Viralrecall - A tool to identify giant viruses integrated into eukaryotic genomes

Abdeali M. Jivaji [1][¶] **and Frank O. Aylward** [1,2]

1 Dept. of Biological Sciences, Virginia Tech **2** Department of Biological Sciences and Center for Emerging, Zoonotic, and Arthropod-borne Pathogens, Virginia Tech, Blacksburg, Virginia, USA ¶ Corresponding author

Summary

Viralrecall 3.0 is a python tool to identify Giant Endogenous Viral Elements (GEVEs) integrated in the genome of eukaryotes. The current version is an update on the original tool by Dr. Aylward and uses the same GVOG HMM database to detect signatures of giant viruses (Aylward & Moniruzzaman, 2021). The key motivation for updating `viralrecall` was to make it more efficient at processing the larger eukaryotic genomes that are being published with the rise in popularity of long-read sequencing.

Statement of Need

This update was needed since the older version (v2.0) was only accessible as a python script that may not be beginner-friendly and easy to use. It was based on an older version of python that required specific and outdated versions of crucial packages that were not easy to setup/troubleshoot. With v3.0, we provide an easy to install option via conda as it is widely used by bioinformaticians worldwide. We provide a direct script to install and setup the database automatically reducing errors. We also include an experimental database of key Mirusvirus proteins to identify these mysterious viruses that share similarities with giant viruses. With the rise of long-read DNA sequencing and the increased abundance of high-quality and complete eukaryotic genomes, this version of the tool has been optimized to be memory-efficient and blazingly fast compared to the v2.0. Specifically in batch mode, where all genome files in a given directory are processed, the tool maximizes the utilization of available cores present in modern high-performance compute clusters to process thousands of genomes in an efficient and quick manner. With this, we hope to bring inexperienced folks and proficient bioinformaticians alike into the field of giant viruses and identify these mysterious viruses throughout the tree of life (Moniruzzaman et al., 2020).

State of the Field

Although several tools exist to identify giant viruses based on DNA sequences, these tools are generally broad-ranged and only include identification of giant viruses as a minor feature. Specifically, the tool genomad is an excellent resource to identify various phages and mobile genetic elements (Camargo et al., 2024). Since giant viruses are significantly different from most previously identified viruses, their dataset required special categories for giant virus identification. While genomad is an excellent resource, our tool is specifically designed to identify giant viruses, either present in metagenomes as independent contigs, or integrated into the genome of eukaryotes as GEVEs. Thus, we provide a more curated database and tool specifically designed to identify these enigmatic viruses.

AI usage disclosure

We used Github copilot for basic autocomplete and the chatbot for suggestions regarding some functions, and for writing tests. Generative AI was not used for wrtiting this manuscript.

Author Contributions

Abdeali M. Jivaji: Methodology, Software, Writing - Original Draft. **Frank O. Aylward:** Review, Conceptualization, Resources, Supervision.

Acknowledgements

We thank the members of the Aylward Lab for their support and comments during the development of this tool.

Aylward, F. O., & Moniruzzaman, M. (2021). ViralRecall—a flexible command-line tool for the detection of giant virus signatures in 'omic data. *Viruses*, *13*(2), 150.

Camargo, A. P., Roux, S., Schulz, F., Babinski, M., Xu, Y., Hu, B., Chain, P. S., Nayfach, S., & Kyrpides, N. C. (2024). Identification of mobile genetic elements with geNomad. *Nature Biotechnology*, *42*(8), 1303–1312.

Moniruzzaman, M., Martinez-Gutierrez, C. A., Weinheimer, A. R., & Aylward, F. O. (2020). Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. *Nature Communications*, *11*(1), 1710.