters) from 12,963 metagenomic viral contigs, at lower abundances. 97% of the low-abundant putative viral sequences were also previously found in marine environments. 2 and 1% of the remaining sequences come from freshwater and nonmarine saline alkaline environments, respectively (**Supplementary Table 2**). Users can retrieve all details of these contigs (metadata, taxonomy, and host association, when detected) from the IMG/VR database[23] (https://img.jgi.doe.gov/vr/) using the corresponding identifiers from the second column (Subject ID) of the tabular output from the Blast file.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS This work was supported by the US Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, under contract no. DE-AC02-05CH11231, and used resources of the National Energy Research Scientific Computing Center, supported by the Office of Science of the US Department of Energy.

AUTHOR CONTRIBUTIONS D.P.-E., N.N.I., and N.C.K. conceived and led the protocol. G.A.P. provided computational and scripting support. All authors wrote and edited the manuscript.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

1. Chen, I.A. *et al.* IMG/M: integrated genome and metagenome comparative data analysis system. *Nucleic Acids Res.* **45**, D507–D516 (2017).
2. Mukherjee, S. *et al.* Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. *Nucleic Acids Res.* **45**, D446–D456 (2017).
3. Angly, F.E. *et al.* The marine viromes of four oceanic regions. *PLoS Biol.* **4**, e368 (2006).
4. Breitbart, M., Miyake, J.H. & Rohwer, F. Global distribution of nearly identical phage-encoded DNA sequences. *FEMS Microbiol. Lett.* **236**, 249–256 (2004).
5. Breitbart, M. & Rohwer, F. Here a virus, there a virus, everywhere the same virus? *Trends Microbiol.* **13**, 278–284 (2005).
6. Marhaver, K.L., Edwards, R.A. & Rohwer, F. Viral communities associated with healthy and bleaching corals. *Environ. Microbiol.* **10**, 2277–2286 (2008).
7. Suttle, C.A., Chan, A.M. & Cottrell, M.T. Use of ultrafiltration to isolate viruses from seawater which are pathogens of marine phytoplankton **57**, 721–726 (1991).
8. Dell'Anno, A., Corinaldesi, C., Magagnini, M. & Danovaro, R. Determination of viral production in aquatic sediments using the dilution-based approach. *Nat. Protoc.* **4**, 1013–1022 (2009).
9. Thurber, R.V., Haynes, M., Breitbart, M., Wegley, L. & Rohwer, F. Laboratory procedures to generate viral metagenomes. *Nat. Protoc.* **4**, 470–483 (2009).
10. Brum, J.R. *et al.* Ocean plankton. Patterns and ecological drivers of ocean viral communities. *Science* **348**, 1261498 (2015).
11. Dinsdale, E.A. *et al.* Functional metagenomic profiling of nine biomes. *Nature* **452**, 629–632 (2008).
12. Mizuno, C.M., Rodriguez-Valera, F., Kimes, N.E. & Ghai, R. Expanding the marine virosphere using metagenomics. *PLoS Genet.* **9**, e1003987 (2013).
13. Roux, S. *et al.* Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**, 689–693 (2016).
14. Akhter, S., Aziz, R.K. & Edwards, R.A. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Res.* **40**, e126 (2012).
15. Fouts, D.E. Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res.* **34**, 5839–5851 (2006).
16. Lima-Mendez, G., Van Helden, J., Toussaint, A. & Leplae, R. Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics* **24**, 863–865 (2008).
17. Arndt, D. *et al.* PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* **44**, W16–W21 (2016).
18. Roux, S., Enault, F., Hurwitz, B.L. & Sullivan, M.B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985 (2015).
19. Grazziotin, A.L., Koonin, E.V. & Kristensen, D.M. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids Res.* **45**, D491–D498 (2017).
20. Paez-Espino, D. *et al.* Uncovering earth's virome. *Nature* **536**, 425–430 (2016).
21. Ivanova, N. *et al.* A call for standardized classification of metagenome projects. *Environ. Microbiol.* **12**, 1803–1805 (2010).
22. Mukherjee, S. *et al.* Genomes OnLine Database(GOLD) v.6: data updates and feature enhancements. *Nucleic Acids Res.* **45**, D446–D456 (2016).
23. Paez-Espino, D. *et al.* IMG/VR: a database of cultured and uncultured DNA viruses and retroviruses. *Nucleic Acids Res.* **45**, D457–D465 (2017).
24. Merchant, N. *et al.* The iPlant Collaborative: cyberinfrastructure for enabling data to discovery for the life sciences. *PLoS Biol.* **14**, e1002342 (2016).
25. Suttle, C.A. Marine viruses—major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801–812 (2007).
26. Edwards, R.A., McNair, K., Faust, K., Raes, J. & Dutilh, B.E. Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiol. Rev.* **40**, 258–272 (2016).
27. Villarroel, J. *et al.* HostPhinder: a phage host prediction tool. *Viruses* **8** http://dx.doi.org/10.3390/v8050116 (2016).
28. Goren, M.G., Yosef, I. & Qimron, U. Programming bacteriophages by swapping their specificity determinants. *Trends Microbiol.* **23**, 744–746 (2015).
29. Salmond, G.P. & Fineran, P.C. A century of the phage: past, present and future. *Nat. Rev. Microbiol.* **13**, 777–786 (2015).
30. Edgar, R.C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
31. Enright, A.J., Van Dongen, S. & Ouzounis, C.A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
32. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
33. Finn, R.D., Clements, J. & Eddy, S.R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–W37 (2011).
34. Chen, I.A. *et al.* IMG/M: integrated genome and metagenome comparative data analysis system. *Nucleic Acids Res.* **45**, D507–D516 (2016).
35. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
36. Dutilh, B.E. *et al.* A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* **5**, 4498 (2014).
37. Aziz, R.K., Dwivedi, B., Akhter, S., Breitbart, M. & Edwards, R.A. Multidimensional metrics for estimating phage abundance, distribution, gene density, and sequence coverage in metagenomes. *Front. Microbiol.* **6**, 381 (2015).
38. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
39. Langdon, W.B. Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. *BioData Min.* **8**, 1 (2015).
40. Finn, R.D. *et al.* HMMER web server: 2015 update. *Nucleic Acids Res.* **43**, W30–W38 (2015).
41. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
42. Li, D., Liu, C.M., Luo, R., Sadakane, K. & Lam, T.W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
43. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
44. Dick, G.J. *et al.* Community-wide analysis of microbial genome sequence signatures. *Genome Biol.* **10**, R85 (2009).
45. Oulas, A. *et al.* Metagenomic investigation of the geologically unique Hellenic volcanic arc reveals a distinctive ecosystem with unexpected physiology. *Environ. Microbiol.* **18**, 1122–1136 (2016).
46. Price, M.N., Dehal, P.S. & Arkin, A.P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
47. Huson, D.H. & Scornavacca, C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Syst. Biol.* **61**, 1061–1067 (2012).