TCGA
DATA PRIMER

Version 1.2

NATIONAL® CANCER INSTITUTE

Center for Biomedical Informatics and Information

November 10, 2009

REVISION HISTORY

The following is the revision history for this document.

Date	*Description*	*Revised By*
12/29/08	Major copy editing throughout document; complete revision of chapter 6	Jill Hadfield
9/28/09	Major document review	Ari Kahn
10/9/09	Incorporation of review comments	Jill Hadfield

Credits and Resources

TCGA Data Coordinating Center (DCC) Credits			
Software Engineering	***Data Modeling***	***Data Primer Writing***	***Program Management***
Robert Sfeir[1]	Ari Kahn[1]	Ari Kahn[1]	Carl Schaefer[2]
David Nassau[1]	Jessica Chen[1]	Laura Jackel[4]	Subhashree Madhavan[2]
Jessica Chen[1]	Robert Sfeir[1]	Lauren Anthone[4]	Matthew Shaker[1]
Larry Feng[1]		Jill Hadfield[2]	Robert Sfeir[1]
David Kane[1]			David Kane[1]
Jessica Chen[1]			Ari Kahn[1]
Ari Kahn[1]			
Systems Engineering		***DCC SOP Writing***	
Gavin Brennan[3]		Ari Kahn[1]	
Ralph Rutherford[3]			
Sriram Kalyanasundaram[3]			
David Smith[1]			
[1] Systems Research and Applications International, Inc. (SRA)		[2] National Cancer Institute Center for Biomedical Informatics and Information Technology (CBIIT)	
[3]Terrapin Systems	[4]Lockheed Martin		

Contacts and Support	
NCICB Application Support	http://ncicb.nci.nih.gov/NCICB/support Telephone: 301-451-4384 Toll free: 888-478-4423 Email: ncicb@pop.nci.nih.gov Telephone support is available Monday to Friday, 8 am - 8 pm, Easter Time, excluding government holidays.

TCGA Resources	
Name	***URL***
TCGA Data Portal	http://tcga-data.nci.nih.gov/tcga/
TCGA Data Access Matrix	http://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm
caBIG® Cancer Molecular Analysis Portal	http://tcga.cancer.gov/dataportal/data/cma/
TCGA - Data Listserv	https://list.nih.gov/archives/tcga-data-l.html

TABLE OF CONTENTS

Revision History i

Credits and Resources iii

Chapter 1
About TCGA Data 1

Overview of the TCGA 1
TCGA Data Flow Overview 1
Data Flow Description 2
Experiment Archives and File Formats 4
Creating and Identifying Biospecimen Analytes 5
Processing Analytes 6
About Aliquot Barcodes 7
Deciphering Analyte Barcodes 7
Deciphering Plate Barcodes 8
Distributing Aliquots to GSCs and CGCCs 9

Chapter 2
Understanding Sequence-Based Genomic Data 11

About Sequence-Based Data Files 11
Data Received by the GSCs 11
Understanding Sequence Trace Files 12
Trace File Format 12
Understanding Trace ID-to-Sample Relationship Files 13
Trace ID-to-Sample Relationship File Format 14
Understanding Mutation Annotation Format Files 14
MAF File Validation 14
Mutation File Format 15
About FASTA Files 16

Chapter 3
Understanding Array-Based Data 17

About Array-Based Data 17
Data Received by Cancer Genomic Characterization Centers 17
MAGE-Based Data 17
MAGE and TCGA Experiments 18
MAGE-TAB Specification 18
About Investigation Description Format Files (IDFs) 18
IDF File Formats 20
IDF Protocols 20
About Sample and Data Relationship Files (SDRFs) 21
SDRF File Format 22
About Array Description Format Files 22
About Raw and Processed Data Files 23

Chapter 4
Categorizing Data 25

About Data Categorization 25
Data Categorization Overview 26
Understanding Data Type/Data Level Relationships 27
Determining the Data Type/Data Level of a Results File 29

Chapter 5
Data Access 35

About Archives 35
Archive Naming Conventions 36
Archive Data Freezes 37
About Data Access 38
Accessing Bulk Downloads 38
Accessing Archives Through TCGA Data Portal 41
Patient Privacy Issues 43
Insuring Data Integrity 44
Data Access...Other DCC Resources 44

Chapter 6
Aggregating and Mapping Data 47

About Aggregating and Mapping Data 47
Aggregating Data Using Aliquot Barcodes 48
Aggregating Sample Data Between Different Centers and Platforms 51
Aggregating Data Using Clinical Metadata 54
Working with XML Files 55

Working with Tab-Delimited Files 61
Mapping Aliquot Barcodes to Assay Result Files 67
Mapping Array-Based Data 67
Mapping Sequence-Based Data 68
Mapping Between File Elements 69

Appendix A
Aliquot Barcode Values 73

Analyte Barcode Values 73
Site ID Values 73
Patient ID Values 74
Sample ID Values 74
Sample Type Values 74
Vial Identifier Values 74
Portion ID Values 74
Portion Code Values 75
Analyte Code values 75
Plate Barcode Values 75
Plate ID Values 75
Center ID Values 75

Appendix B
Platform Codes 77

Appendix C
Glossary 79

Index 81

CHAPTER 1
ABOUT TCGA DATA

This chapter provides a high-level description of The Cancer Genome Atlas (TCGA), including the data that it generates.

Topics in this chapter include:

- *Overview of the TCGA* on this page
- *TCGA Data Flow Overview* on page 1
- *Creating and Identifying Biospecimen Analytes* on page 5
- *About Aliquot Barcodes* on page 7
- *Distributing Aliquots to GSCs and CGCCs* on page 9

Overview of the TCGA

The Cancer Genome Atlas (TCGA), a three-year pilot project of the National Cancer Institute and the National Human Genome Research Institute (NHGRI), is the foundation of 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.

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.

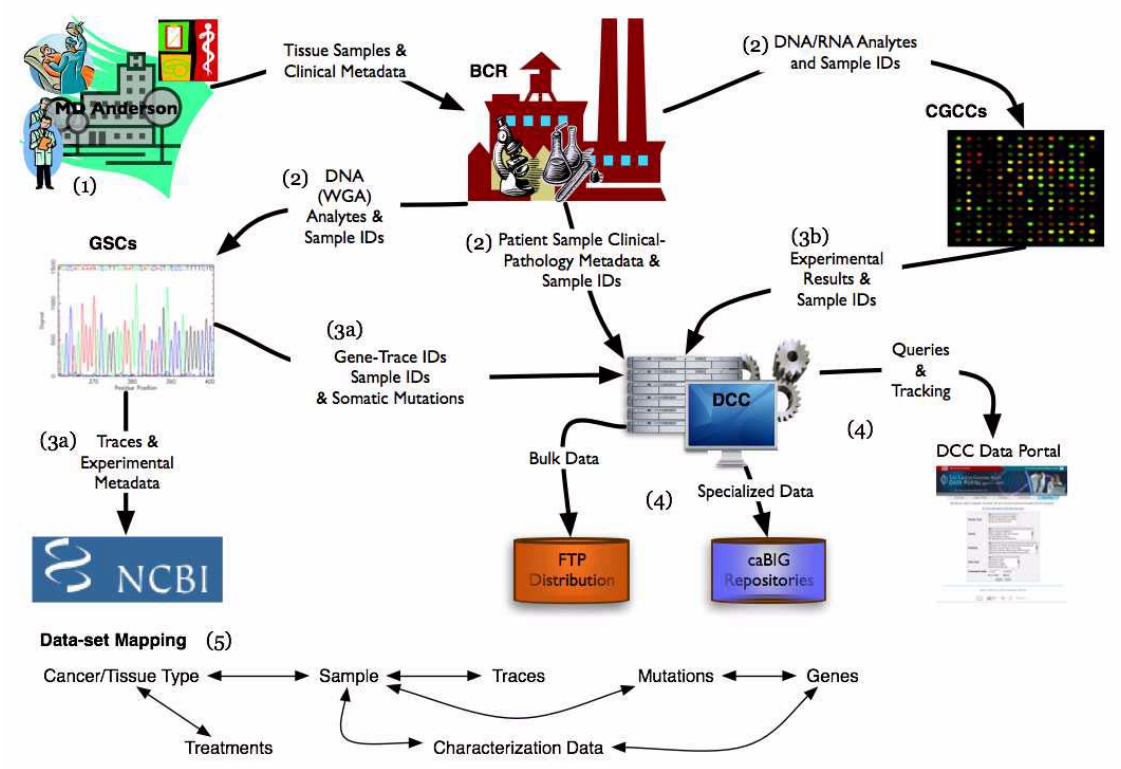


Figure 1.1 Data flow in TCGA

Data Flow Description

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

Com-ponent #	***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.

Table 1.1 Description of the TCGA data flow

Com-ponent #	Data Flow Description
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. Data is modeled and formatted using the MAGE-TAB specification. Additionally, each CGCC may provide an *ADF* file if the platform is non-standard, such as with methylation data. See *About Array Description Format Files* on page 22. For more information, see *Chapter 3, Understanding Array-Based Data,* on page 17.
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, the DCC deposits restricted and unrestricted data 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/.
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_Preparation_and_Transfer_SOP.zip.

Experiment Archives and File Formats

Each center transfers its data to the DCC in digitally compressed archives. All archives include common documents and follow distinct naming conventions. For details about archives, including naming conventions, see *Insuring Data Integrity* on page 44.

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 *Understanding Sequence Trace Files* on page 12.
DCC	Trace ID-to-Sample Relationship files. See *Understanding Trace ID-to-Sample Relationship Files* on page 13.
	Mutation file (MAF). See *Understanding Mutation Annotation Format 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* on page 22.
	Raw and processed array data results. See *About Raw and Processed Data Files* on page 23.

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 BCR assigns aliquot barcodes to the samples and data. BCR aliquot barcodes are the most important type of ID in the TCGA Data Enterprise, as the IDs uniquely identify 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 GSCs and CGCCs for analysis. *Table 1.3* describes the steps in the process.

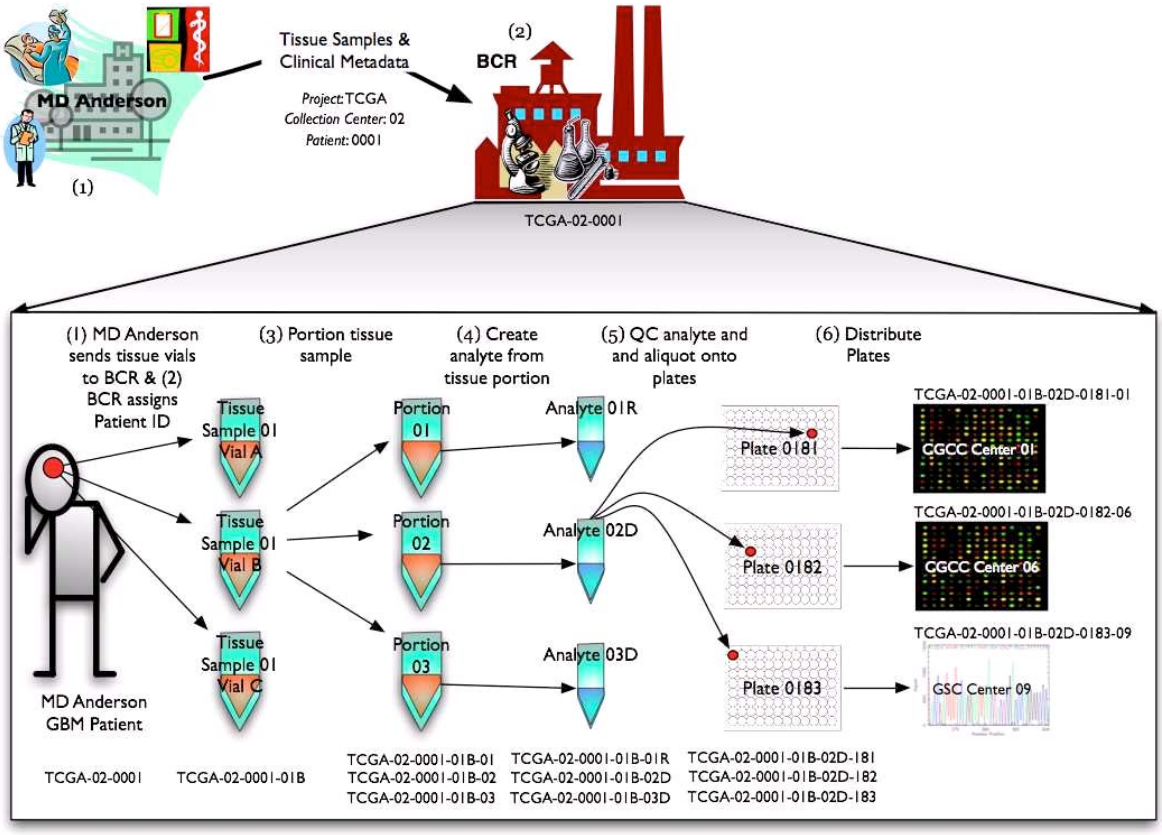


Figure 1.2 BCR's process of creating and coding analytes. See Table 1.3 for details.

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.The BCR codes each sample 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 using caBIG® standards in XML-formatted files. For more information about the available UML model, navigate to caDSR Contexts > caBIG > BiospecimenCoreResource at this location: http://umlmodelbrowser.nci.nih.gov/umlmodelbrowser/). Additionally, the BCR provides an XML schema of the model in each BCR archive.

About Aliquot Barcodes

The BCR creates an *aliquot barcode* to identify and track the distribution of each product from a collection site. 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 8.)

Figure 1.1 and *Table 1.3* provide examples of the constituents of barcodes. *Appendix A, Aliquot Barcode Values* on page 73 provides barcode values.

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 73.

Identifier	***Description***
ProjectName	Current project name. Currently TCGA is the only project.
SiteID	Tissue collection center identifier. See *Site ID Values* on page 73.
PatientID	Patient identifier that is associated with a Site ID. See *Patient ID Values* on page 74.
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 character 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 74.

Table 1.4 Analyte barcode constituent IDs

Identifier	*Description*
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 74.

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 75.

Name	***Description***
PlateID	Identifies individual plates See *Plate ID Values* on page 75.
CenterID	Identifies each of the CGCCs or GSCs See *Center ID Values* on page 75.

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)

Distributing Aliquots 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 *About Sequence-Based Data Files* on page 11 and *About Array-Based Data* on page 17.

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 sequencingDNA and analyzing nucleic acid analytes at the Genomic Sequencing Centers (GSCs).

Topics in this chapter include:

- *About Sequence-Based Data Files* on page 11
- *Understanding Sequence Trace Files* on page 12
- *Understanding Trace ID-to-Sample Relationship Files* on page 13
- *Understanding Mutation Annotation Format Files* on page 14
- *About FASTA Files* on page 16

About Sequence-Based Data Files

This section introduces the types of data distributed by the Biospecimen Core Resource (BCR) to the Genome Sequencing centers (GSCs).

Data Received by the GSCs

As illustrated in *Figure 1.1* on page 2, the BCR distributes DNA and RNA samples, identified by their corresponding barcodes, to GSCs for analysis.

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 *Understanding Sequence Trace Files* on page 12.). GSCs deposit trace files in the National Center for Biotechnology Information (NCBI) Trace.
- **Trace ID-to-sample relationship files** – Data that provide the relationship between NCBI traces and aliquot barcodes (see *Understanding Trace ID-to-Sample Relationship Files* on page 13). GSCs transfer these files to the DCC.
- **Mutation (MAF) files** – Data that annotate mutations in tumor cells (see *Understanding Mutation Annotation Format Files* on page 14). GSCs send MAF files to the DCC.

Understanding 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, a repository within NCBI for raw sequencing data. 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 an excerpt of a DNA sequence trace file opened in a Trace Archive page.

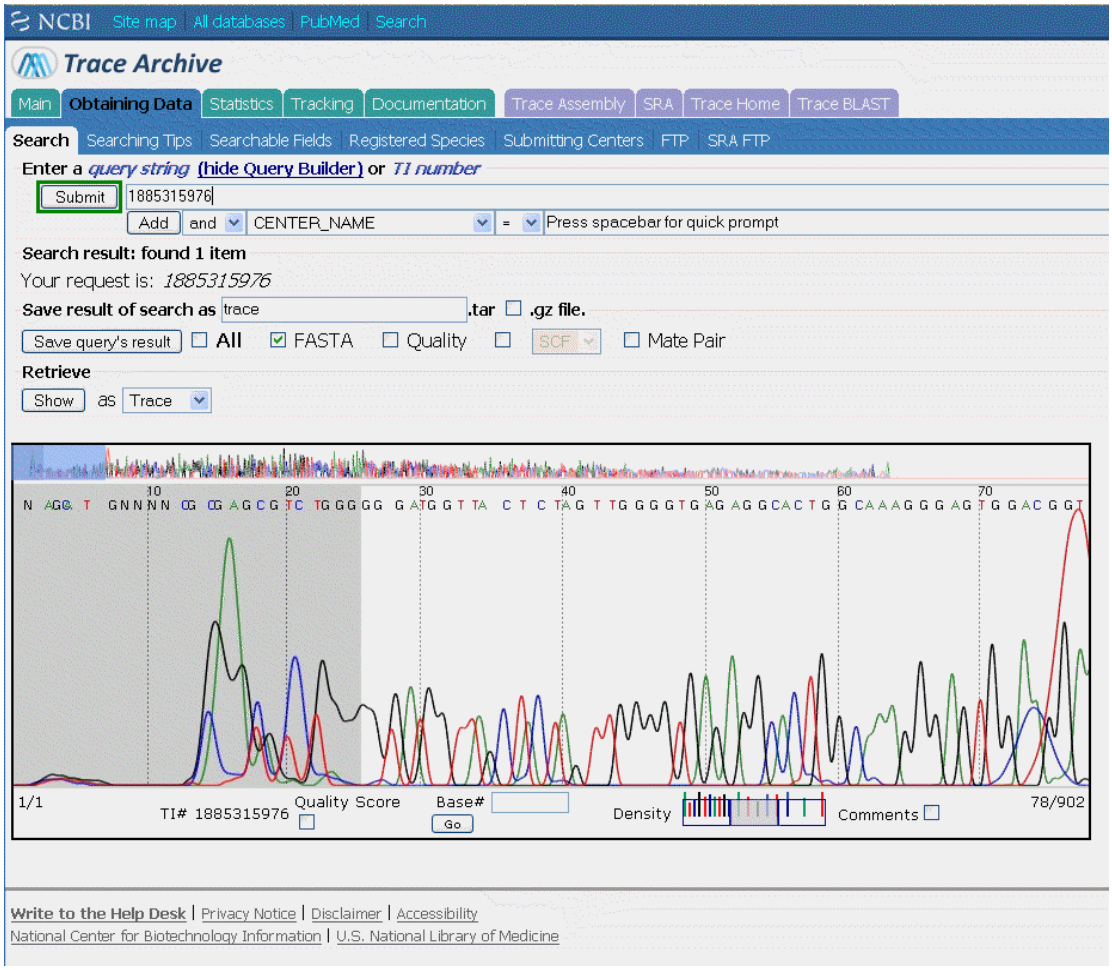


Figure 2.1 Example of trace file metadata

You can download trace files, identified by `{trace}.tar` filenames, from the website.

Understanding Trace ID-to-Sample Relationship Files

Trace ID-to-sample relationship files link NCBI traces to the analyte from which the trace file was derived by pairing trace IDs with their respective aliquot barcodes. The files contain a listing of NCBI trace IDs and TCGA aliquot barcodes. 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 access.

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 trace ID-to-sample relationship files include BCR aliquot barcodes, researchers can track and record DNA information about a specific patient and make connections between their genes, chromosomal coordinates, tumor types, and so forth.

To insure 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.`

Understanding Mutation Annotation Format 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 Aggregating and Mapping Data* on page 47.

Mutations are discovered by aligning DNA sequences derived from tumor samples to sequences derived from normal samples and a reference sequence. A MAF file contains the annotations of those mutations.

To create a MAF file, GSCs compare a patient's normal chromosomal sequence with the tumor chromosomal sequence and a template reference sequence. Any abnormal differences between the three 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 germline (originating from the germline). These can be subcategorized as follows:

Somatic mutations:

- Missense and nonsense
- Splice site, defined as SNP within 2 base pair of the splice junction
- Silent mutations
- Indels which are insertions or deletions 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

The GSCs use an independent (orthogonal) genotyping method to retest each splice site or candidate polymorphism: somatic, missense, nonsense, splice site, and indel. If

the retest confirms the SNP, 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 validated. However, if the validation process reveals that a given candidate's somatic variation event is actually 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
`Tumor_Seq_Allele1`	Tumor sequencing (discovery) allele 1
`Tumor_Seq_Allele2`	Tumor sequencing (discovery) allele 2
`dbSNP_RS`	dbSNP id. *Example:* rs12345

Table 2.1 Mutation annotation column headers

Column Header	*Description*
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_Allele 1	Matched normal genotyping (validation) allele 1
Match_Norm_Validation_Allele 2	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 column headers

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.

FASTA files are outside the scope of this document.

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* on page 17
- *About Investigation Description Format Files (IDFs)* on page 18
- *About Sample and Data Relationship Files (SDRFs)* on page 21
- *About Array Description Format Files* on page 22
- *About Raw and Processed Data Files* on page 23

About Array-Based Data

This section introduces the types of data distributed by the BCR to the Cancer Genomic Characterization Centers (CGCCs).

Data Received by Cancer Genomic Characterization Centers

As illustrated in *Figure 1.1* on page 2, the Biospecimen Core Resource (*BCR*) distributes DNA and RNA samples, identified by their corresponding barcodes to CGCCs for analysis.

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

MAGE-Based Data

In the context of TCGA, array-based data is modeled on 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. Because TCGA mandates that data be made available as soon as possible, there can be multiple MAGE-TAB documents associated with a given 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 Derived (processed from raw) 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* on page 22).
- **Data Matrices** – Raw and processed data files (see *About Raw and Processed Data Files* on page 23).

About Investigation Description Format Files (IDFs)

An Investigation Description Format (*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
3DRF 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/index Metaphrase.html
Term Source Version		1.3.0.1	

Figure 3.1 An example IDF file; from a MAGE-TAB specification document, accessible from [http://www.mged.org/mage-tab/MAGE-TABv1.0.pdf.](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_GBM.HG-CGH-244A.1.idf.txt`

File names are derived from the name of the archive where they are housed.

For example, this IDF file:

`broad.mit.edu_GBM.HT_HG-U133A.1.idf.txt`

...is derived from this archive name:

`broad.mit.edu_GBM.HT_HG-U133A.1.x.y`.

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

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

(*See Insuring Data Integrity* on page 44.)

Note: Archive naming schemes are case sensitive. Names of the MAGE-TAB files they contain must match the same upper and lower case letters. See *Archive Naming Conventions* on page 36.

IDF Protocols

An IDF file records the protocols used in an experiment. They are 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. The naming scheme is as follows:

`Domain:ProtocolType:Platform:Version`

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

Table 3.1 describes each part of a protocol name.

Name	***Description***
`Domain`	Matches a center's internet domain name
`Protocol Type`	Originates from MDEG Ontology subclasses of ProtocolType, for example, Experimental, DataTransformation, HigherLevelAnalysis (http://mged.sourceforge.net/ontologies/MGEDontology.php#ProtocolType).
`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

About Sample and Data Relationship Files (SDRFs)

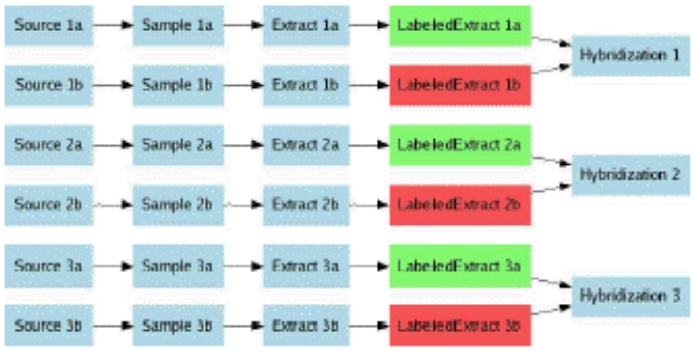
A Sample and Data Relationship (*SDRF*) file is a tab-delimited file that describes the relationships between samples, array, data, and other objects used or produced in the experiment. An SDRF contains one or more column headers for the following main types of metadata:

- **Name** – name of the sources and/or samples used in the array. There can be multiple columns of names.
- **Protocol REF** – provides ID(s) for one or more protocols used in the array and referenced in a corresponding IDF or MAGE document files.
- **File** – one or more columns that list files produced in the investigation. *Examples:* Array Data File, Derived Array Data File
- **Attribute** – values, comments, or characteristics relating to and modifying one of the above kinds of columns. *Examples*: Date, Provider, Performer, Label, Factor Values.

Column headers can be used as many times as necessary to adequately describe 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 protocols (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, accessible from: http://www.mged.org/mage-tab/MAGE-TABv1.0.pdf.

The DAG in *Figure 3.2* illustrates the initial steps in an experiment in which the CGCC processed six source analytes, creating samples. The center then created extracts from the samples, and labeled them with Cy3 and Cy5 signals, producing labeled extracts.They hybridized the labeled extracts to an array, the results of which produced a raw file of the data readings (e.g. Affymetrix .cel file). Subsequently, the CGCC normalized the raw file with the other assays, producing a derived array data matrix file (not represented in the DAG).

Refer to the MAGE-TAB specification PDF, accessible from this website: http://www.mged.org/mage-tab/MAGE-TABv1.0.pdf. for more information.

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. *Figure 3.3* shows excerpts of SDRF files showing how variations of treatments can be addressed in SDRF's.

Note: The **Extract Name** column of an SDRF always contains the *BCR* aliquot barcode for TCGA samples and sometimes other types of IDs for non-TCGA, such as control samples. See *Creating and Identifying Biospecimen Analytes* on page 5. That BCR aliquot barcode maps samples with platforms, experiments, and results files such as probe signal or gene expression files, copy number files, etc. (see *Aggregating Data Using Aliquot Barcodes* on page 48).
The **Hybridization Name** column of an SDRF contains the ID that is referenced in the column header for Level 2 and 3 data matrices. See *Figure 6.20* to see how IDs relate between different files and data.

For more information about using an SDRF file, see the MAGE-Tab specification document: 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

An Array Description Format (*ADF*) file is a tab-delimited file that 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 CGCC uses the Affymetrix HT_HG-U133A array, a well-known standard platform, it does not submit an ADF file. If, on the contrary, the CGCC uses a non-standard platform to derive methylation data, it submits an ADF along with the other corresponding data.

Figure 3.3 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			MAPK8
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			MAPK8
...	...	...	...	...	...	...	...	...	...
4	6	20	20	462020	TCGCTTCGGTTGTCCT	control	control_spike_calibration	MGED Ontology	

Figure 3.3 An example of a simple ADF file; from the MAGE-TAB specification document, accessible from 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.

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

Array analysis software generates files that contain the raw, or unprocessed, data from the assay. Each different type of analysis software generates these array data files in its unique format. For example, the analysis software from Affymetrix generates raw files in its native `.cel` file format.

Raw data may be normalized during the course of an experiment. Data that results from processing, in this case, from normalization, are known as derived array data; the files that contain these data are derived array data files. Subsequent down-stream processing produces more derived data, and therefore more derived array data files.

Files that contain summary data from multiple samples are array data matrix files. The matrix file formats are dictated by the MAGE-TAB specification. Array data matrix files contain summary data from raw files from multiple samples. Derived data matrix files contain summary data from processed files from multiple samples. 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 Data Matrix format. These files are listed under the *File[1] columns in an SDRF file.

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 Biospecimen Core Resource (BCR), Genomic Sequencing Centers (GSCs), and Cancer Genomic Characterization Centers (CGCs).

Topics in this chapter include:

- *About Data Categorization* on this page
- *Understanding Data Type/Data Level Relationships* on page 27

About Data Categorization

The DCC collects and coordinates the data it receives from the BCR, the GSCs, and the CGCCs. *Table 4.1* provides a review of the types of data that each data source submits.

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 receive data 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 by data type and data level.

About Data Type Classification

Each array/platform produces a different type of data as shown in *Table 4.2*. The second column of the table shows examples of data types.

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 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 4.2 TCGA platforms and potential corresponding data types

About Data Level Classification

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 4.3 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 amplified in 10% of TCGA glioma samples.

Table 4.3 Descriptions of TCGA data levels

Understanding Data Type/Data Level Relationships

Each data platform can produce multiple data types. To understand data categorization, it is important to clarify the relationships between data type and data level. *Table 4.2* displays current TCGA platforms and corresponding data types. Note that the SNP platforms have more than one data type.

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 how the BCRs, CGCs, and CGCCs prepare and transfer data to the TCGA DCC.

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

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. (Continued)

Determining the Data Type/Data Level of a Results File

Determining the data type and data level of BCR and GSC data is straightforward because of so few suffixes and data types.

- The BCR uses only one file suffix:.xml.
- All GSCs submit sequence data – trace files, trace-relationship files and MAF files. Trace-relationship (data type) files are always level 1 and have a suffix of tr. Mutation (data type) files are always level 2 or 3 and have a suffix of maf. The diff between the two levels is that level 3 are ONLY validated somatic mutations. Level 3 maf is the only open access (public) mutation file.

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.

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

1. A * Data type-data level file-suffix matrix
2. * Comment Data Type and Data Level columns that modify *File columns in the SDRF

Using the Data Type-Data Level-File Suffix Matrix to Identify Data Types and Levels

The * Data type-data level-file suffix matrix displays TCGA data centers and their corresponding platforms and data type(s). The data level columns display all levels of data that the center can submit to the DCC.

Each platform's data type/data level is associated with a unique file suffix.

To identify the data level and data type from a data file name, center and center type of a file, follow these steps:

1. To determine the data level, in one of the **Level n - Suffixes** columns, locate the matrix cell that contains the same suffix as the data file name suffix. The column header displays the associated data level.

2. To determine the data type, read across the row you located in Step 1, and locate the corresponding data type in the **Data Type - Base** column.

Similarly, if you know the data center that submitted the analyte(s) and the platform used to analyze the data, you can determine the associated data type and data level by finding the corresponding cells on the matrix.

Note: *Figure 4.1* shows a segment of a Data Type Data Level File Suffix Matrix. 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	
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 - Su
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

Figure 4.1 Example of a segment of a Data Type Data Level File Suffix Matrix.

How to Use the Matrix

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 you know the data center that submitted the analyte(s) and the platform used to analyze the data, you can figure out a given data type and data level by finding the appropriate data coordinates on the matrix. Likewise, if you know the suffix for a set of data, and the data center, you 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 a 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 4.3*.

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 CopyNumber-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

3. The second method of determining the data type and data level of a result file is using the Comment Data Type and Data Level columns that modify *File columns in the SDRF.

   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 "`55000240307000721079 89.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	55000240307000721079 89.G03.CEL	Expression-Gene	Level 1
TCGA-02-0002-01A-01R-0177-01	55000240307000721079 89.A09.CEL	Expression-Gene	Level 1
TCGA-02-0003-01A-01R-0177-01	55000240307000721079 89.A10.CEL	Expression-Gene	Level 1
TCGA-02-0006-01B-01R-0177-01	55000240307000721079 89.A11.CEL	Expression-Gene	Level 1
TCGA-02-0007-01A-01R-0177-01	55000240307000721079 89.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 a description of methods for accessing data in TCGA.

Topics in this chapter include the following:

- *About Archives* on this page
- *About Data Access* on this page
- *Insuring Data Integrity* on page 44
- *Insuring Data Integrity* on page 44
- *Data Access...Other DCC Resources* on page 44

About Archives

By definition, an archive is a repository where records are stored, and so describing archives in the context of TCGA is central to understanding its data access. A TCGA archive is a directory containing the experimental results of a set of assays conducted on a set of samples. The results contain experimental assays from the same center, 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. That is, since centers submit data as soon as possible, there are usually many archives submitted by a center and thus many archives per experiment.

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: 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 36)
- Archives are compressed before transfer (refer to “Archive Compression” in *TCGA Data Center Standard Operating Procedures.doc*, located at: 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 *Insuring Data Integrity* on page 44).

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 website 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`**.
`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.`** (All letters of the tumor type must be capitalized.)
`platform`	The platform reflects the assay technology 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 77.
`archiveSerialIndex`	The archive serial index is a serially increasing numeric identifier for the number of archives transferred for a particular platform for a particular tumor type from a particular center.

Table 5.1 Archive name format

Name	Description
`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. When revision of an archive is required, files that are changed or added are captured in `CHANGES.txt` and `ADDITIONS.txt` files, respectively. For detailed information, refer to "*Anatomy of an Archive*" in *TCGA Data Center Standard Operating Procedures.doc*, located: 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 numeric identifier for the separate parts of a large archive that has been split into several smaller entities. A series starts at 1. If an archive is not split into a series of archives, then `series` is 0.

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 periodically creates a list of the archives that are included in a data freeze. The list represents all of the most current, new and revised data up to a certain date. Data freezes are announced to the public through the following mechanisms:

- A TCGA listservs including the public TCGA Data listserv (see *Data Access...Other DCC Resources* on page 44).
- A TCGA Portal news item.
- The public FTP site. Freeze lists are available 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 5.1*

```
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 5.1 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.

About Data Access

As noted previously, the BCR, GSCs, and CGCCs transfer compressed archives of their data files to the DCC. The DCC categorizes data, distributes it via bulk download and deposits it into caBIG®-enabled applications and databases. The data is available to the research community using two methods:

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

For more information about the archives, see *Insuring Data Integrity* on page 44.

Accessing Bulk Downloads

The three URLs shown in numbers 1 & 2a above provide direct data access. In those instances, the user downloads the data as distributed by the DCC.

Each bulk download site has a particular directory structure that classifies distributed data files. *Figure 5.2* provides a pictorial representation of the data distribution process and the URLs constructs for accessing the data. It is a hierarchy for downloading files in bulk.

To download a file using methods 1 and 2a, construct a complete path to the files. The following components, *whose description numbers correspond to the rows of Figure 5.2*, describe in detail how to construct paths to the directories.

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 *Figure 5.2* provides the URL for base URL web access, the root directory.
2. **Access Controls**: Inside the root directory are two directories, represented by the second row of boxes (red) in the figure. One represents *unrestricted/public* access and the other represents *restricted/controlled* access[1]. Append the URL text for either to the root directory URL. This allows you *programmatic access* to either directory. Unless a directory contains a file, you may not see the file using any other method of access.

After the access control level, (row 2), you must append the directory name onto the previous directory's concatenated URL for every level down to the file level (row 10 of *Figure 5.2*).

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)

Two URLs created using this example:

HTTP:

http://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/tcga4yeo/tumor/gbm/cgcc/broad.mit.edu/genome_wide_snp_6/snp/broad.mit.edu_GBM.Genome_Wide_SNP_6.1.5.0/broad.mit.edu_GBM.Genome_Wide_SNP_6.1.sdrf.txt

SFTP:

sftp://caftps.nci.nih.gov//users/tcga4yeo/tumor/gbm/cgcc/broad.mit.edu/genome_wide_snp_6/snp/broad.mit.edu_GBM.Genome_Wide_SNP_6.1.5.0/broad.mit.edu_GBM.Genome_Wide_SNP_6.1.sdrf.txt

See also the legend following *Figure 5.2*.

1. To take advantage of controlled access data in TCGA, you must provide a username and password. To request access, follow the instructions at http://cancergenome.nih.gov/dataportal/data/access/closed/dar/.

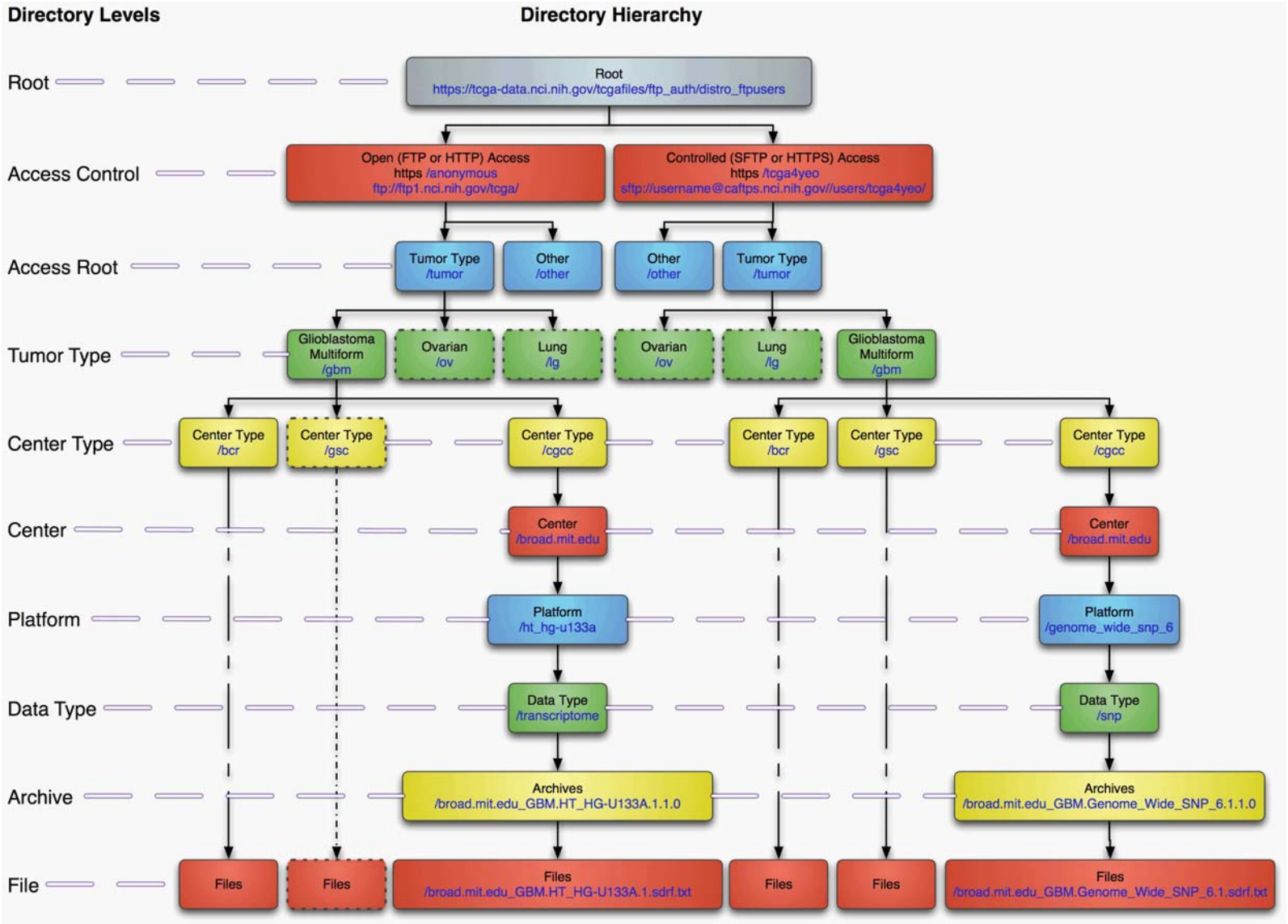


Figure 5.2 Bulk download hierarchy

Figure 5.2 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.
- Arrows point in the direction further down the directory hierarchy.
- Large dashed lines indicate that child directories for each level do exist but they are not shown to save space in the diagram.
- Small dashed 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 you to browse for and locate particular data sets as well as to access the datasets *programmatically*. 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.

Accessing Archives Through TCGA Data Portal

TCGA Data Portal provides user friendly access to the FTP and SFTP sites through archive searches and/or the Data Access Matrix. Data that are considered restricted are placed in TCGA-secure FTP (SFTP) site. Both restricted and unrestricted data are deposited into caBIG®-compatible repositories.

As described in *About Data Access* on page 38, the archives provide for direct data access through which the user downloads what was distributed.

Searching for Archives

An archive search allows you to search for complete archives, as submitted by the contributing center, by selecting archive classification parameters such as cancer type, center, platforms, data type, and submission date. The archive search interface also allows you to search for file names. Search results are always displayed as a list of archives. The results list indicates which data are controlled or open access. You can sort the results by search parameter. Each archive provides a link that allows you to view and download individual files from an archive.

The Archive Search page is located at this website: http://tcga-data.nci.nih.gov/tcga/findArchives.htm. See *Figure 5.3*.

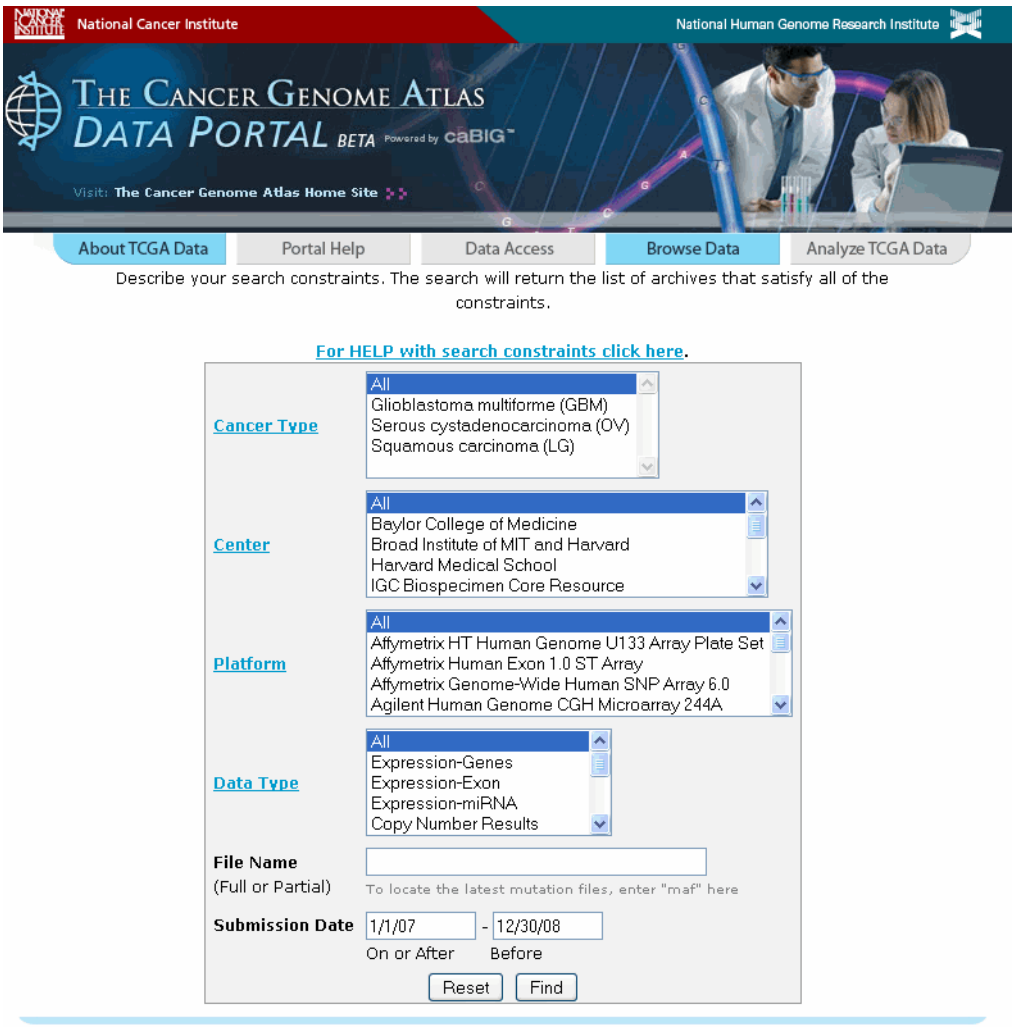


Figure 5.3 TCGA Data Portal Archives Search page

TCGA website provides instructions for searching for and accessing data.

Accessing Data Via the Data Access Matrix

TCGA project provides the cancer research community with access to data from a variety of sources via the Data Access Matrix (the Matrix), available at this location: http://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm, part of the TCGA Data Portal. The Matrix enables researchers to target data sets in TCGA servers through a user-friendly, graphic-based data set selection system. Currently, from the Matrix, you can 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, you 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.

If you choose to download data using the Data Access Matrix, when you launch a download, the system "`tars`" and "`qzips`" your files in a single archive. The processing time depends on the amount of data you requested, but it may take several hours. Although you cannot cancel the process at this point, you may close your browser or navigate away from the processing page. The processing continues in the background. When the process is complete, the process page displays a link to the archive file on the TCGE server. Additionally, the system sends a link to the archive in a message to the email address you provided when you started the download process in the Matrix.

Figure 5.4 displays a page from the Matrix user interface. A Data Access Matrix User's Guide is available http://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/docs/tcga_DataAccessMatrix_UserGuide.pdf.

Figure 5.4 Home page for the Data Access Matrix

For examples of using the TCGA Data Portal and the Data Access Matrix to access data, see *Aggregating Data Using Clinical Metadata* on page 54.

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 bioinformatics methods and/or third-party databases. Because patient privacy protection is paramount to NIH and TCGA, human subject protection and data access policies have been implemented to minimize the risk that the privacy of the donors and the confidentiality of their data will be compromised. As part of this effort, data generated from TCGA are available in the following two tiers:

1. The unrestricted/public data tier is a publicly accessible tier of data that cannot be aggregated to generate a data set 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 tier on the website with protected clinical data and individually unique information that contains data that can potentially identify an individual. 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/. Information about public vs. private clinical data is available in *Aggregating Data Using Clinical Metadata* on page 54.

Insuring Data Integrity

According to best practices, you should insure that each file you download from TCGA has not been corrupted in the process. This is especially important for very large archives you download from the *TCGA Data Portal*.

To use MD5 hash files to confirm the integrity of archived files, follow these steps:

1. Each time you download a TCGA file, download its corresponding MD5 hash file from the same directory.
2. Create your own MD5 hash file from the downloaded TCGA file using either the program md5sum (for Unix and Mac OSX) or md5sums (for Windows).
3. Compare the MD5 you generated to the MD5 file you downloaded from TCGA. If they match, the integrity of the TCGA files is assured.

The command-line software, md5sums (Windows) and md5sum (Unix and Mac OSX), are implementations of the MD5 algorithm for creating MD5 hashes, and are available from `ps-tools.net` at http://www.pc-tools.net/win32/md5sums/, and FreeBSD respectively. The Unix and Mac OSX command line, md5sum `archiveName.tar.gz`, corresponds to the Windows command `md5sums -u archiveName.tar.gz`.

An MD5 hash file name is the same as its compressed archive counterpart, with the addition of the md5 suffix.

For example, this MD5 hash file:

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

...corresponds to this archive:

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

Data Access...Other DCC Resources

The DCC provides three resources that can help you sort, query and interpret TCGA data. These resources are updated regularly.

- A complete list of sample barcodes in an easy-to-use table is available for download from ftp://ftp1.nci.nih.gov/tcga/other/BCR_Biospecimen_Barcodes_Table.tar.gz The table enables you to sort or

query the barcodes using the constituent parts of the code (see *Chapter 6, Aggregating and Mapping Data*).

- A Sample-File Association Matrix is available for download from ftp://ftp1.nci.nih.gov/tcga/other/SampleToFile_AssociationMatrix.tar.gz. It relates a sample barcode to the files that were produced during the characterization assay of a sample (see *About Aggregating and Mapping Data* on page 47).
- A public TCGA Data listserv (tcga-data-l@list.nih.gov) notifies subscribers when new or revised archives are available on the portal. You can register as a subscriber at this website: https://list.nih.gov/archives/tcga-data-l.html.

CHAPTER

6

AGGREGATING AND MAPPING DATA

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

Topics in this chapter include:

- *About Aggregating and Mapping Data*
- *Aggregating Data Using Aliquot Barcodes* on page 48
- *Aggregating Data Using Clinical Metadata* on page 54
- *Mapping Aliquot Barcodes to Assay Result Files* on page 67

About Aggregating and Mapping Data

To analyze TCGA data, you may be required to aggregate samples, then map the aggregated data to result files. Aggregation involves selecting a subset of samples within a data type or between data types and grouping them according to parameters of interest, using either the information contained in the *BCR* aliquot barcode or using the information contained in the clinical metadata. Aggregating the data produces a list of aliquot barcodes that can be mapped to assay result files.

Note: To replicate the steps in this chapter, you need to download data files as described in *Chapter 5, Data Access*.

Aggregating and mapping data is summarized in these general steps, described in detail in this chapter:

1. Produce a list of barcodes that can be mapped to results files, using one of these two methods:

Using aliquot barcode parameters:

   a. Create an aliquot barcode table, or use the table created by the DCC.

b. Sort the data by selected parameters that are part of the barcode. Parameters can be data type, patient IDs, sample IDs, the center that is the source of data, etc.

Using clinical data of interest:

a. Select the clinical data of interest.

b. Download the appropriate BCR file(s) that include the appropriate IDs for the selected clinical data.

c. Find the barcodes associated to your data of interest.

d. Create a list of the corresponding barcodes.

2. In both cases, use aggregated barcodes to map to result files.

To use aliquot barcodes from aggregated data to map to result files, you must first identify the result files that correspond to those barcodes. Once you have identified the files, the method of mapping an aliquot barcode to its assay-result files varies according to the type of data you want to map.

Aggregating Data Using Aliquot Barcodes

The aliquot barcode is the most important ID, or reference point, in the entire TCGA enterprise. As described previously (*About Aliquot Barcodes* on page 7), the barcode indicates that the aliquot from a particular sample and center has been processed as follows:

- The BCR performed quality control testing on the sample and sent it to a data generating center (for example, a CGCC) along with the aliquot barcode.
- The center performed an assay and sent the analysis results (still associated with the aliquot barcode) to the DCC.
- The DCC categorized the data, making links between the clinical data and assay results.

You can aggregate data with aliquot barcodes for the purposes of analysis or comparison. For example, if you are looking for results from a particular type of analyte, you can aggregate barcodes and sort the information accordingly. Any information provided by the aliquot barcodes can form the basis for aggregation.

The main steps for aggregating samples using their aliquot barcodes are as follows:

1. Split all barcodes into their constituent IDs. *Table 6.1* provides an example of splitting one BCR barcode, `TCGA-02-0001-01C-01D-0182-01`, into its constituent parts.

Aliquot Barcode IDs	***Barcode Value***	***Description***
`project`	TCGA	The project

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

Aliquot Barcode IDs	*Barcode Value*	*Description*
collection_center	02	The GBM brain tumor sample from MD Anderson
patient	0001	The first patient from MD Anderson for GBM tumor type
sample_type	01	A solid tumor
sample_sequence	C	The third vial
portion_sequence	01	The first portion of the tumor sample
portion_analyte	D	The analyte is a DNA sample
plate_id	0182	The plate ID within the 96-well plate.
center_id	01	The Broad Institute which is to receive the sample

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

The best way to break barcodes into their constituent IDs is to create an aliquot barcode table or to use one prepared by the DCC. *Table 6.2* and *Figure 6.1* are examples of aliquot barcode tables in which aliquot barcodes are broken down into constituent IDs.

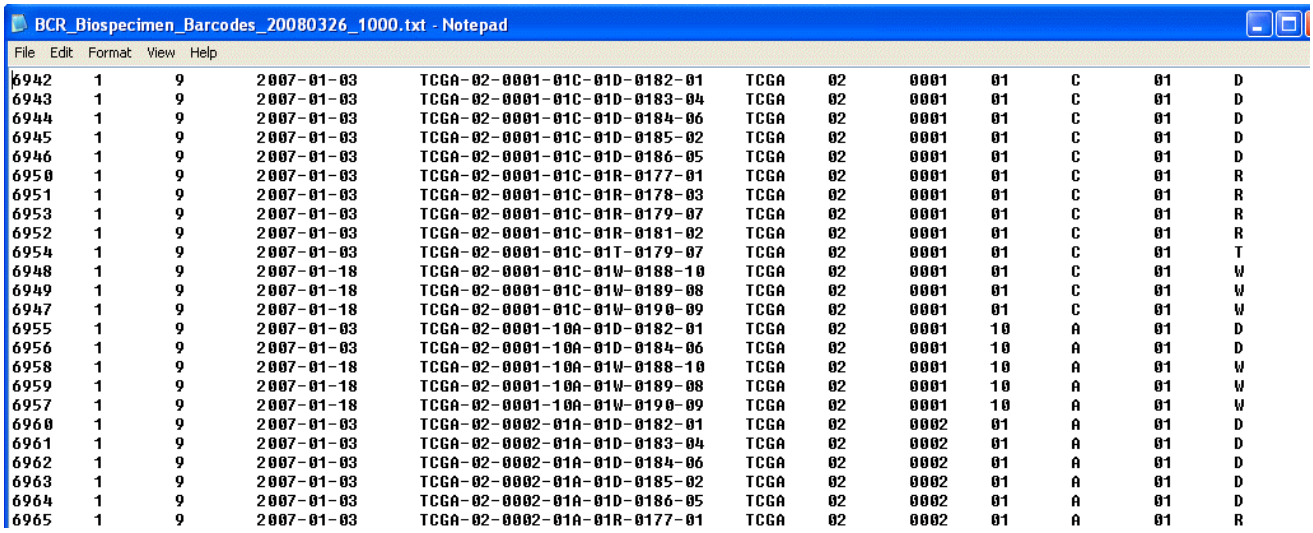
Table 6.2 is an example of aliquot barcodes created by an individual researcher. The researcher has created a list of aliquot bar codes and has split them into their constituent IDs.

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

Table 6.2 Aliquot barcode table

Note: You can download an aliquot barcode table from the DCC: ftp://ftp1.nci.nih.gov/tcga/other/BCR_Biospecimen_Barcodes_Table.tar.gz. The downloaded zip file contains a legend that describes the content of the table columns.

Figure 6.1 is an example of a segment of a DCC-prepared aliquot barcode table, opened as a tab-delimited TXT file in a text editor.



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 6.1 Aliquot barcode table prepared by the DCC; opened as a text file in Notepad

2. Sort the barcodes in the aliquot barcode table by using column headers to aggregate the data by the barcode constituent of interest.

   Note: In Unix operating systems, use the "sort" command. In Windows operating systems, open the file in a spreadsheet application (for example, Excel) and use the sort features provided.

Figure 6.2 is an example of an aliquot barcode table as it appears in Excel before sorting any data.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
763	13955	6	12	1/22/2008	TCGA-02-0070-10A-01G-0294-04	TCGA	2	70	10	A	1	G	294	4
764	13954	6	12	1/22/2008	TCGA-02-0070-10A-01G-0295-06	TCGA	2	70	10	A	1	G	295	6
765	13953	6	12	1/22/2008	TCGA-02-0070-10A-01G-0296-02	TCGA	2	70	10	A	1	G	296	2
766	13959	6	12	1/23/2008	TCGA-02-0070-10A-01W-0317-10	TCGA	2	70	10	A	1	W	317	10
767	13958	6	12	1/23/2008	TCGA-02-0070-10A-01W-0318-08	TCGA	2	70	10	A	1	W	318	8
768	13956	6	12	1/23/2008	TCGA-02-0070-10A-01W-0319-09	TCGA	2	70	10	A	1	W	319	9
769	13957	6	12	1/23/2008	TCGA-02-0070-10A-01W-0320-02	TCGA	2	70	10	A	1	W	320	2
770	7462	2	13	1/30/2007	TCGA-02-0071-01A-01D-0193-01	TCGA	2	71	1	A	1	D	193	1
771	7466	2	13	1/30/2007	TCGA-02-0071-01A-01D-0196-04	TCGA	2	71	1	A	1	D	196	4
772	7465	2	13	1/30/2007	TCGA-02-0071-01A-01D-0197-06	TCGA	2	71	1	A	1	D	197	6
773	7464	2	13	1/30/2007	TCGA-02-0071-01A-01D-0198-02	TCGA	2	71	1	A	1	D	198	2
774	7463	2	13	1/30/2007	TCGA-02-0071-01A-01D-0199-05	TCGA	2	71	1	A	1	D	199	5
775	7471	2	13	1/30/2007	TCGA-02-0071-01A-01R-0194-03	TCGA	2	71	1	A	1	R	194	3

Figure 6.2 Unsorted aliquot barcode table imported into Excel

Note: Column “L” (outlined) in *Figure 6.2* lists the analyte code for each sample in the table. In this example, samples originating from DNA (analyte code “D”) are interspersed with other samples.

By sorting a table you can aggregate aliquot barcodes with common elements. For example, if you sort by analyte (column L in *Figure 6.2* and *Figure 6.3*), you can aggregate all barcodes originating from DNA (analyte code “D”). Example of a sorted aliquot barcode table is shown in *Figure 6.3*.

.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
730	8131	3	11	1/28/2007	TCGA-06-0142-10A-01D-0215-04	TCGA	6	142	10	A	1	D	215	4
731	8132	3	11	1/28/2007	TCGA-06-0142-10A-01D-0216-06	TCGA	6	142	10	A	1	D	216	6
732	8133	3	11	1/28/2007	TCGA-06-0142-10A-01D-0217-02	TCGA	6	142	10	A	1	D	217	2
733	8139	3	11	1/28/2007	TCGA-06-0143-01A-01D-0214-01	TCGA	6	143	1	A	1	D	214	1
734	8140	3	11	1/28/2007	TCGA-06-0143-01A-01D-0215-04	TCGA	6	143	1	A	1	D	215	4
735	8141	3	11	1/28/2007	TCGA-06-0143-01A-01D-0216-06	TCGA	6	143	1	A	1	D	216	6
736	8142	3	11	1/28/2007	TCGA-06-0143-01A-01D-0217-02	TCGA	6	143	1	A	1	D	217	2
737	8143	3	11	1/28/2007	TCGA-06-0143-01A-01D-0218-05	TCGA	6	143	1	A	1	D	218	5
738	8153	3	11	1/28/2007	TCGA-06-0143-10A-01D-0214-01	TCGA	6	143	10	A	1	D	214	1
739	8154	3	11	1/28/2007	TCGA-06-0143-10A-01D-0215-04	TCGA	6	143	10	A	1	D	215	4
740	8155	3	11	1/28/2007	TCGA-06-0143-10A-01D-0216-06	TCGA	6	143	10	A	1	D	216	6
741	8156	3	11	1/28/2007	TCGA-06-0143-10A-01D-0217-02	TCGA	6	143	10	A	1	D	217	2
742	8165	3	11	1/28/2007	TCGA-06-0145-01A-01D-0214-01	TCGA	6	145	1	A	1	D	214	1
743	8161	3	11	1/28/2007	TCGA-06-0145-01A-01D-0215-04	TCGA	6	145	1	A	1	D	215	4

Figure 6.3 Aliquot barcode table sorted by analyte DNA ("D" in column L "analyte").

Note: Column "L" (outlined) in *Figure 6.4* lists the analyte code for each sample in the table. In this example samples originating from DNA (analyte code "D") are grouped together.

Common examples of sorting by barcode elements are as follows:

- Samples of particular types with particular types of analytes
  - tumor samples (01-09) *vs.* normal samples (10-19)) and/or DNA portion analytes (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) from 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)

3. Select aliquot barcodes grouped by the code/ID of interest. The result is a customized list of aggregated barcodes. You can use the list to compare other data types and to map to files of interest. For more information, see *Mapping Aliquot Barcodes to Assay Result Files* on page 67.

Aggregating Sample Data Between Different Centers and Platforms

Given results from two different centers or platforms, how do you match up the results for the same sample? Matching or comparing results from two different 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 and TCGA-02-0021-01A-01D-0186-05) between results of different centers or platforms for the following reasons:

- Aliquot barcodes are specific to a particular center and platforms because the barcodes contain plate barcodes. The plate barcode includes a center ID and plate ID. (See *Deciphering Plate Barcodes* on page 8.)
- Analyte barcodes, which are the equivalent of aliquot barcodes minus their plate barcodes do not contain center- or platform-specific data, by definition. For example, analyte barcode TCGA-02-0021-01A-01D is the aliquot code TCGA-02-0021-01A-01D-**0002-04** minus the plate code **0002-04**. The receiving center ID is “04”, the last two digits of the aliquot barcode.
- Different centers have few biospecimen barcodes in common.

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). Sample barcodes are the best choice for matching results between different centers or platforms.

Aggregating results between different centers or platforms using the sample barcode involves using an aliquot barcode table as an intermediary. You can query an aliquot barcode table to obtain a set of aliquot barcodes that match the sample barcode using the constitutive parts of the sample barcode. (See *Aggregating Data Using Aliquot Barcodes* on page 48. Example barcode tables are shown in *Table 6.2* and *Figure 6.1*.)

Figure 6.4 illustrates how to aggregate data generated by different centers. The goal is to create a list of aliquot barcodes that correspond to the samples used in another experiment. Descriptions follow the illustration.

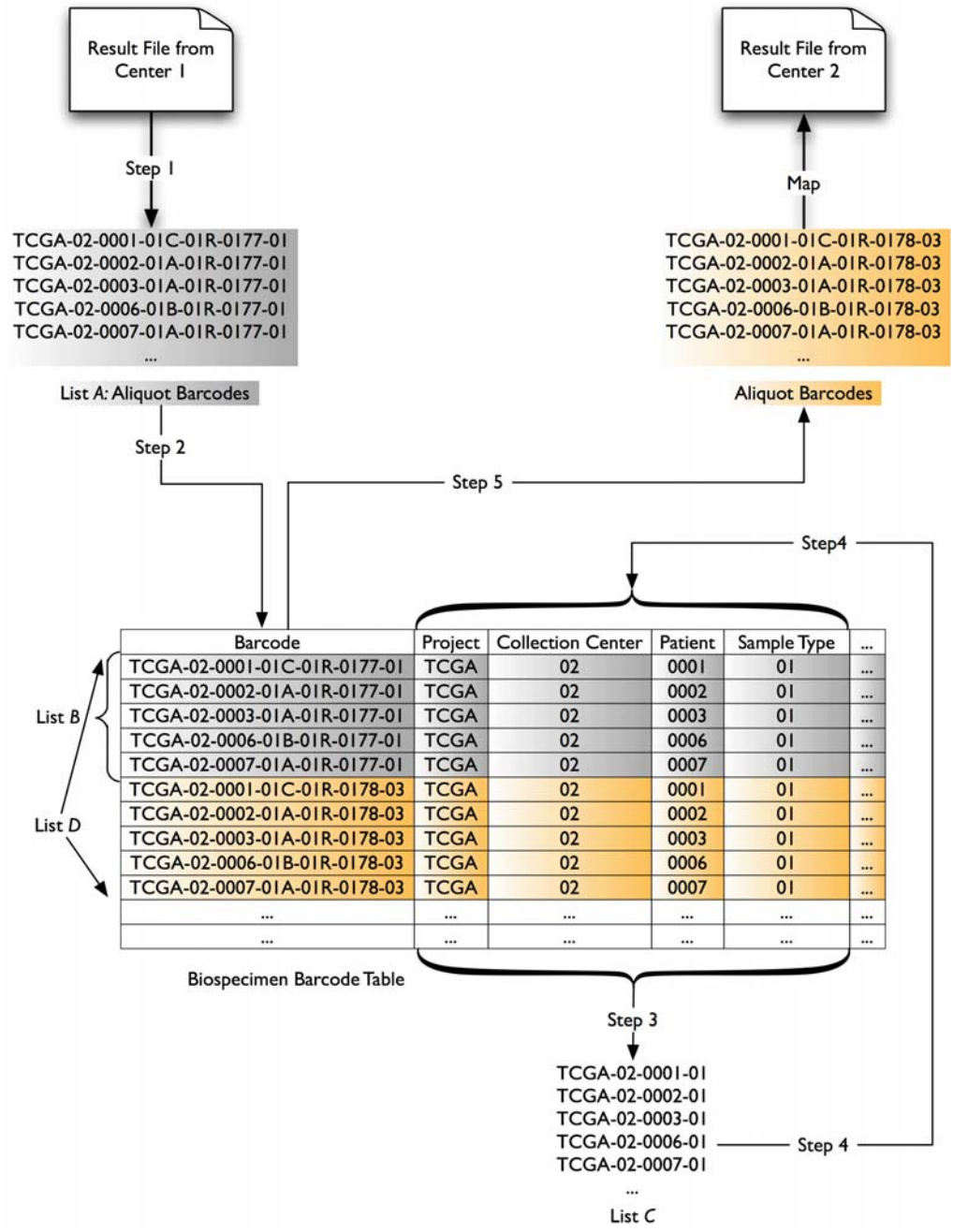


Figure 6.4 Scenario for aggregating data between centers. Steps are described below the figure.

With an aliquot barcode table in hand, (as described in step 1 on page 47), you should follow these steps, which correspond by number to steps shown in *Figure 6.4*.

1. Create a list of aliquot barcodes (A) from the files from one center or platform (Center 1 in *Figure 6.4*).

2. Query an aliquot barcode table using the complete biospecimen barcodes from list A. This creates a list of matches, List B, which is also an aliquot barcode list.
3. Create a unique list of the constituent IDs for each of the matching barcodes in List B. This new list, C, consists of the barcodes' Project Name, Site ID, Patient ID, and the Sample Type portion of the Sample ID.
4. Use list C to initiate a new search on the aliquot-barcode table, querying the columns used to create sample barcodes in step 3 (Project Name, Site ID, Patient ID, and the sample type of the Sample ID). The query results become List D, an aliquot barcode list.
5. Map the aliquot barcodes in list D to result files from the other center (Center 2 in *Figure 6.4*).

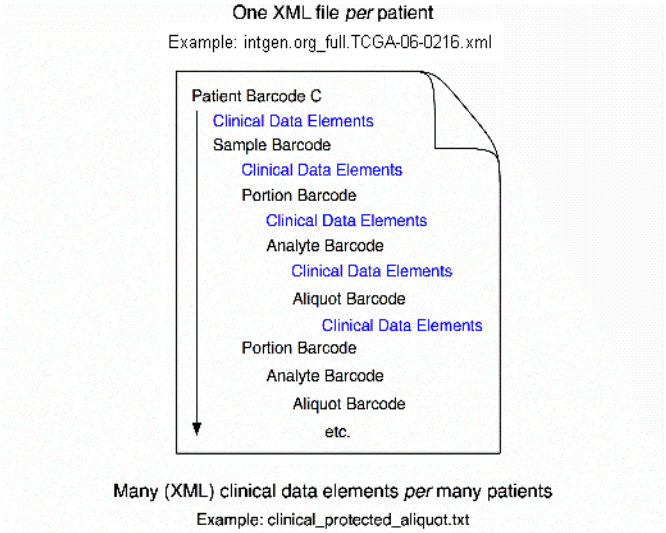
You can use the list of aliquot barcodes you matched in Step 5 as aggregate data that you can map to result files. For more information, see *Mapping Aliquot Barcodes to Assay Result Files* on page 67.

Aggregating Data Using Clinical Metadata

You can use one or both of two types of clinical metadata files submitted to the BCR to aggregate samples using clinical metadata. The two file types (XML and tab-delimited TXT) represent the same data structure in different ways, but both enable you to amass a series of barcodes that correspond to patients that fit within the clinical data types of interest.

You can use either type of file to extract and aggregate aliquot barcodes associated with patients' clinical data. See the following section, *Working with XML Files*, and *Working with Tab-Delimited Files* on page 61.

Each XML file contains data for just one patient. A TXT file contains data for multiple patients. The distinction between XML files and the tab-delimited TXT files is illustrated in *Figure 6.5*.



Patient Barcode	Sample Barcode	Portion Barcode	Analyte Barcode	Aliquot Barcode	Clinical Data Element	Clinical Data Element ...
Patient Barcode A						
Patient Barcode B						
Patient Barcode C						
Patient Barcode C						

Figure 6.5 This figure distinguishes between a clinical data XML file and a clinical data TXT file.

Once you have parsed relevant data/aliquot barcodes from the available XML files or TXT files, you can aggregate samples using clinical data by choosing the clinical data elements of interest and using them as search and/or sort criteria for the data elements of interest. Use the relevant barcodes to aggregate barcode lists in preparation for mapping the data to result files. See *Mapping Aliquot Barcodes to Assay Result Files* on page 67 for more information.

The following sections explain how to accomplish your objective using XML and TXT files.

Working with XML Files

XML files present data in a structured, hierarchical manner, represented in *Figure 6.6*. *Figure 6.7* on page 57 is an example segment of the corresponding XML file itself. Because of the complexity of XML files, and because there is one XML file for each patient rather than for each of the patient's data elements, the DCC parses each XML file into separate comma-separated value (CSV) files, each of which represents a major BCR XML element, with metadata for multiple patients.

Note: You can use the clinical metadata XML file "as is" if you understand the BCR UML model and the BCR XML schema (see *Understanding Array-Based Data* on page 17), and can parse the XML files.

Each major clinical data element, presented as a circled rectangle in *Figure 6.6*, and as a circled item in *Figure 6.7* on page 57, corresponds to a TXT file, identified by element name, for example, `portions_csv.txt`, or `aliquot_csv.txt`. Each TXT file contains, for many patients, barcodes for that element as well as the barcodes it inherits from its parents. This facilitates mapping between elements.

For example, a portion file (`portion_csv.txt`) lists the inherited barcodes for the patient from which the portion was derived (TCGA-02-0021), the sample from which it was derived (TCGA-02-0021-01A), with its own code appended (TCGA-02-0021-01A-01). Similarly an aliquot file (`aliquot_csv.txt`) lists the inherited barcodes for the patient, sample, portion, analyte, and the aliquot itself.

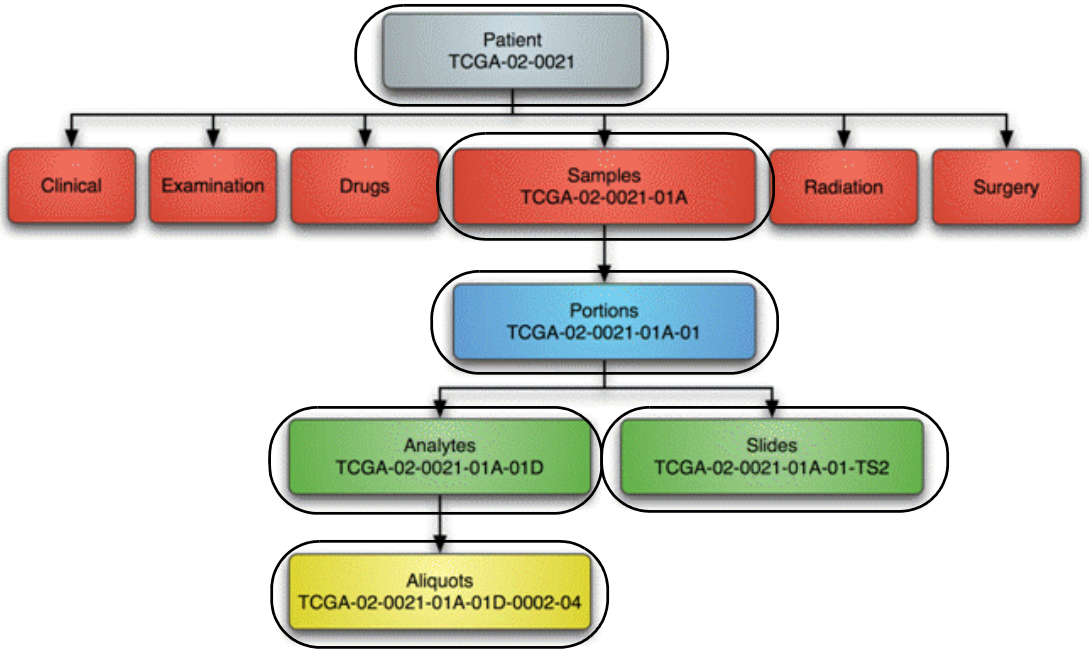


Figure 6.6 Inheritance of major BCR element barcodes in an XML file

The XML file in *Figure 6.7*, `intgen.org_full.TCGA-02-0021.xml`, corresponds to *Figure 6.6*. It contains clinical data for patient TCGA-02-0021 that was submitted to the BCR. Circled items are major BCR clinical data elements, also called “XML elements”. The numbers in the rectangles are the codes associated

with the given element. The elements correspond to the rows circled in *Figure 6.6* on page 56.

Note: The Aliquot row is not displayed in *Figure 6.7*.

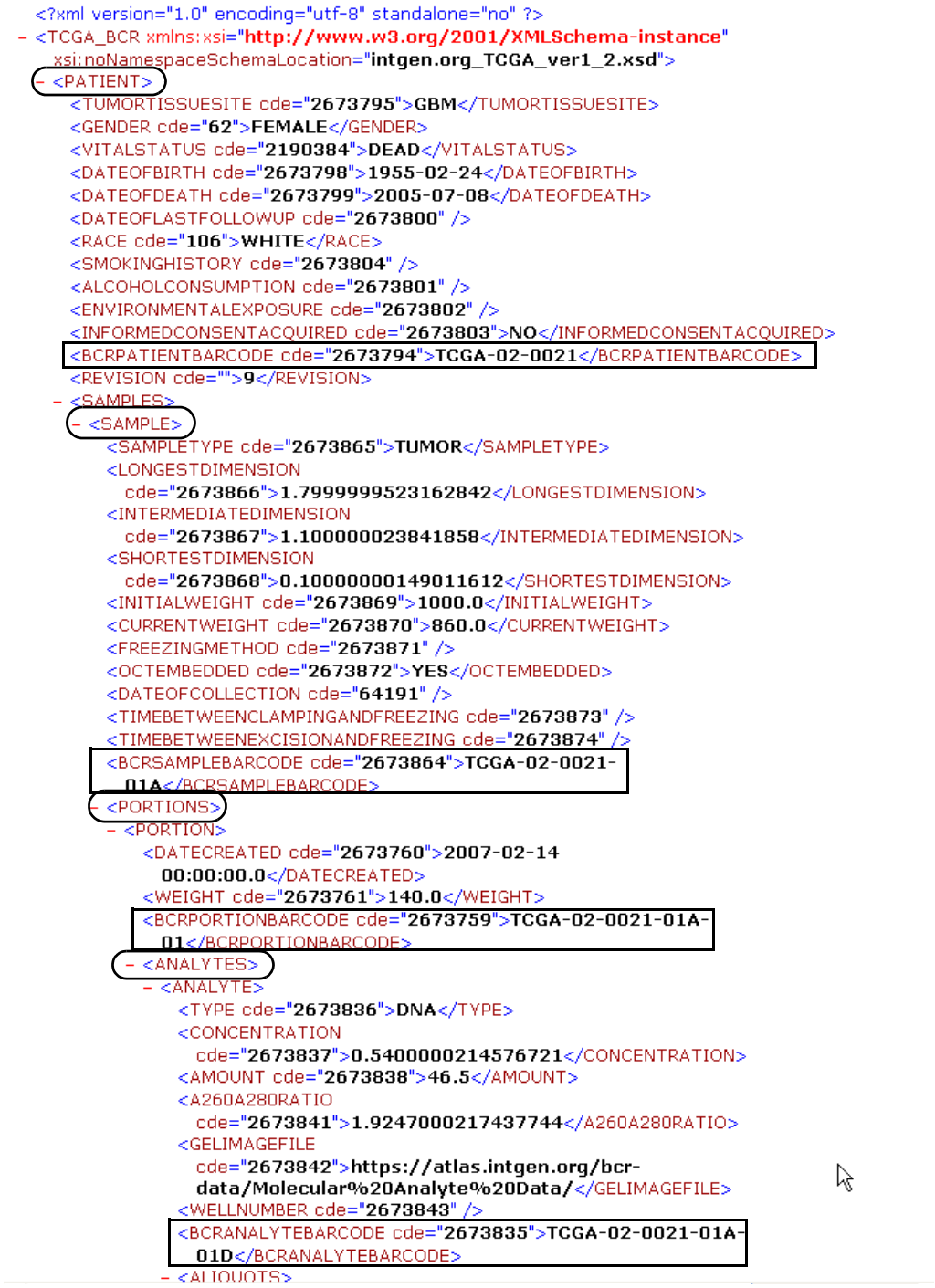


Figure 6.7 XML file example for patient TCGA-02-0021

The code for each succeeding sub-element of the example patient data in the XML file (circled items in *Figure 6.7*) is appended to the barcode it inherits from its parent elements as follows:

Patient barcode = `TCGA-02-0021`

Sample barcode = patient barcode + sample barcode `01A`

Portion barcode = sample barcode + portion code `01`

Analyte barcode = portion barcode + analyte barcode `01D`

Note: You can use the barcodes associated with each subsection (element) of the XML data to aggregate the data for further analysis.

The DCC creates the following clinical data element CSV files for each XML file:

- `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`

Note: You can import the CSV files into spreadsheet applications such as Excel.

Use the extracted barcodes to sort and aggregate barcode lists to map the data to result files.

Note: If you choose to work with XML files, take into consideration that you must work with many files because one XML file represents data from just one patient.

Using XML Files to Aggregate Clinical Metadata

Figure 6.6 on page 56 illustrates an example using Patient "TCGA-02-0021". The figure displays a hierarchical structure of the clinical metadata for this patient. The same hierarchical structure can be followed in the BCR XML file (*Figure 6.8*) that corresponds to this patient.

The following workflow explains how to find and extract data for this patient using XML files.

1. Locate and download data for a patient of interest.
   a. On the TCGA home page, http://tcga-data.nci.nih.gov/tcga/, click the **Search by Archive** link.

Note: You can also navigate to TCGA home page can by clicking the **Browse Data** tab on any page of TCGA portal. (see *About Data Access* on page 38).

2. Do one of the following to locate files for a particular patient.

   a. To search by patient ID, type the patient ID, in the case of this example, ***TCGA-02-0021***, in the **File Name** field at the bottom of the Archive Search page (*Figure 6.8*), and then click **Find**.

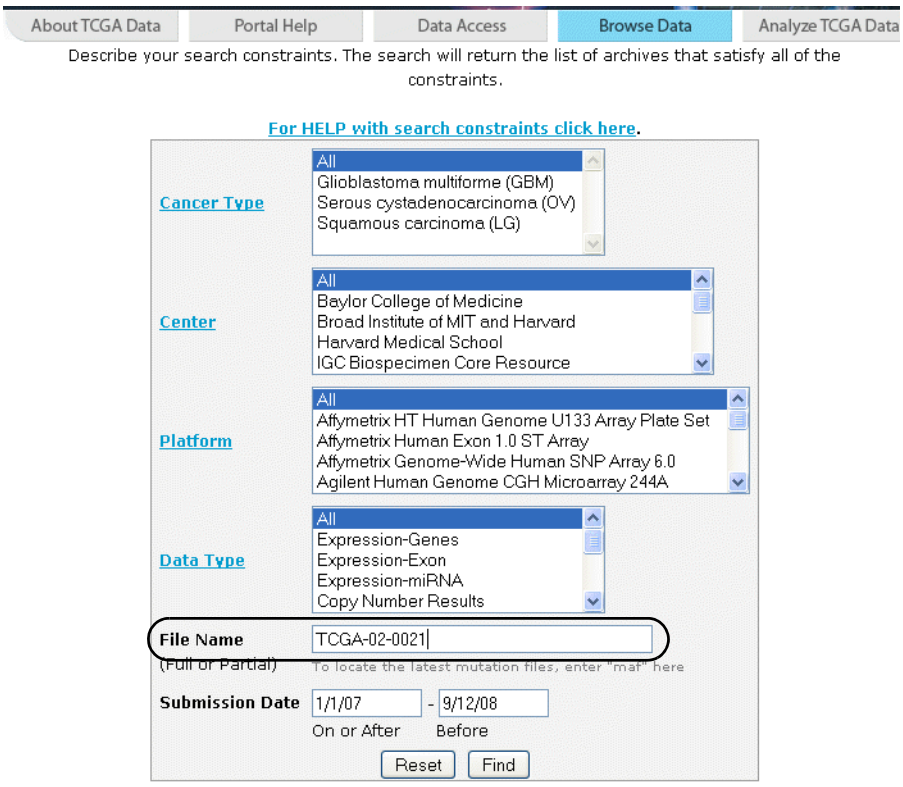


Figure 6.8 Archive Search page with a patient ID entered for the file query

- or -

   b. To search for a patient from the list of centers, on the Archive Search page, rather than entering a patient ID for the search, from the list of Centers, select the **IGC Biospecimen Core Resource** (*Figure 6.9*).

.

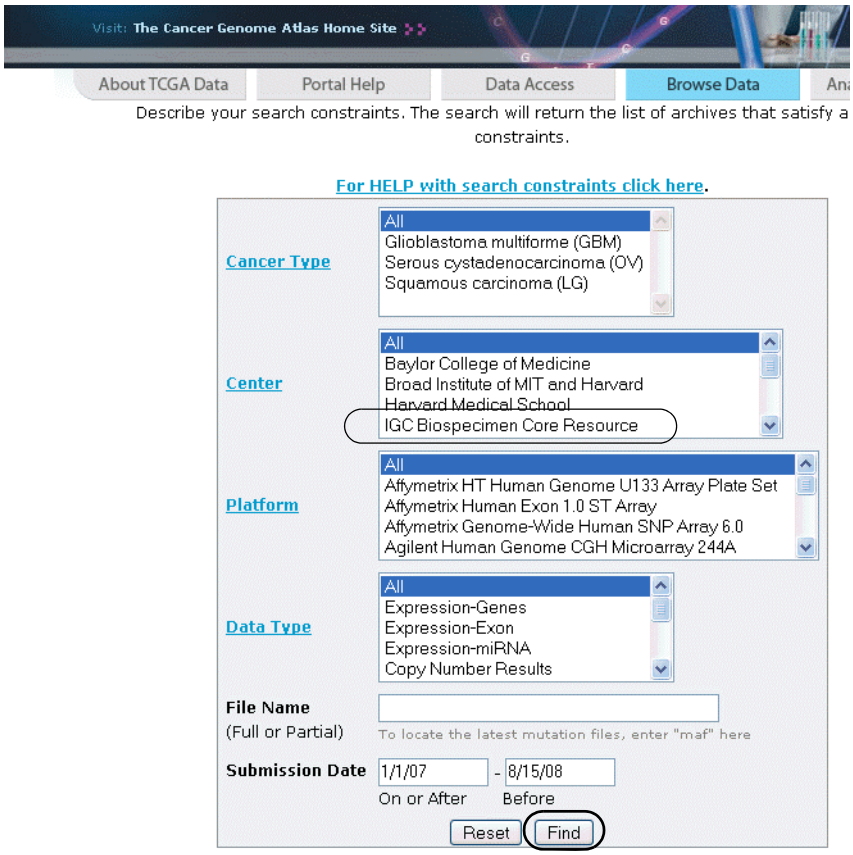


Figure 6.9 TCGA Archives page

- Click the **Find** button at the bottom of the page. This opens the Data Portal Archives page(*Figure 6.10*).

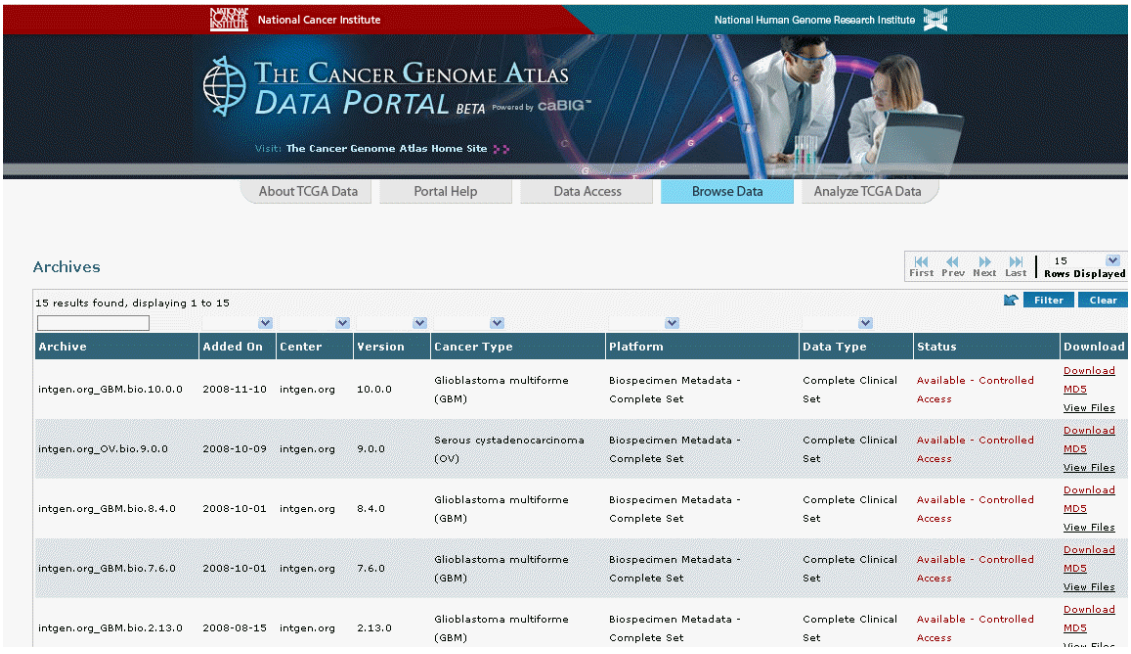


Figure 6.10 Data Portal Archives page

- To see a list of the files submitted to the BCR, click the **View Files** hypertext link. Each XML file is named by a patient ID.

Figure 6.11 displays a list of XML files in `intgen.org_GBM.bio.1.14.0`. The circled file in the list represents the patient of interest used in these examples.

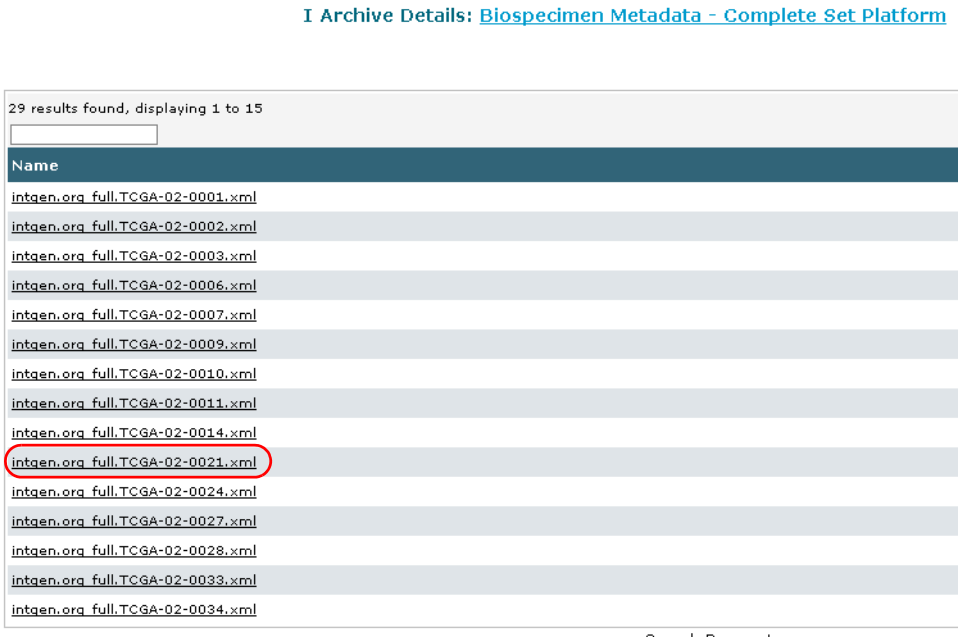


Figure 6.11 Example list of XML files displayed in the TCGA Data Portal

- Continue to view archive files to locate your file(s) of interest.
- To download file, click the **Download** button on the Archives page that corresponds to the appropriate file.

The XML file in this example, `intgen.org_full.TCGA-02-0021.xml`, (*Figure 6.11*) is the one that corresponds to *Figure 6.7* and thus to Patient TCGA-02-0021, noted in the top row of *Figure 6.6*. This file contains all of the clinical data for patient TCGA-02-0021 submitted to the BCR.

Working with Tab-Delimited Files

In tab-delimited text files, you can view clinical metadata in a spreadsheet, where data can easily be sorted and extracted. You can also submit queries and/or map the data using a spreadsheet program (for example, Excel), unix utilities (for example, “sort and join”), or you can load different delimited files directly into a database management system (DBMS, such as mysql). These options enable you to sort and aggregate all patients and extract needed data.

The following list includes the delimited file names representing eleven of the major BCR “XML elements”. The prefix in this list, “clinical_protected”, is constant while the portion of each name immediately preceding the file extension corresponds to a clinical element in XML files. Each major element file would contain barcodes for that element and its parent’s barcodes to facilitate mapping between elements.

- `clinical_protected_aliquot.txt` (See *Figure 6.12*)

- clinical_protected_analyte.txt
- clinical_protected_drug.txt
- clinical_protected_examination.txt
- clinical_protected_portion.txt
- clinical_protected_protocol.txt (See *Figure 6.13*)
- clinical_protected_public.txt
- clinical_protected_radiation.txt
- clinical_protected_sample.txt
- clinical_protected_slide.txt (See *Figure 6.14*)
- clinical_protected_surgery.txt

Examples of these files (opened in Excel) are displayed on page 63. Note the diversity of data types in the columns on the various tables. Note also the corresponding barcode information that can help you to assemble a list of barcodes identifying patients with the clinical status of interest. It might be patients with a smoking history, patients within a specified age range, patients given a specific treatment, patients with a specifically localized form of tumor – whatever the clinical history, the researcher can identify it within these files, note and assemble the corresponding aliquot barcode list accordingly.

BCRPATIENTBARCODE	BCRSAMPLEBARCODE	BCRPORTIONBARCODE	BCRANALYTEBARCODE	BCRALIQUOTBARCODE	AMOUNT	SHIPPING	CONCENTRATION
TCGA-02-0001	TCGA-02-0001-01A	null	null	null	null	null	null
TCGA-02-0001	TCGA-02-0001-01C	TCGA-02-0001-01C-01	TCGA-02-0001-01C-01D	TCGA-02-0001-01C-01D-0182-01	4	4/3/2007	0.5
TCGA-02-0001	TCGA-02-0001-01C	TCGA-02-0001-01C-01	TCGA-02-0001-01C-01D	TCGA-02-0001-01C-01D-0183-04	10	4/3/2007	0.5
TCGA-02-0001	TCGA-02-0001-01C	TCGA-02-0001-01C-01	TCGA-02-0001-01C-01D	TCGA-02-0001-01C-01D-0184-06	6	4/3/2007	0.5
TCGA-02-0001	TCGA-02-0001-01C	TCGA-02-0001-01C-01	TCGA-02-0001-01C-01D	TCGA-02-0001-01C-01D-0185-02	20	4/3/2007	0.5
TCGA-02-0001	TCGA-02-0001-01C	TCGA-02-0001-01C-01	TCGA-02-0001-01C-01D	TCGA-02-0001-01C-01D-0186-05	10	4/3/2007	0.5
TCGA-02-0001	TCGA-02-0001-01C	TCGA-02-0001-01C-01	TCGA-02-0001-01C-01R	TCGA-02-0001-01C-01R-0177-01	10	4/3/2007	0.5
TCGA-02-0001	TCGA-02-0001-01C	TCGA-02-0001-01C-01	TCGA-02-0001-01C-01R	TCGA-02-0001-01C-01R-0178-03	10	4/3/2007	0.5
TCGA-02-0001	TCGA-02-0001-01C	TCGA-02-0001-01C-01	TCGA-02-0001-01C-01R	TCGA-02-0001-01C-01R-0179-07	4	4/3/2007	0.5
TCGA-02-0001	TCGA-02-0001-01C	TCGA-02-0001-01C-01	TCGA-02-0001-01C-01R	TCGA-02-0001-01C-01R-0181-02	6	4/3/2007	0.5
TCGA-02-0001	TCGA-02-0001-01C	TCGA-02-0001-01C-01	TCGA-02-0001-01C-01T	TCGA-02-0001-01C-01T-0179-07	20	4/3/2007	0.1
TCGA-02-0001	TCGA-02-0001-01C	TCGA-02-0001-01C-01	TCGA-02-0001-01C-01W	TCGA-02-0001-01C-01W-0188-10	400	4/18/2007	0.5

Figure 6.12 aliquot text file example

BCRPATIENTBARCODE	BCRSAMPLEBARCODE	BCRPORTIONBARCODE	BCRANALYTEBARCODE	EXPERIMENTALPROTOCOLTYPE	PROTOCOLNAME	PROTOCOLFILENAME
TCGA-02-0001	TCGA-02-0001-01A	null	null	null	null	null
TCGA-02-0001	TCGA-02-0001-01C	TCGA-02-0001-01C-01	TCGA-02-0001-01C-01D	nucleic_acid_extraction	Extraction Procedure for Samples	Extraction Procedure.pdf
TCGA-02-0001	TCGA-02-0001-01C	TCGA-02-0001-01C-01	TCGA-02-0001-01C-01R	nucleic_acid_extraction	Extraction Procedure for Samples	Extraction Procedure.pdf
TCGA-02-0001	TCGA-02-0001-01C	TCGA-02-0001-01C-01	TCGA-02-0001-01C-01T	nucleic_acid_extraction	Extraction Procedure for Samples	Extraction Procedure.pdf
TCGA-02-0001	TCGA-02-0001-01C	TCGA-02-0001-01C-01	TCGA-02-0001-01C-01W	null	null	null
TCGA-02-0002	TCGA-02-0002-01A	TCGA-02-0002-01A-01	TCGA-02-0002-01A-01D	nucleic_acid_extraction	Extraction Procedure for Samples	Extraction Procedure.pdf
TCGA-02-0002	TCGA-02-0002-01A	TCGA-02-0002-01A-01	TCGA-02-0002-01A-01R	nucleic_acid_extraction	Extraction Procedure for Samples	Extraction Procedure.pdf
TCGA-02-0002	TCGA-02-0002-01A	TCGA-02-0002-01A-01	TCGA-02-0002-01A-01T	nucleic_acid_extraction	Extraction Procedure for Samples	Extraction Procedure.pdf
TCGA-02-0003	TCGA-02-0003-01A	TCGA-02-0003-01A-01	TCGA-02-0003-01A-01D	nucleic_acid_extraction	Extraction Procedure for Samples	Extraction Procedure.pdf
TCGA-02-0003	TCGA-02-0003-01A	TCGA-02-0003-01A-01	TCGA-02-0003-01A-01R	nucleic_acid_extraction	Extraction Procedure for Samples	Extraction Procedure.pdf
TCGA-02-0003	TCGA-02-0003-01A	TCGA-02-0003-01A-01	TCGA-02-0003-01A-01T	nucleic_acid_extraction	Extraction Procedure for Samples	Extraction Procedure.pdf
TCGA-02-0003	TCGA-02-0003-01A	TCGA-02-0003-01A-01	TCGA-02-0003-01A-01W	null	null	null
TCGA-02-0006	TCGA-02-0006-01A	null	null	null	null	null
TCGA-02-0006	TCGA-02-0006-01B	TCGA-02-0006-01B-01	TCGA-02-0006-01B-01D	nucleic_acid_extraction	Extraction Procedure for Samples	Extraction Procedure.pdf
TCGA-02-0006	TCGA-02-0006-01B	TCGA-02-0006-01B-01	TCGA-02-0006-01B-01R	nucleic_acid_extraction	Extraction Procedure for Samples	Extraction Procedure.pdf

Figure 6.13 protocol text file example

BCRPATI	BCRPORTIONBARCODE	BCRSLIDEBARCODE	SECTIONLOCATION	NUMBERPROLIFERATINGCELLS	PERCENTTUMORCELLS	PERCENTTUMORNUCLEI	PERCENT	PERCENT	PER
TCGA-02-(	TCGA-02-0001-01C-01	TCGA-02-0001-01C-01-BS1	BOTTOM	1	95-100	95-100	0	0	
TCGA-02-(	TCGA-02-0001-01C-01	TCGA-02-0001-01C-01-TS2	TOP	0	90-94	95-100	0	0	1
TCGA-02-(	null	null	null	null	null	null	null	null	null
TCGA-02-(	TCGA-02-0002-01A-01	TCGA-02-0002-01A-01-BS1	BOTTOM	2	90-94	95-100	0	10-Jun	
TCGA-02-(	TCGA-02-0002-01A-01	TCGA-02-0002-01A-01-TS1	TOP	2	90-94	95-100	0	10-Jun	
TCGA-02-(	TCGA-02-0003-01A-01	TCGA-02-0003-01A-01-BS1	BOTTOM	0	80-84	95-100	0	16-20	
TCGA-02-(	TCGA-02-0003-01A-01	TCGA-02-0003-01A-01-TS1	TOP	4	95-100	95-100	0	5-Jan	
TCGA-02-(	TCGA-02-0006-01B-01	TCGA-02-0006-01B-01-BS1	BOTTOM	3	90-94	95-100	0	10-Jun	
TCGA-02-(	TCGA-02-0006-01B-01	TCGA-02-0006-01B-01-TS2	TOP	0	95-100	95-100	0	5-Jan	
TCGA-02-(	null	null	null	null	null	null	null	null	null
TCGA-02-(	TCGA-02-0007-01A-01	TCGA-02-0007-01A-01-BS1	BOTTOM	0	95-100	95-100	0	5-Jan	
TCGA-02-(	TCGA-02-0007-01A-01	TCGA-02-0007-01A-01-TS2	TOP	2	80-84	95-100	0	10-Jun	
TCGA-02-(	TCGA-02-0009-01A-01	TCGA-02-0009-01A-01-BS1	BOTTOM	0	95-100	95-100	0	5-Jan	
TCGA-02-(	TCGA-02-0009-01A-01	TCGA-02-0009-01A-01-TS1	TOP	0	95-100	95-100	0	5-Jan	
TCGA-02-(	TCGA-02-0010-01A-01	TCGA-02-0010-01A-01-BS1	BOTTOM	3	90-94	95-100	0	10-Jun	

Figure 6.14 slide text file example

Using Tab-delimited Files to Aggregate Clinical Metadata

The following workflow explains how to locate, download and extract data for patient "TCGA-02-0021" using tab-delimited TXT files.

1. Locate the patient sample data of interest using the Data Access Matrix, available from TCGA home page, http://tcga-data.nci.nih.gov/tcga/.

   Note: You can also navigate to the Data Access Matrix by clicking the **Browse Data** tab on any page of TCGA portal.

2. In the Data Access Matrix home page, scroll to patient TCGA-02-0021, shown in *Figure 6.15*. The last two cells of the row corresponding to your patient of choice will pull up clinical data.

   Note: You can also follow the instructions provided in the *Data Matrix Access User Guide* to create a filter for selecting the clinical data for patient number TCGA-02-0021.

Figure 6.15 Data Access Matrix page for browsing data.

You can open the Public data without a Matrix user account. The public data is an unrestricted, albeit sparse subset of the complete clinical data.

3. To access unrestricted data, follow these steps:

   a. Click the cell in the **Public** column.

   b. Click the **Build Archive** button, circled in *Figure 6.15*.

The hierarchical list of results displays all files that contain unrestricted clinical metadata for the identified patient (*Figure 6.17*).

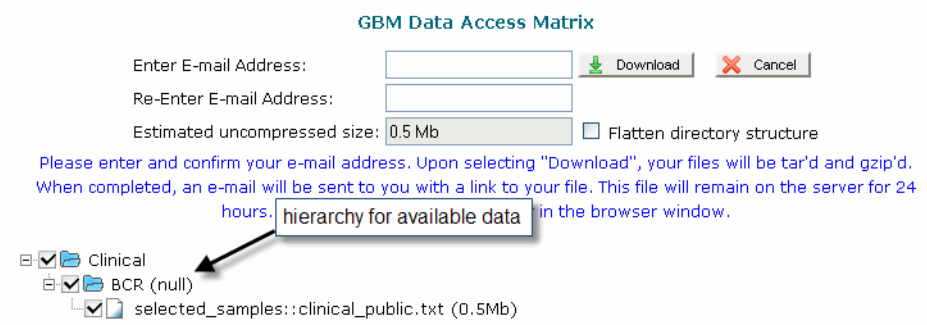


Figure 6.16 Matrix page for accessing public clinical data

c. Check or uncheck the appropriate files you want to access.

d. Click **Download** to launch the process. Click **Cancel** to abort the download.

- or -

4. To access restricted data, follow these steps:

   a. Click the cell in the **All Clinical** column.

   b. Click the **Build Archive** button. On the page that opens, the hierarchical list displays all files that contain data for the identified patient (*Figure 6.17*).

      You can access this restricted data only with a Data Access Matrix account. Link for requesting account?

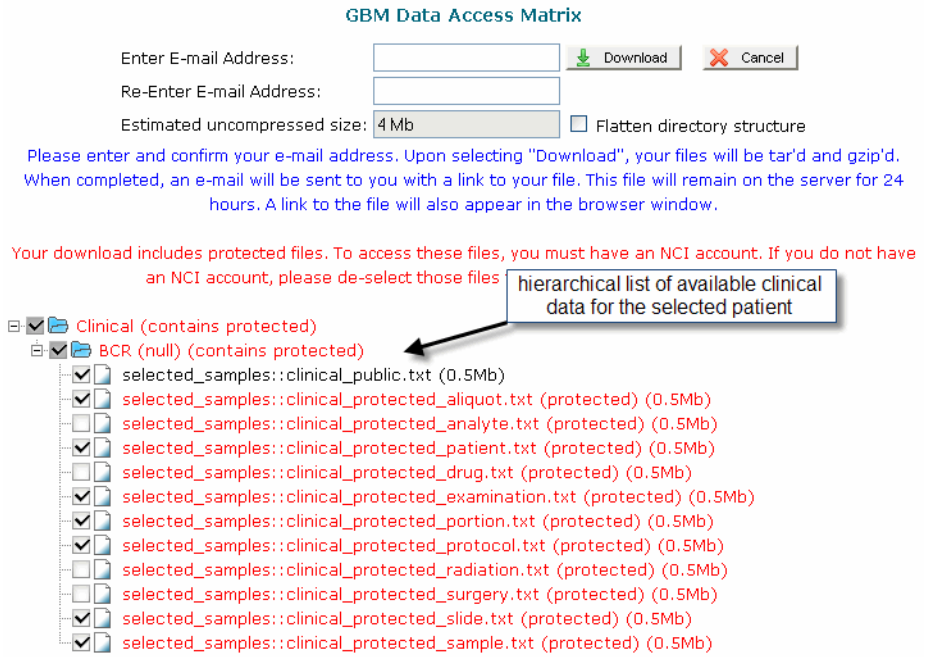
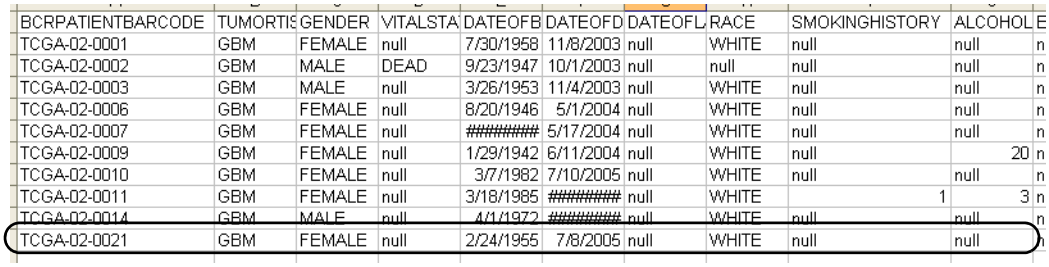


Figure 6.17 Matrix page for accessing all clinical data TXT files

c. Enter and confirm your email address.

d. Click **Download** and extract the compressed files. Click **Cancel** to abort the download.

5. Open one of the files, such as the `clinical_protected_patient.txt` file, in a spreadsheet application such as Excel. This file shows a profile of the patient used for the example, TCGA-02-0021 (circled in *Figure 6.18*).



BCRPATIENTBARCODE	TUMORTIS	GENDER	VITALSTA	DATEOFB	DATEOFD	DATEOFL	RACE	SMOKINGHISTORY	ALCOHOLE
TCGA-02-0001	GBM	FEMALE	null	7/30/1958	11/8/2003	null	WHITE	null	null
TCGA-02-0002	GBM	MALE	DEAD	9/23/1947	10/1/2003	null	null	null	null
TCGA-02-0003	GBM	MALE	null	3/26/1953	11/4/2003	null	WHITE	null	null
TCGA-02-0006	GBM	FEMALE	null	8/20/1946	5/1/2004	null	WHITE	null	null
TCGA-02-0007	GBM	FEMALE	null	########	5/17/2004	null	WHITE	null	null
TCGA-02-0009	GBM	FEMALE	null	1/29/1942	6/11/2004	null	WHITE	null	20
TCGA-02-0010	GBM	FEMALE	null	3/7/1982	7/10/2005	null	WHITE	null	null
TCGA-02-0011	GBM	FEMALE	null	3/18/1985	########	null	WHITE	1	3
TCGA-02-0014	GBM	MALE	null	4/1/1972	########	null	WHITE	null	null
TCGA-02-0021	GBM	FEMALE	null	2/24/1955	7/8/2005	null	WHITE	null	null

Figure 6.18 A segment of the clinical_protected_patient.txt file, opened in Excel.

The TXT file illustrated in *Figure 6.18* shows only the patient barcodes (column 1).

Remember that each delimited text file represents clinical data elements (also called “XML elements”), as identified by the file name, found in the XML files for each individual patient. As you open the delimited files representing the XML elements shown with barcodes (shown in *Figure 6.7*, with samples, portions, analytes, slides, etc. displayed), they reveal more barcode information for the example patient and many other patients, as well. Note the many additional barcodes pertaining to the example patient in a segment of the aliquot TXT file shown in *Figure 6.19*.

TCGA-02-0014	TCGA-02-0014-01A	TCGA-02-0014-01A-02	null	null	null
BCRPATIENTBARCODE	BCRSAMPLEBARCODE	BCRPORTIONBARCODE	BCRANALYTEBARCODE	BCRALIQUOTBARCODE	AMOUI
TCGA-02-0021	TCGA-02-0021-01A	TCGA-02-0021-01A-01	TCGA-02-0021-01A-01D	TCGA-02-0021-01A-01D-0182-01	
TCGA-02-0021	TCGA-02-0021-01A	TCGA-02-0021-01A-01	TCGA-02-0021-01A-01D	TCGA-02-0021-01A-01D-0183-04	
TCGA-02-0021	TCGA-02-0021-01A	TCGA-02-0021-01A-01	TCGA-02-0021-01A-01D	TCGA-02-0021-01A-01D-0184-06	
TCGA-02-0021	TCGA-02-0021-01A	TCGA-02-0021-01A-01	TCGA-02-0021-01A-01D	TCGA-02-0021-01A-01D-0185-02	
TCGA-02-0021	TCGA-02-0021-01A	TCGA-02-0021-01A-01	TCGA-02-0021-01A-01D	TCGA-02-0021-01A-01D-0186-05	
TCGA-02-0021	TCGA-02-0021-01A	TCGA-02-0021-01A-01	TCGA-02-0021-01A-01R	TCGA-02-0021-01A-01R-0177-01	
TCGA-02-0021	TCGA-02-0021-01A	TCGA-02-0021-01A-01	TCGA-02-0021-01A-01R	TCGA-02-0021-01A-01R-0178-03	
TCGA-02-0021	TCGA-02-0021-01A	TCGA-02-0021-01A-01	TCGA-02-0021-01A-01R	TCGA-02-0021-01A-01R-0179-07	
TCGA-02-0021	TCGA-02-0021-01A	TCGA-02-0021-01A-01	TCGA-02-0021-01A-01R	TCGA-02-0021-01A-01R-0181-02	
TCGA-02-0021	TCGA-02-0021-01A	TCGA-02-0021-01A-01	TCGA-02-0021-01A-01T	TCGA-02-0021-01A-01T-0179-07	
TCGA-02-0021	TCGA-02-0021-01A	TCGA-02-0021-01A-01	TCGA-02-0021-01A-01W	TCGA-02-0021-01A-01W-0188-10	
TCGA-02-0021	TCGA-02-0021-01A	TCGA-02-0021-01A-01	TCGA-02-0021-01A-01W	TCGA-02-0021-01A-01W-0189-08	
TCGA-02-0021	TCGA-02-0021-01A	TCGA-02-0021-01A-01	TCGA-02-0021-01A-01W	TCGA-02-0021-01A-01W-0190-09	

Figure 6.19 A segment of the clinical_protected_aliquot.txt file, showing data for patient TCGA-02-0021.

Once you have parsed relevant data/aliquot barcodes from the available XML files or TXT files like those described above, aggregating samples using clinical data is straightforward.

In summary, 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 TXT files. Extract relevant barcodes for aggregating barcode lists, in preparation for mapping the data to result files.

See *Mapping Aliquot Barcodes to Assay Result Files* on page 67 for more information.

Mapping Aliquot Barcodes to Assay Result Files

Once you have aggregated a set of barcodes of interest as discussed in *Aggregating Data Using Aliquot Barcodes* on page 48, you can use them to identify the assay result files associated with them. The method you use for mapping aliquot barcodes to their result files depends on the data type—array vs. sequence-based—and data level of interest.

You must use the SDRF files from the archives that you downloaded data for, or the SDRF files that are included in the DAM download. Alternatively, you can download all SDRFs at once using http://tcga-data.nci.nih.gov/tcga/latestDownloadableResults.htm?filetype=sdrf. SDRF files have the aliquot barcode, so you map between the aliquot barcode in your aggregated list and the barcodes in SDRF files.

HLA files must already have been mapped to SDRF files, as described in *Mapping Data Levels 3 and 4* on page 68.

The level to map to is determined by the level of data you require. For instance, you might want to do all the data processing and analysis yourself starting the basic level 1 files. Or you might want higher level data that has already been processed so you can start doing some statistics or make some conclusions.

Mapping Array-Based Data

Characterization centers' experiments produce many different data files that have particular data types and data levels. Mapping between aliquot barcodes involves MAGE-TAB SDRF files and/or TCGA Higher Level Analysis (HLA) specification files. Because of the variations in data mapping, these are described in separate sections.

Mapping Data Levels 1 and 2

The MAGE-TAB SDRF files provide 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/Data Level Relationships* on page 27).

The SDRF file is like a database describing the relationships between samples and their results. For example, in *Table 6.3*, the **Extract Name** column contains the BCR aliquot barcode for each sample and the **Derived Array Data Matrix File** column lists the result file associated with each barcode. In fact, any file listed in the SDRF in the same row as a barcode is associated with that aliquot's assay result.

Note: One-to-many (i.e. one extract name, or hybridization, or any other experiment entity may be associated with many files) and many-to-many (i.e. many extract names or other experiment entities may be associated with many files) relationships between Extract Names and *File columns can occur. Usually, however, the relationships are one-to-one or many-to-one. *Table 6.3* is an example of the many to one relationship in

that multiple extracts are found in one Derived Array Data Matrix File. Differences in the extract names are highlighted in bold.

Extract Name (Sample ID)	*Derived Array Data Matrix File (MAGE-TAB SDRF)*
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 6.3 MAGE-TAB SDRF—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.

Mapping Data Levels 3 and 4

Currently 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). 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 HLA SDRF files to MAGE-TAB SDRF files via files listed in columns that they have in common.

For example, *Table 6.4* represents two columns of data from an HLA SDRF: a "**Derived HLA Data File**" column and a "**Derived Array Data Matrix**" column, which it has in common with the MAGE-TAB SDRF (*Table 6.3*). If files listed in the **Derived Array Data Matrix File** column in an HLA SDRF (column 2 in *Table 6.4*) correspond to files also listed in a MAGE-TAB SDRF (*Table 6.3*), you can associate the HLA files with a set of sample IDs through the files in the **Derived Array Data Matrix** columns.

Derived HLA Data File	*Derived Array Data Matrix File (HLA-SDRF)*
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	broad.mit.edu_GBM.Genome_Wide_SNP_6.1.ismpolish.data.txt

Table 6.4 HLA SDRF—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 ID-to-sample 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 samples (**Tumor_Sample_Barcode** column) and normal samples (**Matched_Norm_Sample_Barcode** column).

To associate barcodes with GSC-based data, for example sequence trace files or mutations, you must first download the GSC files and then parse them 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 6.20* attempts to create a cohesive view of the relationships between elements, at the same time illustrate mapping between the different elements.

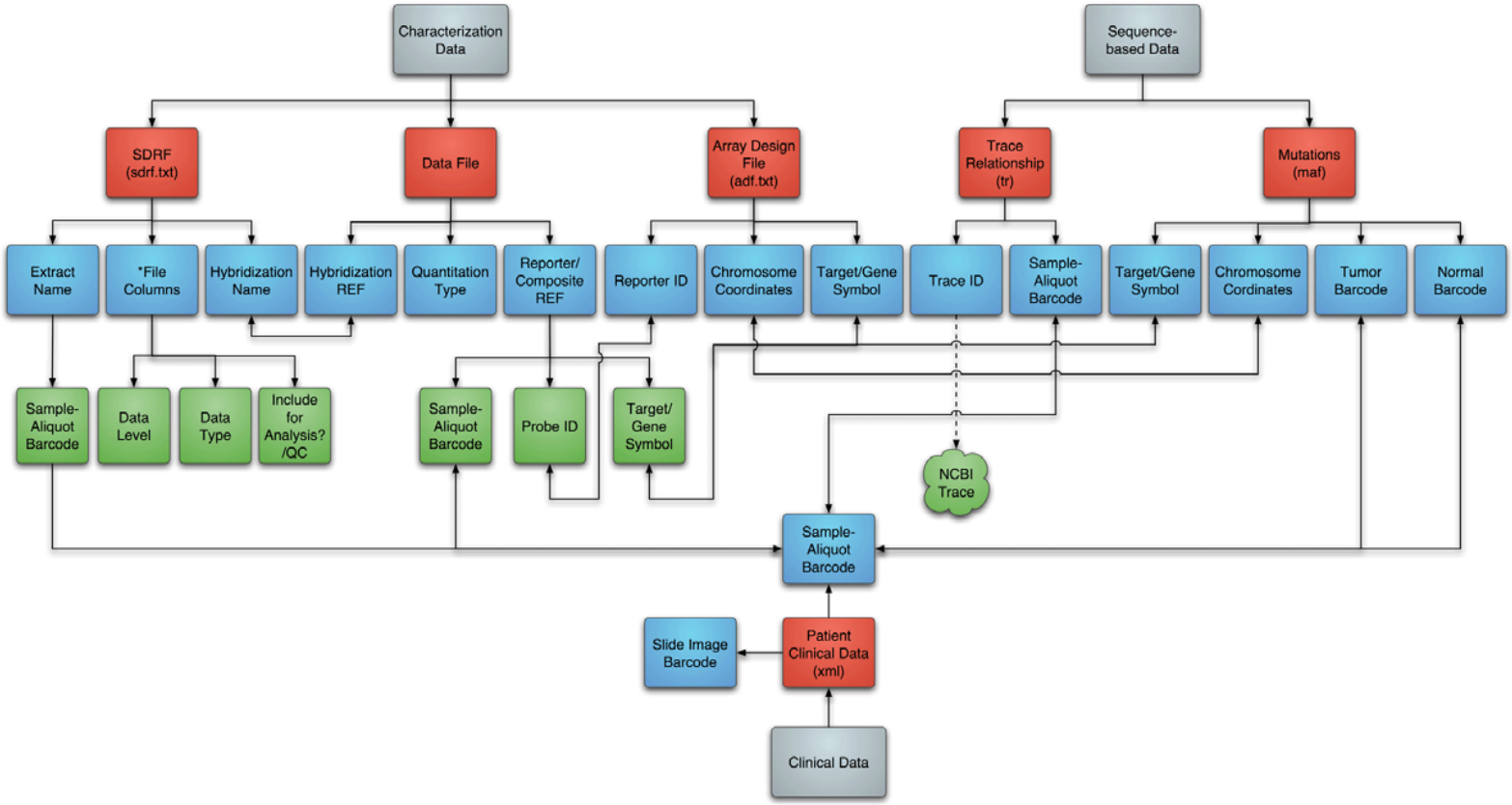


Figure 6.20 Linking between TCGA file elements.

See *Table 6.5* for a key to the elements represented in *Figure 6.20*.

Color and/or Object	Representation
Grey	Classifications of data
Red	Files
Blue	Elements of interest in their parent (red) files[a]
Green squares	Element values of their parent (blue) objects
Green clouds	Mappings of data outside of TCGA
Dashed lines	Mappings of data outside of TCGA
Solid arrowed lines	Mappings between elements or their values

Table 6.5 Key to elements in Figure 6.20

a. Blue objects that do not have a green child object also represent the values of their "missing" child objects.

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 73
- *Plate Barcode Values* on page 75

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, shown in *Table A.1*, represent TCGA centers.

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, examples of which are shown in *Table A.2*, range from `01-09` for tumor types, `10-19` for normal types, and `20-29` for control samples. For the most upt to date list of values, see http://tcga-data.nci.nih.gov/tcga/AllTCGASampleTypes.htm.

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 range from A to 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 composites of portion and analyte identifiers. Solid tumors are divided into a sequence of 100 to 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, as shown in *Table A.3*, represent the types of analytes for which the sample is analyzed.

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 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, as shown in *Table A.4*, represent the CGCCs and GSCs.

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)
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

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 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
aliquot	Portion of a sample
aliquot barcode	An identifier that is a combination of an analyte barcode and a plate barcode separated with a hyphen as follows: `{analyte barcode}-{plate barcode}`
analyte barcode	An identifier that identifies the collection site, patient, sample, and portion ID
astrocytic tumors: astrocytoma	Neoplasms of the brain and spinal cord derived from glial cells. Also known as astrocytomas.
BCR	Biospecimen Core Resource
binary file	Computer file which may contain any type of data encoded in binary form for computer storage and processing purposes. Binary data is usually written in a numeral system that uses two symbols, usually 0 and 1.
bulk download	Transferring files en masse as the data was deposited
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	Data Coordinating Center. Manages TCGA data entered into public databases as it becomes available.

Table C.1 Glossary of genomic analysis terms

Term	*Definition*
experiment	An experiment for an 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's gene symbols. 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	Loss of heterozygosity
MAF	A mutation annotation format file. A MAF annotates mutations discovered by aligning DNA sequences derived from tumor samples to sequences derived from normal samples and a reference sequence.
MIAME	Minimum Information About a Microarray Experiment that is needed to enable the interpretation of the results of the experiment unambiguously and potentially to reproduce the experiment. See http://www.mged.org/Workgroups/MIAME/miame.html.
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 known as oligodendrogliomas.
plate barcode	An identifier that identifies the plate and the GSC or CGCC to which it will be distributed.
programmatic access	Access using a script in a recommended application
restricted/controlled	Protected access; available only through using a username and password requested from the TCGA
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, 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.
unrestricted/public	Open access; available to anyone

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

INDEX

A

ADF 79
  description 22
  example 23
Affymetrix platforms 26, 77
aggregated data
  mapping data levels 1 & 2 67
  mapping data levels 3 & 4 68
aggregating data
  between different centers 51
  between different platforms 51
  by clinical metadata 54
Agilent platforms 26, 77
aliquot 79
aliquot barcode
  constructing 5, 6
  deciphering 7
  definition 7, 79
  values 73
analyte
  coding 5
  collection 5
  processing steps 6
analyte barcode
  analyte code values 75
  deciphering 7
  definition 7, 79
  example 7, 8
  patient ID values 74
  portion code values 75
  portion ID values 74
  sample ID values 74
  sample type values 74
  site ID values 73
  vial identifier values 74
Applied Biosystems platforms 26
archive 79
  confirming integrity 44
  MD5 file in 36
  search 41
  search results 41
  using MD5 hash files 44
archive data
  confirming integrity 44
  description 35
  freezes in 37
  naming conventions 36
Array Design File, see ADF
astrocytic tumors 79
astrocytoma 79

B

barcode
  analyte, deciphering 7
  analyte examples 8
  analyte values 73
  construction 5, 6
  plate, deciphering 8
  plate example 9
BCR 79
  aliquot barcode 7
  data type 25
  description 5
  UML models 6
  XML elements 61
Biospecimen Core Resource, see BCR
Biospecimen Metadata platforms 26, 77
bulk download
  constructing path to directories 38
  definition 79
  directory structure description 38, 41
  directory structure illustration 40
bulk downloads
  controlled access 38
  open access 38
  TCGA data 38

C

cancer centers, role in data flow 2
Cancer Genomic Characterization Center, see CGCC

cancer types in TCGA 36
categorized data, description 26
CGCC
  data types generated 25
  description 17, 79
CGCCs
  role in data flow 3
clinical metadata to aggregate data 54
collection sites
  role in data flow 2
controlled 80
controlled access, bulk downloads 38
CSV file names 61

D

DAG
  example 21
  in SDRFs 21
data
  categorization, overview 26
  file formats 4
  freezes 37
  resources in DCC 44
data access
  bulk downloads 38
  Data Access Matrix 42
  DCC resources 44
  methods for 38
  TCGA Data Portal 41
Data Access Matrix
  downloading data 43
Data Access Matrix, description 42
data archive, naming convention 36
Data Coordination Center, see DCC
data flow
  description 2
  illustration 1
data freezes
  description 37
  lists of 38
data level
  characterization of CGCC data 33
  characterization of GSC data 31
  current as applies to data type 28
  Data Type-Data Level File-Suffix Matrix to determine 31
  data type relationships 27
  descriptions 27
  determining in result file 29, 33
  mapping levels 1 & 2 67
  mapping levels 3 & 4 68
  normalized as applies to data type 28
  types 26
data level 1 27, 28
data level 2 27, 28
data level 3 27, 28
data level 4 27, 28
data type
  characterization of CGCC data 33
  characterization of GSC data 31
  corresponding data level 28
  data level relationships 27
  Data Type-Data Level File-Suffix Matrix to determine 31
  determining in result file 29, 33
Data Type-Data Level File-Suffix Matrix
  example 30
  simplified example 32
  using to determine data types/data levels 31
DCC 79
  bulk data download 38
  data access resources 44
  data freezes 37
  data received by
  data resources 44
  description
distributing data, from BCR 9
document description 1
downloading data
  from Data Access Matrix 43

E

experiment, definition 80
experiment archives, description 4

F

FASTA file, description 16
file formats
  compatible with NCIB & DCC repositories 4
  IDF files 20
  MAF files 15
  SDRF files 22
  trace files 12
  trace ID-to-sample relationship 14

G

Genomic Sequencing Center, see GSC
glossary 79
GSC
  data types in 25
  definition 80
  description 11

H

HLA files, mapping to 68

HLA SDRF 68
HUGO 80

I

IDF file
  definition 80
  description 18
  formats 20
  protocols 20
Illumina platforms 26, 77
Investigation Description Format, see IDF

L

LOH 80

M

MAF 80
MAF files
  column headers 15
  description 14
  format 15
  validation 14
MAGE-based experiments in TCGA 18
MAGE-related references 17
MAGE-TAB
  in TCGA 17
  references 17
  specification 18
mapping
  between TCGA file elements 69
  data levels 1 & 2 67
  data levels 3 & 4 68
  sequence data 68
MD5
  hash files 44
MD5 file in archives 36
MIAME 80
mutation annotation format files, see MAF files

N

naming conventions, data archives 36
NCBI 80
NCICB 80

O

oligodendroglial tumor 80
oligodendroglioma 80
open access, bulk downloads 38

P

patient privacy in TCGA 43
plate barcode 80
  center ID values 75
  deciphering 8
  definition 7
  examples 8
  plate ID values 75
platform codes 77
platforms, supported 26, 77
primer description 1
processing analytes 6
programmatic access 80
protocols, in IDFs 20
public 80

R

restricted 80

S

Sample and Data Relationship Format file, see SDRF files
SDRF
  HLA 68
SDRF files
  description 21
  example 22
  format 22
search
  archive 41
  archive results 41
sequence trace file 12
SNP 80

T

TCGA 80
  cancer types in 36
  data, bulk downloads 38
  data freezes 37
  data type/data level relationships 27
  glossary 79
  MAGE-based experiments 18
  MAGE-TAB in 17
  patient privacy 43
TCGA Data Access Matrix
  data access 38
  link to 38
TCGA Data Portal
  data access 38
  definition 80
  description 41
  link to 38
trace files
  description 12
  format 12

trace ID-to-sample relationship files
description 13
format 14

U

unrestricted 80