NCI Genomic Data Commons FAQ

What is the National Cancer Institute Genomic Data Commons?

The National Cancer Institute (NCI) Genomic Data Commons (GDC) represents a partnership between the NCI and the University of Chicago to solve a pressing issue in cancer research. The GDC is a first-of-its-kind facility that comprehensively stores cancer genomic data into a single repository after it’s generated through NCI-funded research programs. These data, representing thousands of cancer patients and tumors, will be harmonized so that they are compatible, accessible and broadly useful to scientists. The GDC will provide an interactive system for researchers to access data, with the goals of advancing the molecular diagnosis of cancer and suggesting potential therapeutic targets based on genomic information.

Who will lead the Genomic Data Commons?

The head of the GDC is Robert Grossman, PhD, Director of the Center for Data Intensive Science (CDIS) and professor in the Department of Medicine at the University of Chicago.

Why is such a project needed?

Mutations to DNA and other genomic abnormalities play a central role in cancer formation and growth. To identify the relevant genetic markers and pave the way for new diagnostics and treatments, the NCI has funded several large research projects that have collected genomic data on nearly two dozen tumor types from more than 10,000 patients.

However, most researchers cannot fully use this wealth of information. That’s because the data are stored in different locations, with different software and different management systems. They are gathered by different research groups, using different technologies and protocols. Due to the data’s sheer size, disparate formats and dispersed storage locations, a cancer researcher needs millions of dollars, a concerted effort and a dedicated team to set up the necessary infrastructure to store all the data and to analyze it. This impedes research at all but the largest institutions, and stymies collaboration.

Where are the data that will be included in the Genomic Data Commons?

Initially, the GDC will standardize and house data sets including the entirety of large-scale NCI programs such as The Cancer Genome Atlas (TCGA), Therapeutically Applicable Research to Generate Effective Treatments (TARGET) and the Cancer Genome Characterization Initiative (CGCI). These programs represent about 2.2 petabytes of legacy data. Meanwhile, the GDC which will add another petabyte or more of storage each year to accommodate new NCI projects.

How will the Genomic Data Commons help researchers?
By eliminating a major chokepoint and facilitating access to vast quantities of data, the GDC opens new research frontiers in data-driven cancer research. Regardless of institution or laboratory size, researchers from across the country will be able to explore new ideas and engage in previously unfeasible collaborative efforts. For example, large-scale studies will be greatly accelerated, particularly research into cancer genomes used to identify important mutations or genomic patterns, in addition to comparisons of different cancer types across different groups of people.

**What could the Genomic Data Commons mean for patients?**

The GDC is the first step toward developing a knowledge system for cancer, which was originally recommended in a 2011 Institute of Medicine report called “Toward Precision Medicine.” The panel recommended the creation of a single data repository, saying such a system would provide an essential infrastructure for integrating basic biological knowledge with medical histories and health outcomes of individual patients. A primary goal of the GDC is to eventually allow any doctor the ability to upload and compare tumors from individual patients against the thousands of others in the GDC to find the most effective drug or course of treatment for that tumor genome.

**How will the Genomic Data Commons aid cloud-based technological initiatives?**

The GDC will serve as an important foundation for future cloud-based technological initiatives, which will allow researchers to compute and analyze data remotely. These will be most effective when they are linked to a single, centralized repository of harmonized data.

**Can Genomic Data Commons technology be applied toward diseases besides cancer?**

GDC technology and software remains open source so that they can one day be applied to conditions such as neurodegenerative diseases and diabetes. Researchers have gathered enormous volumes of data on many of these disorders, and need the same kind of data harmonization and centralization.

**How will the data be secured?**

The GDC will operate system controls to ensure the integrity, confidentiality and availability of data at a level of compliance called FISMA Moderate. The data will be hosted in a secure, compliant infrastructure and role-based authentication and authorization will be employed.

**Where will the Genomic Data Commons be located?**

The GDC will be located at the University of Chicago Kenwood Data Center. The GDC will have an 80 Gbps connection to the UChicago Science DMZ network, which is currently connected to the Internet2 high-performance research network, the Metropolitan Research and Education Network (MREN) and the Energy Sciences Network (ESnet) at the StarLight facility in Chicago via 100 Gbps connections.
When will the Genomic Data Commons be fully operational?

The GDC will be established in phases and is planned to be fully operational by 2016.

The GDC will be constructed and operated with NCI funding through a subcontract from Leidos Biomedical Research, Inc. at the Frederick National Laboratory for Cancer Research. The Ontario Institute of Cancer Research is developing some of the components of the GDC through a subcontract with the University of Chicago.