Genomic Data Commons provides unprecedented cancer data resource

26 February 2021

Credit: CC0 Public Domain

The National Cancer Institute's Genomic Data Commons (GDC), launched in 2016 by then-Vice President Joseph Biden and hosted at the University of Chicago, has become one of the largest and most widely used resources in cancer genomics, with more than 3.3 petabytes of data from more than 65 projects and over 84,000 anonymized patient cases, serving more than 50,000 unique users each month.

In new papers published Feb. 22 in *Nature Communications* and *Nature Genetics*, the UChicago-based research team shares new details about the GDC, which is funded by the National Cancer Institute (NCI), via subcontract with the Frederick National Laboratory for Cancer Research, currently operated by Leidos Biomedical Research, Inc. One of the papers describes the design and operation of the GDC. The other describes the pipelines used by the GDC for the harmonization of data submitted to the GDC and the generation of datasets used by the GDC research community.

The goal of the GDC is to provide the cancer research community with a data repository of uniformly processed genomic and associated clinical data that enables data sharing and collaborative analysis in the support of precision medicine.

Data production for what would become the GDC began in June 2015 using a private cloud. After just a year, the GDC had analyzed more than 50,000 raw sequencing data inputs. The GDC includes genomic, transcriptomic, epigenomic, proteomic, clinical, and imaging data. The processing pipelines described in the Nature paper have produced more than 1,660 TB of data on more than two dozen types of primary cancers. These data are stored within the GDC Data Portal, where they are available for viewing and downloading.

Along with the data portal, the GDC also offers additional user resources, including the GDC Data Analysis, Visualization, and Exploration (DAVE) Tools for interactive exploration of data by genomic variant or specific alteration; the GDC Data Submission Portal for submitting data; the GDC Data Transfer Tool (DTT) for downloading large genomic datasets; and the GDC data harmonization system, which allows users to run data submitted to the GDC through the harmonizing processing pipelines.

"These data have a critical role to play," said Robert Grossman, Ph.D., principal investigator for the GDC and director of the Center for Translational Data Science at UChicago. "As data accumulates, new signals will become easier to identify as important targets for understanding cancer biology. In addition, the data-sharing infrastructure can serve to inform research studies, providing new insight into genetic variation between individuals and how it may affect cancer patient outcomes."

More information: Zhenyu Zhang et al, Uniform genomic data analysis in the NCI Genomic Data

Commons, *Nature Communications* (2021). DOI: 10.1038/s41467-021-21254-9

Allison P. Heath et al. The NCI Genomic Data Commons, *Nature Genetics* (2021). DOI: 10.1038/s41588-021-00791-5

Provided by University of Chicago Medical Center