NIH NATIONAL CANCER INSTITUTE

# Center for Cancer Genomics | **Genomic Data Commons**

## Next Generation Cancer Knowledge Network

The National Cancer Institute (NCI) Genomic Data Commons (GDC) is a next generation cancer knowledge network established by the Center for Cancer Genomics (CCG) to support the hosting, standardization, and distribution of genomic and clinical data from cancer research programs.

The GDC harmonizes raw sequence data, applies state-of-the-art methods for generating high-level data such as mutation calls and structural variants, and provides scalable downloads and web-based analysis tools.


## Mission & 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.**

The GDC provides 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 that share targeted genetic lesions.

## Resources

As a unified data repository, the GDC provides the community with several resources for retrieving data, submitting data, and visualizing custom analyses. Resources are maintained in a secure data center with provided user support and documentation.

Primary GDC components include:

- **GDC Data Portal** – A robust data-driven platform that allows users to search and download cancer data sets for analysis using web technologies. https://portal.gdc.cancer.gov
- **GDC Data Transfer Tool** – A performance efficient utility for the download and upload of large, high volume data. https://gdc.cancer.gov/access-data/gdc-data-transfer-tool
- **GDC Application Programming Interface (API)** – Programmatic interface to GDC data for consumption by third party applications. https://gdc.cancer.gov/developers/gdc-application-programming-interface-api
- **GDC Data Submission Portal** – A user-friendly web-based tool for submitting clinical, biospecimen, and molecular data. https://gdc.cancer.gov/submit-data
- **GDC Data Analysis, Visualization, and Exploration (DAVE) Tools** – Interactive tools supporting web-based gene- and variant-level analyses. Researchers may build custom cohorts and visualize most frequently mutated genes, genes affected by high impact mutations, overall patient survival by affected gene, and mutations mapped to protein-coding regions without downloading files. https://portal.gdc.cancer.gov/projects

## Data Sets & Data Types

In the establishment of the GDC, standard data types and file formats for the submission of clinical, biospecimen, and molecular data, and the generation of higher-level derived data were developed. The GDC Data Dictionary provides details of GDC supported data sets and data types: https://docs.gdc.cancer.gov/Data_Dictionary/viewer

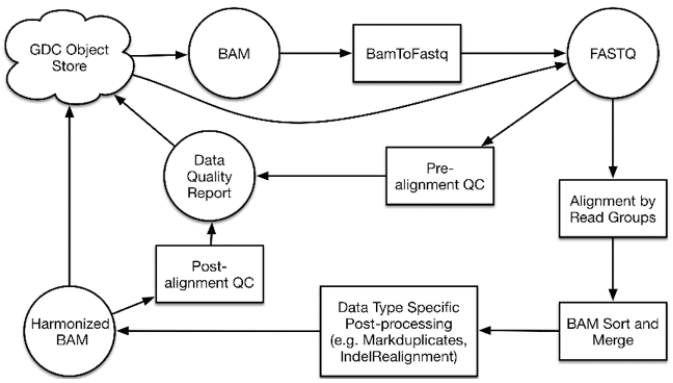
| Core Data Type | File Format |
|---|---|
| Clinical and Biospecimen | JSON and tab-delimited |
| Sequencing (DNA, mRNA, miRNA) | BAM / FASTQ (raw), NCBI SRA 1.5 (metadata) |
| Variants and Mutations | VCF / MAF |
| Expression (Gene, Exon, miRNA) | .txt |

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, **FMI** - Foundation Medicine, and **CCLE** – Cancer Cell Line Encyclopedia.

## Bioinformatics Pipelines

The GDC uses submitted FASTQ or BAM formatted sequence and microarray data to generate derived analysis data. This includes analyses such as tumor sequence variant calls, RNA-Seq gene expression quantification values, and copy-number segmentation values. GDC bioinformatics pipelines support sequence alignment to the latest reference genome to produce harmonized genomic data.




Details of the GDC bioinformatics pipelines are available at: https://gdc.cancer.gov/about-data/data-harmonization-and-generation/genomic-data-harmonization-0

## Benefits

The GDC provides the research community with the following benefits:

- Access to high-quality clinical, biospecimen, and molecular data
- Resources supporting the performance efficient download and upload of 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 in support of data sharing
- Interactive data analysis, visualization, and exploration tools promoting development of a true cancer genomics knowledge base
- 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, and expression analyses
- Programmatic interfaces supporting data retrieval, download, and submission, and BAM slicing by 3rd party apps
- 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 Website: https://gdc.cancer.gov or contact GDC Support: support@nci-gdc.datacommons.io

For information on CCG and CCG supported programs, please visit the CCG Website: https://www.cancer.gov/ccg.

Follow @NCIGDC_Updates