GA4GH streaming API htsget a bridge to the future for modern genomic data processing

June 22 2018

The Large Scale Genomics Work Stream of the Global Alliance for Genomics and Health (GA4GH) has announced eight new implementations of its htsget protocol, a standard released in October 2017 for accessing large-scale genomic sequencing data online without using file transfers. The protocol and interoperability testing are reported in a paper released online this week in the journal *Bioinformatics*.

The cornerstone of solving common diseases such as cancer and diabetes is to be able to compare the genome sequences of thousands of individuals to identify recurring genetic variants. Since no single institution can amass such a dataset on its own, it is critical for organizations to share information across traditional boundaries.

Historically, this has been done through the use of standardised file formats: a file generated at one institution can be downloaded and integrated with files at another institution because they use the same format.

This has worked well since the late 2000s, when these formats were developed as part of the international 1000 Genomes Project and they have enabled a global ecosystem of interoperable sequence analysis tools and pipelines.

But the field is changing. Genomics is shifting from a research endeavor to one more broadly implemented in routine clinical care; datasets will be so large that the current model of institutionally siloed file systems

will not be sufficient to enable global sharing and collaboration.

"Datasets containing hundreds of millions—rather than hundreds of thousands—of sequences will be available within the next five years and sharing files of that size is simply not realistic," said Ewan Birney, Director of EMBL-EBI and Chair of GA4GH. "Users would have to download terabyte-sized files just to access data on a small subset of the genome sequence."

At the same time, the world is changing—from film to financial data, myriad domains are shifting from traditional file-based approaches for storing and processing data to more modern, big-data, cloud-based approaches. Genomics will have to follow suit, but not without sacrificing current standards that make data interoperable.

"We're not attempting to replace the existing file formats," said Thomas Keane, Team Leader of EGA and the Archive Infrastructure at EMBL-EBI and co-chair of the GA4GH Large Scale Genomics Work Stream and its htsget task team. "Doing so would require adaptation of every single bioinformatics tool for processing data that is currently compatible with those formats."

Instead, htsget provides a consistent protocol for researchers to access data stored in different repositories—whether based in big public clouds or in more traditional infrastructure. It also includes a robust security and authentication mechanism, which is key for sensitive data.

It can be operated efficiently for very large datasets, and, because it uses the existing standards for transmitting data, it can be readily integrated into current pipelines and analytical methods. Users can employ htsget to download only the subsection of a genome sequence in which they are interested rather than the whole file, or they can download the entire genome as a series of "data slices" distributed across multiple disparate

machines.

"We've thought of this as a bridge to the future," said Mike Lin, specification maintainer for the GA4GH htsget team. "It's a gradual path to upgrade current file-based pipelines and repositories to a more interoperable, API-based architecture—which has always been a foundational vision of GA4GH."

Lin will lead a webinar introducing the protocol and answering questions about implementation for the broad community on July 24. Anyone interested in learning more about htsget and how to implement it in their bioinformatics pipelines is invited to attend. Register here.

More information: Jerome Kelleher et al, htsget: a protocol for securely streaming genomic data, *Bioinformatics* (2018). DOI: 10.1093/bioinformatics/bty492

Provided by Global Alliance for Genomics and Health