Working with TCGA Data

Accessing Data TCGA Data Primer

This page describes methods of uniquely identifying TCGA data. It also identifies applications and analysis tools for working with the data and generating TCGA data reports.

The following topics are included in this section.

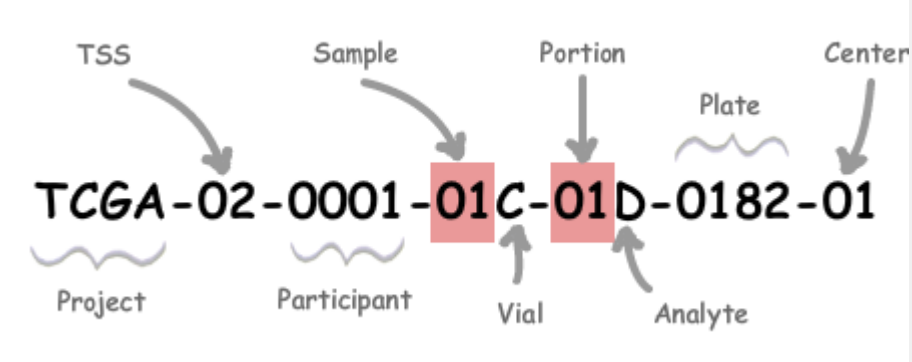
Understanding TCGA Barcodes

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.

Reading Barcodes

A TCGA barcode is composed of a collection of identifiers. Each specifically identifies a TCGA data element. Refer to the following figure for an illustration of how metadata identifiers comprise a barcode. An aliquot barcode, an example of which shows in the illustration, contains the highest number of identifiers.



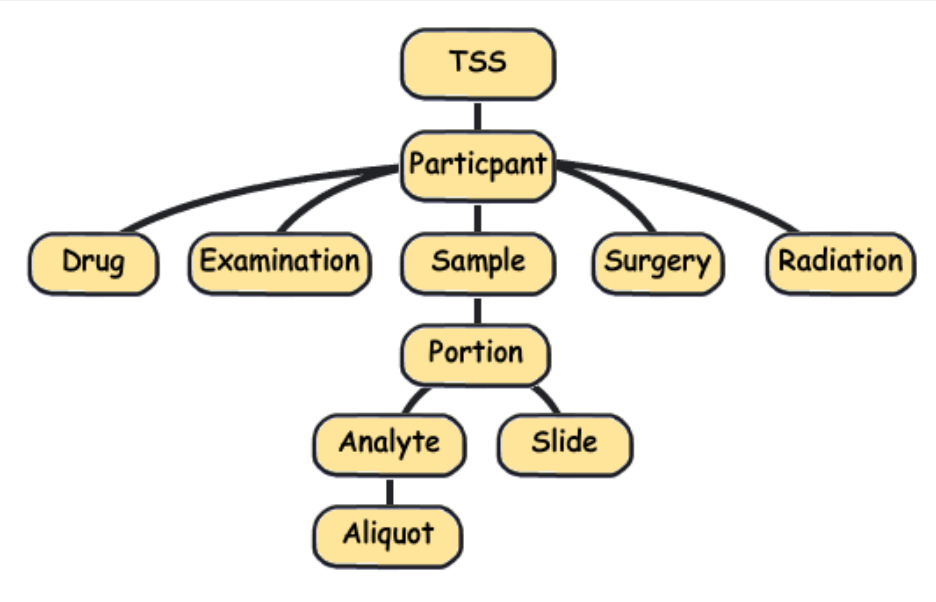
This figure of an aliquot barcode shows how it can be broken down into its components and translated into its metadata. The barcode metadata are further described in the following table.

Label	Identifier for	Value	Value description	Possible values
Project	Project name	TCGA	TCGA project	TCGA
TSS	Tissue source site	02	GBM (brain tumor) sample from MD Anderson	See Code Tables Report
Participant	Study participant	0001	The first participant from MD Anderson for GBM study	Any alpha-numeric value
Sample	Sample type	01	A solid tumor	Tumor types range from 01 - 09, normal types from 10 - 19 and control samples from 20 - 29. See Code Tables Report for a complete list of sample codes
Vial	Order of sample in a sequence of samples	C	The third vial	A to Z
Portion	Order of portion in a sequence of 100 - 120 mg sample portions	01	The first portion of the sample	01-99
Analyte	Molecular type of analyte for analysis	D	The analyte is a DNA sample	See Code Tables Report
Plate	Order of plate in a sequence of 96-well plates	0182	The 182nd plate	4-digit alphanumeric value
Center	Sequencing or characterization center that will receive the aliquot for analysis	01	The Broad Institute GCC	See Code Tables Report

Barcode Types

Barcodes can also be visualized hierarchically, with TSS barcodes at the top of the tree and aliquot barcodes at the bottom. A parent barcode prefixes any of its descendent barcodes, reflecting the

derivation of one biospecimen type from another. For example, samples are collected from a participant and so the corresponding sample barcodes contain the participant barcode from which they were derived.



Hierarchy of biospecimen elements. Barcodes are used to represent all biospecimen elements in this diagram.

Using the aliquot barcode example from the figure in Reading Barcodes, the following table displays a possible set of related barcodes at each level of the hierarchy:

Level	Barcode	Comment
TSS	TCGA-02	
Participant	TCGA-02-0001	
Drug	TCGA-02-0001-C1	Drug ID is 'C','D','H','I' or 'T' followed by a number
Examination	TCGA-02-0001-E3124	Examination ID is 'E' followed by a number
Surgery	TCGA-02-0001-S145	Surgery ID is 'S' followed by a number
Radiation	TCGA-02-0001-R2	Radiation ID is 'R' followed by a number
Sample	TCGA-02-0001-01	
Portion	TCGA-02-0001-01C-01	
Shipped Portion	TCGA-CM-5341-01A-21-1933-20	Used in the platform of MDA_RPPA_CORE only
Slide	TCGA-02-0001-01C-01-TS1	Tissue slide ID can be 'TS' ('Top Slide'), 'BS' ('Bottom Slide') or 'MS' ('Middle slide'), followed by a number or letter to indicate slide order
Analyte	TCGA-02-0001-01C-01D	Analytes of W and X both refer to analytes derived from whole genome amplification

Aliquot	TCGA-02-0001- 01C-01D-0182-01

TCGA Applications

This list describes applications used by TCGA centers for identifying, processing, annotating, validating, organizing, and searching TCGA data.

Annotations Manager

The Annotations Manager is an application that allows authorized TCGA team members to add annotations about TCGA participants and samples down to the aliquot level. This application is available on TCGA Data Portal.

Bulk Download

Bulk Download is an application that performs file-based searches on the latest versions of TCGA archives for download.

Data Matrix

The Data Matrix or Data Access Matrix (DAM) is an application that provides TCGA data users the option to build and download custom data archives by selecting samples represented graphically in a matrix. The Data Matrix only provides the latest revision of each archive; older revisions are available through bulk download or HTTP access. Also, it does not allow for querying across multiple disease studies.

TCGA Client Side Validator

TCGA Client Side Validator is a Java application of QCLive. This application is run locally by all data submission centers to validate data archives before submitting them to the DCC.

Biospecimen Metadata Browser

The Biospecimen Metadata Browser is an application that provides a means to search and browse TCGA biospecimen metadata. The Biospecimen Metadata Browser allows users to search and browse UUIDs/barcodes and the metadata that they represent.

TCGA File Search

The TCGA File Search is a new interface designed to allow users to access TCGA data based on simple biological concepts. File Search accesses the same information as the Data Matrix and Bulk Download, but does so in a more easily accessible manner. The File Search is capable of including files from multiple diseases.

Publication MAF Search

The Publication MAF Search (Beta) is a new TCGA application that allows users to search the contents of publication MAF files without opening the original files. It includes key filters such as gene name, barcode, disease name, and others. This interactive tool allows users to search the contents of publication MAF files across multiple diseases. It can also export search output and download custom MAF files.

Web Services

Contents

Annotations

The Annotations Web Services are a set of REST services that allow for programmatic querying and entering of annotations data, as produced by the Annotations Manager.

Data Matrix

The Data Matrix Web Service is a REST service that allows for programmatic generation of archives of data from the Data Matrix web application.

Data Reports

The Data Reports Web Services are a set of REST services that allow for programmatic querying of three Data Reports:

- Sample Count Summary Report
- Aliquot ID Breakdown Report
- Latest Archive Report

See Data Reports Web Services for the web service user guide for these reports.

DCC Data

The DCC Data Web Service is a REST service that allows for programmatic querying of the DCC's database, which contains project metadata such as center and platform codes and submitted archive names.

Barcode-UUID Mapping

The Barcode-UUID Mapping Web Service is a REST service that allows for programmatic querying of the mapping between TCGA barcodes and their corresponding UUIDs. This web service is primarily used to facilitate the project-wide primary identifier transition from TCGA barcodes to UUID.

Biospecimen Metadata

The Biospecimen Metadata Web Services are a set of REST services that allow for programmatic querying of TCGA biospecimen metadata, as produced by the Biospecimen Metadata Browser.

Data Reports and Dashboards

Contents of this Page

Data Reports

The TCGA Data Portal provides a number of reports that provide detailed information about TCGA project progress and TCGA workflow. They are described below under two report catagories: General Interest and Aliquot Reports. To access these reports see TCGA Data Reports.

General Interest

Data Statistics Dashboard

The Data Statistics Dashboard provides high-level statistics describing TCGA data content and use.

Project Case Overview Dashboard

The Project Case Overview Dashboard (PCOD) is a visual progress report created by the DCC on cases processed (for each disease study) by the BCRs, GCCs and GSCs.

BCR Pipeline Report

The BCR Pipeline Report provides a graphical representation of the current state and workflow at the BCR-level of the project.

Sample Counts for TCGA Data Report

The Sample Counts for TCGA Data Report provides a count summary of Sample IDs by Disease Study, Center, Analyte Type, and Platform.

Latest Archive Report

The Latest Archive Report lists the most recent version of all data archives submitted to the DCC.

Code Tables Report

The Code Tables Report is an application that displays the collection of codes and abbreviations used within TCGA in an interactive table.

CGHub Summary Statistics Report

The CGHub Summary Statistics report summarizes the number of genomic sequences available from CGHub.

Aliquot Reports

Experiment Aliquot Report

The Experiment Aliquot Report provides a summary for all existing Aliquot IDs by listing their disease study, batch number, receiving center, platform, and Data Levels.

Aliquot ID Breakdown Report

The TCGA Aliquot ID Breakdown Report lists each existing aliquot barcode broken up by its components to clearly display the metadata captured within the barcode.

Accessing Data	TCGA Data Primer