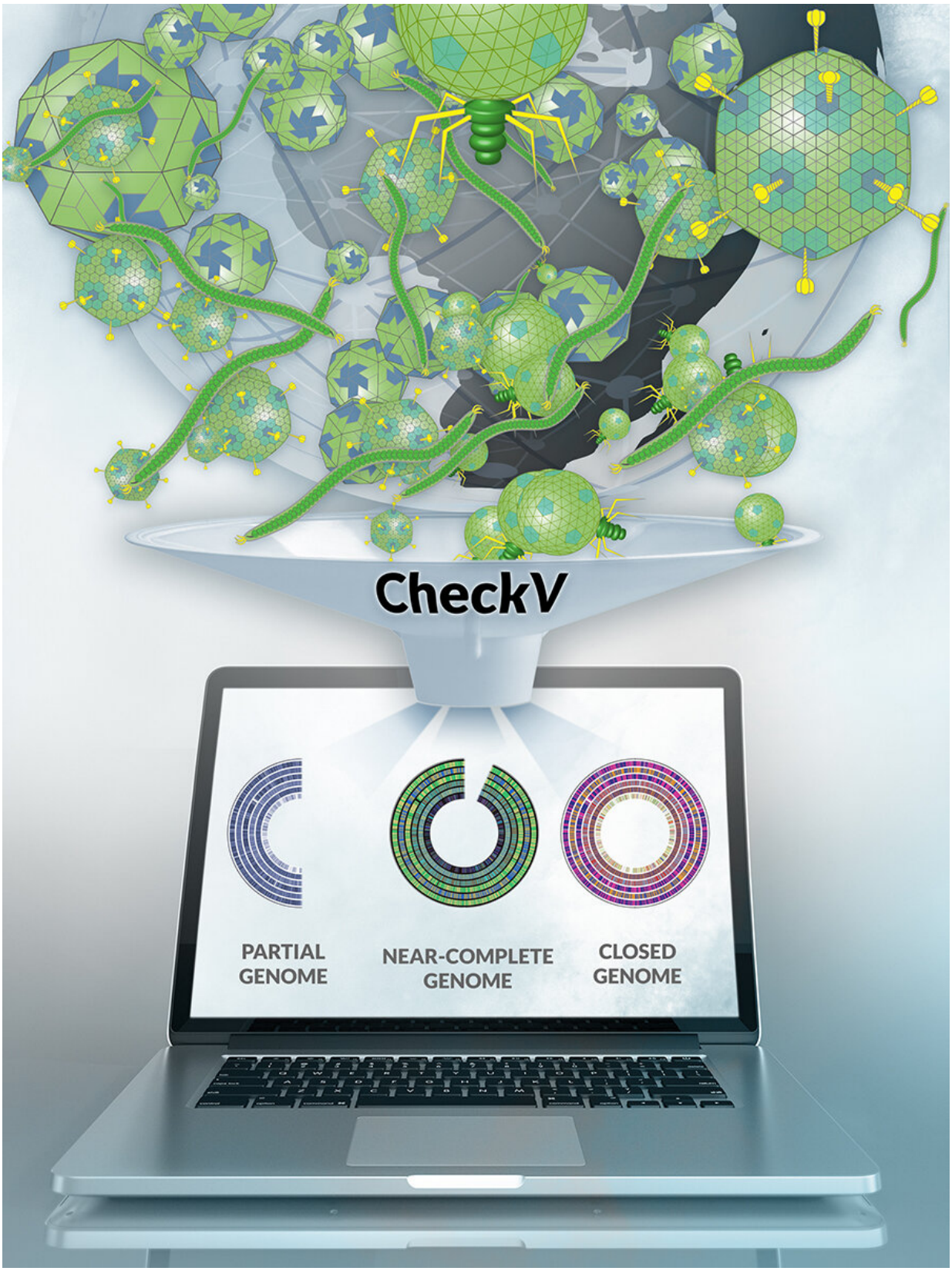
An automated tool for assessing virus data quality

December 22 2020


CheckV
PARTIAL GENOME
NEAR-COMPLETE GENOME
CLOSED GENOME

Artistic interpretation of CheckV assessing virus genome sequences from environmental samples. Credit: Zosia Rostomian, Berkeley Lab

Through advances in sequencing technologies and computational approaches, more and more virus sequences are being recovered and identified from environmental samples (metagenomes). However, the quality and completeness of metagenome-assembled virus sequences vary widely. In a previous effort, an international consortium recommended specific guidelines and best practices for characterizing uncultivated viruses. Following up on those guidelines, JGI researchers have now developed CheckV (pronounced "Check-Vee") to help researchers assess and improve the quality of metagenome-assembled viral genomes.

The microbes that play key roles in cycling nutrients such as carbon, nitrogen and sulfur are themselves regulated by viruses in their environments. Environmental DNA sequencing can help scientist to recover the genomes of these viruses and associate them with their microbial hosts. However, assembling viral genomes from metagenomes is challenging and often results in highly fragmented data, which limits the ability of researchers to accurately perform functional assessment, host prediction, and phylogenetic analysis. The development of CheckV helps researchers to assess the completeness of these sequences and complements a community effort to develop guidelines and best practices for defining virus data quality.

Characterizing viral genome fragments can be difficult, akin to the story of the blind men who encounter an elephant for the first time. Based on the single body part each blind man touches—a tusk, the ear, or the tail—they individually decide that the elephant is either dangerous, akin to a carpet, or a harmless piece of rope. Similarly, genome fragments can

provide an incomplete picture of a virus, and for viruses that have integrated into the host genome, these sequences may be tainted by the presence of non-viral genes.

Up to this point, there has been a lack of fast and accurate tools for researchers to assess the quality of metagenome-assembled viral genomes, including estimation of genome completeness and removal of contamination from the host organism. As reported in *Nature Biotechnology*, a team from the U.S. Department of Energy (DOE) Joint Genome Institute (JGI), a DOE Office of Science User Facility located at Lawrence Berkeley National Laboratory (Berkeley Lab), has developed a command-line tool called CheckV that can automatically do both. The work was led by research scientist Stephen Nayfach, the study's first author in the Microbiome Data Science group led by Nikos Kyrpides.

To demonstrate its utility, Nayfach applied CheckV to sequences of uncultivated viruses (from environmental metagenome samples) from IMG/VR, a database that is part of the Integrated Microbial Genomes & Microbiomes (IMG/M) suite, as well as sequences from the Global Ocean Virome 2.0 dataset based on open ocean samples. CheckV identified a total of 44,652 complete or near-complete viral genomes across both datasets, separating these from the vast majority of other sequences that were incomplete fragments. Additionally, CheckV was able to identify just over 17,000 contiguous sequences (contigs) of proviruses flanked on one or both sides by genes from the host organism. With the virus-host boundary clearly defined using functional annotation methods, it was possible to distinguish between metabolic genes found in the viral genome versus those from the host organism. Without this prediction step, numerous genes for antibiotic resistance and secondary metabolism would have been incorrectly attributed to viruses.

The tool can be broadly utilized by the research community to gauge

virus data quality and will help researchers to follow best practices and guidelines for providing the minimum amount of information for an uncultivated virus genome. CheckV has already been applied to over 2.4 million viral genomes available in the latest release of IMG/VR.

More information: CheckV is freely available for download at: bitbucket.org/berkeleylab/CheckV Stephen Nayfach et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes, *Nature Biotechnology* (2020). DOI: 10.1038/s41587-020-00774-7

Provided by DOE/Joint Genome Institute