VirBot
an RNA viral contig detector for metagenomic data
Chen, Guowei; Tang, Xubo; Shi, Mang; Sun, Yanni

Published in:
Bioinformatics

Published: 01/03/2023

Document Version:
Final Published version, also known as Publisher's PDF, Publisher's Final version or Version of Record

License:
CC BY

Publication record in CityU Scholars:
Go to record

Published version (DOI):
10.1093/bioinformatics/btad093

Publication details:

Citing this paper
Please note that where the full-text provided on CityU Scholars is the Post-print version (also known as Accepted Author Manuscript, Peer-reviewed or Author Final version), it may differ from the Final Published version. When citing, ensure that you check and use the publisher's definitive version for pagination and other details.

General rights
Copyright for the publications made accessible via the CityU Scholars portal is retained by the author(s) and/or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights. Users may not further distribute the material or use it for any profit-making activity or commercial gain.

Publisher permission
Permission for previously published items are in accordance with publisher's copyright policies sourced from the SHERPA RoMEO database. Links to full text versions (either Published or Post-print) are only available if corresponding publishers allow open access.

Take down policy
Contact lbscholars@cityu.edu.hk if you believe that this document breaches copyright and provide us with details. We will remove access to the work immediately and investigate your claim.
Genome analysis

VirBot: an RNA viral contig detector for metagenomic data

Guowei Chen 1, Xubo Tang 1, Mang Shi 2 and Yanni Sun 1, *

1 Department of Electrical Engineering, City University of Hong Kong, Kowloon, Hong Kong SAR, China and 2 School of Medicine, Shenzhen Campus of Sun Yat-sen University, Shenzhen, China

* To whom correspondence should be addressed.

Received on October 19, 2022; revised on January 12, 2023; editorial decision on February 11, 2023; accepted on February 15, 2023

Abstract

Summary: Without relying on cultivation, metagenomic sequencing greatly accelerated the novel RNA virus detection. However, it is not trivial to accurately identify RNA viral contigs from a mixture of species. The low content of RNA viruses in metagenomic data requires a highly specific detector, while new RNA viruses can exhibit high genetic diversity, posing a challenge for alignment-based tools. In this work, we developed VirBot, a simple yet effective RNA virus identification tool based on the protein families and the corresponding adaptive score cutoffs. We benchmarked it with seven popular tools for virus identification on both simulated and real sequencing data. VirBot shows its high specificity in metagenomic datasets and superior sensitivity in detecting novel RNA viruses.

Availability and implementation: https://github.com/GreyGuoweiChen/RNA_virus_detector

Contact: yannisun@cityu.edu.hk

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

RNA viruses have high diversity and can infect a large variety of eukaryotes from planktons to humans. With an estimated number of over 100 million eukaryotic viruses, <0.1% RNA viruses are recorded (Zhang et al., 2019). Both metagenomic and viral metagenomic sequencing data are primary sources for virus discovery (Grenning, 2018). However, both types of data can be loaded with organisms of different origins. To identify RNA viruses from the heterogeneous data, generic alignment tools and targeted virus detection tools have been adopted in different applications (Buchfink et al., 2015; Guo et al., 2021; Menzel et al., 2016; Ren et al., 2017; Steinegger and Söding, 2017; Wood et al., 2019). But, their performance deteriorates on detecting short RNA viral contigs from new habitats.

In this work, we introduce VirBot, an easy-to-use yet effective RNA virus detection tool for metagenomic data. It takes assembled contigs as input and detects ones from RNA viruses. In addition, it outputs the taxonomic labels for the detected RNA viruses. We validated VirBot in various scenarios and benchmark it with other RNA virus detection tools. In samples that only contain novel RNA viruses, VirBot achieves higher recall than other tools. And VirBot also demonstrates high specificity in metagenomic data that only contains a small number of RNA viruses.

2 Materials and methods

VirBot detects RNA viruses using our self-constructed protein domain families and their customized score cutoffs. As shown by a recent analysis (Yuan et al., 2022), RNA viruses have a better conservation in their protein sequences than their nucleic acid sequences. Some proteins, including RdRps, capsid proteins and envelope proteins, have been widely applied to study the evolution and phylogeny of RNA viruses (Shi et al., 2018). Our previous work leveraging RdRp (Tang et al., 2022) has contributed to sensitive viral read identification. However, using RdRp itself is limited because short contigs may not overlap with RdRp genes. Thus, we construct our own RNA viral protein families by clustering 22,281 non-redundant RNA viral proteins derived from available RNA virus and determine the family-specific cut-off using 814,991 non-RNA viral proteins from Uniprot and RefSeq (Fig. 1a). Finally, there are 1384 protein families and 3927 singletons, which only contain one protein sequence. We encode all the families using pHMMs to enhance remote homology search and provide a preliminary taxonomic prediction based on the known label of each sequence in the protein family. Unlike previous tools, we did not use public protein domain database such as Pfam because it is not optimized for RNA viruses. For example, RdRp is an essential protein responsible for RNA replication. When searching with the keyword ‘rdrp’ in Pfam (35.0), it returns 73 pHMMs, and only 13 are constructed from viral RdRp. The limited number of RdRp families specifically constructed for RNA viruses does not comply with the diversity and abundance of RNA viruses. The details about the protein family database construction, the determination of the adaptive cutoffs, and the taxonomic label assignment for each family can be found in Supplementary Figure S1 and Supplementary Section S2. VirBot’s query stage is sketched in Figure 1b, where an input contig will be
classified based on the result of majority vote on its predicted proteins against our protein family database and the singletons.

3 Results

We validated VirBot in five scenarios and compared its performance against seven popular tools. The commands and parameters of the tools are summarized in Supplementary Tables S1 and S2. More details about the experiments and the taxonomic assignment evaluation can be found in Section Result of the Supplementary File.

Simulated metagenomic data. We first evaluated VirBot on two semi-synthetic marine metagenomic samples (NCBI SRA ID: ERR1992810 and ERR2185279) constructed from 82 eukaryotes, 365 prokaryotes and DNA/RNA viruses (Mitchell et al., 2019). As shown in Figure 1c, VirBot showed the highest precision and F1 scores in these samples. More details about the data and the other results on contigs of different lengths are shown in Supplementary Table S3 and Supplementary Figures S2 and S3.

Identifying new RNA virus. To assess the ability of VirBot in identifying new RNA viruses, we used three datasets that contain newly sequenced RNA viruses. One comprises 8849 RNA phages that were barely detected before, while eukaryotic viruses dominate known RNA viruses. The second dataset is an RNA virome sample sequenced from marine water containing 114 139 RNA viral sequences. The third is made up of 5504 non-redundant sequences known RNA viruses. The second dataset is an RNA virome sample classified based on the result of majority vote on its predicted proteins against our protein family database and the singletons.

Real metagenomic data. We tested VirBot on 21 human clinical samples for pathogen detection. VirBot successfully retrieved all the reported RNA viral pathogens and achieved an average recall of 99.3% and precision of 99.2% among different groups. The result is shown in Supplementary Table S5.

4 Conclusion

In this work, we constructed a comprehensive protein family database for RNA viruses and derived the adaptive bit score cutoff using a large number of non-RNA viral proteins. We developed VirBot based on these RNA viral pHMMs for fast and more sensitive RNA virus detection in metagenomic data. Our pHMMs are designed specifically for RNA virus and show higher sensitivity than other HMM-based tools (Fig. 1d, Supplementary Fig. S4 and Supplementary Table S4). Compared with using a fixed cut-off (recommended E-value), VirBot improved the specificity from 8.8% to 97.4% in the simulated data, which is shown in Supplementary Table S6.

Funding

This work was supported by Hong Kong Research Grants Council (RGC) General Research Fund (GRF) [11206819] and Hong Kong Innovation and Technology Fund (ITF) [MRP/071/20X].

Conflict of Interest: The authors declare that there is no conflict of interest.

Data availability

The source code of VirBot and the data are available at https://github.com/GreyGuoWeiChen/RNA_virus_detector.

References


