AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Resource Announcements

DATABASES AND SOFTWARE

virMine 2.0: Identifying Viral Sequences in Microbial Communities

Genevieve Johnson,[a] **Catherine Putonti**[a,b]

[a]Bioinformatics Program, Loyola University Chicago, Chicago, Illinois, USA
[b]Department of Biology, Loyola University Chicago, Chicago, Illinois, USA

ABSTRACT Here, we present virMine 2.0, the next generation of the virMine software tool. virMine 2.0 uses an exclusion technique to remove nonviral data from sequencing reads and scores the remaining data based on relatedness to viral elements, eliminating the sole dependency on homology identification.

In contrast to the wealth of data available for cellular organisms, the viral diversity on Earth is underrepresented in sequence databases. As a result, homology-based identification of viral sequences is limited. Furthermore, viruses contain a high degree of genetic diversity, and it can be nearly impossible to distinguish conserved genes among viruses (1). Bioinformatic approaches for *de novo* viral identification employ homology-based, nucleotide usage, or coverage analyses or combinations thereof (for a review, see reference 2).

Previously, we introduced a tool called virMine (3), which utilizes the wealth of sequence data for cellular organisms to identify likely viral sequences in metagenomes. The tool takes either (i) short reads, either single-end or paired-end fastq file(s), or (ii) a long-read or assembled-sequence fasta file. In the former case, read quality control is conducted, followed by assembly. Three methods for assembly are included in virMine, namely, SPAdes, metaSPAdes, and MEGAHIT; alternatively, the user can select the all3 option, in which all three assembly methods are executed and the assembly with the greatest N_{50} value is selected for further analysis. Contigs (either those assembled by virMine or those from the supplied long-read or assembled-sequence fasta file) can be filtered. This step is optional, and virMine filters include minimum and/or maximum contig length, minimum contig coverage, and the presence of sequences of interest. Next, virMine preforms gene prediction. Contigs are scored based on their gene content's origin, i.e., cellular, viral, or unknown. virMine has successfully identified prophages and viral sequences, both homologous to known viruses and novel, from synthetic data sets and environmental samples from freshwater, the gut, and urine (3, 4).

virMine 2.0, presented here, follows the same methodology as its predecessor while incorporating updated versions of the underlying tools and databases. These updates include Python v.3.9, BBMap v.38.94 (the tool used to compute coverage statistics for the coverage filter [https://sourceforge.net/projects/bbmap]), and SPAdes v.3.15.3 (5). Furthermore, a new script to generate the virMine databases is included in this release. The script retrieves the latest bacterial Clusters of Orthologous Genes (COG) database, released in 2020 (6); it then removes all sequences of viral origin (category X) and formats the database for virMine sequence comparisons. The script also generates a viral database from the latest collection of RefSeq eukaryotic viral and phage genomes (ftp://ftp.ncbi.nlm.nih.gov/genomes/Viruses/all.faa.tar.gz) (7).

Source code and a Docker image are available at https://github.com/putonti/virmine. To use the Docker image, the user must first install the Docker application itself (https://www.docker.com). The Dockerfile builds the necessary environment with all dependencies.

Editor Irene L. G. Newton, Indiana University, Bloomington

Address correspondence to Catherine Putonti, cputonti@luc.edu.

The authors declare no conflict of interest.

Received 6 February 2022
Accepted 10 April 2022
Published 2 May 2022

Once Docker is installed, the virMine repository can be cloned locally. The viral and bacterial database files can be generated using the virmine_make_dbs.py script or can be substituted with personal database files. The database files and the input files must then be transferred to the input folder within the local cloned repository. The GitHub repository provides example commands for analyses using either paired-end reads or assembled contigs. Results from the runs are saved locally within the cloned repository directory. Test data and sample output files are provided through the GitHub repository. The Docker image of virMine is also available at https://hub.docker.com/repository/docker/genevievej16/virmine; using the Docker Hub image eliminates the need to build the Docker image locally from the cloned repository.

Data availability. virMine and its Docker image are located online at https://github.com/putonti/virmine and https://hub.docker.com/repository/docker/genevievej16/virmine, respectively. Also included in the repository is a script to generate the updated bacterial and viral databases. Additional documentation, including setup, walkthroughs, and example commands, and test data are available in the GitHub repository.

ACKNOWLEDGMENTS

We thank Thomas Hatzopoulos and Andrea Garretto, prior contributors to this code.
This work is supported by the National Science Foundation (award 1661357, to C.P.).

REFERENCES

1. Hatfull GF. 2008. Bacteriophage genomics. Curr Opin Microbiol 11:447–453. https://doi.org/10.1016/j.mib.2008.09.004.
2. Kieft K, Anantharaman K. 2022. Virus genomics: what is being overlooked? Curr Opin Virol 53:101200. https://doi.org/10.1016/j.coviro.2022.101200.
3. Garretto A, Hatzopoulos T, Putonti C. 2019. virMine: automated detection of viral sequences from complex metagenomic samples. PeerJ 7:e6695. https://doi.org/10.7717/peerj.6695.
4. Garretto A, Thomas-White K, Wolfe AJ, Putonti C. 2018. Detecting viral genomes in the female urinary microbiome. J Gen Virol 99:1141–1146. https://doi.org/10.1099/jgv.0.001097.
5. Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. 2020. Using SPAdes de novo assembler. Curr Protoc Bioinformatics 70:e102. https://doi.org/10.1002/cpbi.102.
6. Galperin MY, Wolf YI, Makarova KS, Vera Alvarez R, Landsman D, Koonin EV. 2021. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. Nucleic Acids Res 49:D274–D281. https://doi.org/10.1093/nar/gkaa1018.
7. 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 Res 44:D733–745. https://doi.org/10.1093/nar/gkv1189.