New cloud-based platform opens genomics data to all

January 12 2022

Credit: CC0 Public Domain

Harnessing the power of genomics to find risk factors for major diseases or search for relatives relies on the costly and time-consuming ability to analyze huge numbers of genomes. A team co-led by a Johns Hopkins

University computer scientist has leveled the playing field by creating a cloud-based platform that grants genomics researchers easy access to one of the world's largest genomics databases.

Known as AnVIL (Genomic Data Science Analysis, Visualization, and Informatics Lab-space), the new platform gives any researcher with an internet connection access to thousands of analysis tools, patient records, and more than 300,000 genomes. The work, a project of the National Human Genome Institute (NHGRI), appears today in *Cell Genomics*.

"AnVIL is inverting the model of genomics data sharing, offering unprecedented new opportunities for science by connecting researchers and datasets in new ways and promising to enable exciting new discoveries," said project co-leader Michael Schatz, Bloomberg Distinguished Professor of Computer Science and Biology at Johns Hopkins.

Typically genomic analysis starts with researchers downloading massive amounts of data from centralized warehouses to their own data centers, a process that is not only time-consuming, inefficient, and expensive, but also makes collaborating with researchers at other institutions difficult.

"AnVIL will be transformative for institutions of all sizes, especially smaller institutions that don't have the resources to build their own data centers. It is our hope that AnVIL levels the playing field, so that everyone has equal access to make discoveries," Schatz said.

Genetic risk factors for ailments such as cancer or cardiovascular disease are often very subtle, requiring researchers to analyze thousands of patients' genomes to discover new associations. The raw data for a single human genome comprises about 40GB, so downloading thousands of genomes can take takes several days to several weeks: A single genome requires about 10 DVDs worth of data, so transferring thousands means

moving "tens of thousands of DVDs worth of data," Schatz said.

In addition, many studies require integrating data collected at multiple institutions, which means each institution must download its own copy while ensuring that patient-data security is maintained. This challenge is expected to become even greater in the future, as researchers embark on ever-larger studies requiring the analysis of hundreds of thousands to millions of genomes at once.

"Connecting to AnVIL remotely eliminates the need for these massive downloads and saves on the overhead," Schatz says. "Instead of painfully moving data to researchers, we allow researchers to effortlessly move to the data in the cloud. It also makes sharing datasets much easier so that data can be connected in new ways to find new associations, and it simplifies a lot of computing issues, like providing strong encryption and privacy for patient datasets."

AnVIL also provides researchers with several major analysis tools, including Galaxy, developed in part at Johns Hopkins, along with other popular tools such as R/Bioconductor, Jupyter notebooks, WDLs, Gen3, and Dockstore to support both interactive analysis and large-scale batch computing. Collectively, these tools allow researchers to tackle even the largest studies without having to build out their own computing environments.

Researchers from all over the world currently use the platform to study a variety of genetic diseases, including autism spectrum disorders, cardiovascular disease, and epilepsy. Schatz's team, part of the Telomere-to-Telomere Consortium, used it to reanalyze thousands of human genomes with the new reference genome to discover more than 1 million new variants.

Already, the AnVIL team has collected petabytes of data from several of

the largest NHGRI projects, including hundreds of thousands of genomes from the Genotype-Tissue Expression (GTEx), Centers for Mendelian Genetics (CMG), and Centers for Common Disease Genomics (CCDG) projects, with plans to host many more projects in the near future.

The AnVIL team includes researchers from Johns Hopkins University, the Broad Institute of MIT and Harvard, Harvard University, Vanderbilt University, the University of Chicago, Oregon Health and Sciences University, Yale University School of Medicine, the University of California, Santa Cruz, Roswell Park Comprehensive Cancer Institute, the Pennsylvania State University, the City University of New York, the Carnegie Institute, and Washington University in St. Louis.

More information: Michael C. Schatz, Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL), *Cell Genomics* (2022). DOI: 10.1016/j.xgen.2021.100085. www.cell.com/cell-genomics/ful … 2666-979X(21)00106-3

Provided by Johns Hopkins University