TCGA Data Overview

Xiaofei Zhou, Han Zhang and Meng Wang

The Ohio State University

February 17, 2016

Background

- Cancer is one of the biggest causes of deaths in the world.
- In 2012, about 8.2 million people globally died of cancer, accounting for 14.6 % all human deaths.
- In the US, one person dies from cancer every minute, that is, 1,500 deaths each day.
- Our goal is to control and conquer cancer.

The Cancer Genome Atlas (TCGA)

- Aims at increasing the understanding of the molecular basis of cancer through genome analysis technologies.
- Is supported by NCI and NHGRI.
- We have already learned:
  - certain regions of the genome are linked with several types of cancers.
  - these regions are usually contain genes involved in the pathways of cell apoptosis.
  - signatures – specific changes in human genome – allow us to tell one type of cancer from another. These signatures help doctors with diagnosis, treatments and/or prognosis of cancer.

TCGA Cancer Selection Criteria

TCGA selected cancers based on:

- overall public health impact
- poor prognosis
- availability of human tumor and matched-normal tissue samples that meet TCGA standards for
  - patient consent
  - quality
  - quantity

There are 34 types of cancer currently available under TCGA program, click this link to see what they are.

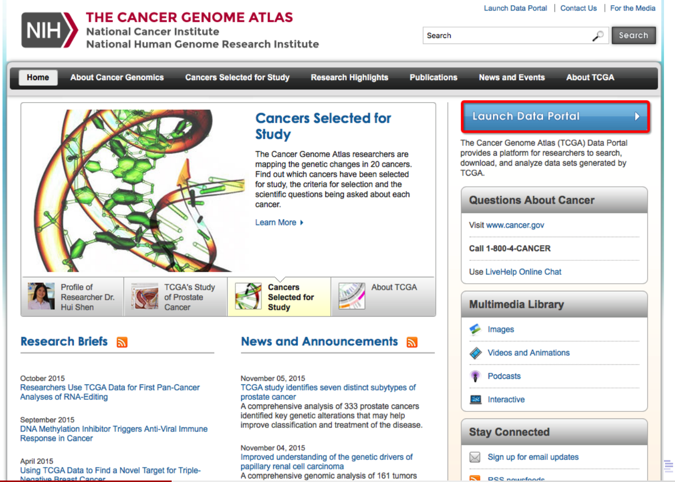
Accessing TCGA Data

TCGA data are available in two data repositories:

- The TCGA Data Portal:
  - provides access to almost all TCGA datasets, for example, SNP, copy number, methylation, expression, sequencing and so on.
- The Cancer Genomics Hub (CGHub):
  - links to TCGA primary sequence data.

Accessing TCGA Data

To view more general information about the TCGA program, please visit http://cancergenome.nih.gov/



Launch Data Portal | Contact Us | For the Media

NIH THE CANCER GENOME ATLAS
National Cancer Institute
National Human Genome Research Institute

Search [Search]

Home | About Cancer Genomics | Cancers Selected for Study | Research Highlights | Publications | News and Events | About TCGA

Cancers Selected for Study

The Cancer Genome Atlas researchers are mapping the genetic changes in 20 cancers. Find out which cancers have been selected for study, the criteria for selection and the scientific questions being asked about each cancer.

Learn More ▸

Profile of Researcher Dr. Hui Shen | TCGA's Study of Prostate Cancer | Cancers Selected for Study | About TCGA

Launch Data Portal ▸

The Cancer Genome Atlas (TCGA) Data Portal provides a platform for researchers to search, download, and analyze data sets generated by TCGA.

Questions About Cancer

Visit www.cancer.gov

Call 1-800-4-CANCER

Use LiveHelp Online Chat

Multimedia Library

- Images
- Videos and Animations
- Podcasts
- Interactive

Stay Connected

- Sign up for email updates
- RSS newsfeeds

Research Briefs

October 2015
Researchers Use TCGA Data for First Pan-Cancer Analyses of RNA-Editing

September 2015
DNA Methylation Inhibitor Triggers Anti-Viral Immune Response in Cancer

April 2015
Using TCGA Data to Find a Novel Target for Triple-Negative Breast Cancer

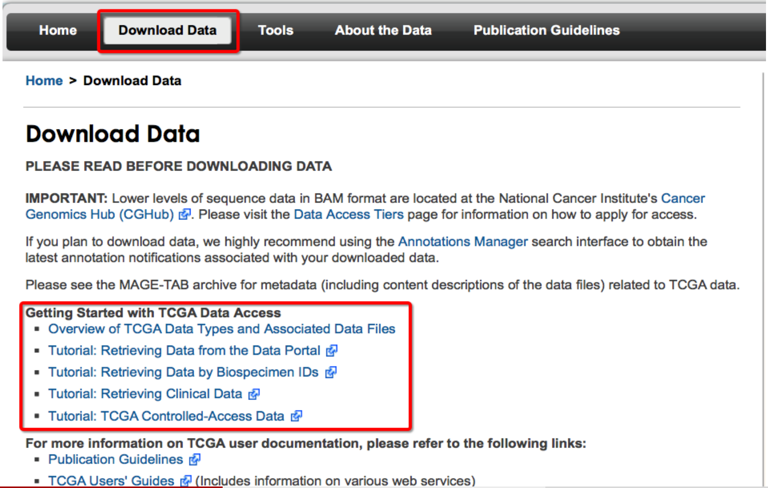
News and Announcements

November 05, 2015
TCGA study identifies seven distinct subtypes of prostate cancer
A comprehensive analysis of 333 prostate cancers identified key genetic alterations that may help improve classification and treatment of the disease.

November 04, 2015
Improved understanding of the genetic drivers of papillary renal cell carcinoma
A comprehensive genomic analysis of 161 tumors

Accessing TCGA Data

The tutorials under the `Download Data Section` in TCGA Data Portal can help you get familiar with methods of retrieving data.



Download Data

PLEASE READ BEFORE DOWNLOADING DATA

IMPORTANT: Lower levels of sequence data in BAM format are located at the National Cancer Institute's Cancer Genomics Hub (CGHub). Please visit the Data Access Tiers page for information on how to apply for access.

If you plan to download data, we highly recommend using the Annotations Manager search interface to obtain the latest annotation notifications associated with your downloaded data.

Please see the MAGE-TAB archive for metadata (including content descriptions of the data files) related to TCGA data.

Getting Started with TCGA Data Access

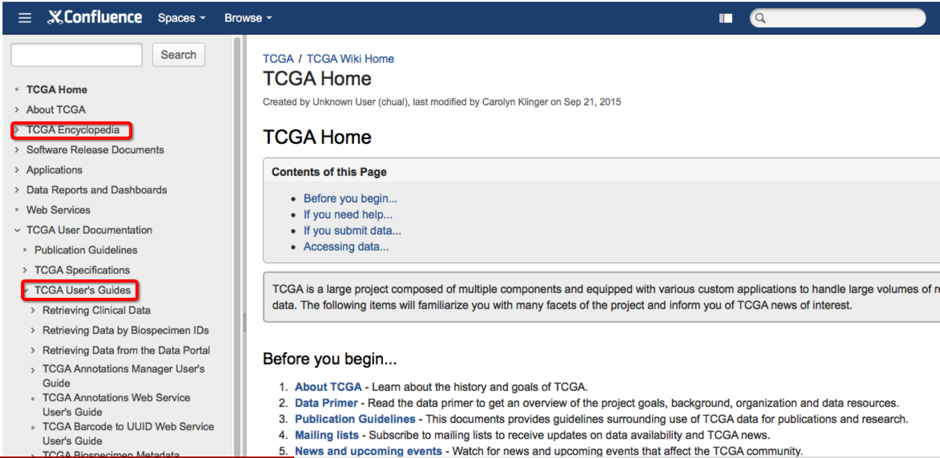
- Overview of TCGA Data Types and Associated Data Files
- Tutorial: Retrieving Data from the Data Portal
- Tutorial: Retrieving Data by Biospecimen IDs
- Tutorial: Retrieving Clinical Data
- Tutorial: TCGA Controlled-Access Data

For more information on TCGA user documentation, please refer to the following links:

- Publication Guidelines
- TCGA Users' Guides (Includes information on various web services)

TCGA Wiki

If you encounter unfamiliar terms when you access the data, go to https://wiki.nci.nih.gov/display/TCGA/TCGA+Home



- TCGA Home
- About TCGA
- TCGA Encyclopedia
- Software Release Documents
- Applications
- Data Reports and Dashboards
- Web Services
- TCGA User Documentation
  - Publication Guidelines
  - TCGA Specifications
  - TCGA User's Guides
    - Retrieving Clinical Data
    - Retrieving Data by Biospecimen IDs
    - Retrieving Data from the Data Portal
    - TCGA Annotations Manager User's Guide
    - TCGA Annotations Web Service User's Guide
    - TCGA Barcode to UUID Web Service User's Guide
    - TCGA Biospecimen Metadata

TCGA / TCGA Wiki Home

TCGA Home

Created by Unknown User (chual), last modified by Carolyn Klinger on Sep 21, 2015

TCGA Home

Contents of this Page

- Before you begin...
- If you need help...
- If you submit data...
- Accessing data...

TCGA is a large project composed of multiple components and equipped with various custom applications to handle large volumes of r data. The following items will familiarize you with many facets of the project and inform you of TCGA news of interest.

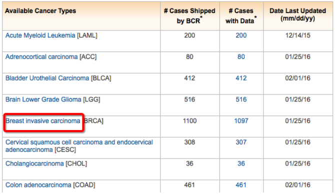
Before you begin...

1. **About TCGA** - Learn about the history and goals of TCGA.
2. **Data Primer** - Read the data primer to get an overview of the project goals, background, organization and data resources.
3. **Publication Guidelines** - This documents provides guidelines surrounding use of TCGA data for publications and research.
4. **Mailing lists** - Subscribe to mailing lists to receive updates on data availability and TCGA news.
5. **News and upcoming events** - Watch for news and upcoming events that affect the TCGA community.

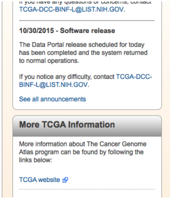
Accessing TCGA Data

Quick View on Available Cancer Data

First click here to see available cancer types. Then click on the cancer you are interested in.



Available Cancer Types	# Cases Shipped by BCR*	# Cases with Data*	Date Last Updated (mm/dd/yy)
Acute Myeloid Leukemia [LAML]	200	200	12/14/15
Adrenocortical carcinoma [ACC]	80	80	01/25/16
Bladder Urothelial Carcinoma [BLCA]	412	412	02/01/16
Brain Lower Grade Glioma [LGG]	516	516	01/25/16
Breast invasive carcinoma [BRCA]	1100	1097	01/25/16
Cervical squamous cell carcinoma and endocervical adenocarcinoma [CESC]	308	307	01/25/16
Cholangiocarcinoma [CHOL]	36	36	01/25/16
Colon adenocarcinoma [COAD]	461	461	02/01/16



If you have any questions or concerns, contact TCGA-DCC-BINF-L@LIST.NIH.GOV.

10/30/2015 - Software release

The Data Portal release scheduled for today has been completed and the system returned to normal operations.

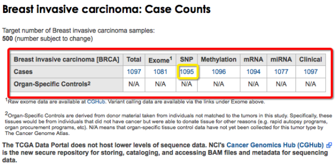
If you notice any difficulty, contact TCGA-DCC-BINF-L@LIST.NIH.GOV.

See all announcements

More TCGA Information

More information about The Cancer Genome Atlas program can be found by following the links below:

TCGA website



Breast invasive carcinoma: Case Counts

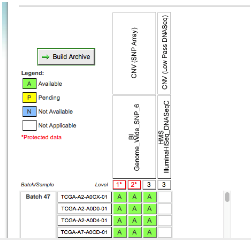
Target number of Breast invasive carcinoma samples:
500 (number subject to change)

Breast invasive carcinoma [BRCA]	Total	Exome[1]	SNP	Methylation	mRNA	miRNA	Clinical
Cases	1097	1081	1095	1096	1094	1077	1097
Organ-Specific Controls[2]	N/A	N/A	N/A	N/A	N/A	N/A	N/A

[1] Raw exome data are available at CGHub. Variant calling data are available via the links under Exome above.

[2] Organ-Specific Controls are derived from donor material taken from individuals not matched to the tumors in this study. Specifically, these tissues would be from individuals that did not have cancer but were able to donate tissue for other reasons (e.g. rapid autopsy programs, organ procurement programs, etc). N/A means that organ-specific tissue control data have not yet been collected for this tumor type by The Cancer Genome Atlas.

The TCGA Data Portal does not host lower levels of sequence data. NCI's Cancer Genomics Hub (CGHub) is the new secure repository for storing, cataloging, and accessing BAM files and metadata for sequencing data.



Legend:
- A Available
- P Pending
- N Not Available
- □ Not Applicable

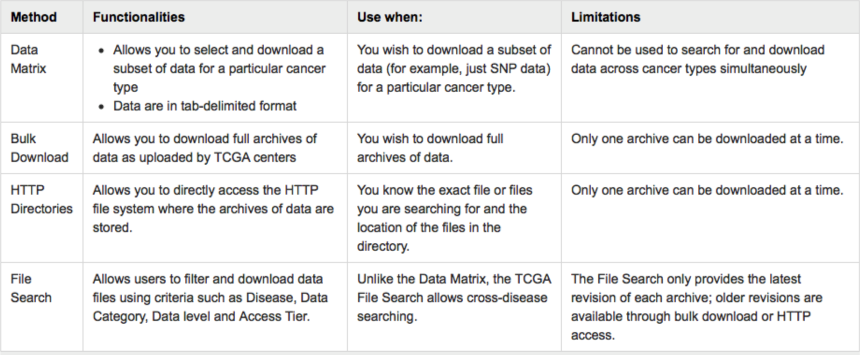
*Protected data

		CNV (SNP Array)			CNV (Low Pass DNASeq)
		BI Genome_Wide_SNP_6			HMS IlluminaHiSeq_DNASeqC
Batch/Sample	*Level*	1*	2*	3	3
Batch 47	TCGA-A2-A0CX-01	A	A	A	
	TCGA-A2-A0D0-01	A	A	A	
	TCGA-A2-A0D4-01	A	A	A	
	TCGA-A7-A0CD-01	A	A	A	

Accessing TCGA Data

TCGA Data Portal

We have 4 major methods to retrieve data through the TCGA Data Portal:



Method	Functionalities	Use when:	Limitations
Data Matrix	- Allows you to select and download a subset of data for a particular cancer type - Data are in tab-delimited format	You wish to download a subset of data (for example, just SNP data) for a particular cancer type.	Cannot be used to search for and download data across cancer types simultaneously
Bulk Download	Allows you to download full archives of data as uploaded by TCGA centers	You wish to download full archives of data.	Only one archive can be downloaded at a time.
HTTP Directories	Allows you to directly access the HTTP file system where the archives of data are stored.	You know the exact file or files you are searching for and the location of the files in the directory.	Only one archive can be downloaded at a time.
File Search	Allows users to filter and download data files using criteria such as Disease, Data Category, Data level and Access Tier.	Unlike the Data Matrix, the TCGA File Search allows cross-disease searching.	The File Search only provides the latest revision of each archive; older revisions are available through bulk download or HTTP access.

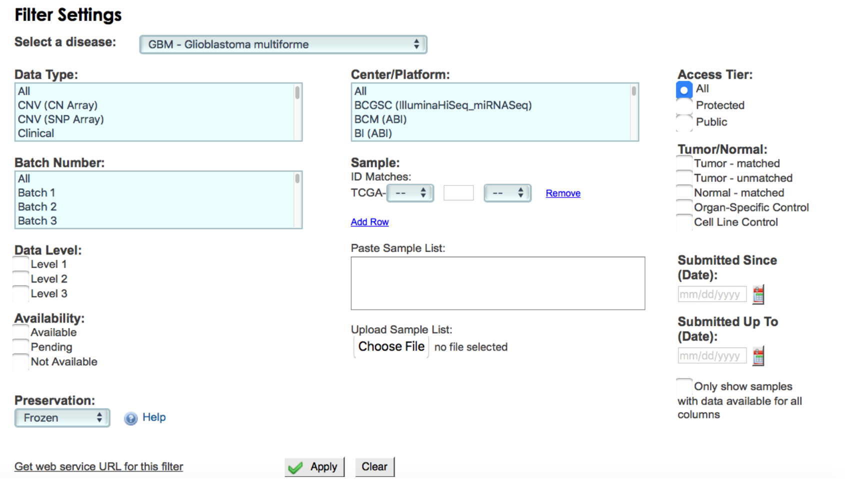
Accessing TCGA Data

TCGA Data Portal – Retrieving Data Using Data Matrix

- Data Matrix allows user to search for a subset of data of a particular cancer type.
- Limitation:
    - cannot search across multiple cancer types simultaneously.
    - only returns the most recent data files.
- `TCGA Data Portal Data Matrix Access link`
- `TCGA Data Portal Data Matrix Tutorial link`

Accessing TCGA Data

TCGA Data Portal – Retrieving Data Using the Data Matrix



Filter Settings

Select a disease: GBM - Glioblastoma multiforme

Data Type:
- All
- CNV (CN Array)
- CNV (SNP Array)
- Clinical

Batch Number:
- All
- Batch 1
- Batch 2
- Batch 3

Data Level:
□ Level 1
□ Level 2
□ Level 3

Availability:
□ Available
□ Pending
□ Not Available

Preservation: Frozen
Help

Center/Platform:
- All
- BCGSC (IlluminaHiSeq_miRNASeq)
- BCM (ABI)
- BI (ABI)

Sample:
ID Matches:
TCGA- -- -- Remove

Add Row

Paste Sample List:

Upload Sample List:
Choose File no file selected

Access Tier:
● All
○ Protected
○ Public

Tumor/Normal:
□ Tumor - matched
□ Tumor - unmatched
□ Normal - matched
□ Organ-Specific Control
□ Cell Line Control

Submitted Since (Date): mm/dd/yyyy

Submitted Up To (Date): mm/dd/yyyy

□ Only show samples with data available for all columns

Get web service URL for this filter

Accessing TCGA Data

TCGA Data Portal – Terminology

- **Data Type**: a label to categorize the many forms of platform data within TCGA Network.
- **Batch**: a set of related analytes from the same disease, or serial index.
- **Data Level:** 1 (Raw Data), 2 (Processed Data), 3 (Segmented or Interpreted Data) and 4 (Region of Interest Data).
- **Preservation**: the method used to preserve the sample after it has been removed from a participant.
- **Center**: a single or a collection of institutions and research centers that perform the same function within TCGA.
- **Platform**: a vendor-specific technology for assaying or sequencing.

Accessing TCGA Data

TCGA Data Portal – Terminology

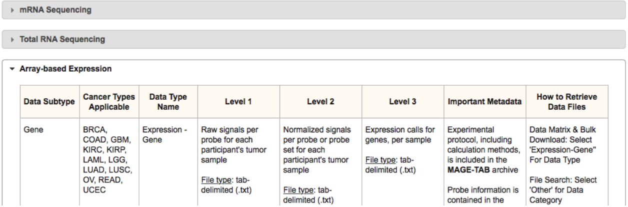
- **Sample**: use Biospecimen IDs to find data (See later slide).
- **Access Tier**: whether the data are open or control-access.
- **Tumor/Normal**
  - **Tumor - matched**: data for a tumor tissue for which matched normal tissue exists.
  - **Tumor - unmatched**: data for a tumor tissue for which there is no matched normal tissue.
  - **Normal - matched**: data for normal tissue for which matched tumor tissue exists.
  - **Organ-specific Control**: data for normal tissue from a participant who does not have cancer.
  - **Cell Line Control**: data for cell-line controls.

Find the detailed definitions by using TCGA Wiki or Data Selection Tutorial.

Accessing TCGA Data

TCGA Data Portal – Data Types

To see all available data types under TCGA Data Portal, click here.



▸ mRNA Sequencing

▸ Total RNA Sequencing

▾ Array-based Expression

Data Subtype	Cancer Types Applicable	Data Type Name	Level 1	Level 2	Level 3	Important Metadata	How to Retrieve Data Files
Gene	BRCA, COAD, GBM, KIRC, KIRP, LAML, LGG, LUAD, LUSC, OV, READ, UCEC	Expression - Gene	Raw signals per probe for each participant's tumor sample File type: tab-delimited (.txt)	Normalized signals per probe or probe set for each participant's tumor sample File type: tab-delimited (.txt)	Expression calls for genes, per sample File type: tab-delimited (.txt)	Experimental protocol, including calculation methods, is included in the **MAGE-TAB** archive Probe information is contained in the	Data Matrix & Bulk Download: Select 'Expression-Gene" For Data Type File Search: Select 'Other' for Data Category

Note that non-genomic info are available under clinical data. (see later slides)
Also note that primary DNA and RNA sequence data are only available through CGHub, not TCGA Data Portal.

Accessing TCGA Data

The TCGA Data Portal - more useful links

Data Download Page

- Data Matrix Tutorial
- Bulk Download Tutorial
- HTTP Download Tutorial
- File Search Download Tutorial

Accessing TCGA Data

Understanding Biospecimen IDs

Biospecimen IDs exist in 2 forms:

- Universally Unique Identifier (UUID)
- Barcode

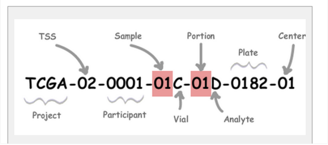
UUID: is NOT human readable.
e.g. ebf3e73f-41a0-4ca5-b608-fe1c629e16de

Accessing TCGA Data

Understanding Biospecimen IDs

Barcode: consists of many identifiers and is human readable.

e.g.



- **Project:** Project name, such as TCGA
- **TSS:** A Tissue Source Site collects samples and clinical metadata
- **Participant:** someone who contributes one or more samples to a TSS
- **Sample:** Tumor types range from 01 to 09; normal types range from 10 to 19

Accessing TCGA Data

Understanding Biospecimen IDs

To see how to retrieve a sample or a set of samples from Biospecimen IDs, click here.
To see how to get one type of Biospecimen ID from another type of Biospecimen ID, click here.
To see how to find the corresponding participants or aliquots from Biospecimen IDs, click here.

Access the TCGA Data

Clinical Data

TCGA clinical data include:

- Clinical information about the participant
- Information about the how participant samples (biospecimens) were processed

Note that clinical data contains a lot of non-genomic information such as the participant's age, gender, and race.

Clinical data are contained in four types of files:

- **Biotab** Files: contain clinical and biospecimen information for a set of participants.
- **XML** Files: Each TCGA participant has a separate clinical and biospecimen XML file.
- **svs** Files: Tissue slide images.
- **pdf** Files: Original pathology reports.

Access the TCGA Data

Clinical Data

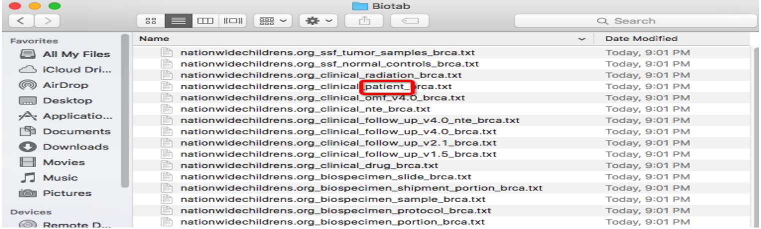
Note that clinical data are retrieved differently from other data types.

Data Download Link

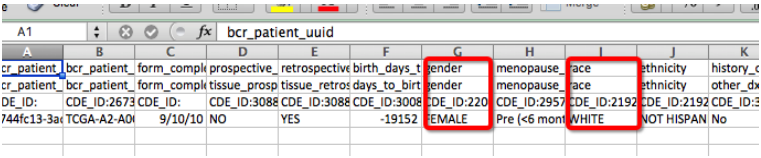
- Clinical Data: Data Matrix Tutorial
- Clinical Data: Bulk Download Tutorial
- Clinical Data: File Search Tutorial
- Image and Pathology Reports Retrieval Tutorial

Accessing TCGA Data

Clinical Data - Obtaining non-genomic information



Name	Date Modified
nationwidechildrens.org_ssf_tumor_samples_brca.txt	Today, 9:01 PM
nationwidechildrens.org_ssf_normal_controls_brca.txt	Today, 9:01 PM
nationwidechildrens.org_clinical_radiation_brca.txt	Today, 9:01 PM
nationwidechildrens.org_clinical_patient_brca.txt	Today, 9:01 PM
nationwidechildrens.org_clinical_omf_v4.0_brca.txt	Today, 9:01 PM
nationwidechildrens.org_clinical_nte_brca.txt	Today, 9:01 PM
nationwidechildrens.org_clinical_follow_up_v4.0_nte_brca.txt	Today, 9:01 PM
nationwidechildrens.org_clinical_follow_up_v4.0_brca.txt	Today, 9:01 PM
nationwidechildrens.org_clinical_follow_up_v2.1_brca.txt	Today, 9:01 PM
nationwidechildrens.org_clinical_follow_up_v1.5_brca.txt	Today, 9:01 PM
nationwidechildrens.org_clinical_drug_brca.txt	Today, 9:01 PM
nationwidechildrens.org_biospecimen_slide_brca.txt	Today, 9:01 PM
nationwidechildrens.org_biospecimen_shipment_portion_brca.txt	Today, 9:01 PM
nationwidechildrens.org_biospecimen_sample_brca.txt	Today, 9:01 PM
nationwidechildrens.org_biospecimen_protocol_brca.txt	Today, 9:01 PM
nationwidechildrens.org_biospecimen_portion_brca.txt	Today, 9:01 PM



A1 bcr_patient_uuid

A	B	C	D	E	F	G	H	I	J	K
cr_patient_	bcr_patient_	form_comple	prospective_	retrospective	birth_days_t	gender	menopause_	race	ethnicity	history_c
cr_patient_	bcr_patient_	form_comple	tissue_prosp	tissue_retros	days_to_birt	gender	menopause_	race	ethnicity	other_dx
DE_ID:	CDE_ID:2673	CDE_ID:	CDE_ID:3088	CDE_ID:3088	CDE_ID:3008	CDE_ID:220	CDE_ID:2957	CDE_ID:2192	CDE_ID:2192	CDE_ID:3
744fc13-3a	TCGA-A2-A0	9/10/10	NO	YES	-19152	FEMALE	Pre (<6 mont	WHITE	NOT HISPAN	No

Accessing TCGA Data

Understanding the Downloaded Data Files

- **Metadata:** the summary data about the downloaded data. Often describe biospecimen-related elements.
- **idf** file: provides general information about the investigation and experiment.
- **sdrf** file: encapsulates a succession of processes applied to samples and the multiple states it takes on as a result of the processes.

Thank you!