Introduction to TCGA

TCGA Data Primer TCGA Data Overview

This page provides a high level description of TCGA and the data that it generates. The following topics are included in this section.

Overview of TCGA

The Cancer Genome Atlas (TCGA) is a comprehensive and coordinated effort to accelerate the understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing.The overarching goal of TCGA is to improve our ability to diagnose, treat and prevent cancer. To achieve this goal in a scientifically rigorous manner, the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI)] used a phased-in strategy to launch TCGA. A pilot project developed and tested the research framework needed to systematically explore the entire spectrum of genomic changes involved in more than 20 types of human cancer.

See The Cancer Genome Atlas for more information on the project.

Privacy Policy

The TCGA project produces large volumes of genomic information derived from human tumor specimens collected from participants. It also collects significant amounts of clinical information associated with these specimens. The aggregated data generated is unique to each individual and, despite the lack of any direct identifying information within the data, there is a risk of individual re-identification by bioinformatics methods and/or third-party databases.

Because participant privacy protection is of paramount concern to NIH, NCI, and TCGA, human subject protection and data access policies have been implemented to minimize the risk that the privacy of the donors and the confidentiality of their data will be compromised. As part of this effort, data generated from TCGA are available in two tiers:

Open access - Houses data that cannot be aggregated to generate a data set unique to an individual. This tier does not require user certification for data access.

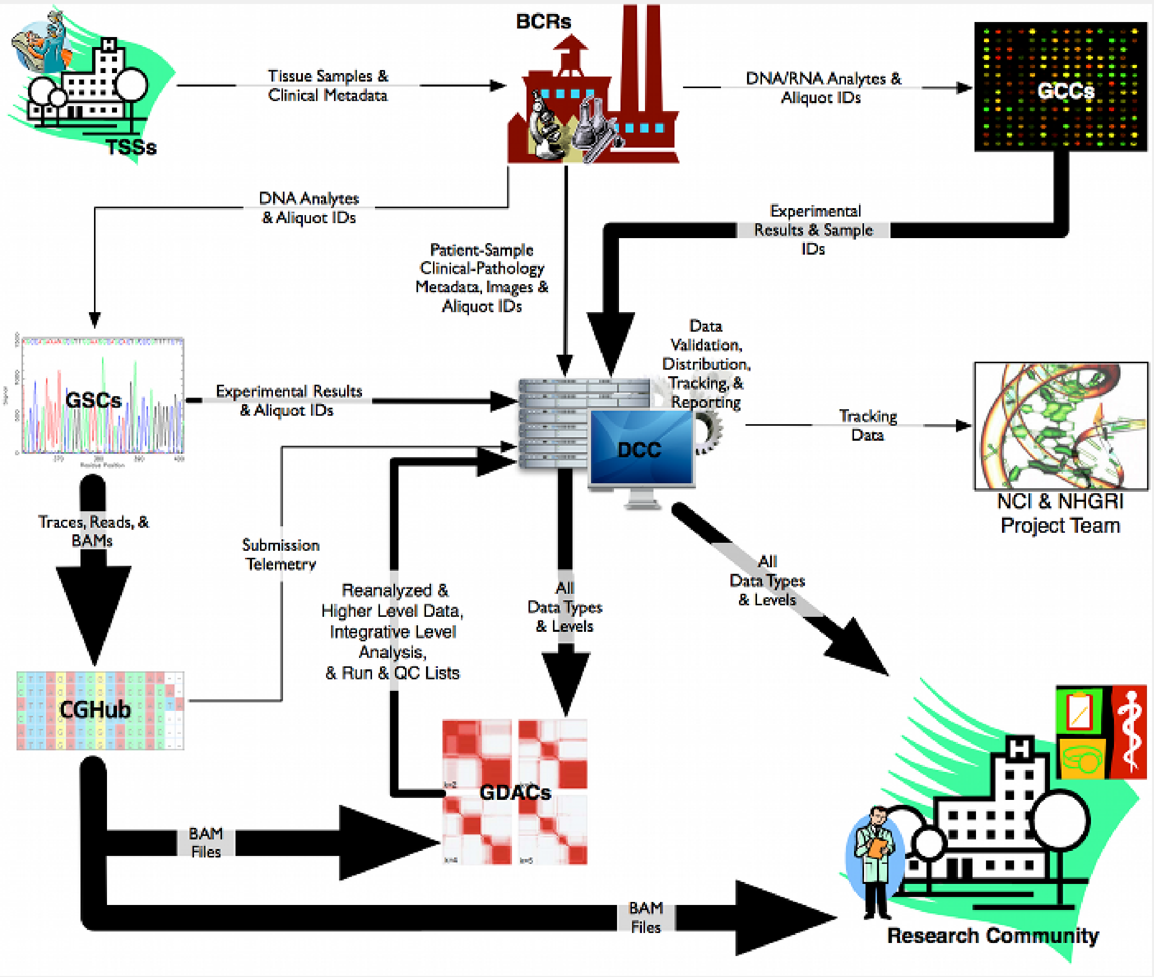
Controlled access - Houses individually-unique information that could potentially be used to identify an individual. This tier requires user certification for data access.

See Access Control Policy for more information on data tiers.

TCGA Data Flow

The following steps, illustrated in the accompanying figure, summarize data flow through the TCGA pipeline:

1. Tissue samples along with clinical data are collected by Tissue Source Sites (TSS) and sent to the Biospecimen Core Resources (BCRs).
2. The BCRs submit clinical data and metadata to the Data Coordinating Center (DCC) and analytes to the Genome Characterization Centers (GCCs) and Sequencing Centers (GSCs), where mutation calls are generated and then submitted to the DCC.
3. GSCs submit trace files, sequences and alignment mappings to the Cancer Genomics Hub (CGHub) as well.
4. Data submitted to the DCC and CGHub are made available to the research community and Genome Data Analysis Centers (GDACs).
5. Analysis pipelines and data results produced by GDACs are served to the research community via the DCC.



The flow of TCGA data and biospecimen products. Arrow thickness depicts the relative volume of data transferred between TCGA centers/groups.

The following table provides a quick overview of the different centers and groups that form TCGA. For more information on a specific center/group, click on the corresponding "Center/Group" label:

Center/Group	Description
TSS	A Tissue Source Site (TSS) collects samples (tissue, cell or blood) and clinical metadata, which are then sent to a BCR. A TSS is identified by its TSS ID.
BCR	A Biospecimen Core Resource (BCR) is a TCGA center where samples are carefully catalogued, processed, quality-checked and stored along with participant clinical information. Analytes are aliquoted and assigned aliquot barcodes by the BCR before they are sent to the other centers.
GCC	A Genome Characterization Center (GCC) is a TCGA center that uses high-throughput technologies to analyze genomic changes involved in cancer. The genomic changes that are identified will be further studied by the GSCs. GCCs transfer the files of experimental results of characterization assays in data archives to the DCC.
GSC	A Genome Sequencing Center (GSC) is a TCGA center that uses high-throughput methods to identify changes to DNA sequences that are associated with specific cancer types. GSCs sequence analytes (provided by BCRs) and analyze them for putative somatic and germline mutations. Sequencing results are sent to the Cancer Genomics Hub and mutation results are sent to the DCC.

DCC	The Data Coordinating Center (DCC) is the central provider of TCGA data. The DCC standardizes data formats and validates submitted data. The DCC receives and validates data from the BCRs, GCCs and GSCs before making it available to the research community through applications hosted via the TCGA Data Portal.
GDAC	A Genome Data Analysis Center (GDAC) is a TCGA center that provides novel informatics tools to the research community while also providing analysis results using TCGA data. At present, the DCC does not accept any GDAC data submissions through the automated validation and deployment system. One GDAC, the Broad Institute GDAC, currently uploads analysis data to the DCC for provisional deployment via controlled access.
CGHub	The Cancer Genomics Hub (CGHub) is a secure repository for storing, cataloguing, and accessing cancer genome sequences, alignments, and mutation information from the Cancer Genome Altas (TCGA) consortium and related projects. CGHub is managed by the University of California, Santa Cruz (UCSC), under a subcontract from SAIC-Frederick. GSCs submit trace files, short read sequences and BAM files to CGHub.
Project Team	A project team coordinates TCGA and is comprised of individuals from NCI and NHGRI.

TCGA Primary Identifiers

Historically, the BCR received participant samples and their associated metadata from TSSs. The BCR then assigned human-readable IDs, referred to as TCGA barcodes, representing the metadata of the participants and their samples. TCGA barcodes were used to tie together data that spans the TCGA network, since the IDs uniquely identify a set of results for a particular sample produced by a particular data-generating center (i.e. GCC, GSC or GDAC). The constitutive parts of this barcode provided metadata values for a sample. Currently the BCR is assigning both a TCGA barcode and a UUID to samples. The UUID is the primary identifier.

See TCGA Barcodes for more information about barcodes.

TCGA Data Primer TCGA Data Overview