MetaviralSPAdes—New assembler for virus genomes

May 27 2020

When a new virus emerges, biologists rush to reconstruct its genome—a prerequisite for future diagnostic and vaccine development. The challenge with viral sequencing during an outbreak is that a sample from a patient, like saliva from a COVID-19 patient that was used for the very first SARS-COV-2 coronavirus sequencing effort, contains genomes of many other, often harmless, viruses. Not to mention hundreds of much larger bacterial genomes that live in our mouth and make it difficult to find the viral sequences among them.

It raises the challenge of metagenome sequencing, reading hundreds of genomes at once, a more difficult computational problem than sequencing a single genome. Metagenome sequencing results in 1000s sequences representing pieces of various viral and bacterial genomes and it remains unclear which of these pieces represent the genome of the pathogen. The next task, known as metavirome sequencing, is to identify various viral sequences (hidden among much longer bacterial sequences!) and stitch them together into a complete viral genome of a pathogen that caused the outbreak.

There was no specialized viral metagenome assembler until recently. But the joint team of Russian and US researchers from Saint-Petersburg State University and University of California at San Diego just released the metaviralSPAdes assembler (published in journal *Bioinformatics* on May 16) that turns the analysis of the metavirome sequencing results into an easy task.

Biologists still cannot read the entire genome in the same way we read a book from the beginning to the end—instead, they read small snippets of a genome. Genome assembly is not unlike putting together a puzzle from a million such pieces, and it is often viewed as one of the most difficult algorithmic problems in bioinformatics. Nevertheless, the most widely used genome assembler today, called SPAdes, was used in nearly 9000 papers. It was used to analyze pathogens causing MERS in Saudi Arabia (Cotten et al., *The Lancet* 2013), Ebola in Congo (Maganga et al., *New England Journal of Medicine*, 2013), gonorrhoea in England (Chisholm et al., *Sex Transm Infect*, 2016), meningitis in Ghana (Kwambana-Adams et al., *BMC Infectious Diseases* 2016), dengue in Sumatra (Sasmono et al., *Am. J. Trop. Med. Hyg*. 2017), and dozens other outbreaks in the last eight years since SPAdes was released.

However, metagenome assembly of a 1000 genomes is much more difficult than assembly of a single genome. It is not unlike putting together a 1000 puzzles (each with a million pieces!) at once, from a huge pile of a billion pieces from all these puzzles, all mixed together in a single bag. However, three years ago, the same Russian-US team that developed SPAdes, also developed metaSPAdes to address these challenges. metaSPAdes has quickly become a leading metagenomic assembler that has already been applied in over 500 metagenomic projects.

However, it is not trivial to extract viral sequences from a huge metaSPAdes output, as a virus genome is hiding among thousands of pieces of reconstructed bacterial genomes. New metaviralSPAdes assembler not only finds such pieces of viral sequences but also stitches them together into the completed genome.

The COVID-19 pandemic is a wake-up call for biologists studying transmission of viruses from animal to humans. It emphasizes the importance of viral surveillance in various animal hosts such as going to

caves where bats live, collecting their fecal samples, and studying the huge repertoire of bat viruses BEFORE rather than AFTER a pandemic strikes. Bats have an unparalleled immune system that allows them to co-exist with a multitude of viruses that can kill humans. However, in addition to the logistics of collecting all such samples, building a census of various viral genomes across various animals is a difficult computational problem.

With metaviralSPAdes at hand, biologists can now reconstruct these viral genomes in bats or any other potential sources of future pandemics.

 More information: *Bioinformatics* (2020). DOI: 10.1093/bioinformatics/btaa490

Provided by St. Petersburg State University