

TCGA DATA PORTAL

User's Guide



**NATIONAL[®]
CANCER
INSTITUTE**

Center for Biomedical
Informatics and Information

CONTENTS

Chapter 1

Using the TCGA Data Portal	3
Getting Started: Search Parameters	3
Basics of Performing a Search	6
Selecting Multiple Query Terms Within a Search Parameter	8
Retrieving Query Results	9
Data Availability within the Portal	11
Cite Data	11
Additional Help	11

CHAPTER 1

USING THE TCGA DATA PORTAL

Getting Started: Search Parameters

The Cancer Genome Atlas (TCGA) Data Portal Search Engine allows users to choose five different parameters to retrieve data. The five parameters are:

1. Cancer Type

TCGA Pilot Project is studying three cancers:

- brain cancer, listed as glioblastoma multiforme (GBM),
- lung squamous adenocarcinoma, abbreviated as squamous carcinoma (LG) and
- ovarian serous cystadenocarcinoma, abbreviated as serous cystadenocarcinoma (OV).

2. Center

There are ten TCGA research centers that are depositing data into the Data Portal.

Those ten centers are:

- International Genomics Consortium and Translational Genomics Research Institute Biological Core Resource, denoted as IGC Biological Core Resource
- Brigham and Women's Hospital of Harvard Medical School and Dana Farber Cancer Institute, denoted as Harvard Medical School
- The Eli and Edythe L. Broad Institute of the Massachusetts Institute of Technology (MIT) and Harvard University and the Dana Farber Cancer Institute, denoted as Broad Institute of MIT and Harvard
- Lawrence Berkeley National Laboratory

- Johns Hopkins University and University of Southern California joint group, denoted as Johns Hopkins/University of Southern California
- Memorial Sloan-Kettering Cancer Center
- Stanford University School of Medicine, denoted as Stanford University
- Baylor College of Medicine, Human Genome Sequencing Center, denoted as Baylor College of Medicine
- Washington University School of Medicine, Genome Sequencing Center, denoted as Washington University School of Medicine
- University of North Carolina, Lineberger Comprehensive Cancer Center, denoted as University of North Carolina

3. Platform

In the TCGA Pilot Project, data is being generated by genomic characterization and sequencing platforms, described below:

Full Platform Name
Affymetrix HT Human Genome U133 Array Plate Set
Affymetrix Human Exon 1.0 ST Array
Affymetrix Genome-Wide Human SNP Array 6.0
Agilent 8 x 15K Human miRNA-specific Microarray
Agilent Human Genome CGH Microarray 244A
Agilent Whole Human Genome Microarray Kit, 4 x 44K
Illumina DNA Methylation OMA002 Cancer Panel I
Illumina 550K Infinium® HumanHap550 SNP Chip
Biospecimen Metadata – Complete Set
Biospecimen Metadata – Minimal Set
Applied Biosystems Sequence data

Table 1.1 Genomic Characterization and Sequencing Platforms in the TCGA Pilot Project

4. Data Type

Users can also search the Data Portal based on the different types of data that are generated by TCGA research centers. Clinical and genomic data available within the database are outlined below:

Data Type	Content	Associated Platforms
All	All data types employed by TCGA centers	All
Complete Clinical Set	Detailed clinical phenotype and outcome data	N/A

Table 1.2 Clinical and Genomic Data Available in the TCGA Data Portal

Data Type	Content	Associated Platforms
Minimal Clinical Set	Clinical Diagnosis, Histologic Diagnosis, Pathologic Status, Tissue Anatomic Site	N/A
Expression-Exon	Exon Expression Profiling	Affymetrix Human Exon 1.0 ST Array
Expression-miRNA	miRNA Expression Profiling	Agilent Human Genome CGH Microarray 244A
Methylation	DNA methylation patterns within the genome	Illumina DNA Methylation OMA002 Cancer Panel I
SNP	Single Nucleotide Polymorphisms, raw genotype calls, computed genotype frequencies	Illumina 550K Infinium® HumanHap550 SNP Chip, Affymetrix Genome-Wide Human SNP Array 6.0
Copy Number Results	Interpreted copy number and loss of heterozygosity data	Affymetrix Genome-Wide Human SNP Array 6.0, Illumina 550K Infinium® HumanHap550 SNP Chip, Agilent Whole Human Genome Microarray Kit 4 x 44K, Agilent Human Genome CGH Microarray 244A
Somatic Mutations	Somatic variants compiled from Genomic Sequencing Centers	Applied Biosystems Sequence data
Trace-Gene-Sample Relationship	Trace files with NCBI-required annotation linked to gene-sample identifier	Applied Biosystems Sequence data
Sequencing Trace	Trace files with NCBI-required annotations	Applied Biosystems Sequence data
SNP Frequencies	Frequencies of SNPs over many samples and/or patients	Illumina 550K Infinium® HumanHap550 SNP Chip, Affymetrix Genome-Wide Human SNP Array 6.0
Expression-Genes	Transcription profiling based upon genes and some unidentified transcripts	Affymetrix HT Human Genome U133 Array Plate Set, Agilent Whole Human Genome Microarray Kit 4 x 44K

Table 1.2 Clinical and Genomic Data Available in the TCGA Data Portal (Continued)

5. Data Submitted After

The “Date Submitted After” search parameter allows users to search and retrieve data based upon a date. A query that includes a date entered in this field will return all data that was deposited into the Portal from the entered date to the present date. Users should enter a date in the following format: mm/dd/yyyy; for example: 07/04/2007.

Basics of Performing a Search

Users can select to query one or all five of these search parameters (i.e.: cancer type, center, platform, data type, data submitted after). For searches where more than one search parameter is selected, the query will produce results that are the intersection of the selected search parameters.

For example, when a user selects: Cancer type: Glioblastoma multiforme (GBM), the query will return all data for glioblastoma multiforme.

Describe your search constraints. The search will return the list of archives that satisfy all of the constraints.

[For HELP with search constraints click here.](#)

Cancer Type
 All
 Glioblastoma multiforme (GBM)
 Serous cystadenocarcinoma (OV)
 Squamous carcinoma (LG)

Center
 All
 Baylor College of Medicine
 Broad Institute of MIT and Harvard
 Harvard Medical School
 IGC Biospecimen Core Resource

Platform
 All
 Affymetrix HT Human Genome U133 Array Plate Set
 Affymetrix Human Exon 1.0 ST Array
 Affymetrix Genome-Wide Human SNP Array 6.0
 Agilent Human Genome CGH Microarray 244A

Data Type
 All
 Expression-Genes
 Expression-Exon
 Expression-miRNA
 Copy Number Results

Submission Date
 1/1/07 - 8/14/07
 On or After Before

Reset Find

Figure 1.1 Search Data dialog box with Cancer Type selected

When a user selects the following search parameters:

Tumor Type: Glioblastoma multiforme (GBM)

Center: Lawrence Berkeley National Laboratories

Data Type: Expression - Exon,

Overview Types of Data Cite Data Data Access **Search Data**

Describe your search constraints. The search will return the list of archives that satisfy all of the constraints.

[For HELP with search constraints click here.](#)

Cancer Type
 All
 Glioblastoma multiforme (GBM)
 Serous cystadenocarcinoma (OV)
 Squamous carcinoma (LG)

Center
 Harvard Medical School
 IGC Biospecimen Core Resource
 Johns Hopkins / University of Southern California
 Lawrence Berkeley National Laboratory
 Memorial Sloan-Kettering Cancer Center

Platform
 All
 Affymetrix HT Human Genome U133 Array Plate Set
 Affymetrix Human Exon 1.0 ST Array
 Affymetrix Genome-Wide Human SNP Array 6.0
 Agilent Human Genome CGH Microarray 244A

Data Type
 All
 Expression-Genes
 Expression-Exon
 Expression-miRNA
 Copy Number Results

Submission Date 1/1/07 - 8/14/07
 On or After Before

Reset Find

Figure 1.2 Search Data dialog box with Cancer Type, Center, and Data Type selected

the query will return exon-based expression data generated by Lawrence Berkeley National Laboratories on glioblastoma multiforme samples.

Overview Types of Data Cite Data Data Access **Search Data**

Archives

2 results found, displaying 1 to 2

Archive	Added On	Center	Version	Cancer Type	Platform	Data Type	Status	Download
lbl.gov_GBM.HuEx-1_0-st-v2.1.1.0	2007-07-17	lbl.gov	1.1.0	Glioblastoma multiforme (GBM)	Affymetrix Human Exon 1.0 ST Array	Expression-Exon	Available - Open Access	Download MDS View Files
lbl.gov_GBM.HuEx-1_0-st-v2.2.1.0	2007-07-17	lbl.gov	2.1.0	Glioblastoma multiforme (GBM)	Affymetrix Human Exon 1.0 ST Array	Expression-Exon	Available - Open Access	Download MDS View Files

Search Parameters:

Date Range Start On: 1/1/07 End On: 8/6/07

Tumors * Glioblastoma multiforme (GBM)

Centers * Lawrence Berkeley National Laboratory

Data Types * Expression-Exon

Figure 1.3 Query results

Selecting Multiple Query Terms Within a Search Parameter

Within each of the five search parameters, users can select more than one option within each category by keeping the *Control* key depressed while clicking on the multiple options of their choice.

An example is illustrated below:

The screenshot shows the 'Search Data' tab of the TCGA Data Portal. The search area contains the following fields:

- Cancer Type:** A dropdown menu with 'All' selected. Other visible options are Glioblastoma multiforme (GBM), Serous cystadenocarcinoma (OV), and Squamous carcinoma (LG).
- Center:** A dropdown menu with 'Baylor College of Medicine' selected. Other visible options are Broad Institute of MIT and Harvard, Harvard Medical School, IGC Biospecimen Core Resource, and Johns Hopkins / University of Southern California.
- Platform:** A dropdown menu with 'All' selected. Other visible options are Affymetrix HT Human Genome U133 Array Plate Set, Affymetrix Human Exon 1.0 ST Array, Affymetrix Genome-Wide Human SNP Array 6.0, and Agilent Human Genome CGH Microarray 244A.
- Data Type:** A dropdown menu with 'All' selected. Other visible options are Expression-Genes, Expression-Exon, Expression-miRNA, and Copy Number Results.
- Submission Date:** Two date input fields with '1/1/07' and '8/14/07' entered. Below them are radio buttons for 'On or After' and 'Before'.

At the bottom of the search area are 'Reset' and 'Find' buttons. A link for help is also present: 'For HELP with search constraints click here.'

Figure 1.4 Selecting Multiple query terms within a search parameter

Retrieving Query Results

After selecting the search options, click **Find** to retrieve query results. Query results are returned in a tabular format, with each row containing separate data files. An example is shown below.

Added On	Center	Version	Cancer Type	Platform	Data Type	Status	Download
G-U133A.1.1.0	2007-07-17 broad.mit.edu	1.1.0	Glioblastoma multiforme (GBM)	Affymetrix HT Human Genome U133 Array Plate Set	Expression-Genes	Available - Open Access	Download MDS View Files
G-U133A.2.0.0	2007-07-17 broad.mit.edu	2.0.0	Glioblastoma multiforme (GBM)	Affymetrix HT Human Genome U133 Array Plate Set	Expression-Genes	Available - Open Access	Download MDS View Files
3-CGH-244A.1.0.0	2007-07-17 hms.harvard.edu	1.0.0	Glioblastoma multiforme (GBM)	Agilent Human Genome CGH Microarray 244A	Copy Number Results	Available - Open Access	Download MDS View Files
v2.1.1.0	2007-07-17 ibi.gov	1.1.0	Glioblastoma multiforme (GBM)	Affymetrix Human Exon 1.0 ST Array	Expression-Exon	Available - Open Access	Download MDS View Files
v2.2.1.0	2007-07-17 ibi.gov	2.1.0	Glioblastoma multiforme (GBM)	Affymetrix Human Exon 1.0 ST Array	Expression-Exon	Available - Open Access	Download MDS View Files
244A.2.0.0	2007-07-17 mskcc.org	2.0.0	Glioblastoma multiforme (GBM)	Agilent Human Genome CGH Microarray 244A	Copy Number Results	Available - Open Access	Download MDS View Files
v2550K.2.0.0	2007-07-17 stanford.edu	2.0.0	Glioblastoma multiforme (GBM)	Illumina 550K Infinium HumanHap550 SNP Chip	SNP	Available - Controlled Access	Download MDS View Files
3-CGH-244A.2.0.0	2007-07-17 hms.harvard.edu	2.0.0	Glioblastoma multiforme (GBM)	Agilent Human Genome CGH Microarray 244A	Copy Number Results	Available - Open Access	Download MDS View Files

Figure 1.5 Retrieving query results

For each data file returned, nine parameters are denoted for the file. The parameters are described in the table below:

Parameter	Description
Archive	Name of the data file
Added on	Date the data file was added to the Portal
Center	Name of the institution that provided the data file; e.g. "Broad Institute of MIT and Harvard"

Table 1.3 Parameters for a Data File

Parameter	Description
Version	<p>Version of the data file; e.g.: 2.1.0</p> <p>The first number (from left to right) is a serially increasing index that represents the number of files transferred to the Portal.</p> <p>The second number is for revisions. If a center decided to make changes (add files, correct files, etc.) to a previously deposited file, they would increase the revision.</p> <p>The revision starts at zero.</p> <p>The third number represents the series and starts at zero. If a file is very large, a center could separate a file into many parts (or series) for transfer and then users would download all of the series to re-compile the entire file. In this case, the entire file would be represented by the same first 2 digits and the third digit (the series) would increase depending on how many parts the file was separated into.</p>
Cancer Type	Cancer type from which the data was generated; e.g., "Glioblastoma multiforme (GBM)"
Platform	Technology platform from which the data file was generated
Data Type	Specific type of genomic, clinical, or genetic characterization data contained in data file
Status	<p>Availability status of file for download.</p> <ul style="list-style-type: none"> • Available: The file was submitted, has passed quality control and is available for users to download. • In review: The data file has been submitted, but is not ready to download. The text for files "In review" will be yellow. • Open-Access: Files can be downloaded by all users. • Controlled-Access: To download file, users must agree to TCGA Data Use Certification and become authorized users through the Data Access Request process^a. • Text in red denotes Controlled-Access data.
Download	<p>Users can choose to download the file, download the associated MD5 file^b or view the file.</p> <p>Note:</p> <ul style="list-style-type: none"> • If "Download" and "MD5" text is red, the file contains controlled-access data files. • If "Download" and "MD5" text is yellow, the files are "In Review" and are not yet available for download. • If "Download" and "MD5" text is black, then files are open-access and available to download.

Table 1.3 Parameters for a Data File (Continued)

- a. More information on data access can be found at: <http://cancergenome.nih.gov/dataportal/data/access>

- b. MD5 files are used to verify the integrity of a transferred file. Users can do that by downloading the MD5 and the corresponding file of interest. A user would then create their own MD5 and compare it with the one downloaded. If they are the same, the file has maintained its integrity.

Data Availability within the Portal

Note: Data is being continually added to this database. Please check back frequently for updates and additional data submissions.

Cite Data

It is the intent of NCI and NHGRI to promote the dissemination of analyses of TCGA Dataset(s) as widely as possible. Approved Users are strongly encouraged to publish their results in peer-reviewed journals, but are asked to apply the normal standards of scientific etiquette when deciding to publish results based substantially on unpublished TCGA data.

NCI and NHGRI intend that TCGA genomic data are released as rapidly as possible after they are produced, with no restrictions on use. However, Approved Users will agree that any publication of results or analyses derived from the use of TCGA Datasets will acknowledge TCGA with the following sentence:

“The results published here are in whole or part based upon data generated by The Cancer Genome Atlas pilot project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions who constitute the TCGA research network can be found at <http://cancergenome.nih.gov>”.

Additional Help

For additional help, please contact the National Cancer Institute Center for Bioinformatics and Information Technology Application Support:

Website: <http://ncicb.nci.nih.gov/support>

Email: ncicb@pop.nci.nih.gov

Telephone: 301-451-4384 or toll free: 888-478-4423

