TCGA
DATA PRIMER

Version 1.0

NATIONAL® CANCER INSTITUTE

Center for Bioinformatics

July 15, 2008

TABLE OF CONTENTS

CHAPTER 1
ABOUT TCGA DATA

This chapter provides a high-level description of The Cancer Genome Atlas (TCGA).

Topics in this chapter include:

About this Document

The Cancer Genome Atlas (TCGA), a three-year pilot project of the National Cancer Institute and the National Human Genome Research Institute (NHGRI), is a large-scale collaborative effort to understand the genomic changes that occur in cancer. For more information, see http://cancergenome.nih.gov/dataportal/index.asp.

The goal of TCGA is to accelerate our understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing. A better understanding of the molecular basis of cancer will, in turn, lead to improvements in the diagnosis, treatment, and prevention of cancer.

This purpose of this *TCGA Data Primer* is to provide individual researchers with a description of the TCGA data enterprise, that is, what kinds of data are available, how the data is created and distributed, how the data is modeled and formatted, how they can access that data, and how they can make use of the data.

TCGA data, which meets the highest standards for protection and respect of the research participants, is made available through public databases supported by the TCGA Data Portal, as well as through the NCI's cancer Biomedical Informatics Grid™ (caBIG™) and the National Library of Medicine's National Center for Biotechnology Information (NCBI).

TCGA Data Flow Overview

Figure 1.1 illustrates the flow of data and products from one TCGA group to another (1-3), the distribution of those data into publicly accessible databases (3-4) and mapping between data sets (5). *Table 1.1* describes in detail the components depicted.

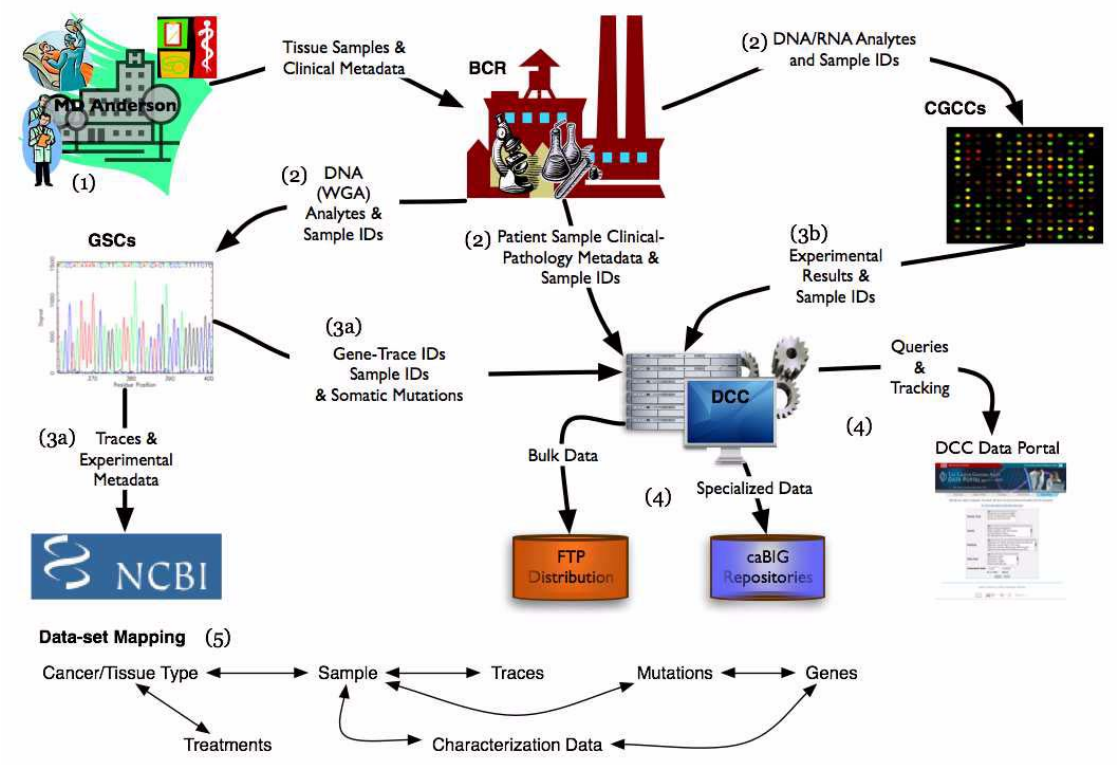


Figure 1.1 Data flow in TCGA. See Table 1.1 for a description of the elements of this figure.

Data Flow Description

Table 1.1 describes each component in *Figure 1.1*. Component numbers in the table correspond to those in the illustration.

Compo-nent #	Data Flow Description
1	Collection sites, for example, the University of Texas MD Anderson Cancer Center, send tissue samples and clinical metadata to the Biospecimen Core Resource (BCR). To learn more about BCR, see *Creating and Identifying Biospecimen Analytes* on page 5.
2	The BCR prepares aliquots of the samples and assigns unique identifiers (IDs) to them as follows: After extracting and plating biospecimen analytes and processing clinical pathology metadata, the BCR assigns each product an aliquot barcode. The barcode identifies the particular patient and sample, from a particular center, the particular tumor type, the center that will receive the aliquot from the BCR and the result of an analyte on a particular platform. For more information about these IDs and barcodes, see *About Aliquot Barcodes* on page 7. The BCR transfers the products to the appropriate TCGA center types as follows: • **Genomic Sequencing Centers** (*GSCs*): Plated Whole Genome Amplified (WGA) DNA analytes and corresponding aliquot barcodes. • **Cancer Genomic Characterization Centers** (*CGCCs*): Plated DNA/RNA analytes and corresponding aliquot barcodes. • **Data Coordinating Centers** (*DCCs*): Patient-sample clinical pathology metadata and corresponding aliquot barcodes.
3a	GSCs sequence the analytes and transfer the following data as files in compressed *archives* to the appropriate repositories: • **NCBI**: Trace files • **DCC**: Trace ID-to-sample relationship files and mutation files (For a complete list of file formats that are compatible with NCBI and DCC repositories, see *Table 1.2* on page 4. For more information, see *Chapter 2, Understanding Sequence-Based Genomic Data,* on page 11.)
3b	*CGCCs* transfer experimental results of characterization assays in compressed archives to the *DCC*. These files can include results of the following assays: gene expression, copy number variation, and methylation, in the following formats: MAGE-TAB IDF, SDRF, and raw and processed files. Additionally, each CGCC provides an ADF file if the platform is non-standard, such as with methylation data. See *About Array Description Format Files (ADFs)* on page 23. For more information, see *Chapter 3, Understanding Array-Based Data,* on page 17.

Table 1.1 Description of the TCGA data flow

Compo-nent #	Data Flow Description
4	The *DCC* validates all data it receives and transfers data that is considered unrestricted to the TCGA public FTP site and data that is considered restricted to a TCGA secure FTP (SFTP) site. In addition, restricted and unrestricted data are deposited into caBIG™-compatible repositories. The *TCGA Data Portal* provides user-friendly access to the FTP and SFTP sites. The Portal is available at this website: http://tcga-data.nci.nih.gov/tcga/findArchives.htm
5	The DCC maps and maintains relationships between all the data types, samples, and treatments; and tracks metrics and ultimate locations of all data.

Table 1.1 Description of the TCGA data flow

Note: For detailed technical information about data and archive formats, refer to *TCGA Data Preparation and Transfer SOP (Data Preparation SOP)* at this location: https://gforge.nci.nih.gov/docman/view.php/265/5004/Data_Preperation_and_Transfer_SOP.tar.gz, and *TCGA Higher-level Analysis Data Format Specification (HLA Specification)* at this location: https://gforge.nci.nih.gov/docman/view.php/265/8841/HLA_SOP.zip.

Experiment Archives and File Formats

Each center transfers its data to the DCC in compressed *archives*. All archives include common documents and follow distinct naming conventions. For details about archives, including naming conventions, see *About Archives* on page 40.

In the context of TCGA, an *experiment* is a complete study from a given center that consists of all the assays from a particular platform for all the samples of a particular tumor type. An experiment is likely to be represented by many archives.

Table 1.2 lists file formats that are compatible with NCBI and DCC repositories.

Data Repository	Compatible Data File Format
NCBI	Trace file. See *Sequence Trace Files* on page 12.
DCC	Trace ID-to-Sample Relationship files. See *Trace ID-to-Sample Relationship Files* on page 13.
	Mutation file (MAF). See *Mutation Annotation Format (MAF) Files* on page 14.
	Investigation Description Format (IDF) file. See *About Investigation Description Format Files (IDFs)* on page 18.
	Sample and Data Relationship Format (SDRF) file. See *About Sample and Data Relationship Files (SDRFs)* on page 21.
	Array Description Format File (ADF). See *About Array Description Format Files (ADFs)* on page 23.
	Raw file. Array data. See *About Raw and Processed Data Files* on page 24.

Table 1.2 Compatible file formats by repository

Creating and Identifying Biospecimen Analytes

The BCR collects biological tissue samples and the clinical and biological information (metadata) associated with those samples from collection sites. During the process, the samples and data are assigned BCR aliquot barcodes, the most important type of ID in the TCGA Data Enterprise. Each barcode uniquely identifies a set of results for a particular sample produced by a particular cancer genomic center (CGC). Additionally, the constitutive parts of a barcode provide clinical values for that sample.

Figure 1.2 illustrates how the BCR Aliquot Barcode is constructed. It shows the processes that the BCR uses to create and code biospecimen analytes, such as DNA and RNA, for distribution to *GSC*s and *CGCC*s for analysis. *Table 1.3* describes the steps in the process.

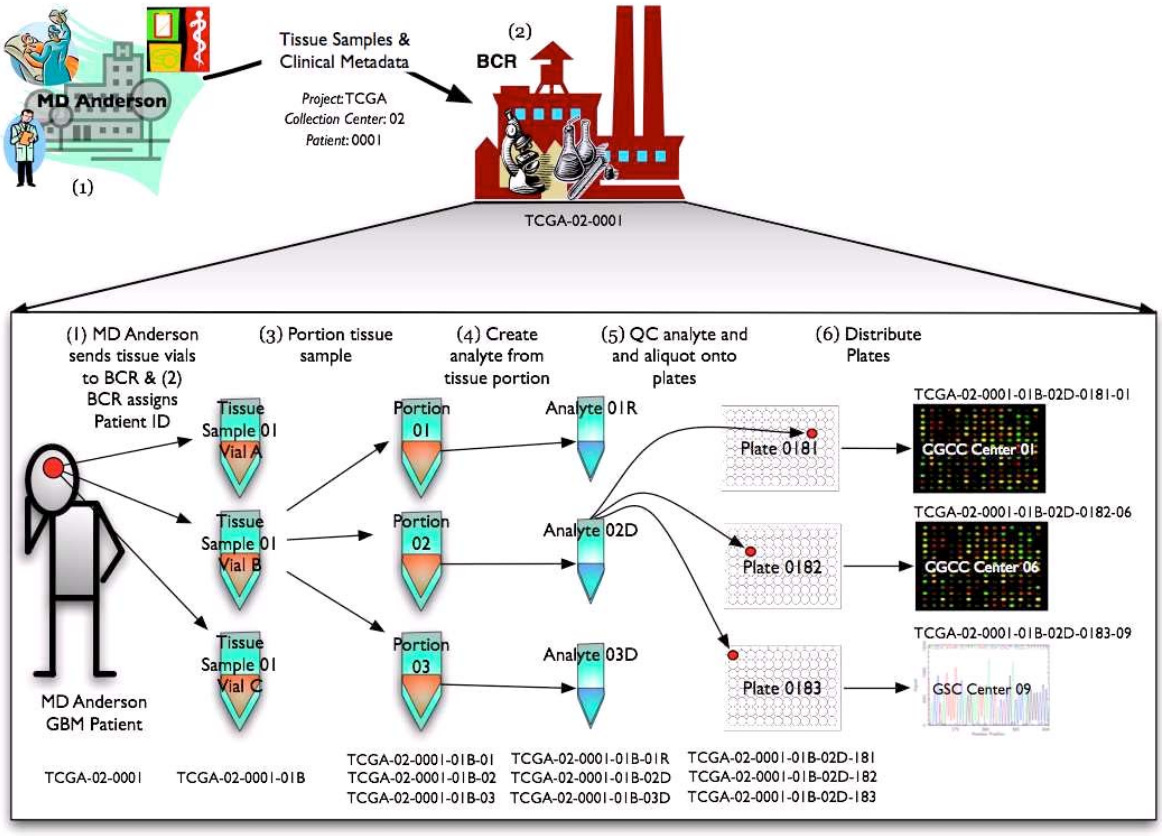


Figure 1.2 BCR's process of creating and coding analytes. See Table 1.3 for details.

Processing Analytes

Table 1.3 describes the steps in *Figure 1.2*. Step numbers in the table correspond to those in the center row of the illustration. For more information, see *About Aliquot Barcodes* on page 7.

Step	Process and Identification Descriptions
1	A tissue collection site, such as the example MD Anderson Cancer Center, sends tissue samples in vials to the BCR. Each sample is coded by the BCR using a project name and collection center ID, in this case, `02`. (To date the only project name is TCGA.) `TCGA-02` indicates project TCGA and the tissue collection site MD Anderson Cancer Center Brain Bank (02).
2	The BCR appends a patient ID, a sample-type code, and vial number code to the barcode in Step 1. `TCGA-02-0001-01B` indicates a solid tumor sample (01) in the second vial (B) from the first patient (0001) from the MD Anderson Cancer Center Brain Bank (02). Note that there can be many samples and/or vials per patient.
3	The BCR apportions the samples and appends a portion ID to the barcode in Step 2. `TCGA-02-0001-01B-02` indicates the second portion (02) of a solid tumor sample (01) in the second vial (B) from the first patient (0001) from the MD Anderson Cancer Center Brain Bank (02).
4	The BCR creates analytes from the portions and appends an analyte code to the barcode in Step 3. `TCGA-02-0001-01B-02D` indicates the second (02) DNA analyte (D) of a solid tumor sample (01) in the second vial (B) from the first patient (0001) from the MD Anderson Cancer Center Brain Bank (02).
5	The BCR plates the analytes that pass quality control, and appends a plate ID. `TCGA-02-0001-01B-02D-0182` indicates plate 0182 of the second DNA analyte (02D) of a solid tumor sample (01) in the second vial (B) from the first patient (0001) from the MD Anderson Cancer Center Brain Bank (02).
6	The BCR appends a center ID to the barcode in Step 5 that indicates which center will receive the aliquot. The plate containing the barcoded aliquot is distributed to its respective center. `TCGA-02-0001-01B-02D-182-06` indicates plate 0182 of the second DNA analyte (02D) of a solid tumor sample (01) in the second vial (B) from the first patient (0001) from the MD Anderson Cancer Center Brain Bank (02) distributed to Stanford (06).

Table 1.3 BCR process and identification descriptions

Note: The BCR models TCGA biological and clinical data as XML in the `.xml` file format using caBIG standards. A UML document describing that model is available by navigating to caDSR Contexts > caBIG > BiospecimenCoreResource at this location: http://umlmodelbrowser.nci.nih.gov/umlmodelbrowser/). An XML schema of that model is also available.

About Aliquot Barcodes

BCR creates an *aliquot barcode* to identify and track the distribution of each product from a collection site uniquely. The barcode persists with the experimental results of its associated analyte throughout downstream processing.

The aliquot barcode format is a combination of an analyte barcode and a plate barcode separated with a hyphen as follows:

```
{analyte barcode}-{plate barcode}
```

- **Analyte barcode** – Identifies the collection site, patient, sample, and portion ID. (For more information, see *Deciphering Analyte Barcodes* on page 7.)
- **Plate barcode** – Identifies the plate and the *GSC* or *CGCC* to which it will be distributed. (For more information, see *Deciphering Plate Barcodes* on page 9.)

An example of the process deriving the barcode is provided in *Figure 1.1* and *Table 1.3*. Barcode values are provided in *Appendix A,* on page 53.

Deciphering Analyte Barcodes

An analyte barcode is a series of unique IDs that, when combined, identify each individual analyte. Identifiers that compose the analyte barcode appear in a sequence with the following convention:

```
{ProjectName}-{SiteID}-{PatientID}-{SampleID}-{PortionID}
```

Table 1.4 describes the analyte barcode's constituent IDs. Example analyte barcodes are provided on page 8. Also see *Analyte Barcode Values* on page 53

Identifier	***Description***
ProjectName	Current project name. Currently TCGA is the only project
SiteID	Tissue Collection center identifier. See *Site ID Values* on page 53.
PatientID	Patient identifier that is associated with a Site ID. See *Patient ID Values* on page 54.

Table 1.4 Analyte barcode constituent IDs

Identifier	*Description*
SampleID	Sample type and sample vial indentifier, as follows: • **Sample type** – Two digit number that represents a biospecimen type. IDs `01 - 09` indicate tumor types. (For example, `01` is a solid tumor.) IDs `10 - 19` indicate normal types IDs `20 - 29` indicate control samples • **Vial identifier** – Alphabetic letter that represents portions of a sample from an individual patient. For example: The sample ID `01A` represents the first vial (`A`) containing a sample from a solid tumor (`01`) of a given patient. The sample ID `01B` represents the second vial (`B`) containing a sample of the same tumor (`01`) from the same patient. See *Sample ID Values* on page 54.
PortionID	Individual 100 mg – 120 mg section of a sample. Consists of a portion code and an analyte code, as follows: • **Portion code** – Two-digit number that identifies the portion. Range is from 01 to as many as 99 for larger tissue samples. • **Analyte code** – Alphabetic code that represents an analyte type. (For example, code `D` represents DNA, and `R` represents RNA.) For example: The portion ID `14D` represents the 14th portion of DNA (`D`). The portion ID `25R` represents the 25th portion of RNA (`R`). **Note:** A normal sample or buccal smear is not divided and is considered one portion. See *Portion ID Values* on page 55.

Table 1.4 Analyte barcode constituent IDs (Continued)

Examples of Analyte Barcodes

Following are two examples of analyte barcodes and what they represent:

- `TCGA-02-0021-01B-01D`

  where,

  - `TCGA` is the project name (projectID)
  - `02` is the MD Anderson Cancer Center Brain Bank (site ID)
  - `0021` is the 21st patient from the MD Anderson Cancer Center Brain Bank (patient ID)
  - `01B` is a solid tumor (sample type) from the 2nd vial of tissue (vial identifier) from patient 0021
  - `01D` is the first portion (portion code) of DNA (analyte code)

- `TCGA-02-0034-10A-03R`

where,

- `TCGA` is the project name
- `02` is the MD Anderson Cancer Center Brain Bank (site ID)
- `0034` is the 34th patient from the MD Anderson Cancer Center Brain Bank (patient ID)
- `10A` is normal blood (sample type) from the 1st vial (A) tissue from patient 0034
- `03R` is the 3rd portion (portion ID) of RNA (analyte code)

Deciphering Plate Barcodes

A plate barcode provides a unique ID for each 96-well plate. The plate barcode is a composite of a PlateID and a CenterID as follows:

`{PlateID}-{CenterID}`

Table 1.5 describes the plate barcode's constituent IDs. Example plate barcodes are provided after the table. Also see *Plate Barcode Values* on page 55.

Name	***Description***
PlateID	Identifies individual plates See *Plate ID Values* on page 55.
CenterID	Identifies each of the *CGCCs* or *GSCs* See *Center ID Values* on page 55.

Table 1.5 Plate barcode constituent IDs

Examples of Plate Barcodes

Following are two examples of plate barcodes and what they represent:

`0002-04`

- `0002` is the second plate in a sequence of 96-well plates (plate ID)
- `04` is the Memorial Sloan-Kettering center (center ID)

`0010-07`

- `0010` is the tenth plate in a sequence of 96-well plates (plate ID)
- `07` is the Broad Institute (center ID)

Distributions to GSCs and CGCCs

Once the BCR has prepared the sample aliquots and assigned unique identifiers, it transfers them to appropriate TCGA centers, as described in *Chapter 2, Understanding Sequence-Based Genomic Data* and *Chapter 3, Understanding Array-Based Data*.

CHAPTER

2

UNDERSTANDING SEQUENCE-BASED GENOMIC DATA

This chapter provides an introduction to the types of data and data files that are produced during the processes of sequencing DNA and analyzing nucleic acid analytes at the Genomic Sequencing Centers (GSCs).

Topics in this chapter include:

About Sequence-Based Data Files

Data Received by the GSCs

As illustrated in *Figure 1.1* on page 2, the Biological Collection Resource (BCR) distributes DNA and RNA samples, identified by their corresponding barcodes, to Genomic Sequencing Centers for analysis.

Genomic Sequencing Centers (GSCs) sequence the DNA they receive from the BCR. They use the sequence data to identify germline and somatic mutations, insertions, and deletions (collectively called sequence polymorphisms) in genes and other loci. For example, a GSC could compare the sequences of a blood sample and the sequences of a tumor sample to determine abnormal variations between the two.

The GSCs generate the following sequence-based data files:

- **Sequence Trace files** – Raw data produced by a DNA sequencing instrument (see *Sequence Trace Files* on page 12.). Trace files are deposited in the NCBI.
- **Trace ID-to-sample relationship files** – Trace-ID to sample relationship data (see *Trace ID-to-Sample Relationship Files* on page 13). These files are transferred to the DCC.
- **Mutation files** – Mutation Annotation Format (MAF) files (see *Mutation Annotation Format (MAF) Files* on page 14). MAF files are sent to the DCC.

Sequence Trace Files

Sequence trace files contain the raw data output from automated sequencing instruments. GSCs submit these files, with the associated experimental metadata, directly to *NCBI* Trace for dissemination to the public. NCBI Trace assigns each sequence trace record a trace ID and provides that ID back to the submitting GSC.

For more information, see NCBI Trace: http://www.ncbi.nlm.nih.gov/Traces/trace.cgi

Trace files themselves are NOT submitted to the DCC, but the GSCs do transfer to the DCC the trace ID-to-sample relationship files that contain only the NCBI trace ID, (`trace_id`), and the aliquot barcode associated with the trace file submissions.

Trace File Format

Trace files are binary files that have the file extension `.scf` (sequence chromatogram format).

This image from the NCBI Trace website, (*Figure 2.1*), displays a trace file opened in a Trace Archive page.

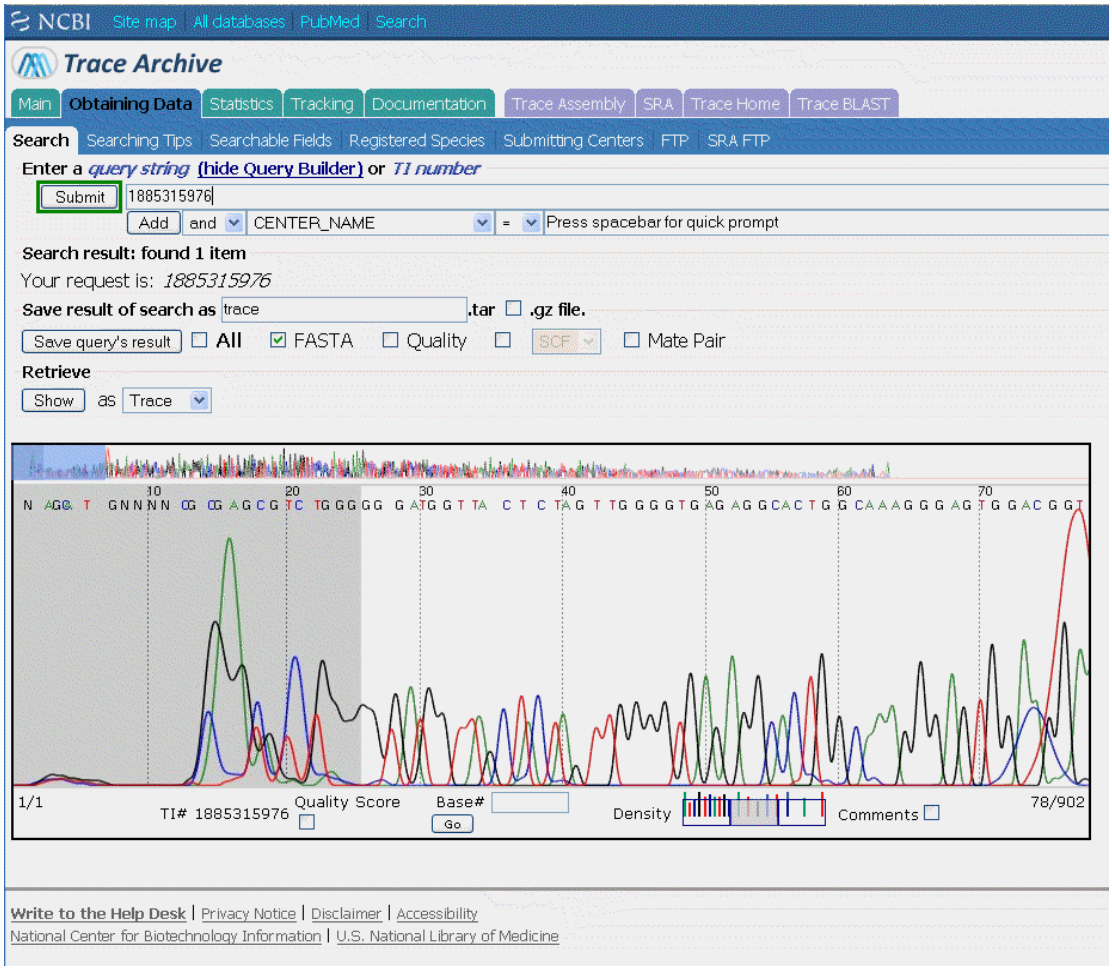


Figure 2.1 Examples of trace file metadata

You can download trace files from the website, which will have a {trace}.tar filename.

Trace ID-to-Sample Relationship Files

Trace ID-to-sample relationship files provide the relationship between NCBI traces and aliquot barcodes. The files contain a listing of NCBI trace IDs and TCGA aliquot barcodes. These trace IDs are IDs assigned to sequence traces submitted by GSCs to NCBI Trace. The TCGA aliquot barcodes are assigned by the BCR to each aliquoted analyte produced from a particular sample and patient. This combination of data (trace IDs and aliquot barcodes) enables researchers to associate sample IDs/aliquot barcodes with assay results. GSCs transfer these trace ID-to-sample relationship files to the DCC for inclusion in the TCGA data repository for public use via FTP.

This data also enables the DCC to query NCBI Trace for additional metadata and to relate this metadata to other experimental results by mapping to BCR biospecimen barcodes.

Because the relationship files include the *BCR* aliquot barcode, researchers can track and record DNA information about a specific patient and make connections between their genes, chromosomal coordinates, tumor types, and so forth. At the same time, trace ID-to-sample relationship files provide patient privacy. The DCC secures the trace relationship/aliquot barcode data in a separate data repository that is accessible to registered research organizations only via a secure FTP (SFTP) site.

Trace ID-to-Sample Relationship File Format

Trace ID-to-sample relationship files have the file extension `.tr`. The data in a trace ID to-sample relationship file is tab-delimited, with no leading spaces.

The files are modeled using the following ordered data elements as column headers:

- `trace_id` (NCBI Trace is `ti`)
- `biospecimen_barcode` (see *About Aliquot Barcodes* on page 7)

Example trace ID-to-sample relationship file name:
`broad.mit.edu_GBM.ABI.1.tr.`

Mutation Annotation Format (MAF) Files

As with trace ID-to-sample relationship files, mutation annotation format (MAF) files contain aliquot barcodes. Those barcodes enable researchers to associate sample IDs with assay results. For more information, see *About Mapping Data* on page 49.

The MAF files annotate mutations discovered by aligning DNA sequences derived from tumor samples to sequences derived from normal samples and a reference sequence.

To create an MAF file, GSCs compare a patient's normal chromosomal sequence with the tumor chromosomal sequence. Any abnormal differences between the two sequences are captured in the mutation file.

GSCs transfer mutation annotation data to the *DCC*. A MAF file identifies, for each sample, the discovered putative or validated mutations and categorizes those mutations (SNP, deletion, or insertion) as somatic (originating in the tissue) or germ-line (originating from the germ-line). These can be further described as follows:

Somatic mutations:

- Missense and nonsense
- Splice site, defined as SNP within 2 bp of the splice junction
- Silent mutations
- Indels that overlap the coding region or splice site of a gene or the targeted region of a genetic element of interest.

SNPs:

- Any germline SNP with validation status "unknown" is included. SNPs already validated in dbSNP are not included since they are unlikely to be involved in cancer.

MAF File Validation

All candidate somatic missense, nonsense, splice site and indels are retested by an independent (orthogonal) genotyping method. If the SNP is confirmed by an independent method, it is deemed valid. Silent mutations may be validated for the purpose of calculating the background mutation rate. No germline (SNP or indel) candidates are processed through validation. However, if the validation process reveals a given candidate somatic variation event to be germline or loss of heterozygosity, those validated data are reported in the validation file.

Mutation File Format

MAF files have the file extension `.maf` (mutation annotation format).

Example: `broad.mit.edu_GBM.ABI.1.maf`.

Table 2.1 lists MAF format column headers.

Column Header	***Description***
`Hugo_Symbol`	HUGO/HGNC symbol for the gene. *Example:* EGFR
`Entrez_Gene_Id`	Entrez Gene ID. *Example:* 1956
`GSC Center`	The genome sequencing center reporting the variant. Either `broad.mit.edu`, `hgsc.bcm.edu` or `genome.wustl.edu`
`NCBI_Build`	NCBI build number; currently build 36 is used by all centers *Example:* 36.1)
`Chromosome`	Chromosome number without prefix. *Example*: X, 1, 2
`Start_position`	Mutation start coordinate. (1-based coordinate system)
`End_position`	Mutation end coordinate; inclusive, 1-based coordinate system
`Strand`	Either + or -
`Variant_Classification`	One of: • Missense_Mutation • Nonsense_Mutation • Silent • Splice_Site_SNP • Frame_Shift_Ins • Frame_Shift_Del • In_Frame_Del • In_Frame_Ins • Splice_Site_Indel
`Variant_Type`	One of: SNP, Ins, or Del
`Reference_Allele`	The plus strand reference allele at this position

Table 2.1 Mutation annotation file fields

Column Header	Description
Tumor_Seq_Allele1	Tumor sequencing (discovery) allele 1
Tumor_Seq_Allele2	Tumor sequencing (discovery) allele 2
dbSNP_RS	dbSNP id. *Example:* rs12345
dbSNP_Val_Status	dbSNP validation status. For example, by_frequency
Tumor_Sample_Barcode	Tumor sample identifier in the BCR aliquot barcode; that is, TCGA-SiteID-PatientID-SampleID-PortionID-PlateID-CenterID. *Example:* TCGA-02-0021-01A-01D-0002-04
Matched_Norm_Sample_Barcode	Normal sample identifier in the BCR aliquot barcode. *Example*: TCGA-02-0021-**10**A-01D-0002-04 as opposed to TCGA-02-0021-**01**A-01D-0002-04
Match_Norm_Seq_Allele1	Matched normal sequencing allele 1
Match_Norm_Seq_Allele2	Matched normal sequencing allele 2
Tumor_Validation_Allele1	Tumor genotyping (validation) allele 1
Tumor_Validation_Allele2	Tumor genotyping (validation) allele 2
Match_Norm_Validation_Allele1	Matched normal genotyping (validation) allele 1
Match_Norm_Validation_Allele2	Matched normal genotyping (validation) allele 2
Verification_Status	One of Valid, wildtype, unknown
Validation_Status	One of Valid, wildtype, unknown
Validation Method	The assay platform used for the validation call
Mutation_Status	One of Germline, somatic, LOH, or unknown

Table 2.1 Mutation annotation file fields (Continued)

About FASTA Files

A FASTA file is a text-based format used to represent either nucleic acid sequences or peptide sequences, in which base pairs or amino acids are represented using single-letter codes. FASTA files are embedded in the Trace files submitted to NCBI. NCBI Trace extracts the FASTA files and makes them available for download.

Because FASTA files are not generated directly by TCGA, this section does not discuss FASTA file extensions.

CHAPTER

3

UNDERSTANDING ARRAY-BASED DATA

This chapter provides an introduction to the types of data and data files that are produced from array characterization assays at the Cancer Genomic Characterization Centers (CGCCs).

Topics in this chapter include:

About Array-Based Data

Data Received by CGCCs

As illustrated in *Figure 1.1* on page 2, the Biological Collection Resource (*BCR*) distributes DNA and RNA samples, identified by their corresponding barcodes to Cancer Genomic Characterization Centers (CGCCs) for analysis.

CGCCs run those analytes in array-based assays to produce genome characterization results such as gene expression, copy number variation, and methylation assays. The results of those assays are transferred to the Data Coordinating Center (DCC).

MAGE-Based References

In the context of TCGA, array-based data is modeled using the Microarray and Gene Expression (MAGE) Object Model (OM). The MAGE-TAB specification is used to represent the MAGE-OM. MAGE-based documents usually represent a *TCGA experiment*.

For more on the MAGE-OM, MAGE-TAB or related information, refer to these sources:

- MIAME
  http://www.mged.org/Workgroups/MIAME/miame.html
- MIAME and MAGE-OM
  http://www.mged.org/Workgroups/MIAME/miame_mage-om.html
- MAGE-OM
  http://www.mged.org/Workgroups/MAGE/mage.html
- MAGE-TAB

  http://www.mged.org/mage-tab/
- MAGE-TAB publication
  http://www.biomedcentral.com/1471-2105/7/489
- MAGE-TAB specification
  http://www.mged.org/mage-tab/
- MGED Ontology
  http://mged.sourceforge.net/ontologies/MGEDontology.php

MAGE and TCGA Experiments

MAGE-based documents usually represent an experiment consisting of many assays, and that experiment usually represents a complete study. In the case of TCGA, an *experiment* for a particular center is composed of all the assays of a particular platform for all the samples of a particular tumor type. Since TCGA mandates that data be made available as soon as possible, many MAGE-TAB documents may be created for an experiment (one *per* archive transferred).

MAGE-TAB Specification

The MAGE-TAB specification discusses four different types of MAGE-TAB files: IDF, SDRF, ADF, and Raw and Processed data files.

CGCCs transfer the following four types of MAGE-TAB files to the DCC:

- **IDF** – Investigation Description Format (see *About Investigation Description Format Files (IDFs)* on page 18).
- **SDRF** – Sample Data and Relationship Format (see *About Sample and Data Relationship Files (SDRFs)* on page 21).
- **ADF** – Array Description Format (see *About Array Description Format Files (ADFs)* on page 23).
- **Data Matrices** - Raw and processed data files (see *About Raw and Processed Data Files* on page 24).

About Investigation Description Format Files (IDFs)

An IDF file is a tab-delimited file that provides general information about the investigation and experiment, including its name, a brief description, the investigator's contact details, bibliographic references, and text descriptions of the protocols used in the investigation.

Figure 3.1 provides an example of an IDF file.

Investigation Title	University of Heidelberg H sapiens TK6		
Experimental Design	genetic_modification_design	time_series_design	
Experimental Factor Name	Genetic Modification	Incubation Time	
Experimental Factor Type	genetic_modification	time	
Experimental Factor Term Source REF	MGED Ontology	MGED Ontology	
Person Last Name	Maier	Fleckenstein	Li
Person First Name	Patrick	Katharina	Li
Person Email	patrick.maier@radonk.ma.uni-heidelberg.de		
Person Phone	+496213833773		
Person Address	Theodor-Kutzer-Ufer 1-3		
Person Affiliation	Department of Radiation Oncology, University of Heidelberg		
Person Roles	submitter; investigator	investigator	investigator
Person Roles Term Source REF	MGED Ontology	MGED Ontology	MGED Ontology
Quality Control Type	biological_replicate		
Quality Control Term Source REF	MGED Ontology		
Replicate Type	biological_replicate		
Replicate Term Source REF	MGED Ontology		
Date of Experiment	2005-02-28		
Public Release Date	2006-01-03		
PubMed ID	12345678		
Publication Author List	Patrick Maier; Katharina Fleckenstein; Li Li; Stephanie Laufs; Jens Zeller; Stefan Fruehauf; Carsten Herskind; Frederik Wenz		
Publication Status	submitted		
Experiment Description	Gene expression of TK6 cells transduced with an oncoretrovirus expressing MDR1 (TK6MDR1) was compared to untransduced TK6 cells and to TK6 cell transduced with an oncoretrovirus expressing the Neomycin resistance gene (TK6neo). Two biological replicates of each were generated and the expression profiles were determined using Affymetrix Human Genome U133 Plus2.0 GeneChip microarrays. Comparisons between the sample groups allow the identification of genes with expression dependent on the MDR1 overexpression.		
Protocol Name	GROWTHPRTCL10653	EXTPRTCL10654	TRANPRTCL10656
Protocol Type	grow	nucleic_acid_extraction	bioassay_data_transformation
Protocol Description	TK6 cells were grown in suspension cultures in RPMI 1640 medium supplemented with 10% horse serum (Invitrogen, Karlsruhe, Germany). The cells were routinely maintained at 37 C and 5% CO2.	Approximately 10^6 cells were lysed in RLT buffer (Qiagen).Total RNA was extracted from the cell lysate using an RNeasy kit (Qiagen).	Mixed Model Normalization with SAS Micro Array Solutions (version 1.3).
Protocol Parameters	media; time	Extracted Product; Amplification	
Protocol Term Source REF	MGED Ontology	MGED Ontology	MGED Ontology
SDRF File	e-mexp-428_tab.txt		
Term Source Name	Cell Type Ontology	MGED Ontology	NCI Metathesaurus
Term Source File	http://obo.sourceforge.net/cgi-bin/detail.cgi?cell	http://mged.sourceforge.net/ontologies/MGEDontology.php	http://ncimeta.nci.nih.gov/indexMetaphrase.html
Term Source Version		1.3.0.1	

Figure 3.1 An example IDF file; from the MAGE-TAB specification document, page 4, accessible from this site. http://www.mged.org/mage-tab/MAGE-TABv1.0.pdf.

All values of attributes in an IDF document remain constant throughout a TCGA experiment with the exception of the following attributes:

- All attributes relating to **Person** *can* change depending on roles
- **Date of Experiment** should change
- **Public Release Date** should change
- **Protocols** *may* change (see *IDF Protocols* on page 20)
- **SDRF Files** should change

IDF File Formats

The following file names are current examples of IDF formats:

- `broad.mit.edu_GBM.HT_HG-U133A.1.idf.txt`
- `mskcc.org_DryRun.AgilentHGCGH244K.1.idf.txt`

File names are derived from the name of archive in which they are contained.

For example, this IDF file:

`broad.mit.edu_GBM.HG-U133_Plus_2.1.idf.txt`

...is derived from this archive name:

`broad.mit.edu_GBM.HG-U133_Plus_2.1.`

In turn, archive names are derived from the following combination of IDs:

```
Domain.domain_tumorType.platform.archiveSerialIndex.
revision.series
```

(*See About Archives on page 40.*)

Note: Archives that include Array Description Format (ADF) files follow a different naming scheme (see *About Array Description Format Files (ADFs)* on page 23). ADF names do not include the index, archive revision, and series indicators.

Archive naming schemes are case sensitive. Names of the files they contain must match the same upper and lower case letters. See *Archive Naming Conventions* on page 40.

IDF Protocols

Protocols are entered in the *IDF* file. They are presented in all submissions for a particular platform; they are also referenced in the corresponding SDRF files (see *About Sample and Data Relationship Files (SDRFs)* on page 21) and in online databases. A protocol name is used as an ID for a protocol, using the following format:

`Domain:ProtocolType:Platform:Version`

For example, `broad.mit.edu:hybridization:HT_HG-U133A:01`.

Table 3.1 describes each part of a protocol entry.

Name	***Description***
`domain`	Matches a center's internet domain name
`protocolType`	Originates from MDEG Ontology subclasses of ProtocolType, for example, Experimental, DataTransformation, HigherLevelAnalysis (http://mged.sourceforge.net/ontologies/MGEDontology.php#ProtocolType).
`protocolTerm`	The source REF (usually MDEG Ontology)
`protocolDescription`	A brief description of the protocol, or its URL.

Table 3.1 Protocol entry formats

Name	Description
protocolParameters	Multiple parameters separated by semicolons. Parameters reflect the “Parameter Value [*]” entries in the SDRF.
platform	Matches the Array Design platform
version	Allows for changes or optimizations of protocol parameters. If a protocol is modified, then the version is incremented

Table 3.1 Protocol entry formats (Continued)

About Sample and Data Relationship Files (SDRFs)

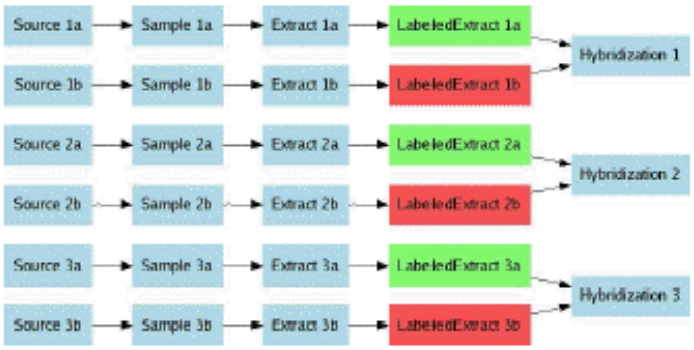
An SDRF file is a tab-delimited file that describes the relationships between samples, array, data, and other objects used or produced in the experiment. An SDRF includes. four kinds of columns:

1. Name--name of the sources and/or samples used in the array
2. Protocol Ref--provides ID(s) for one or more protocols used in the array and referenced in a corresponding IDF or other MAGE document files.
3. File--one or more columns that list files produced in the investigation. *Examples:* Array Data File; Array Derived Data File
4. Attribute--values or characteristics relating to the array. *Examples*: Date, Provider, Performer, Label, Factor Values.

Column types can be used as many times as necessary to adequately describes the use and interaction of materials in the experiment. For more information, see page 34 & 35 of the MAGE-TAB specification document: http://www.mged.org/mage-tab/MAGE-TABv1.0.pdf.

An SDRF is a text-based representation of a directed acyclic graph (DAG). A DAG illustrates what procedures (listed in Protocol REF columns) were used to process a set of samples that produced the resulting experimental results. It shows the relationships between samples, arrays, data, and other objects used or produced in the investigation, and provides all MIAME information that is not provided elsewhere.

An example of a DAG and its corresponding SDRF is shown in *Figure 3.2*



(a) Investigation design graph

Source Name	Sample Name	Extract Name	Labeled Extract Name	Label	Hybridization Name
Source 1a	Sample 1a	Extract 1a	LabeledExtract 1a	Cy3	Hybridization 1
Source 1b	Sample 1b	Extract 1b	LabeledExtract 1b	Cy5	Hybridization 1
Source 2a	Sample 2a	Extract 2a	LabeledExtract 2a	Cy3	Hybridization 2
Source 2b	Sample 2b	Extract 2b	LabeledExtract 2b	Cy5	Hybridization 2
Source 3a	Sample 3a	Extract 3a	LabeledExtract 3a	Cy3	Hybridization 3
Source 3b	Sample 3b	Extract 3b	LabeledExtract 3b	Cy5	Hybridization 3

Figure 3.2 A DAG and the corresponding SDRF. The colors in the DAG represent Cy3 (green) and Cy5 (red) labels in the labeled extracts. From the MAGE-TAB specification document, page 8, accessible from this site. http://www.mged.org/mage-tab/MAGE-TABv1.0.pdf.

As another example, an aliquot of analyte X was labeled, then the labeled extract was hybridized to an array producing a raw file, as in Affymetrix .CEL files. The raw file was normalized with the other assays producing a derived array data matrix file.

The examples describe processes illustrated in DAGs. Each is represented in a corresponding SDRF.

Refer to the MAGE-TAB specification pdf, accessible from this website: http://www.mged.org/mage-tab/MAGE-TABv1.0.pdf. for more information. (Note that protocols are not shown in MAGE-TAB documents.)

The SDRF file is often the most important part of the experiment description due to the complex relationships which are possible between samples and their respective hybridizations. Construction of simple experiment designs are straightforward, but even complex experimental designs can be expressed in an SDRF.

Note: The **Extract Name** column of an SDRF contains the *BCR* aliquot barcode (see *Creating and Identifying Biospecimen Analytes* on page 5). That BCR aliquot barcode maps samples with platforms and experiments (see *Aggregating Data Using the Aliquot Barcode* on page 46).
The **Hybridization Name** column of an SDRF contains the ID that is referenced in the column header for a data matrix.

An example of part of an SDRF file is shown in *Figure 3.3*.

Source Name	Sample Name	Extract Name	Labeled Extract Name	Hybridization Name	Array Data File	Derived Array Data Matrix File
Source 1	Sample 1	Extract 1	LabeledExtract 1	Hybridization 1	Data1.CEL	FGEM.txt
Source 2	Sample 2	Extract 2	LabeledExtract 2	Hybridization 2	Data2.CEL	FGEM.txt
Source 3	Sample 3	Extract 3	LabeledExtract 3	Hybridization 3	Data3.CEL	FGEM.txt
Source 4	Sample 4	Extract 4	LabeledExtract 4	Hybridization 4	Data4.CEL	FGEM.txt

Figure 3.3 An example of part of an SDRF; from the MAGE-TAB specification document, page 8, accessible from this site. http://www.mged.org/mage-tab/MAGE-TABv1.0.pdf.

SDRF File Format

The following file name is an example of the SDRF format.

`broad.mit.edu_GBM.HT_HG-U133A.1.`**`sdrf`**`.txt`

About Array Description Format Files (ADFs)

An ADF is a tab-delimited file in spreadsheet format defining each array type used. An ADF file describes the design of an array, for example, the sequence located at each position on an array and the annotation of this sequence. If the investigation uses arrays for which a description has been previously provided, such as a standard commercial array, cross-references to entries in a public repository (for example, an Array Express accession number) can be included instead of explicit array descriptions.

CGCCs submit ADF files with the associated characterization data to the DCC only if the platform used was non-standard. For example, if the Affymetrix HT_HG-U133A array, a well-known standard platform, was used, no ADF is submitted to the DCC. If, on the contrary, methylation data, for which a non-standard platform was used, is being submitted, an ADF is submitted with the other corresponding data.

Figure 3.4 shows an example of a simple ADF file.

Block Column	Block Row	Column	Row	Reporter Name	Reporter Sequence	Reporter Group [role]	Control Type	Control Type Term Source REF	Composite Element Name
1	1	1	1	R1	ATGGTTGGTTACGTGT	experimental			PTEN
1	1	1	2	R2	CCGCGTTGCCCCGCC	experimental			PAX2
1	1	1	3	R3	CGTAGCTGATCGATGA	experimental			WWOX
1	1	1	4	R4	GGTTGGCTGAGATCGT	experimental			MAPKB
1	1	2	1	R1	ATGGTTGGTTACGTGT	experimental			PTEN
1	1	2	2	R2	CCGCGTTGCCCCGCC	experimental			PAX2
1	1	2	3	R3	CGTAGCTGATCGATGA	experimental			WWOX
1	1	2	4	R4	GGTTGGCTGAGATCGT	experimental			MAPKB
...	...	...	...	...	...	...	...	...	...
4	6	20	20	462020	TCGCTTCGGTTGTCGT	control	control_spike_calibration	MGED Ontology	

Figure 3.4 An example of a simple ADF file; from the MAGE-TAB specification document, page 21, accessible from this site. http://www.mged.org/mage-tab/MAGE-TABv1.0.pdf.

Note: An Array Description File column in an SDRF file contains an ADF file name that is included in the experiment's files. An Array Description REF column in an SDRF file contains an ID that references an array design in some other database that contains the design. The SDRF lists either the ADF as a column or the AD REF as a column, but not both.

Many array designs are available from the platform's vendor, or are deposited in an array database under the listed ID (for example, caArray or ArrayExpress). The specific format is customarily spelled out in an accompanying SOP.

About Raw and Processed Data Files

As described in *About Sample and Data Relationship Files (SDRFs)* on page 21, a series of procedures are used to process a set of samples. The raw and processed files are produced as a procedure is completed. A raw file (array data file) is the initial file produced from an array. Array data files are in the native format of the array-processing software. Array data files may be summarized for one or many assays in an array data matrix file or the raw file may be normalized with other assays to produce a derived array data file. Each procedure further derives the data. MAGE-TAB data matrices (that is, an array data matrix file or a derived array data matrix file) have specified formats.

Raw and processed data files can be ASCII or binary files, typically in their native formats. Alternatively, data may also be provided in the specially-defined tab-delimited format Data Matrix. These files are listed under the ***File**[1] columns in an SDRF file

- Array data files and derived array data files denote raw and processed files in their native formats respectively.
- Array data matrix files and derived array data matrix files denote raw and processed files that summarize the data of the native formats.

1. "***File**" refers to any column header that ends with the word "File". *Examples:* Array Data File; Derived Array Data File; Array Data Matrix File; Image File

CHAPTER 4
CATEGORIZING DATA

This chapter provides an overview of the processes that the Data Coordination Center (DCC) follows to categorize data it receives from the BCR, GSCs, and CGCs.

Topics in this chapter include:

About Data Categorization

Data Received by the DCC

The Data Coordination Center collects and coordinates the data it receives from the BCR, GSCs, and CGCCs. *Table 4.1* provides a description of the types of data submitted by each data source.

Data Source	***Kind of Data Transferred***
BCR	Patient-sample clinical pathology metadata and sample IDs.
GSCs	Trace ID-to-sample relationship files that provide a relationship between the original aliquot barcode and DNA sequencing data, as well as mutation sequence data.
CGCCs	Experimental results for characterization assays, such as gene expression, copy number variation, and methylation.

Table 4.1 Experimental data sources

The DCC archive-processing system processes TCGA data archives automatically when they are received from each data source center. The DCC then makes this processed data available on the bulk distribution site and the TCGA Data Portal (http://tcga-data.nci.nih.gov). The portal provides a user-friendly and searchable view of the

FTP and SFTP sites. For more information about using the data portal, see http://cancergenome.nih.gov/dataportal/contact/tcga_portal_help.asp.

For additional information, see *Chapter 5, Data Access*.

Data Categorization Overview

The DCC classifies data using data type and data level.

- Data type is the kind of data produced by a platform. The second column of *Table 4.3* shows examples of data types.
- Data level attempts to segregate raw data from derived data, from higher-level analysis or interpreted results for each data type, platform, and center. This is designed to make it easier for researchers to locate and access their data of interest. The four category levels range from raw data (level 1) to "region of Interest" (ROI) data (level 4).

Table 4.2 lists and describes each data level.

Data Level	Level Type	Description	Example
1	Raw	Low-level data for a single sample, not normalized across samples, and not interpreted for the presence or absence of specific molecular abnormalities.	Sequence trace file; Affymetrix.CEL file
2	Processed	Data for a single sample that has been normalized and interpreted for the presence or absence of specific molecular abnormalities.	Mutation call for a single sample; amplification/deletion/LOH call for a probed locus in a sample; expression of a splice variant.
3	Segmented/ Interpreted	For genomic copy-number assays, segmented data is processed data for a single sample that has been further analyzed to aggregate individual probed loci into larger contiguous regions.	Amplification/deletion/LOH call for a region in a sample.
4	Summary Finding (ROI)	A quantified association, across classes of samples, among two or more specific molecular abnormalities, sample characteristics, or clinical variables.	A finding that a particular genomic region (a "region of interest") is found to be amplified in 10% of TCGA glioma samples.

Table 4.2 Descriptions of TCGA data levels

Data Type/Data Level Relationships

Understanding Data Type and Data Level Relationships

Each data platform can potentially produce multiple data types.

For example, *Table 4.3* displays current TCGA platforms and corresponding data types. Note that the SNP platforms have more than one data type.

Platform	Data Type
Affymetrix Human Exon 1.0 ST Array	Expression-Gene
Affymetrix Human Exon 1.0 ST Array	Expression-Exon
Affymetrix Genome-Wide Human SNP Array 6.0	SNP
Affymetrix Genome-Wide Human SNP Array 6.0	Copy Number Results
Affymetrix Genome-Wide Human SNP Array 6.0	LOH
Illumina DNA Methylation OMA002 Cancer Panel 1	DNA Methylation
Illumina DNA Methylation OMA003 Cancer Panel 1	DNA Methylation
Illumina 550K Infinium HumanHap550 SNP Chip	SNP
Illumina 550K Infinium HumanHap550 SNP Chip	Copy Number Results
Illumina 550K Infinium HumanHap550 SNP Chip	LOH
Biospecimen Metadata - Complete Set	Complete Clinical Set
Biospecimen Metadata - Minimal Set	Minimal Clinical Set
Agilent Human Genome CGH Microarray 244A	Copy Number results
Agilent Whole Human Genome Microarray Kit, 4 x 44K	Expression-Genes
Agilent 8 x 15K Human miRNA-specific microarray	Expression-miRNA
Agilent Human Genome CGH Microarray 44K	Copy Number Results
Agilent Whole Human Genome, 1 x 44K	Expression-Genes
Agilent Human miRNA Microarray	Expression-miRNA
Agilent 244K Custom Gene Expression G4502A-07-1	Expression-Genes
Applied Biosystems Sequence data	Trace-Gene-Sample Relationship
Applied Biosystems Sequence data	Mutations

Table 4.3 TCGA platforms and potential corresponding data types

For the most up to date list of current TCGA data types, see the TCGA Data Center Standard Operating Procedures document in the.zip file available at this site: https://gforge.nci.nih.gov/docman/view.php/265/5004/Data_Preparation_and_Transfer_SOP.zip. This document also discusses the preparation and transfer of data to the TCGA Data Coordinating Center.

In TCGA, for each data type, data levels further segregate raw data from derived data originating from higher-level analysis or interpreted results. Each center and platform may have a slightly different concept of data level depending on their data types, platforms, and the algorithms used for analysis.

Table 4.4 displays a current and normalized list of data levels as they apply to each data type. Data types are the same as those listed in *Table 4.3*. Descriptions of general data levels are provided in *Table 4.2*.

Data Type (Base-Specific	***Level 1 (Raw Data)***	***Level 2 (Normalized/ Processed)***	***Level 3 (Interpreted Segmented)***	***Level 4 (Summary Finding/ROI)***
Clinical-Complete Set	Clinical data for 1 patient	NA	NA	NA
Clinical-Minimal Set	Clinical data for 1 patient	NA	NA	NA
Copy Number Results-CGH	Raw signals per probe	Normalized signals for copy number alterations of aggregated regions, per probe or probe set	Copy number alterations for aggregated/ segmented regions, per sample	Regions with statistically significant copy number changes across samples
Copy Number Results-SNP	NA	*Copy number alterations per probe or probe set*	Copy number alterations for aggregated regions, per sample	Regions with statistically significant copy number changes across samples
LOH-SNP	NA	*LOH Calls per probe set*	Aggregation of regions of LOH per sample	Statistically significant LOH across samples
SNP	Raw signals per probe	Normalized signals per probe or probe set and allele calls	NA	NA
DNA Methylation	Raw signals per probe	Normalized signals per probe or probe set	Methylated sites/ genes per sample	Statistically significant Methylated sites/ genes across samples
Expression-Exon	Raw signals per probe	Normalized signals per probe set	Expression calls for Exons/Variants per sample	Genes with statistically significant alternative splicing across samples
Expression-Gene	Raw signals per probe	Normalized signals per probe or probe set	Expression calls for Genes per sample	Genes of interest across samples

Table 4.4 Data types and corresponding data level descriptions. Italics indicate that some centers do not produce that data level for its corresponding data type and platform.

Data Type (Base-Specific	*Level 1 (Raw Data)*	*Level 2 (Normalized/ Processed)*	*Level 3 (Interpreted Segmented)*	*Level 4 (Summary Finding/ROI)*
Expression-miRNA	Raw signals per probe	Normalized signals per probe or probe set	Expression calls for miRNAs per sample	miRNAs of interest across samples
Trace-Gene-Sample Relationship	Trace file	NA	NA	NA
Mutations	NA	Putative mutations	Validated Somatic mutations	Statistically significant Mutations across samples

Table 4.4 Data types and corresponding data level descriptions. Italics indicate that some centers do not produce that data level for its corresponding data type and platform.

Determining the Data Type/Data Level of a Results File

Determining the data type and data level of GSC and BCR data is straightforward since there are so few data types and files.

Each TCGA result file, which is any file listed in the SDRF in a *Data* File column (such as the **Array Data File** or **Derived Array Data Matrix File** columns in *Figure 4.1*), has a data type and corresponding data level.

Source Name	Sample Name	Extract Name	Labeled Extract Name	Hybridization Name	Array Data File	Derived Array Data Matrix File
Source 1	Sample 1	Extract 1	LabeledExtract 1	Hybridization 1	Data1.CEL	FGEM.txt
Source 2	Sample 2	Extract 2	LabeledExtract 2	Hybridization 2	Data2.CEL	FGEM.txt
Source 3	Sample 3	Extract 3	LabeledExtract 3	Hybridization 3	Data3.CEL	FGEM.txt
Source 4	Sample 4	Extract 4	LabeledExtract 4	Hybridization 4	Data4.CEL	FGEM.txt

Figure 4.1 A representation of an SDRF file from the MAGE-TAB Specification Document, http://www.mged.org/mage-tab/MAGE-TABv1.0.pdf, page 6.

There are currently two methods of identifying the data type and data level of a result file.

1. The first method uses a Data Type-Data Level File-Suffix Matrix. The Matrix table displays TCGA data centers and their corresponding platforms and data type(s). The data level columns display levels of data that the center can submit to the DCC, as well as provides for the mapping of file suffixes to data type and data level in that matrix. Those mappings can be used to identify a file's data type and data level.

   Note: *Figure 4.2* shows a segment of a Data Type Data Level File Suffix Matrix. however, the complete matrix is too large to fit on a page. You can download the matrix in its entirety at this site: https://gforge.nci.nih.gov/frs/download.php/4153/DataType_DataLevel_Matrix.xls.zip.

	A	B	C	D	E	F	G	H	I
1	**Center**		**Platform**	**Data Type**		**Level 1 (Raw)**			
2	**Center- Type**	**Center - Domain**	**Platform - Name**	**Data Type - Base**	**Data Type - Specific**	**Level 1 - Description**	**Level 1 - Suffixes**	**Level 1 - Suffixes - Examples**	**Level 1 - Suffixes - Ac**
3	BCR	intgen.org	Biospecimen Metadata - Complete Set	Clinical	Complete Set	Clinical data for 1 patient	_full.TCGA-[0-9]{2}-[0-9]{4}.xml	_full.TCGA-02-0001.xml	NA
4	BCR	intgen.org	Biospecimen Metadata - Minimal Set	Clinical	Minimal Set	Clinical data for 1 patient	_min.TCGA-[0-9]{2}-[0-9]{4}.xml	_min.TCGA-02-0001.xml	NA
5	CGCC	broad.mit.edu	Affymetrix Genome-Wide Human SNP Array 6.0	SNP	Copy Number Results	Raw signal values per probe	.CEL		
6	CGCC	broad.mit.edu	Affymetrix Genome-Wide Human SNP Array 6.0	SNP	LOH	Raw signal values per probe	.CEL		
7	CGCC	broad.mit.edu	Affymetrix Genome-Wide Human SNP Array 6.0	SNP	SNP	Raw signal values per probe	.CEL		
8	CGCC	broad.mit.edu	Affymetrix HT Human Genome U133 Array Plate Se	Expression	Gene	Raw signal values per probe	.CEL		
9	CGCC	hms.harvard.edu	Agilent Human Genome CGH Microarray 244A	CGH	Copy Number Results	Raw signal values per probe	.txt	TCGA-02-0039-01A-01G-0326-02_US23502331_251469337030_S01_CGH-v4_95_Feb07.txt	
10	CGCC	jhu-usc.edu	Illumina DNA Methylation OMA002 Cancer Panel	DNA Methylation	DNA Methylation	Raw calls per probe	.cy3-cy5-value.txt, .detection-p-value.txt		
11	CGCC	jhu-usc.edu	Illumina DNA Methylation OMA003 Cancer Panel	DNA Methylation	DNA Methylation	Raw calls per probe	.cy3-cy5-value.txt, .detection-p-value.txt		
12	CGCC	lbl.gov	Affymetrix Human Exon 1.0 ST Array	Expression	Exon	Raw calls per probe	.CEL		
13	CGCC	lbl.gov	Affymetrix Human Exon 1.0 ST Array	Expression	Gene	NA	NA	NA	NA
14	CGCC	mskcc.org	Agilent Human Genome CGH Microarray 244A	CGH	Copy Number Results	Raw calls per probe	_CGH-v4_91.txt		
15	CGCC	stanford.edu	Illumina 550K Infinium HumanHap550 SNP Chip	SNP	Copy Number Results	NA	NA	NA	NA
16	CGCC	stanford.edu	Illumina 550K Infinium HumanHap550 SNP Chip	SNP	LOH	NA	NA	NA	NA
17	CGCC	stanford.edu	Illumina 550K Infinium HumanHap550 SNP Chip	SNP	SNP	Raw calls per probe	.idat, .XandYintensity.txt, .Genotypes.txt, .B_allele_freq.txt		
18	CGCC	unc.edu	Agilent 244K Custom Gene Expression G4502A-07-	Expression	Gene	Raw calls per probe	.txt		
19	CGCC	unc.edu	Agilent 244K Custom Gene Expression G4502A-07-	Expression	Gene	Raw calls per probe	.txt		
20	CGCC	unc.edu	Agilent 8 x 15K Human miRNA microarray	Expression	miRNA	Raw calls per probe	.txt		
21	CGCC	unc.edu	Agilent Human Genome CGH Microarray 44K	CGH	Copy Number Results	Raw calls per probe			
22	GSC	broad.mit.edu	Applied Biosystems Sequence data	Mutations	Mutations	NA	NA	NA	NA
23	GSC	broad.mit.edu	Applied Biosystems Sequence data	Trace-Gene-Sample Relationshi	Trace-Gene-Sample Relationshi	Trace File IDs	.tr	broad.mit.edu_GBM.ABI.1.tr	
24	GSC	genome.wustl.ed	Applied Biosystems Sequence data	Mutations	Mutations	NA	NA	NA	NA
25	GSC	genome.wustl.ed	Applied Biosystems Sequence data	Trace-Gene-Sample Relationshi	Trace-Gene-Sample Relationshi	Trace File IDs	.tr	genome.wustl.edu_GBM.ABI.52.tr	
26	GSC	hgsc.bcm.edu	Applied Biosystems Sequence data	Mutations	Mutations	NA	NA	NA	NA
27	GSC	hgsc.bcm.edu	Applied Biosystems Sequence data	Trace-Gene-Sample Relationshi	Trace-Gene-Sample Relationshi	Trace File IDs	.tr	hgsc.bcm.edu_GBM.ABI.2.tr	

Figure 4.2 Example of a segment of a Data Type Data Level File Suffix Matrix.

How the Matrix is used:

The process of determining the data type and data level of GSC-based data is straightforward because all the GSCs submit the same data types using the file suffixes.

If a user knows the data center that submitted the analyte(s) and the platform used to analyze the data, (s)he can figure out a given data type and data level by finding the appropriate data coordinates on the matrix. Likewise, if the user knows the suffix for a set of data, and the data center, (s)he can find the tile with that suffix on the matrix, then determine the data type and data level that correspond to that suffix.

Examples of using the matrix to determine data types/data levels:

Example 1 — Refer to *Table 4.5* on page 32 for this example. *Table 4.5* provides a simplified, partial version of complete Data File Data Type File-Suffix Matrix. It demonstrates a comprehensive list of TCGA data type/data levels and data file suffixes for the contributing TCGA institution included in the table. Descriptions of general data levels are provided in *Table 4.2*.

The Broad Institute at MIT (broad.mit.edu) (*column A*) produces data for TCGA using two platforms: Affymetrix Genome-Wide SNP 6.0 (Genome_Wide_SNP_6) and Affymetrix HT Human Genome U133 Array Plate Set (HT_HG-U133A). (*column B*). By finding these on the matrix, you'll learn that HT_HG-U133A (*row 4*) produces one data type that has four data levels. Each data level for HT_HG-U133A has its own file suffix (*rows E, G, I, and K*). A file with a suffix `level2.txt` from broad.mit.edu for platform HT_HG-U133A (*cell 4G*) has a data type of "Expression-Gene" (*cell 4C*) and a data level of 2, "Normalized signal per probe set" (*cell 4F*).

Example 2 — Refer to *Table 4.5* on page 32 for this example.

The Genome_Wide_SNP_6 platform (*column B*) produces three data types (SNP, Copy Number-SNP, and LOH SNP). Notice that data levels 3 and 4 (*columns H-K*) are not applicable for the SNP data type, and data levels 1 and 2 (*columns D-G*) are not applicable to the Copy Number-SNP and LOH-SNP data types. For each data type there is a different file suffix if that data level is applicable. A file suffix of ".seg.txt" from broad.mit.edu for platform Genome_Wide_SNP_6 (*cell 1I*) has a data type of "Copy Number-SNP" (*cell 1C*) and a data level of 3, "Copy number alterations for aggregated regions, per sample" (*cell 1H*).

	A	B	C	D	E	F	G	H	I	J	K
	Center	**PLATFORM**	**Data Type**	**Level 1**	**Level 1 Suffixes**	**Level 2**	**Level 2 Suffixes**	**Level 3**	**Level 3 Suffixes**	**Level 4**	**Level 4 Suffixes**
1	broad.mit.edu	GENOME_WIDE_SNP_6	Copy Number Results-SNP	NA	NA	Copy number alterations per probe set	.ismpolish.txt	Copy number alterations for aggregated regions, per sample	.seg.txt	Regions with statistically significant copy number changes across samples	.ROI
2	broad.mit.edu	GENOME_WIDE_SNP_6	LOH-SNP	NA	NA	LOH Calls per probe set	.loh.txt	Aggregation of regions of LOH per sample	.loh.seg.txt	Statistically significant LOH across samples	TBD
3	broad.mit.edu	GENOME_WIDE_SNP_6	SNP	Raw signals per probe	.CEL	Normalized signal per probe set, and allele calls	.birdseed.txt	NA	NA	Statistically significant regions across samples	TBD
4	broad.mit.edu	HT_HG-U133A	Expression-Gene	Raw calls per probe	.CEL	Normalized signal per probe set	level2.txt	Calls for Genes per sample	level3.txt	Genes of interest across samples	.ROI
5	genome.wustl.edu	ABI	Trace-Gene-Sample Relationship	Trace File IDs	.tr	NA	NA	NA	NA	NA	NA
	genome.wustl.edu	ABI	Mutations	NA	NA	Putative mutations	.maf	Validated Somatic mutations	TBD	Statistically significant Mutations across samples	TBD
	jhu-usc.edu	ILLUMINADNAMETHYLATION_OMA002_CPI	DNA Methylation	Raw calls per probe	.cy3-cy5-value.txt, .detection-p-value.txt	Normalized calls per probe	.beta-value.txt	Methylated sites/genes per sample	TBD	Statistically significant Methylated sites/genes across samples	TBD

Table 4.5 TCGA Data type data level file-suffix matrix, a simplified and partial version

2. The second method of determining the data type and data level of a result file is associated with characterization of CGCC data using only the documents that CGCCs submit and the complexity of the data type to data level mapping.

   Example — Refer to *Table 4.6* for this example. CGCCs submit MAGE-TAB SDRF files containing the columns **(Comment [TCGA Data Level]** and **Comment [TCGA Data Type]**) that identify the data type and data level of the previous file column. An example of those SDRF columns is provided in *Table 4.6*. To identify the data type and data level of a file, look up that file in the SDRF and then look at the Data Type and Data Level columns that come after that column. For example, the data type and data level of the file "`550002403070007210798 9.G03.CEL`" (the first row in *Table 4.6*) are "Expression-Gene" and "Level 1" respectively.

Scan Name	***Array Data File***	***Comment [TCGA Data Type]***	***Comment [TCGA Data Level]***
TCGA-02-0001-01C-01R-0177-01	550002403070007210798 9.G03.CEL	Expression-Gene	Level 1
TCGA-02-0002-01A-01R-0177-01	550002403070007210798 9.A09.CEL	Expression-Gene	Level 1
TCGA-02-0003-01A-01R-0177-01	550002403070007210798 9.A10.CEL	Expression-Gene	Level 1
TCGA-02-0006-01B-01R-0177-01	550002403070007210798 9.A11.CEL	Expression-Gene	Level 1
TCGA-02-0007-01A-01R-0177-01	550002403070007210798 9.G10.CEL	Expression-Gene	Level 1

Table 4.6 Example of a portion of an SDRF file showing data type and data level columns

CHAPTER 5
DATA ACCESS

This chapter provides an description of methods for accessing data in TCGA.

Topics in this chapter include the following:

About Data Access

As noted previously, the BCR, GSCs, and CGCCs transfer compressed archives of their data files to the DCC. Data that is categorized by the DCC is distributed by bulk download to caBIG™-enabled applications and databases. The data is available to public or other users using three methods:

1. Bulk downloads
   a. Open Access (ftp://ftp1.nci.nih.gov/tcga/)
   b. Controlled Access (sftp://caftps.nci.nih.gov)
2. TCGA Data Portal (http://tcga-data.nci.nih.gov)
3. TCGA Data Access Matrix (“the Matrix”); http://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm)

For more information about the archives, see *About Archives* on page 40.

Bulk Downloads

The three URLs shown in numbers 1 & 2 above provide for direct data access. In other words, the user downloads what was distributed. The Data Access Matrix (#3) is different in that it recreates files.

Bulk download sites have a particular directory structure that classifies distributed data files. *Figure 5.1* provides a pictorial representation of how data is distributed and how to construct URLs to access that data. It is a hierarchy for downloading files in bulk:

To download a file, the user must construct a complete path to the files. The following components, *whose description numbers correspond to the rows of the figure*, describe in detail how paths to the directories are to be constructed.

Note: Blue text in objects represents the part of the directory path that should be concatenated onto the Root URL in the case of using the http or https protocol or onto the Access Control URL in the case of using the FTP or SFTP protocols.

1. **Root**: The top row of the illustration provides the URL for base URL web access, the root directory.
2. **Access Controls**: Inside the root directory are two directories, the second row of boxes (red) in the figure. One represents open access and the other represents controlled access. Append the URL text for either to the root directory URL. This allows any user to programmatically access either directory. Unless a directory contains a file, a user may not see the file using any other method of access.

   After the access control level, (row 2), the user must, for every level down to the file level (row 10 of *Figure 5.1*), append the directory name onto previous concatenated URL.
3. **Access Root**: (Example /gbm).
4. **Tumor type** (*Example:* GBM).
5. **Center type** (*Example:* CGCC)
6. **Center** (*Example:* broad.mit.edu)
7. **Platform** (*Example:* genome_wide_snp_6)
8. **Data Type** (*Example:* snp)
9. **Archive** (*Example:* broad.mit.edu_GBM.Genome_Wide_SNP_6.1.0.0).

See also the legend following *Figure 5.1*.

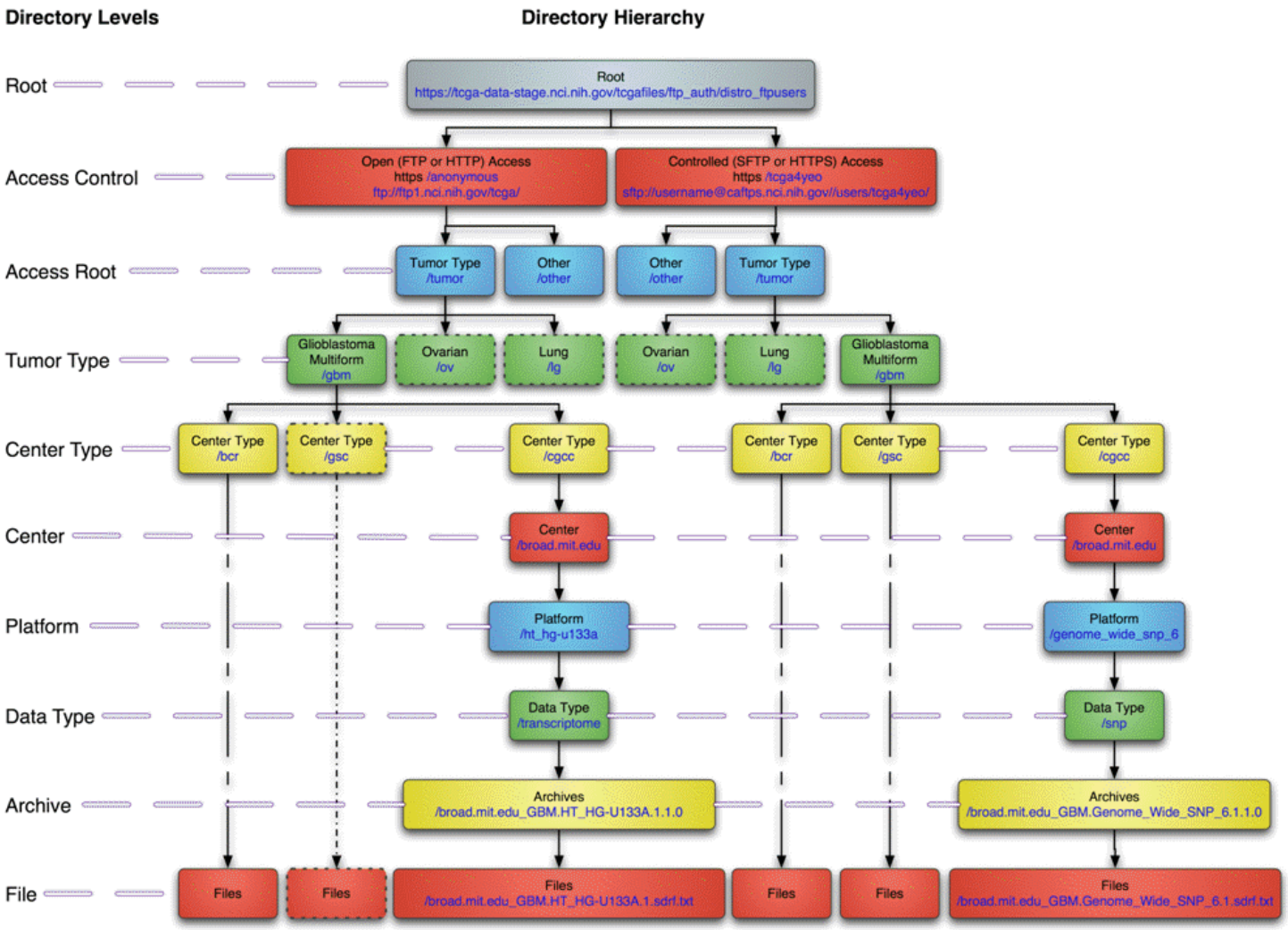


Figure 5.1 Bulk download hierarchy

Figure 4.1 Legend

- Each rectangular object in the illustration represents a directory of a particular type except **File**, which represents data files and the leaves of the hierarchy.
- Each level in the hierarchy represents a level in the directory structure.
- Colors are only meant to distinguish a level from its parent and children levels.
- Objects with dashed outlines represent planned directories.
- Arrowed lines represent the direction further down the hierarchy.
- Large dashed-arrowed lines indicate that child directories for each level do exist but they are not shown to save space in the diagram.
- Small dashed-arrowed lines indicate that child directories for each level are planned.
- Wide-horizontal dashed lines indicate the directory level across objects.

Note: Blue text in objects represent the part of the directory path that should be concatenated onto the Root URL in the case of using the http or https protocol or onto the Access Control URL in the case of using the FTP or SFTP protocols.

These classifications allow perusing for and locating particular datasets as well as programmatic access to the datasets. Downloading of multiple archives or data sets is possible if a FTP/SFTP smart client is used. For example, all the characterization archives can be downloaded in one queue by downloading the CGCC directory. This classification facilitates programmatic download of data by using a consistent directory structure and naming process.

The TCGA Data Portal

The *TCGA Data Portal* provides user friendly access to the FTP and SFTP sites. Data that are considered restricted are placed in the TCGA secure FTP (SFTP) site. In addition, restricted and unrestricted data are deposited into caBIG™-compatible repositories.

TCGA Data Access Matrix

The TCGA project provides the cancer research community with access to data from a variety of sources via a single portal, the Data Access Matrix (the Matrix). Currently, users are able to download entire archives of data as submitted to the Data Coordination Center (DCC) by various Cancer Genome Curation Centers (CGCCs), Genome Sequencing Centers (GSCs), and a Biospecimen Core Resource Center (BCR).

Alternatively, in the Matrix, a user can download specific data sets by cross-selecting a combination of center, platform, data type, data level, or batches of samples. The result of those selections is a subset of TCGA data files specific to those selections.

The Matrix application enables researchers to target data sets in TCGA servers through a user-friendly graphic-based data set selection system.

Figure 5.2 displays a page from the Matrix user interface. A Data Access Matrix User's Guide is available at this site http://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/tcga4yeo/other/TCGA_Portal_Data_Access_User_Guide.pdf.

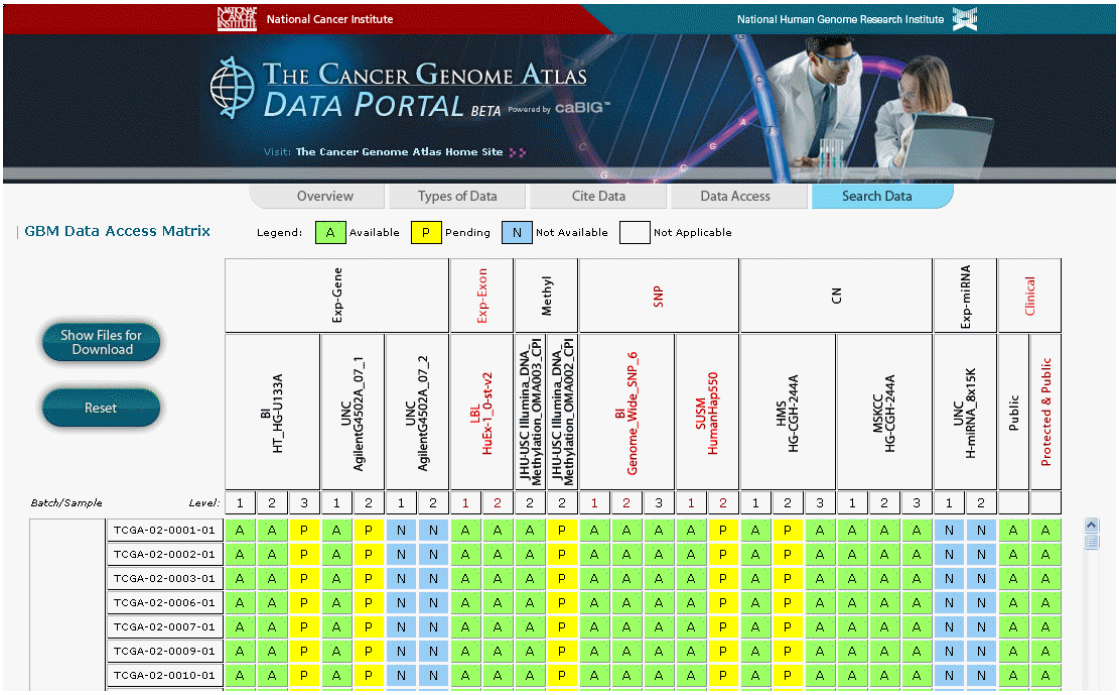


Figure 5.2 A page from the Data Access Matrix

Patient Privacy Issues

The TCGA pilot project produces large volumes of genomic information derived from human tumor specimens collected from patient populations. It also grants access to 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 bioinformatic methods and/or third-party databases. Because patient privacy protection is paramount to NIH and TCGA, human subjects protection and data access policies are being 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 the following two tiers.

1. The open-access data tier is a publicly accessible tier of data that cannot be aggregated to generate a dataset unique to an individual. The open-access data tier does not require user certification for data access.
2. The controlled-access data tier is a controlled-access tier with clinical data and individually unique information. This tier requires user certification for data access.

For more information about these tiers, see http://cancergenome.nih.gov/dataportal/data/access/. To learn how to gain access to the Controlled-Access data see http://cancergenome.nih.gov/dataportal/data/access/closed/.

About Archives

An archive is a directory containing the experimental results of a set of assays conducted on a set of samples. They contain experimental assays from the same platform and tumor type.

Note: A *TCGA experiment* may be represented by many archives because the experiment is defined as the sum of the results of assays for a particular platform from a particular center for all the samples of a particular tumor type.

The following list describes important archival concepts:

- Archives have specific contents and structure (refer to “Anatomy of an Archive” in *TCGA Data Center Standard Operating Procedures.doc*, located at this website: https://gforge.nci.nih.gov/docman/view.php/265/5004/Data_Preparation_and_Transfer_SOP.zip top and bottom bullets
- Archive names require a specific format (see *Archive Naming Conventions* on page 40)
- Archives are compressed before transfer (refer to “Archive Compression” in *TCGA Data Center Standard Operating Procedures.doc*, located at this website: https://gforge.nci.nih.gov/docman/view.php/265/5004/Data_Preparation_and_Transfer_SOP.zip top and bottom bullets.
- Archives are accompanied by a corresponding MD5 file to assure the integrity of an archive (see *Delimiters, such as an underscore and a period, are explicit and intentional. An underscore at the beginning of the name separates the domain name from the rest of the archive name, while periods separate parts of the center’s domain or parts of the rest of the archive name. Archive naming schemes are case sensitive. Names of the files they contain must match the same upper and lower case letters.* on page 41).

Archive Naming Conventions

In the pilot phase of TCGA, samples and data are derived from the following cancer types:

- **GBM** – Brain cancer (glioblastoma multiforme)
- **OV** – Ovarian serous cystadenocarcinoma
- **LG** – Lung squamous adenocarcinoma

Archives are named using the following convention:

```
Domain.domain_n_tumorType.platform.archiveSerialIndex.revision.series
```

Table 5.1 describes each part of an archive name.

Name	*Description*
`Domain.domain`	The domain is the Web site of the center that created the data. For example, the domain for Memorial Sloan-Kettering Cancer Center would be **`mskcc.org`**; the domain for the Broad Institute at MIT would be **`broad.mit.edu`**.

Table 5.1 Archive name format

Name	Description
tumorType	The tumor type is an alphabetical identifier of the tumor being investigated. For example, the abbreviation for Glioblastoma multiforme is **GBM**, Ovarian cancer is **OV**, and Lung cancer is **LG**. (The tumor type is in all caps.)
platform	The platform reflects the assay platform used for investigation. For example, the Broad Institute is using Affymetrix HT HG-U133A to investigate the transcriptome for glioblastoma; therefore, **HT_HG-U133A** is the platform code. For a complete list of platform codes, see *Platform Codes* on page 57.
archiveSerialIndex	The archive serial index is a serially increasing index for the number of archives transferred for a particular platform for a particular tumor type from a particular center.
revision	The revision number indicates the number of times an archive has been revised. The index starts at zero. If an archive is revised and transferred again, revision is incremented. There are situations where the revision of an archive is required. Files that are changed or added to an archive are represented in CHANGES.txt and respectively. For detailed information, refer to "*Anatomy of an Archive*" in *TCGA Data Center Standard Operating Procedures.doc*, located at this website: https://gforge.nci.nih.gov/docman/view.php/265/5004/Data_Preparation_and_Transfer_SOP.zip top and bottom bullets
series	The series number is a serially increasing index for the separate pieces of a large archive that is split into several smaller entities. A series starts at one. If an archive is not split into a series of archives, then series is zero.

Table 5.1 Archive name format (Continued)

Note: Delimiters, such as an underscore and a period, are explicit and intentional. An underscore at the beginning of the name separates the domain name from the rest of the archive name, while periods separate parts of the center's domain or parts of the rest of the archive name.

Archive naming schemes are case sensitive. Names of the files they contain must match the same upper and lower case letters.

Archive Data Freezes

The DCC will from time to time create a list of the archives that comprise a data freeze. That list represents all of the most current, new and revised data up to a certain date.

- Notification of a data freeze is posted to all TCGA listservs including the public TCGA Data listserv (see *Data Access...Other DCC Resources* on page 43).
- A TCGA Portal news item regarding the freeze will be available.
- The freeze lists are always published in the public FTP site under the **Other** directory (*e.g.* ftp://ftp1.nci.nih.gov/tcga/other/

TCGA_Data_Freeze_20080311.txt). A portion of a freeze list is shown in *Figure 5.3*

```
archive_name    date_added      url
broad.mit.edu_GBM.ABI.1.9.0     2008-03-10 00:43:20.514 http://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/dist
broad.mit.edu_GBM.Genome_Wide_SNP_6.1.1.0       2008-03-10 15:06:33.434 http://tcga-data.nci.nih.gov/tcgafil
broad.mit.edu_GBM.Genome_Wide_SNP_6.2.1.0       2008-03-10 17:34:47.06  http://tcga-data.nci.nih.gov/tcgafil
broad.mit.edu_GBM.Genome_Wide_SNP_6.3.1.0       2008-03-10 17:37:31.601 http://tcga-data.nci.nih.gov/tcgafil
broad.mit.edu_GBM.Genome_Wide_SNP_6.4.1.0       2008-03-10 18:34:15.321 http://tcga-data.nci.nih.gov/tcgafil
broad.mit.edu_GBM.Genome_Wide_SNP_6.5.1.0       2008-03-10 20:03:23.519 http://tcga-data.nci.nih.gov/tcgafil
broad.mit.edu_GBM.Genome_Wide_SNP_6.6.0.0       2008-03-10 19:37:16.368 http://tcga-data.nci.nih.gov/tcgafil
broad.mit.edu_GBM.HT_HG-U133A.1.2.0     2008-03-10 00:26:47.569 http://tcga-data.nci.nih.gov/tcgafiles/ftp_a
broad.mit.edu_GBM.HT_HG-U133A.2.1.0     2008-03-10 00:28:22.422 http://tcga-data.nci.nih.gov/tcgafiles/ftp_a
broad.mit.edu_GBM.HT_HG-U133A.3.1.0     2008-03-10 00:30:05.73  http://tcga-data.nci.nih.gov/tcgafiles/ftp_a
broad.mit.edu_GBM.HT_HG-U133A.4.1.0     2008-03-10 00:31:58.465 http://tcga-data.nci.nih.gov/tcgafiles/ftp_a
broad.mit.edu_GBM.HT_HG-U133A.5.1.0     2008-03-10 00:33:47.561 http://tcga-data.nci.nih.gov/tcgafiles/ftp_a
broad.mit.edu_GBM.HT_HG-U133A.6.1.0     2008-03-10 00:35:12.652 http://tcga-data.nci.nih.gov/tcgafiles/ftp_a
broad.mit.edu_GBM.HT_HG-U133A.7.1.0     2008-03-10 00:36:43.241 http://tcga-data.nci.nih.gov/tcgafiles/ftp_a
```

Figure 5.3 A segment of a freeze list

Freeze lists are always labeled by the date of the freeze. The contents of a freeze file are tab-delimited and contain the following columns: archive_name, data_added (to the DCC Bulk Distribution site), and url (a direct URL to download the corresponding archive). Although data will continue to be submitted and distributed, the freeze list should be used as a reference for conducting analysis on a common data set. The list can be referenced in publications using the date of the freeze.

Insuring Data Integrity

When a TCGA file is downloaded, it is wise to confirm that it is not corrupt. This is especially important for very large archives downloaded from the *TCGA Data Portal*.

MD5 hash files can be used to confirm the integrity of archived files. Each TCGA file has a unique and corresponding MD5 hash file that should be downloaded when the selected TCGA file is downloaded. The researcher should also independently create a MD5 hash file from the downloaded TCGA file, and compare the locally-created MD5 file to the MD5 file downloaded from TCGA. If they match, the integrity of the TCGA files is assured.

Table 5.2 lists the programs that can be used to create the MD5 hash locally.

The programs `md5sum` (Unix and Mac OSX) and `md5sums` (Windows) are implementations of the MD5 algorithm for creating MD5 hashes. A MD5 hash file name is the same as its compressed archive counterpart except the suffix is `md5`. For example:

- `broad.mit.edu_GBM.HT_HG-U133A.1.0.0.tar.gz`
- `broad.mit.edu_GBM.HT_HG-U133A.1.0.0.tar.gz.md5.`

Example Type	***Program name***
Unix and Mac OSX	The BSD program `md5sum` creates MD5 hashes. `md5sum archiveName.tar.gz`

Table 5.2 Creating a hash locally

Example Type	Program name
Windows	The Windows program `md5sums` (http://www.pc-tools.net/win32/md5sums/) creates MD5 hashes. `md5sums -u archiveName.tar.gz`

Table 5.2 Creating a hash locally

Data Access...Other DCC Resources

The DCC provides three resources that help make better use of TCGA data. These resources are updated regularly.

- ftp://ftp1.nci.nih.gov/tcga/other/BCR_Biospecimen_Barcodes_Table.tar.gz - A complete list of sample barcodes in an easy-to-use table. The table allows sorting or querying of the barcodes using the constituent parts of the code (see *Chapter 6, Aggregating Data*).

- ftp://ftp1.nci.nih.gov/tcga/other/SampleToFile_AssociationMatrix.tar.gz - The Sample-File Association Matrix relates a sample barcode to the files that were produced during the characterization assay of a sample (see *About Mapping Data* on page 49).

- A public TCGA Data listserv (tcga-data-l@list.nih.gov) is available so that subscribers can be notified of new or revised archives available on the portal. Register at this website: https://list.nih.gov/archives/tcga-data-l.html.

CHAPTER

6

AGGREGATING DATA

This chapter provides details for using aggregation methods to map and analyze TCGA data.

Topics in this chapter include:

About Aggregating Data

A common procedure conducted when analyzing TCGA data is aggregating samples. This involves selecting a subset of samples within a data type or between data types, using either the information contained in the *BCR* aliquot barcode (see *Aggregating Data Using the Aliquot Barcode* on this page) or using the information contained in the clinical metadata (see *Aggregating Data Using Clinical Metadata*).

Aggregating Data Using the Aliquot Barcode

The aliquot barcode is the most important ID, or reference point, in the entire TCGA enterprise. In summary, the barcode, previously described in this document (*About Aliquot Barcodes* on page 7) indicates that the aliquot from particular sample and center was QCed by the BCR and sent with its aliquot barcode to a CGCC, for example, and an assay was performed on it. The center that performed the analysis sent their results (still associated with the aliquot barcode) to the DCC. The DCC categorizes the data, making links between the clinical data and assay results.

With aliquot barcodes, a researcher can aggregate data for the purposes of analysis or comparison. For example, if a researcher is looking for results with particular type of analyte, (s)he can aggregate barcodes and sort the information accordingly. Any information provided by the aliquot barcodes can form the basis for aggregation.

Using an aliquot barcode to aggregate data involves identifying the result files associated with those bar codes. Once you have those files, the method of mapping an

aliquot barcode to its assay-result files varies according to the type of data you want to map.

Aggregation of samples using the aliquot barcode can involve the following processes:

- Splitting all barcodes into their individual IDs to produce an aliquot barcode table.
- Sorting the barcodes using a set of IDs that match parameters of interest.
- Selecting aliquot barcodes matching parameters of interest to compare results with other data types.

Table 6.1 lists the constitutive IDs that comprise the example BCR aliquot barcode, `TCGA-02-0001-01C-01D-0182-01`.

Aliquot Barcode IDs	***Example***
`project`	TCGA is the project.
`collection_center`	02 represents the GBM brain tumor sample from MD Anderson
`patient`	0001 is the first patient from MD Anderson for GBM tumor type
`sample_type`	01 indicates a solid tumor
`sample_sequence`	C is the third vial
`portion_sequence`	01 is the portion code
`portion_analyte`	D indicates a DNA sample
`plate_id`	0182 indicates the plate ID within the 96-well plate.
`center_id`	01 indicates the Broad Institute which is to receive the sample

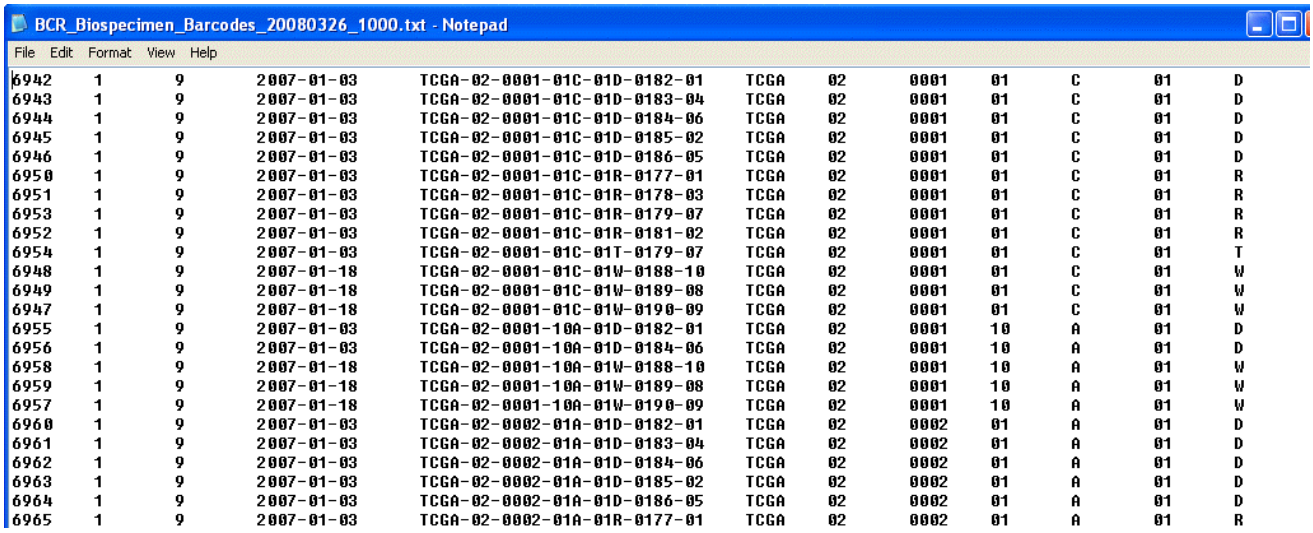
Table 6.1 IDs constitutive to aliquot barcode TCGA-02-0001-01C-01D-0182-01

An aliquot barcode table, which breaks down aliquot barcodes by components, can be used to aggregate aliquots for identifying result files for analysis. The DCC can provides such a table or the researcher can download data and make one locally. A locally-made example is provided in *Figure 6.1*.

BCR Aliquot Barcode	Project Name	Site ID	Patient ID	Sample ID		Portion ID		Plate Barcode	
				sample type	sample sequence	portion sequence	portion analyte	plate ID	center ID
TCGA-02-0001-01C-01R-0177-01	TCGA	2	1	1	C	1	R	177	1
TCGA-02-0002-01A-01R-0177-01	TCGA	2	2	1	A	1	R	177	1
TCGA-02-0003-01A-01R-0177-01	TCGA	2	3	1	A	1	R	177	1
TCGA-02-0006-01B-01R-0177-01	TCGA	2	6	1	B	1	R	177	1
TCGA-02-0007-01A-01R-0177-01	TCGA	2	7	1	A	1	R	177	1
TCGA-02-0009-01A-01R-0177-01	TCGA	2	9	1	A	1	R	177	1
TCGA-02-0010-01A-01R-0177-01	TCGA	2	10	1	A	1	R	177	1

Figure 6.1 Aliquot barcode table

An example of a segment of an aliquot barcode table prepared by the DCC is shown in *Figure 6.2*. When you download an aliquot barcode table from this site, ftp://ftp1.nci.nih.gov/tcga/other/BCR_Biospecimen_Barcodes_Table.tar.gz, a legend describing table column contents is part of the zip file.


BCR_Biospecimen_Barcodes_20080326_1000.txt - Notepad

File Edit Format View Help

6942	1	9	2007-01-03	TCGA-02-0001-01C-01D-0182-01	TCGA	02	0001	01	C	01	D
6943	1	9	2007-01-03	TCGA-02-0001-01C-01D-0183-04	TCGA	02	0001	01	C	01	D
6944	1	9	2007-01-03	TCGA-02-0001-01C-01D-0184-06	TCGA	02	0001	01	C	01	D
6945	1	9	2007-01-03	TCGA-02-0001-01C-01D-0185-02	TCGA	02	0001	01	C	01	D
6946	1	9	2007-01-03	TCGA-02-0001-01C-01D-0186-05	TCGA	02	0001	01	C	01	D
6950	1	9	2007-01-03	TCGA-02-0001-01C-01R-0177-01	TCGA	02	0001	01	C	01	R
6951	1	9	2007-01-03	TCGA-02-0001-01C-01R-0178-03	TCGA	02	0001	01	C	01	R
6953	1	9	2007-01-03	TCGA-02-0001-01C-01R-0179-07	TCGA	02	0001	01	C	01	R
6952	1	9	2007-01-03	TCGA-02-0001-01C-01R-0181-02	TCGA	02	0001	01	C	01	R
6954	1	9	2007-01-03	TCGA-02-0001-01C-01T-0179-07	TCGA	02	0001	01	C	01	T
6948	1	9	2007-01-18	TCGA-02-0001-01C-01W-0188-10	TCGA	02	0001	01	C	01	W
6949	1	9	2007-01-18	TCGA-02-0001-01C-01W-0189-08	TCGA	02	0001	01	C	01	W
6947	1	9	2007-01-18	TCGA-02-0001-01C-01W-0190-09	TCGA	02	0001	01	C	01	W
6955	1	9	2007-01-03	TCGA-02-0001-10A-01D-0182-01	TCGA	02	0001	10	A	01	D
6956	1	9	2007-01-03	TCGA-02-0001-10A-01D-0184-06	TCGA	02	0001	10	A	01	D
6958	1	9	2007-01-18	TCGA-02-0001-10A-01W-0188-10	TCGA	02	0001	10	A	01	W
6959	1	9	2007-01-18	TCGA-02-0001-10A-01W-0189-08	TCGA	02	0001	10	A	01	W
6957	1	9	2007-01-18	TCGA-02-0001-10A-01W-0190-09	TCGA	02	0001	10	A	01	W
6960	1	9	2007-01-03	TCGA-02-0002-01A-01D-0182-01	TCGA	02	0002	01	A	01	D
6961	1	9	2007-01-03	TCGA-02-0002-01A-01D-0183-04	TCGA	02	0002	01	A	01	D
6962	1	9	2007-01-03	TCGA-02-0002-01A-01D-0184-06	TCGA	02	0002	01	A	01	D
6963	1	9	2007-01-03	TCGA-02-0002-01A-01D-0185-02	TCGA	02	0002	01	A	01	D
6964	1	9	2007-01-03	TCGA-02-0002-01A-01D-0186-05	TCGA	02	0002	01	A	01	D
6965	1	9	2007-01-03	TCGA-02-0002-01A-01R-0177-01	TCGA	02	0002	01	A	01	R

Figure 6.2 Aliquot barcode table prepared by the DCC

An aliquot barcode table can be used for common barcode aggregations. A researcher would sort or select aliquot barcodes from the table using selected or favorite headers. For example, a person could sort by analyte of type D to get all the barcodes for DNA. Alternatively, a person could sort or analyze the data, using any of the following analyses:

- Samples of particular types with particular types of analytes - sample type (tumor samples (01-09) *vs.* normal samples (10-19)) and/or portion analytes (DNA (D) *vs.* RNA (R) *vs.* whole genome amplified DNA (W or G))
- Results from particular centers for particular analytes – center ID (01 … 10) and portion analytes (*e.g.* broad.mit.edu CGCC (01) SNP-based (DNA) data (D))
- Batch studies
  - Differences between samples - sample ID (*e.g.* 01A … 01Z vials)
  - Differences between portions - portion ID (*e.g.* 01D … 99Z portion analytes)
  - Differences between plates - plate ID (*e.g.* 0001-04 … 9999-04)

For more information, see *Aggregating Data Using the Aliquot Barcode* on page 45.

Aggregating Data Using Clinical Metadata

Aggregation of samples using clinical metadata involves understanding the BCR UML model (see *Understanding Array-Based Data* on page 17) and BCR XML schema, and parsing the clinical XML. That is not a trivial task. For example, one could parse all the XML files into comma-separated value (CSV) files representing the major BCR XML elements of interest. (See *Figure 6.3*.) Each major element file would contain barcodes for that element and its parent's barcodes to facilitate mapping between elements. For example, a portion file would list the barcodes for portion, sample, and patient; an aliquot file would list the barcodes for aliquot, analyte, portion, sample, and patient. The

DCC plans to create and distribute files representing the major clinical data elements, such as `aliquot_csv.txt` or `analyte_csv.txt`.

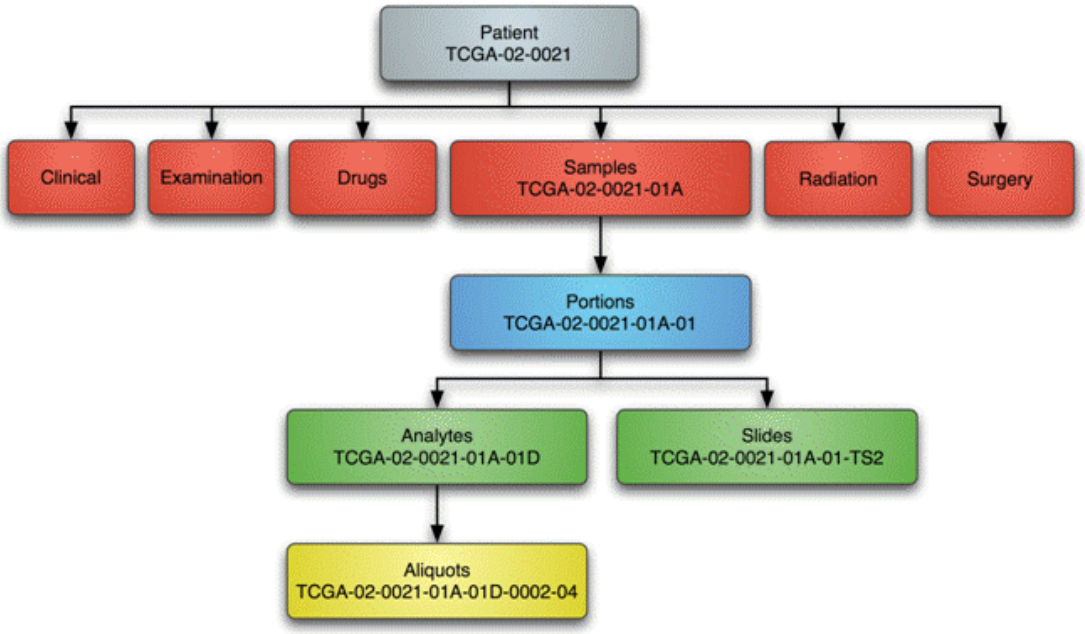


Figure 6.3 Major BCR XML elements. Elements with associated barcodes display an example barcode

The following are the CSV file names representing eleven of the major BCR XML elements:

- `aliquot_csv.txt`
- `analyte_csv.txt`
- `clinical_csv.txt`
- `drug_csv.txt`
- `examination_csv.txt`
- `portion_csv.txt`
- `protocol_csv.txt`
- `radiation_csv.txt`
- `sample_csv.txt`
- `slide_csv.txt`
- `surgery_csv.txt`

The prefix of each file name represents a corresponding major clinical data element.

Once you have parsed available XML files or CSV files like those described above, aggregating samples using clinical data is straightforward. Choose the clinical data elements of interest and apply those as parameters in an XML search or sort the data using a spreadsheet for the data elements of interest in the CSV files.

The objective of aggregation is to produce a list of aliquot barcodes for mapping to result files, a process described in *Mapping Data Levels 1 and 2* on page 49 and *Mapping Data Levels 3 and 4* on page 50.

Aggregating Data Between Different Centers and Platforms Using Sample Data

Matching or comparing results from two difference centers or platforms on the same sample presents a unique challenge.

It is not possible to *match* aliquot barcodes (*for example,* TCGA-02-0021-01A-01D-0002-04) between results of different centers or platforms. Aliquot barcodes are specific to a particular center and platforms because the barcodes contain plate barcodes. The plate barcode (see *Deciphering Plate Barcodes* on page 9) includes a center ID and plate ID. The biospecimen barcode (removing the plate barcode from the aliquot barcode; *e.g.* TCGA-02-0021-01A-01D) is not center or platform specific, however few biospecimen barcodes are in common between centers' data. Sample barcodes are the best choice for matching results between different centers or platforms.

A sample barcode is composed of the Project Name, Site ID, Patient ID, and the sample type of the Sample ID (*for example,* TCGA-02-0021-01). It is possible to query an aliquot barcode table (see *Aggregating Data Using the Aliquot Barcode* on page 45) to obtain a set of aliquot barcodes that match the sample barcode using the constitutive parts of the sample barcode. Aggregating results between different centers or platforms using the sample barcode involves using an aliquot barcode table as an intermediary. To do this, follow these steps.

1. Create a list of aliquot barcodes from the files from one center or platform
2. Query an aliquot barcode table using the list you created in step 1. The query is on the complete aliquot barcode column.
3. Create a unique list of constituent IDs for a sample barcode for each of the matches in step 2.
4. Create a list of aliquot barcodes using the list created in step 3 as a query on the aliquot barcode table. The query is on the corresponding columns in step 3.
5. Use the aliquot barcode list in step 4 to match to result files (see the following section).

About Mapping Data

Mapping Array-Based Data

Aggregated CGCC data can be mapped for characterizing the information. Mapping involves MAGE-TAB SDRF files and TCGA Higher Level Analysis (HLA) specification files. Because of the variations in data mapping between data levels 1 and 2 versus 3 and 4, these are described in separate sections.

Mapping Data Levels 1 and 2

The MAGE-TAB SDRF files provides mapping between aliquot barcodes, result files, and those files' data types and data levels. For a review, see *About Sample and Data*

Relationship Files (SDRFs) on page 21 and *Understanding Data Type and Data Level Relationships* on page 26).

The SDRF file is essentially a database describing the relationships between samples and their results. Refer to *Table 6.2* for this example.

The **Extract Name** column contains the BCR aliquot barcode for each sample. The ***File** columns, such as the Derived Array Data Matrix File column in , provide a listing of the result files associated with an aliquot barcode. If a file is listed in the same row as a barcode, then that file is associated with that aliquot's assay result.

Note: It is possible to have one-to-many and many-to-many relationships between Extract Names and *File columns. Usually, however, the relationships are one-to-one or many-to-one..

Extract Name	*Derived Array Data Matrix File*
TCGA-02-0001-01C-01D-00182-01	broad.mit.edu_GBM.Genome_Wide_SNP_6.1.ismpolish.data.txt
TCGA-02-0001-10A-01D-00182-01	broad.mit.edu_GBM.Genome_Wide_SNP_6.1.ismpolish.data.txt
TCGA-02-0002-01A-01D-00182-01	broad.mit.edu_GBM.Genome_Wide_SNP_6.1.ismpolish.data.txt
TCGA-02-0002-10A-01D-00182-01	broad.mit.edu_GBM.Genome_Wide_SNP_6.1.ismpolish.data.txt

Table 6.2 Example of Data Levels 1-2 Sample ID-to-Result File mapping for CGCC data

The DCC provides a sample-to-file association matrix (ftp://ftp1.nci.nih.gov/tcga/other/SampleToFile_AssociationMatrix.tar.gz) that contains the relationships described above for data levels 1 and 2. A similar matrix for levels 3 and 4 is planned.

Mapping Data Levels 3 and 4

Some TCGA data for levels 3 and 4 are described by TCGA Higher Level Analysis (HLA) specification (https://gforge.nci.nih.gov/docman/view.php/265/8841/HLA_SOP.zip). Refer to *Table 6.3* for this example.

HLA files are listed under **Derived HLA Data File** columns in the HLA SDRF. Sample ID mapping for data levels 3 and 4 requires mapping between HLA SDRF files and MAGE-TAB SDRF files.

For example, if an HLA SDRF file contains a **Derived Array Data Matrix File** column, the files listed in that column are also found in a MAGE-TAB SDRF. Therefore, the HLA files are associated with a set of sample IDs through the files in the **Derived Array Data Matrix** column..

Derived Array Data Matrix File	*Derived HLA Data File*
broad.mit.edu_GBM.Genome_Wide_SNP_6.1.ismpolish.data.txt	broad.mit.edu_GBM.Genome_Wide_SNP_6.1.seg.txt
broad.mit.edu_GBM.Genome_Wide_SNP_6.1.ismpolish.data.txt	broad.mit.edu_GBM.Genome_Wide_SNP_6.1.seg.txt
broad.mit.edu_GBM.Genome_Wide_SNP_6.1.ismpolish.data.txt	broad.mit.edu_GBM.Genome_Wide_SNP_6.1.seg.txt
broad.mit.edu_GBM.Genome_Wide_SNP_6.1.ismpolish.data.txt	broad.mit.edu_GBM.Genome_Wide_SNP_6.1.seg.txt

Table 6.3 Example of Data Levels 3-4 Sample ID-to-Result File mapping for CGCC data

Mapping Sequence-Based Data

There is no equivalent to an SDRF file for sequence-based (GSC) data, and the DCC does not yet provide GSC mapping data in the sample-to-file association matrix. GSC data (see *Understanding Sequence-Based Genomic Data* on page 11) does provide two types of files containing aliquot barcodes: trace relationship (tr) files and mutation (maf) files. Trace ID-to-Sample relationship files contain aliquot barcodes (**biospecimen_barcode** column) and NCBI trace IDs. Mutation files contain aliquot barcodes for tumor (**Tumor_Sample_Barcode** column) and normal samples (**Matched_Norm_Sample_Barcode** column). Therefore, to associate barcodes with GSC files, the files must be downloaded and parsed for each barcode.

Mapping Between File Elements

TCGA data is complex in that a data file contains references to elements that may be external to the file, the data type itself, or even external to TCGA. *Figure 6.4* on page 52 attempts to illustrate mapping between those different elements.

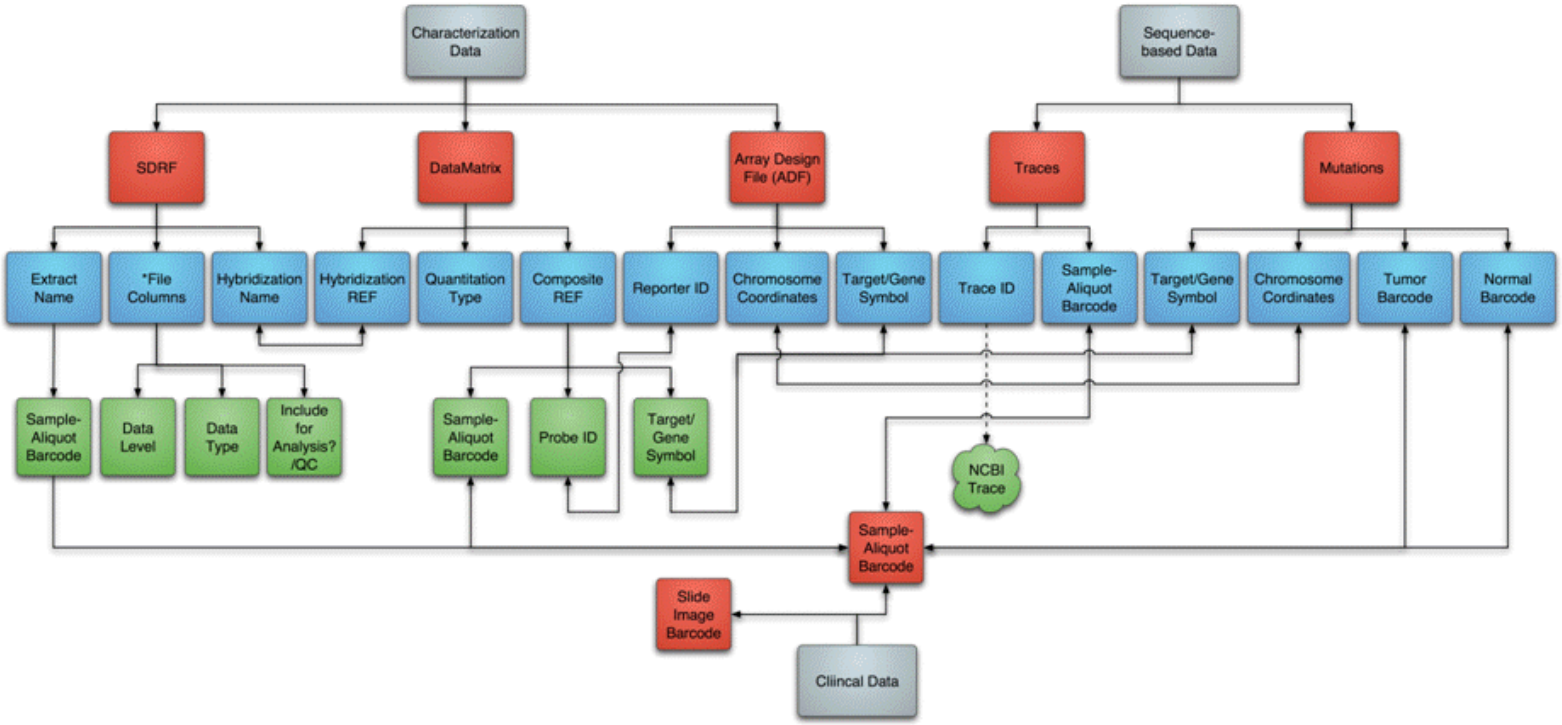


Figure 6.4 Linking between TCGA file elements. Grey objects represent classifications of data. Red objects represent files. Blue objects represent elements of interest in those files. Green-square objects represent the element values contained within the blue elements. If there is no green object as a child of blue element-object, then those blue objects also represent the values of that object. Cloud objects and dashed-arrowed lines represent mappings of data outside of TCGA. Solid-arrowed lines represent mappings between elements or their values.

APPENDIX

A

ALIQUOT BARCODE VALUES

This appendix provides values for unique identifiers (IDs) and codes that compose TCGA barcodes. For detailed information, see *About Aliquot Barcodes* on page 7.

Identifiers and codes in this appendix include:

- *Analyte Barcode Values* on page 53
- *Plate Barcode Values* on page 55

Analyte Barcode Values

Analyte barcodes consist of several unique identifiers, including the following:

- Site ID
- Patient ID
- Sample ID
- Portion ID

Site ID Values

SiteIDs represent TCGA centers. *Table A.1* provides values of all current TCGA sites.

Site ID	***Value***
01	International Genomics Consortium
02	MD Anderson Cancer Center - Brain Bank
03	Lung Cancer Tissue Bank of CALGB
04	Gynecologic Oncology Cancer Group
05	National Cancer Institute
06	Henry Ford Hospital

Table A.1 Site identifiers

Site ID	Value
07	Cell Lines
08	UCSF - Brain Bank
09	UCSF - Ovarian Bank
10	MD Anderson - Ovarian Bank
11	MD Anderson - Lung Bank
12	Duke University - Brain Bank

Table A.1 Site identifiers (Continued)

Patient ID Values

Patient IDs range from 0001 to 9999 per collection site. That is each site (siteID) can have up to 9999 patients.

Sample ID Values

Sample IDs are composites of sample type and vial identifiers. For example, the ID `01A` is the first vial (`A`) of a solid tumor (`01`), and `01B` is the second vial of a solid tumor from the same patient. Values for sample types and vials are provided in *Sample Type Values* and *Vial Identifier Values*.

Sample Type Values

Sample type values range from `01-09` for tumor types, `10-19` for normal types, and `20-29` for control samples.

Table A.2 provides values for sample types

Sample Type	Value
01	solid tumor
10	normal blood
11	normal tissue
12	buccal smear
20	cell line

Table A.2 Sample Type values

Vial Identifier Values

Vial counts pertain to an individual patient-sample. Values are A-Z.

For example:

`A` is the first vial from a given sample from a given patient

`B` is the second vial from the same sample and same patient

Portion ID Values

Portion IDs are a composite of portion and analyte identifiers. Solid tumors are divided into a sequence of 100 - 120 mg sections called portions. Each portion has a two-digit ID.

For example:

PortionID `15D` is the 15th portion (portion code) of a sample for DNA (analyte code) analysis.

Portion Code Values

Portion code values range from 01 to 99. They identify the section of a tissue sample.

Analyte Code values

Analyte codes represent the type of analyte for which the sample is analyzed. *Table A.3* provides analyte code values.

Analyte Code	***Values***
D	DNA
R	RNA
T	Total RNA (contains small RNA and is used mainly for mRNA assays)
W	Whole Genome Amplified (WGA) DNA produced by Qiagen
G	WGA DNA produced by Rubicon Genomics using GenomePlex

Table A.3 Analyte Code values

Plate Barcode Values

Plate barcodes are a composites of plate and center identifiers. Values for plate and centers are provided in *Plate ID Values* and *Center ID Values*.

Plate ID Values

Plate IDs range from 0001 - 9999 (up to 9999 96 well plates).

Center ID Values

Center IDs represent the CGCCs and GSCs. *Table A.4* provides Center ID values.

Center ID	***Plate Recipient Values***
01	CGCC - Broad (broad.mit.edu)
02	CGCC - Harvard (hms.harvard.edu)
03	CGCC - Lawrence Berkeley (lbl.gov)
04	CGCC - Memorial Sloan-Kettering (mskcc.org)
05	CGCC - Sidney Kimmel Baylor (jhu-usc.edu)
06	CGCC - Stanford (stanford.edu)

Table A.4 Center ID values

Center ID	Plate Recipient Values
07	CGCC - UNC (unc.edu)
08	GSC - Broad (broad.mit.edu)
09	GSC - Washington Univ (genome.wustl.edu)
10	GSC - Baylor College of Medicine (hgsc.bcm.edu)

Table A.4 Center ID values (Continued)

APPENDIX B
PLATFORM CODES

Table B.1 lists all the platforms used in TCGA and their assigned abbreviation.

Platform Name	*Platform Code*
Affymetrix HT Human Genome U133 Array Plate Set	HT_HG-U133A
Affymetrix Human Exon 1.0 ST Array	HuEx-1_0-st-v2
Affymetrix Genome-Wide Human SNP Array 6.0	Genome_Wide_SNP_6
Illumina DNA Methylation OMA002 Cancer Panel I	IlluminaDNAMethylation_OMA002_CPI
Illumina DNA Methylation OMA003 Cancer Panel I	IlluminaDNAMethylation_OMA003_CPI
Illumina 550K Infinium HumanHap550 SNP Chip	HumanHap550
Biospecimen Metadata - Complete Set	bio
Biospecimen Metadata - Minimal Set	minbio
Agilent Human Genome CGH Microarray 244A	HG-CGH-244A
Agilent Whole Human Genome Microarray Kit, 4 x 44K	WHG-4x44K_G4112F
Agilent 8 x 15K Human miRNA-specific microarray	H-miRNA_8x15K
Agilent Human Genome CGH Microarray 44K	WHG-CGH_4x44B
Agilent Whole Human Genome, 1 x 44K	WHG-1x44K_G4112A
Agilent Human miRNA Microarray	H-miRNA_G4470A

Table B.1 TCGA Platform codes

APPENDIX C
GLOSSARY

Acronyms, objects, tools and other terms referred to in the chapters and appendixes of this document are described in this glossary.

Term	***Definition***
Archive	A directory containing files from the experimental results of a set of assays conducted on a set of samples.
ADF	Array Description Format file
Astrocytic tumors: Astrocytoma	Neoplasms of the brain and spinal cord derived from glial cells. Also called an astrocytoma.
BCR	Biospecimen Core Resource
CGCC	Cancer Genome Characterization Centers use advanced, complementary analysis technologies to strategically characterize genomic changes for brain (glioblastoma multiforme), lung (squamous cell), and ovarian serous cancer.
DCC	TCGA Data Coordinating Center (DCC) manages data entered into public databases as it becomes available.
experiment	An experiment for a individual center consists of all the assays of a particular platform for all the samples of a particular tumor type.
GSC	Genomic Sequencing Centers perform high-throughput genomic sequencing.
HUGO [gene symbol] [Human Genome Organization]	HUGO is an international organization of scientists involved in human genetics. Established in 1989 by a collection of the world's leading human geneticists. Promotes and sustains international collaboration in the field of human genetics.
IDF	Investigation Description Format file
LOH	

Table C.1 Glossary of genomic analysis terms

Term	***Definition***
NCBI	National Center for Biotechnology Information
NCI	National Cancer Institute
NCICB	National Cancer Institute Center for Bioinformatics
Oligodendroglial tumor: Oligodendroglioma	Rare, slow-growing tumor that grows in the oligodendrocytes (brain cells that provide support and nourishment for nerve cells). Also called an oligodendroglioma.
SDRF	Sample Data and Relationship File
SNP	Single nucleotide polymorphisms or SNPs (pronounced “snips”) are DNA sequence variations that occur when a single nucleotide (A,T,C or G) in the genome sequence is altered.
TCGA	The Cancer Genome Atlas
TCGA Data Portal	Stores all data generated from the TCGA Pilot Project and serves as the access point for the datasets.

Table C.1 Glossary of genomic analysis terms (Continued)

INDEX