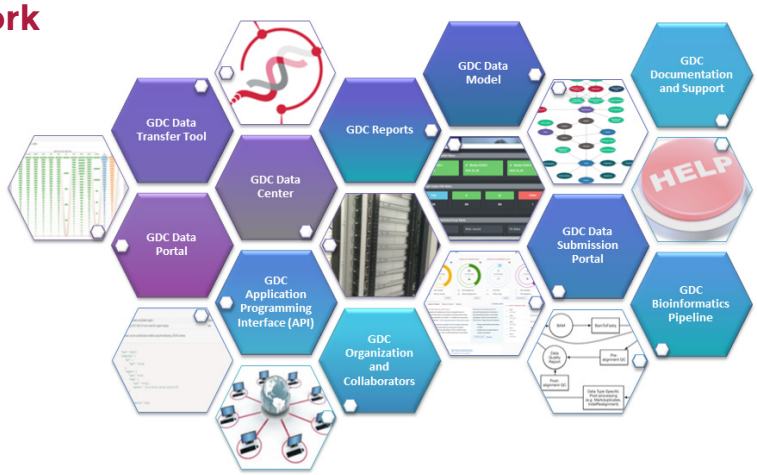
NIH NATIONAL CANCER INSTITUTE

Center for Cancer Genomics
Genomic Data Commons

Next Generation Cancer Knowledge Network



National Cancer Institute's (NCI's) Genomic Data Commons (GDC) is a next generation cancer knowledge network that supports the hosting and standardization of genomic and clinical data from cancer research programs, the harmonization of raw sequence data, and the application of state-of-the art methods for generating high level data (e.g. mutation calls, structural variants, etc.). The NCI Center for Cancer Genomics (CCG) of the National Institutes of Health (NIH) established the GDC to provide the cancer research community with a data service supporting the receipt, quality control, integration, storage, and redistribution of standardized cancer genomic data sets derived from cancer studies.

Mission and Goals

The mission of the GDC is to provide the cancer research community with a unified data repository that enables data sharing across cancer genomic studies in support of precision medicine.

Working towards this mission, the GDC aims to provide a cancer knowledge network that enables the identification of both high- and low-frequency cancer drivers, assists in defining genomic determinants of response to therapy, and informs the composition of clinical trial cohorts sharing targeted genetic lesions.

Resources

To achieve the development of a unified data repository, the GDC provides the community with several resources for retrieving data from the GDC, submitting data to the GDC, and processing data through the GDC bioinformatics pipelines. Resources are maintained in a secure data center with provided user support and documentation.

Primary GDC resources include:

- GDC Data Portal – A robust data-driven platform that allows users to search and download cancer data sets for analysis using web technologies. It also provides access to GDC Reports on data statistics.
- GDC Data Transfer Tool – A performance efficient utility for the download and upload of large, high volume data.
- GDC Application Programming Interface (API) - Programmatic interface to GDC data for consumption by third party applications.
- GDC Data Submission Portal – A user-friendly web-based tool for submitting clinical, biospecimen, and small volumes of molecular data.
- GDC Bioinformatics Pipelines - Pipelines supporting the harmonization of DNA and RNA sequence data and the generation of high level data for DNA sequence variants, mutation analyses, SNP chip genotypes, and expression analyses.
- GDC Data Model – Graphical representation of GDC data elements (e.g. sample → portion → analyte → aliquot) and their relationships.
- GDC Data Center – Secure FISMA Moderate operations supporting the storage, processing, and distribution of GDC data. The GDC Data Center maintains NIH Trusted Partner status and leverages dbGaP data access policies.

GDC resources were developed by several organizations with valuable contributions from community bioinformatics leaders.

Data Sets and Data Types

GDC established standard data types and file formats for the submission of core clinical, biospecimen, and molecular data and the generation of higher-level derived data.

Core Data Type	File Format
Clinical and Biospecimen	JSON and tab-delimited
Sequencing (DNA, mRNA, miRNA)	BAM / FASTQ (raw), NCBI SRA 1.5 (metadata)
Variants and Mutations	VCF / MAF
Expression (Gene, Exon, miRNA)	.txt

The GDC supports the retrieval of these and several other auxiliary data types from cancer programs such as: **TCGA** - *The Cancer Genome Atlas*, **TARGET** - *Therapeutically Applicable Research to Generate Effective Treatments*, and **CGCI** - *Cancer Genome Characterization Initiative*, and several other NCI and non-NCI programs that contribute genomic data for cancer research.

Benefits

The GDC provides the research community with the following benefits:

- Access to high-quality standardized clinical, biospecimen, and molecular data
- Resources supporting the performance efficient download and upload of the GDC data
- Web-based tools supporting fine-grained queries, advanced visualization, smart search technologies, and personalized download facilities
- User-friendly data submission tools for validating and submitting data into the GDC in support of data sharing
- Data harmonization pipelines supporting DNA and RNA sequence harmonization against a common reference genome
- Data generation pipelines supporting the high level data generation of DNA sequence variants, mutation analyses, SNP chip genotypes and expression analyses
- Programmatic interfaces supporting data retrieval, download, and submission by third party applications
- Interfaces to eRA Commons and dbGaP for secure access to controlled data sets

Additional Information

For additional information on the GDC, please visit the GDC web site:
https://gdc.cancer.gov

For information on the CCG and CCG supported programs, please visit the CCG Programs Site:
http://www.cancer.gov/aboutnci/organization/ccg/programs