National Cancer Institute
National Human Genome Research Institute

The Cancer Genome Atlas Data:
Navigating the Data Portal and the Cancer Genomics Hub

The Cancer Genome Atlas

http://cancergenome.nih.gov/

The Cancer Genome Atlas (TCGA) is a large-scale, collaborative effort led by the National Institutes of Health to map the genomic changes that occur in over 30 types of human cancer, including nine rare tumors. Its goal is to support new discoveries and accelerate the pace of research aimed at improving the diagnosis, treatment, and prevention of cancer.

TCGA is a community resource project. The information generated by TCGA is centrally managed and entered into databases as it becomes available, making the data rapidly accessible to the entire research community. By January 2014, TCGA had generated one petabyte of data for about 10,000 cases of tumor and matching normal tissue samples.

TCGA data are available in two data repositories: the TCGA Data Portal and the Cancer Genomics Hub. All data can be accessed directly from the TCGA Data Portal regardless of which repository houses the data file.
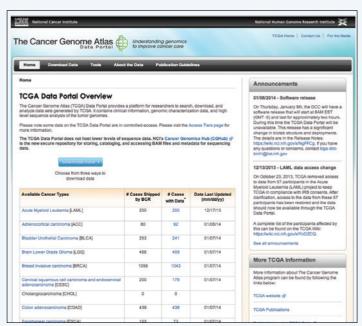
The TCGA Data Portal

A publicly available platform for researchers to search and download datasets generated by TCGA.
https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp

Data Types

Data from both tumor and matching normal samples are available. Data types include:

- Exome (variant analysis)
- SNP
- Methylation
- mRNA
- miRNA
- Deidentified patient clinical information

Data Levels

Data are classified according to level of analysis. The three main data levels are:

- ❍ **Level 1** – Raw Data: raw data for individual samples
- ❍ **Level 2** – Normalized/Processed Data: normalized data for individual samples
- ❍ **Level 3** – Interpreted Data: aggregate of processed data for individual samples

Data Access

There are two tiers of data access:

- ❍ **Open Access** – public data not unique to an individual. This tier does not require user certification.
- ❍ **Controlled Access** – data that may be unique to an individual. This tier requires user certification to be reviewed by TCGA's Data Access Committee. New users should review information available at the National Center for Biotechnology Information's (NCBI) Database of Genotypes and Phenotypes (dbGaP): https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?login=&page=login

Downloading the Data

There are three ways to download data:

- ❍ **Data Matrix:** Users select and download subsets of data based on specific criteria
- ❍ **Bulk Download:** Users search and download archives of data as uploaded by the TCGA centers
- ❍ **HTTP Directories:** Users access the HTTP directories directly where the data archives are stored

Analytical Tools

TCGA also provides analytical tools to assist in overcoming the limitations of current methods for analyzing large, complex datasets. These tools include:

- ❍ The Cancer Imaging Archive
- ❍ Integrative Genomics Viewer
- ❍ cBioPortal for Cancer Genomics
- ❍ UCSC Cancer Genomics Browser
- ❍ Berkeley Morphometric Visualization and Quantification from H&E sections
- ❍ Broad GDAC Firehose
- ❍ MD Anderson GDAC MBatch
- ❍ Regulome Explorer

Responsible Data Use and Publication Guidelines

Users are strongly encouraged to visit the data use policy and publication guidelines for information on how to use TCGA data in publications.

- ❍ **Responsible Data Use Policy:**
  http://cancergenome.nih.gov/abouttcga/policies/responsibleuse

- ❍ **Publication Guidelines:**
  http://cancergenome.nih.gov/publications/publicationguidelines

The Cancer Genomics Hub (CGHub)

A secure repository for storing, cataloging, and accessing TCGA's lower-level sequence data
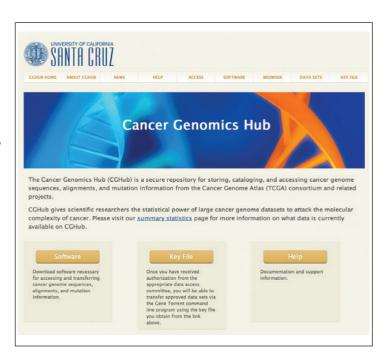https://cghub.ucsc.edu/

Data Types

Data from both tumor and matching normal samples are available. Data types include primary genome sequences and alignments for:

- ❍ DNA
- ❍ RNA
- ❍ miRNA

Data Access

All of the data generated from TCGA samples in CGHub are under Controlled Access. Metadata are the exception, which are Open Access. New users must apply for authorized access through NCBI's Database of Genotypes and Phenotypes (dbGaP).

Downloading the Data

There are two ways to download the data:

- ❍ **Data Browser:** This web application has a user interface that works like an online store. Users can search for data using filters, view subsets of data, and add desired items to a "shopping cart" for downloading.

- ❍ **Key File**: This web service allows users to transfer approved datasets via the Gene Torrent command line program.

CGHub is located at the University of California, Santa Cruz.

TCGA continues to work towards fulfilling its promise to the cancer research community:

- ❍ **Providing the highest quality cancer genomics data as a community resource**
- ❍ **Improving user access and experience**

Stay Connected!

Sign up for

- ❑ **TCGA updates**: http://cancergenome.nih.gov/abouttcga/peoplecontacts/emailsignup
- ❑ **Data Portal updates**: TCGA-Data-L@list.nih.gov
  (https://list.nih.gov/cgi-bin/wa.exe?A0=TCGA-DATA-L)
- ❑ **CGHub updates**: support@cghub.ucsc.edu

Find Out More about TCGA

For more information about
The Cancer Genome Atlas and accessing TCGA data, please contact:

The Cancer Genome Atlas
http://cancergenome.nih.gov/
National Cancer Institute at NIH
31 Center Drive
Building 31, Suite 3A20
Bethesda, MD 20892
Phone: (301) 594-9831
E-mail: tcga@mail.nih.gov

For additional information about the sponsoring institutes,
please visit the following websites:

National Cancer Institute
http://cancer.gov

National Human Genome Research Institute
http://genome.gov

U.S. Department of Health and Human Services ❘ National Institutes of Health

National Cancer Institute ❘ National Human Genome Research Institute