TCGA Portal Data Access Matrix

User's Guide v. 2.3

Center for Biomedical Informatics
and Information Technology

This is a U.S. Government work.

August 24, 2009

TABLE OF CONTENTS

About This Guide 1

Purpose 1
Audience 1
    Typical User 1
    Prerequisites 1
Topics Covered 2
Additional References 2
Text Conventions Used 2
Credits and Resources 3

Chapter 1
Getting Started With the Data Access Matrix 5

Overview 5
    Major Steps For Accessing Data Sets 5
Accessing the Data Access Matrix 6
Navigating The Data Access Matrix 7
    Options Menu Features 10
    Key to Graphic Indicators In the Matrix 11

Chapter 2
Selecting Data Sets 15

Main Steps For Selecting Data Sets 15
Data Filtering Techniques 16
    Modifying the Display of Data 17
    Limiting Data by Sample 19
    Techniques For Filtering Data 20
    Color-Coding Data Sets 24
Data Selection Techniques 25
    Creating a Union of Data Sets 27
    Intersecting Data Sets 28
    Subtracting Data Sets 29

Chapter 3
Downloading and Retrieving TCGA Data 31

Navigating Through the Data Access Download Page 31
    Understanding Data Access Download Components 31
Main Steps For Downloading Your Data Files 35
Selecting All Data Files For Download 37

Selecting a Subset of Data Files For Download 37
Retrieving Your Data Archive 39

Appendix A
Data Set Selection – Examples 43

Selecting Specific Data Sets 43
Selecting All Data From a Batch 44
Selecting Data From a Platform Type 45
Selecting a Union of Data Sets 47
Selecting All Data From a Specific Center 48
Selecting All Data From a Single Center In a Single Batch 49
Selecting All Data From a Data Level 50

Appendix B
Accessing Data From an External Application or Website 51

Integrating TCGA Portal Data Access Service 51
Filtering Data From an External Application 51
    Using the Matrix to Visualize and Modify Your Filter 56
Valid Arguments For Data Set Filters 56
    Numeric Constants For Platforms, Platform Types, and Centers 57

Appendix C
Rules For the Visual Display of Data in the Matrix 59

Rules For Displaying Sample Barcodes 59
Rules For Displaying Cells 60
    Rules For Center Cells 60
    Rules For Clinical Cells 61
Rules For Displaying Data Sets 62
Rules For Displaying Batches 62
Origins of Orphaned Barcodes 62
    CGCC File Generation Errors 62
    CGCC File Submission Error 62
    BCR Barcode Exclusion 62
Rules For Displaying Orphaned Barcodes 63
    Rules for CGCC Orphans 63
    Rules for BCR Orphans 63

Index 65

About This Guide

This chapter introduces you to the Data Access Matrix User's Guide. It includes the following topics:

- *Purpose* on this page
- *Audience* on this page
- *Topics Covered* on page 2
- *Additional References* on page 2
- *Text Conventions Used* on page 2
- *Credits and Resources* on page 3

Purpose

This guide provides an overview of the Data Access Matrix. It explains how to use the Matrix to select and download specific data sets that are of interest to the researcher.

Note: In this Guide, the term "TCGA Portal" may be referred to as "the Portal." Similarly, "the Data Access Matrix" may be referred to as "the Matrix."

Audience

Typical User

This guide is designed for a broad cross-section of the cancer research community, including basic and clinical researchers, clinicians, and patient advocates.

Prerequisites

To get the most out of this guide, you should be familiar with the following topics:

- Clinical, genomic characterization, and gene sequencing data
- Internet and associated terminology

Note: For optimal viewing, use a Firefox browser to access the Matrix.

Topics Covered

If you are new to the Data Access Matrix, read this brief overview, which explains what you will find in each chapter.

- *Chapter 1* introduces you to the Data Access Matrix and provides details about the user interface that will help you to select and download data sets.
- *Chapter 2* provides instructions for selecting specific data sets for your study.
- *Chapter 3* provides instructions for downloading the archives the contain your selected data sets.
- *Appendix A* provides examples of data selection techniques.
- *Appendix B* provides instructions for using the external interface to filter, archive, and retrieve data sets directly from external Web applications.
- *Appendix C* provides the rules that dictate the display of data in the Matrix.

Additional References

For detailed information about the data that is available in the TCGA Data Access Matrix, see the *The Cancer Genome Atlas Data Primer*, available for download at: http://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/other/TCGA_Data_Primer.pdf.zip

Text Conventions Used

This section explains conventions used in this guide. The various typefaces represent interface components, keyboard shortcuts, toolbar buttons, dialog box options, and text that you type.

Convention	***Description***	***Example***
Bold	Highlights names of option buttons, check boxes, drop-down menus, menu commands, command buttons, or icons.	Click **Search**.
URL	Indicates a Web address.	http://domain.com
text in SMALL CAPS	Indicates a keyboard shortcut.	Press ENTER.
text in SMALL CAPS + text in SMALL CAPS	Indicates keys that are pressed simultaneously.	Press SHIFT + CTRL.
Italics	Highlights references to other documents, sections, figures, and tables.	See *Figure 4.5*.
Italic boldface monospaced type	Represents text that you type.	In the **New Subset** text box, enter ***Proprietary Proteins.***
Note:	Highlights information of particular importance	**Note:** This concept is used throughout the document.

Convention	Description	Example
{ }	Surrounds replaceable items.	Replace {last name, first name} with the Principal Investigator's name.

Credits and Resources

The following people contributed to the development of this document.

Data Access Matrix Development and Management Teams		
Development	**Documentation**	**Project and Product Management**
David Nassau[3]	Lauren Anthone[2]	Carl Schaefer[1]
Robert Sfeir[3]		Matthew Shaker[3]
Erin Hedlund[3]		
Larry Feng[3]		
David Kane[3]		
Jessica Chen[3]		
Silpa Nanan[3]		
[1] National Cancer Institute Center for Bioinformatics (NCICB)	[2] Lockheed Martin	[3] SRA International, Inc.

Contacts and Support	
NCICB Application Support	http://ncicb.nci.nih.gov/NCICB/support Telephone: 301-451-4384 Toll free: 888-478-4423

CHAPTER

1

GETTING STARTED WITH THE DATA ACCESS MATRIX

This chapterchapter introduces you to the Data Access Matrix and provides details about the user interface that will help you to select and download data sets.

Topics in this Chapter

- *Overview* on this page
- *Accessing the Data Access Matrix* on page 6
- *Navigating The Data Access Matrix* on page 7

Overview

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

The Data Access Matrix (the Matrix) application enables researchers to select and download data sets from TCGA servers through a user-friendly graphic-based data set selection system.

Major Steps For Accessing Data Sets

Table 1.1 describes the major steps for selecting and downloading specific data.

Step	Action
1	Navigate to the Matrix

Table 1.1 Major steps for accessing data sets

Step	Action
2	Select the disease type of interest. Your choices include the following disease types: • GBM – Glioblastoma Multiforme • OV – Ovarian Serous Cystadenocarcinoma **Note:** A limited number of data sets for OV tissue are available in this release.
3	Select the data sets of interest
4	Confirm/select the files that contain the data of interest
5	Process the files for downloading
6	Retrieve and download the data archive

Table 1.1 Major steps for accessing data sets (Continued)

Accessing the Data Access Matrix

How to Navigate to the Data Access Matrix

1. Do one of the following:

   - Launch your internet browser and type http://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm directly in the address field.

   - or-

   - If you are already working in TCGA Portal, click the **Browse Data** tab on the top of any Portal page to display the Get TCGA Data pane, and then continue to the next step.

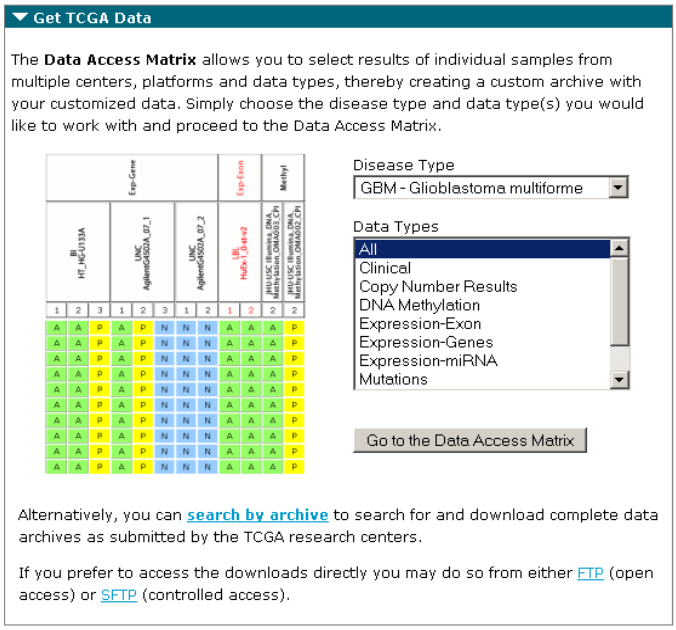

Figure 1.1 TCGA Portal Page – Get TCGA Data Pane

2. Under **Disease Type**, click the selector (down arrow) and then select the disease type of interest.

   Note: Data for ovarian cancer tissue (OV) is limited in this release.

3. In the **Data Types** list, select the types of data you want to include in your research.

   Note: Refer to the TCGA Data Primer for information about data types.

4. Click **Go to the Data Access Matrix**.

As a result, the Matrix appears in its default state—with no data selected. Depending on the choices you made in Steps 2 and 3, the Matrix may appear different than the one illustrated in *Figure 1.2*.

Navigating The Data Access Matrix

To use the Matrix efficiently, it is important that you familiarize yourself with the user interface.

Figure 1.2 illustrates the Data Matrix search page. The elements are described in *Table 1.2*.

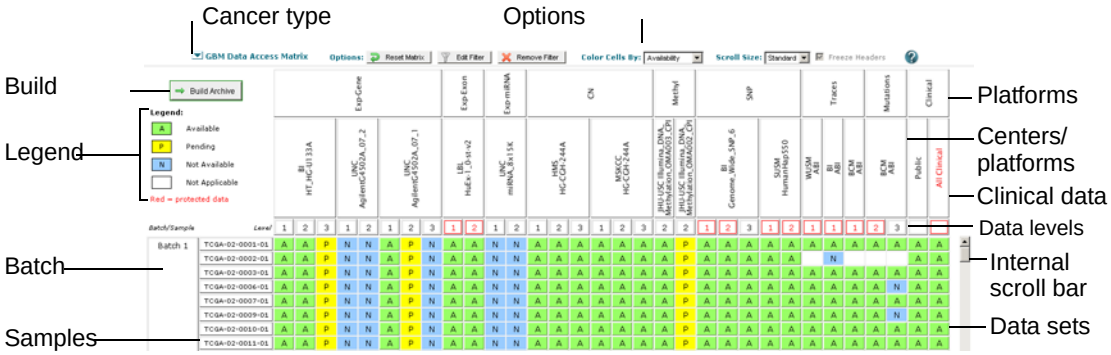

Figure 1.2 Matrix Features

Table 1.2 provides a description for each of the features.

Feature	***Description or Function***
Cancer type selection	Enables you to select the tumor type of interest; displays the tumor type abbreviation. Available tumor types include: • Glioblastoma Multiforme (GBM) • Ovarian Serous Cystadenocarcinoma (OV) **Note:** A limited number of data sets for OV tissue are available in this release.
Options menu	Displays options for filtering and viewing data. For details about the Options menu, see *Options Menu Features* on page 10.

Table 1.2 Description of Matrix features

Feature	***Description or Function***
Build Archive button	Creates an archive of files from selected data sets.
Batch button	Selects the entire batch (group) of samples. The samples included in each batch are displayed to the right of the **Batch** button. See *Data Selection Techniques* on page 25. A batch is a set of samples that the BCR processed and distributed to GSCs and CGCCs for analysis on particular platforms. The batch number corresponds to the BCR archive serial number.
Samples	In this document, and in the context of the Matrix, a sample consists of the following elements: ● Patient ID ● Sample Type Code ● Tissue Sample Count ● Portion Sequence See *Data Selection Techniques* on page 25. For further information about samples, refer to *The TCGA Data Primer*, available for download at: http://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/other/TCGA_Data_Primer.pdf.zip

Table 1.2 Description of Matrix features (Continued)

Feature	*Description or Function*
Legend	Provides a color key to data sets The following codes are displayed in the legend when cells are colored by **Availability**: • **A** – Available data set. Indicates that data is available for research. • **P** – Pending data set. Indicates that a sample has been submitted to the DCC, but the center has not yet processed or published the associated data. • **N** – Not available. Indicates that the BCR has sent a sample to the CGCC, but the CGCC has not submitted analysis result data to the DCC. • Blank (Not Applicable) – Indicates that the BCR did not send a sample to a given CGCC because the sample was incompatible with that center. Therefore no pertinent data exists for that sample. The following codes are displayed in the legend when cells are colored by **Tumor/Normal**: • **TN** (Tumor, matched normal) – Tumor tissue data for which corresponding normal tissue data exists • **T** (Tumor, no matched normal.) – Tumor tissue data for which there is no corresponding normal tissue data • **NT** (Normal, matched tumor) – Normal tissue data for which corresponding tumor tissue data exists. • **N** (Normal, no matched tumor) – Normal tissue data for which there is no corresponding tumor tissue data • Blank – No data set is available. When colored by Tumor/Normal, the Matrix filters out all data sets that are either Pending or Not Available. All colored cells, therefore, represent data sets that are available.
Platform types	Characterization and sequencing technology platforms used to derive genomic and clinical data. • **Exp-Gene** – Expression gene • **Exp-Exon** – Expression exon • **Exp-miRNA** – Expression mitochondrial RNA • **CN** – Copy number • **Methyl** – DNA Methylation • **SNP** – Single nucleotide polymorphism • **Traces** – Trace-Relationship • **Mutations** – This data will be available in a subsequent release of the Data Access Matrix. • **Clinical** See *Data Selection Techniques* on page 25. For platform type details, see *Table 1.5* on page 13.

Table 1.2 Description of Matrix features (Continued)

Feature	Description or Function
Centers/platforms	CGCC cells. A specific platform name is provided when a center has used more than one platform type. Red text indicates restricted, or controlled data. See *Data Selection Techniques* on page 25. The following TCGA research centers provide data for the Matrix as follows: • **BI** (Broad Institute of MIT and Harvard) – The Eli and Edythe L. Broad Institute of the Massachusetts Institute of Technology (MIT) and Harvard University and the Dana Farber Cancer Institute • **HMS** (Harvard Medical School) – Brigham and Women's Hospital of Harvard Medical School and Dana Farber Cancer Institute • **JHU USC** – Johns Hopkins University and University of Southern California joint group • **LBL** – Lawrence Berkeley National Laboratory • **MSKCC** – Memorial Sloan-Kettering Cancer Center • **UNC** (University of North Carolina) – University of North Carolina, Lineberger Comprehensive Cancer Center • **SUSM** – Stanford University School of Medicine
BCR clinical data	Biospecimen Core Resource – Clinical pathology metadata and associated aliquot barcode, which uniquely identifies each product from a collection site. **Note:** Red text indicates controlled/protected data. To obtain access to protected TCGA data sets, please follow the instructions at: http://cancergenome.nih.gov/data/access/closed/ See *Data Selection Techniques* on page 25.
Data levels	Level 1, 2, and 3 data are available for study. See *Data Selection Techniques* on page 25. The concept of TCGA data level segregates raw data from derived data from higher-level analysis or interpreted results for each data type, platform, and center. You can find detailed information on data levels in *The TCGA Data Primer*, available for download at: http://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/other/TCGA_Data_Primer.pdf.zip
Internal scroll bar	Enables you to move up and down through the data sets

Table 1.2 Description of Matrix features (Continued)

Refer to these figures and tables as necessary as you select, and then download, your data sets. See *Selecting Data Sets* on page 15 and *Downloading and Retrieving TCGA Data* on page 31.

Options Menu Features

Figure 1.3 illustrates Options menu features. The features are described in *Table 1.3* on page 11.

Options menu, when no filters have been set

Filter options, when one or more filters have been

Figure 1.3 Options Menu

Feature	***Description or Function***
Reset Matrix button	Removes all filters and deselects all selected cells.
Filter buttons	• **Set** – Launches the filter dialog box with default settings (no filters) • **Edit** – Launches the filter window with your current settings so that you can modify your filter criteria • **Remove** – Clears all filter settings
Color Cells By drop-down list	Enables you to color-code the data sets by availability or tumor/normal data. See *Color-Coding Data Sets* on page 24.
Scroll Size drop-down list	Enables you to display the Matrix in the following sizes to suit your monitor configuration: • **None** – Hides the internal scroll bar so that you can use your browser's scroll bar to scroll through the Matrix. • **Small** • **Medium** • **Large** The Matrix remembers your preference for display size from session to session if you have done the following: • Configured your browser to accept cookies • Submitted at least one request to the server, e.g., you have selected a row or have set a filter
Freeze Headers check box	When selected, allows you to view all the column headers (data levels, platform types, etc.) while you scroll through the sample sets. **Note:** The **Freeze Headers** check box only applies when the Scroll Size None is selected. This feature is not available in Internet Explorer 6.

Table 1.3 Options menu features

Key to Graphic Indicators In the Matrix

Table 1.4 provides a key to graphic elements, colors, and symbols used in the Matrix. See *Table 1.2* on page 7 for further descriptions.

Tip: Every cell and column or row header in the Matrix functions like a toggle button. Click a cell once to activate/select it, and then click it again to deactivate/deselect it.

Appendix C, Rules For the Visual Display of Data in the Matrix, on page 59 provides the rules that dictate the display of data in the Matrix. It explains the classifications of data as Available, Not Available, Not Applicable, Pending, and orphaned.

Element	***Description***
A	Cell (as in a table cell) – Represents data, no matter what color it is. Each cell represents the following: • For CGCC data – each cell represents a data set for a sample at a specific data level from a specific center and platform • For BCR data – each cell represents the clinical data at the specified access level for a particular sample To enhance the visual color cue, each cell is lettered. In this example, "A" indicates data that is "available." For further information about CGCC and BCR data, refer to *The TCGA Data Primer*, available for download at: http://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/other/TCGA_Data_Primer.pdf.zip
BI HT_HG-U133A UNC Agilent-G4502A_07_2 1 2 3 1 2	Selected column header – A highlighted graphic element indicates that an element has been selected, or activated. In this example, the center, "BI," and its data levels 1,2, and 3 are highlighted.
1	Red text and/or red outline– Indicates protected, or controlled data. To obtain access to protected TCGA data sets, please follow the instructions at http://cancergenome.nih.gov/data/access/closed/
A A	Selected cell – A highlighted cell indicates that the data set has been selected, or activated. In this example, the cell on the left is highlighted. **Note: You can select only "available" data sets.**
A	Available sample data, when the data sets are colored by Availability.
P	Pending sample data, when the data sets are colored by Availability.
N	Sample data that is Not Available, when the data sets are colored by Availability.
	Blank cell (data does not exist), when the data sets are colored by Availability.
T	Tumor tissue sample data, when the data sets are colored by Tumor/Normal.
TN	Tumor and matched normal sample data, when the data sets are colored by Tumor/Normal.
N	Normal tissue sample data, when the data sets are colored by Tumor/Normal.
NT	Normal and matched tumor tissue data, when the data sets are colored by Tumor/Normal.

Table 1.4 Key to graphic elements in the matrix

For further information about the graphic elements, see *Legend* on page 9 and *Color-Coding Data Sets* on page 24.

Table 1.5 lists and describes the platforms used to provide TCGA data.

Platform	***Description***	***Associated Data Type***
Affymetrix HT Human Genome U133 Array Plate Set	High-throughput expression profile of approximately 40,0000 transcripts and variants	Expression-Genes
Affymetrix Human Exon 1.0 ST Array	Contains approximately 1million predicted and confirmed exon transcripts	Expression-Exon
Affymetrix Genome-Wide Human SNP Array 6.0	Allows detection of copy number variation with more than 906,600 single nucleotide polymorphisms (SNPs) and over 946,000 probes	Copy Number Results, SNP, SNP Frequencies
Agilent 8 x 15K Human miRNA-specific Microarray	Highly specific and sensitive microRNA expression profiling system	Expression-miRNA
Agilent Human Genome CGH Microarray 244A	Allows analysis of ~240,000 coding and non-coding sequences for chromosomal DNA alterations	Copy Number Results
Agilent Whole Human Genome Microarray Kit, 4 x 44K	High-density profiling analysis tool that covers over 41,000 unique human genes and transcripts	Copy Number Results, Expression-Genes
Illumina DNA Methylation OMA002 Cancer Panel I	First array-based high-throughput, high multiplexing, single-site CpG resolution platform; contains 1,505 CpG loci selected from 807 genes	Methylation
Illumina 550K Infinium® HumanHap550 SNP Chip	Covers 550,000 SNP loci across genome	SNP
Biospecimen Metadata – Complete Set	Detailed clinical phenotype and outcome data	Complete Clinical Set
Biospecimen Metadata – Minimal Set	Clinical Diagnosis, Histologic Diagnosis, Pathologic Status, Tissue Anatomic Site	Minimal Clinical Set
Applied Biosystems Sequence Data	High-throughput Sanger/di-deoxy technology sequencing	Sequence Data

Table 1.5 Platform types used to provide TCGA data

Chapter 2

Selecting Data Sets

This chapter provides instructions for selecting specific data sets for your study. Refer to the *TCGA Data Primer* for further details about samples, centers, platform types, levels, and data availability. It is available for download at: http://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/other/TCGA_Data_Primer.pdf.zip

Topics in this Chapter

- *Main Steps For Selecting Data Sets* on this page
- *Data Filtering Techniques* on page 16
- *Data Selection Techniques* on page 25

Main Steps For Selecting Data Sets

Table 2.1 provides the main steps to follow to select your data sets.

Step	Action
1	Filter the data to narrow your query. This step is optional. See *Data Filtering Techniques* on page 16.

Table 2.1 Main steps for selecting data sets

Step	Action
2	Select data sets of interest by activating any combination of the following buttons as appropriate: • Batch • Sample • Platform • Center/Platform • Data level • Patient sample • Data set (individual cell) See *Data Selection Techniques* on page 25.
3	Review the data set you created. You can expand or limit your selection as necessary.
4	Click the **Build Archive** button to create a data archive.
5	Retrieve and download your data.

Table 2.1 Main steps for selecting data sets (Continued)

Tip: Click the **Reset** button at any time to clear all cells and start your selection over again.

Data Filtering Techniques

The Matrix filter feature enables you to modify the visual display of data sets so that you can narrow your search for, and selection of, data of interest to you.

You can filter TCGA data sets by the following criteria:

- Platform type
- Batch number
- Data level
- Availability
- Center/platform
- Sample ID
- Status (controlled or open-access data)
- Tumor/normal
- Data submission date

For instructions, see *Modifying the Display of Data* on page 17.

Additionally, you can color code matched and unmatched tumor/normal samples. For instructions, see *Color-Coding Data Sets* on page 24

Other features of the Filter Settings window enable you to upload existing data files, and to enter your own data manually.

Modifying the Display of Data

The Matrix provides sets of criteria by which you can restrict the display of data. For example, you can limit the display so that only data sets for SNPs from Batch 1 appear, and all other data sets are hidden.

To filter data, follow these steps:

1. On the **Options** menu, click the **Set Filter** button (Set Filter).

   The Filter Settings window appears.

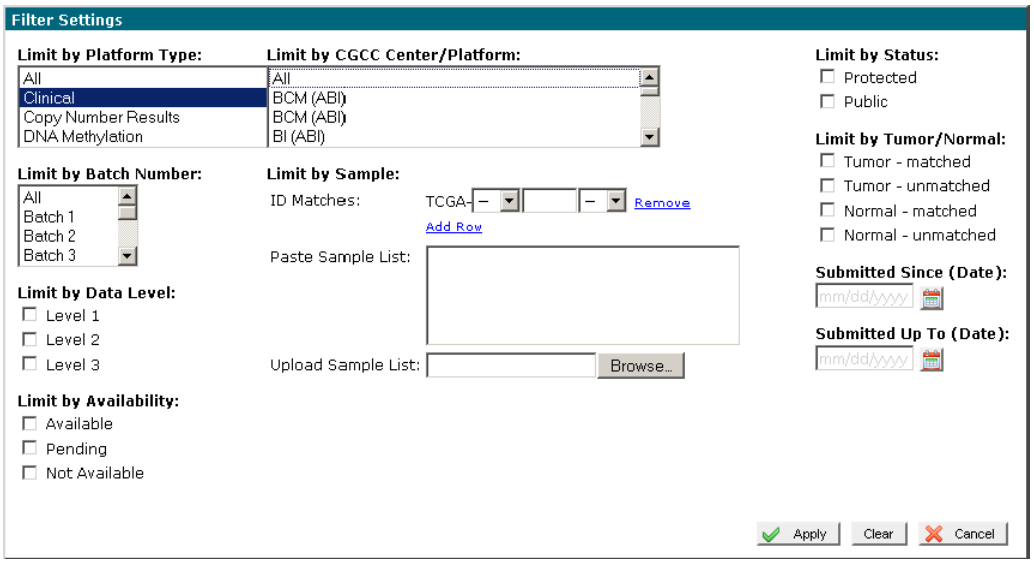

Figure 2.1 Filter Settings Window

2. In the **Filter Settings** window, select the criteria for the Matrix display. The criteria for restricting the display of data are listed in *Table 2.2*.

Limit by...	***Description or Function***
Platform Type	Select one or more platform names to display only the types of platforms of interest. Available platform types are as follows: • **All** – Selects all platform types; clears any previously selected cells • **Copy Number Results** • **Clinical** • **DNA methylation** • **Expression-Gene** • **Expression-Exon** • **Expression-miRNA** • **SNP** • **Trace- Sample Relationship** You can select multiple platform types of pressing and holding the CTRL key while clicking subsequent types.

Table 2.2 Description of filter criteria

Limit by...	***Description or Function***
Batch Number	Select one or more batch numbers to display only the batch of samples of interest. • **All** – Selects all batches; clears any previously selected batches
Data Level	Select the check box beside one or more data levels to display only the data levels of interest. Available data levels are 1, 2, and 3.
Availability	Select the check box beside one or more states of data to display only the states of interest. Valid states are as follows: **Available** – Indicates that a sample has been processed or published and submitted to the DCC **Pending** – Indicates that a sample has been submitted to the DCC, but the center has not yet processed or published the associated data **Not Available** – Indicates that data set is not available for a given platform type, center, and level
Center/Platform	Select one or more center names to display only the types of data from a center and platform of interest. Centers that submitted TCGA data are as follows: • **All** – Selects all center/platform types; clears any previously selected cells • **BI** – Broad Institute of MIT and Harvard • **HMS** – Harvard Medical School) • **JHU USC** – Johns Hopkins University and University of Southern California joint group • **LBL** – Lawrence Berkeley National Laboratory • **MSKCC** – Memorial Sloan-Kettering Cancer Center • **UNC** – University of North Carolina, Lineberger Comprehensive Cancer Center • **SUSM** – Stanford University School of Medicine
Sample	Limits the display to one or more components of the sample barcode: sites, patients, and/or sample types. For instructions on limiting data by sample, see *Limiting Data by Sample* on page 19.
Status	Select the check box beside one of the states to display only the types of data of interest. Available states are as follows: • **Protected** – Controlled-access data • **Public** – Open-access data **Note:** To take advantage of controlled-access data in the Cancer Genome Atlas (TCGA) you must provide a username and password. To request access, follow the instructions at http://cancergenome.nih.gov/dataportal/data/access/closed/dar/

Table 2.2 Description of filter criteria (Continued)

Limit by...	*Description or Function*
Tumor/Normal	Select the check box beside one or more tissue samples to display only the matched and unmatched tumor and normal data sets of interest. Available data sets are as follows: • Tumor matched • Tumor unmatched • Normal matched • Normal unmatched For definitions, see "Legend" in *Table 1.2* on page 7.
Submitted Since (date)	Type a date (***mm/dd/yyyy***) or click the **Calendar** icon ( ) to display only TCGA data that was submitted on or after a given date.
Submitted Up To (date)	Type a date (***mm/dd/yyyy***) or click the **Calendar** icon ( ) to display only TCGA data that was submitted before a given date.

Table 2.2 Description of filter criteria (Continued)

3. To display only the data that matches the criteria you selected, click **Apply**.

Limiting Data by Sample

The Filter Settings window provides a set of criteria by which you can restrict the display of data by selecting the following components of the sample barcode: project name, site ID, patient ID, and sample type ID. See *Filtering Data by Sample Barcodes* on page 22.

Note: Currently TCGA (The Cancer Genome Atlas) is the only project available.

To limit the data by sample, follow these steps:

1. On the **Options** menu, click the **Set Filter** button (Set Filter).

   The Filter Settings window appears, which includes all the criteria for limiting data by sample.

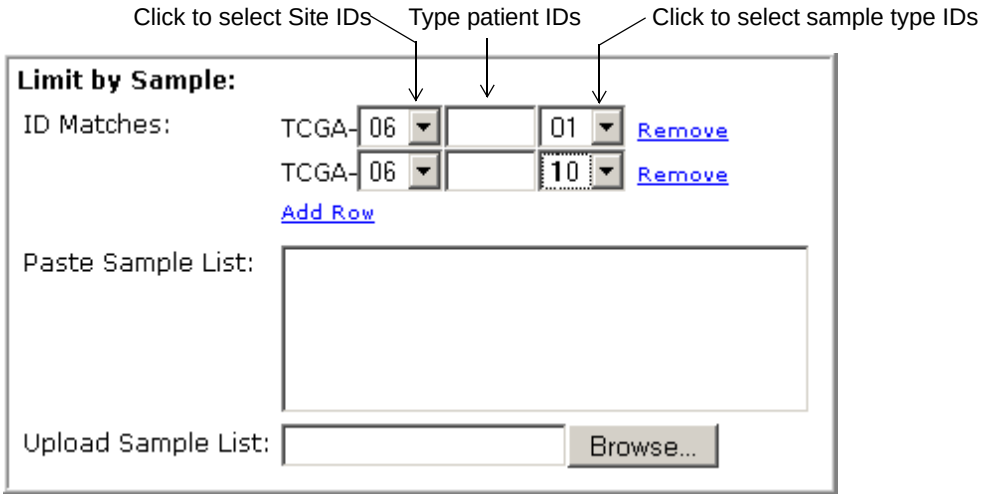

Figure 2.2 Filter Settings Window – Filter by Sample Criteria

2. To filter by site, select a valid site ID from the drop-down list on the left.
3. To filter by patient ID, type a valid 4-digit patient ID in the patient ID box.
4. To filter by sample type, select a valid sample type ID from the drop-down list on the right.

   Note: It is acceptable to leave out the sample type ID. This makes it possible to search for patients without adding a wildcard. For example either TCGA-02-0001-* or TCGA-02-0001 will find all samples matching that patient.

5. To add another set of IDs, click **Add Row**, or, to remove a set of IDs, click **Remove**.

   Another row of sample criteria is added.

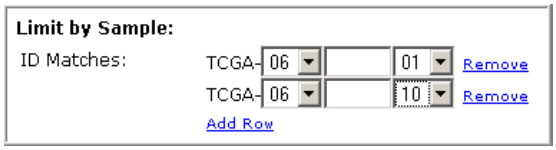

Figure 2.3 Limit by Sample

   Note: Blank fields are equivalent to wildcards (*), and the addition of rows is equivalent to adding a logical OR between the rows. Another way to express the filter in *Figure 2.3, Limit by Sample* above is TCGA-06-*-01 OR TCGA-06-*-10. See *Filtering Data by Sample Barcodes* on page 22.

6. To use a pre-defined list of sample IDs, type or paste the samples of interest in the **Paste Sample List**.
7. To use all pre-defined samples in a file, click **Browse**, and then navigate to the file.

Techniques For Filtering Data

Figure 2.10 provides a key to the filter selection diagrams in this chapter. The main components of the key are as follows:

- Default display – Section of the matrix as it is displayed without filters
- Filters – One or more sets of criteria by which you can limit the data displayed
  - Gray filters – Selected filters
  - White filters – Available filters, not selected
  - Yellow arrows – Flow of data through a filter
- Filtered display – Section of the matrix as it is displayed after applying filters

For detailed information about the display, see *Figure 1.2, Matrix Features,* on page 7.

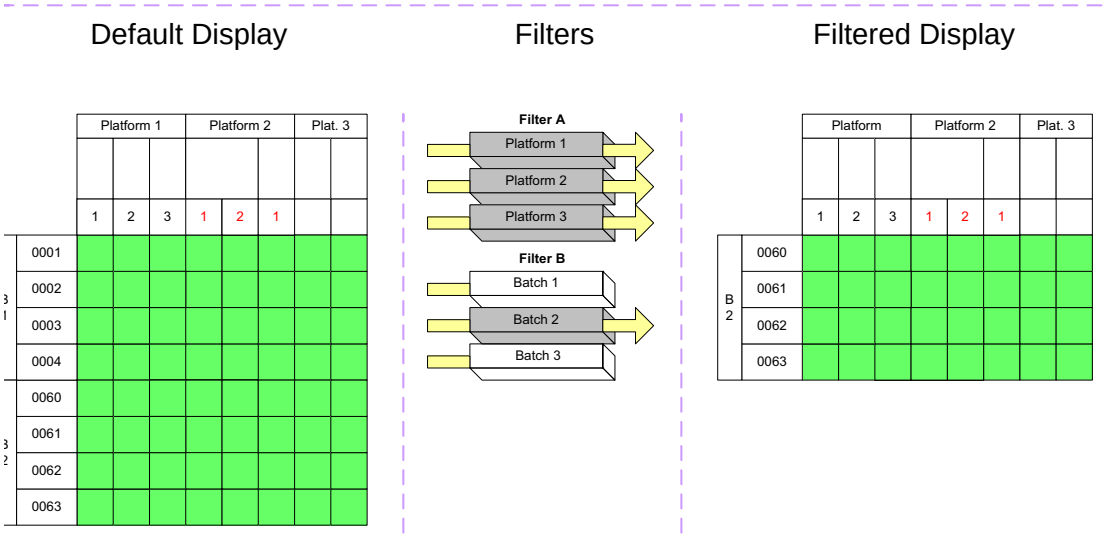

Figure 2.4 Key to Data Access Matrix Filter Diagrams

Figure 2.5 provides two examples of limiting the display of data as follows:

- Example A displays the following samples:
  - Platforms: Exp-Gene, SNP, Clinical
  - Batch 2
  - Data levels: 1, 2, 3
- Example B displays the following samples:
  - Platforms: Exp-Gene, SNP
  - Batch 2
  - Data levels: 1, 2

Each additional filter applied to the Matrix limits the display further.

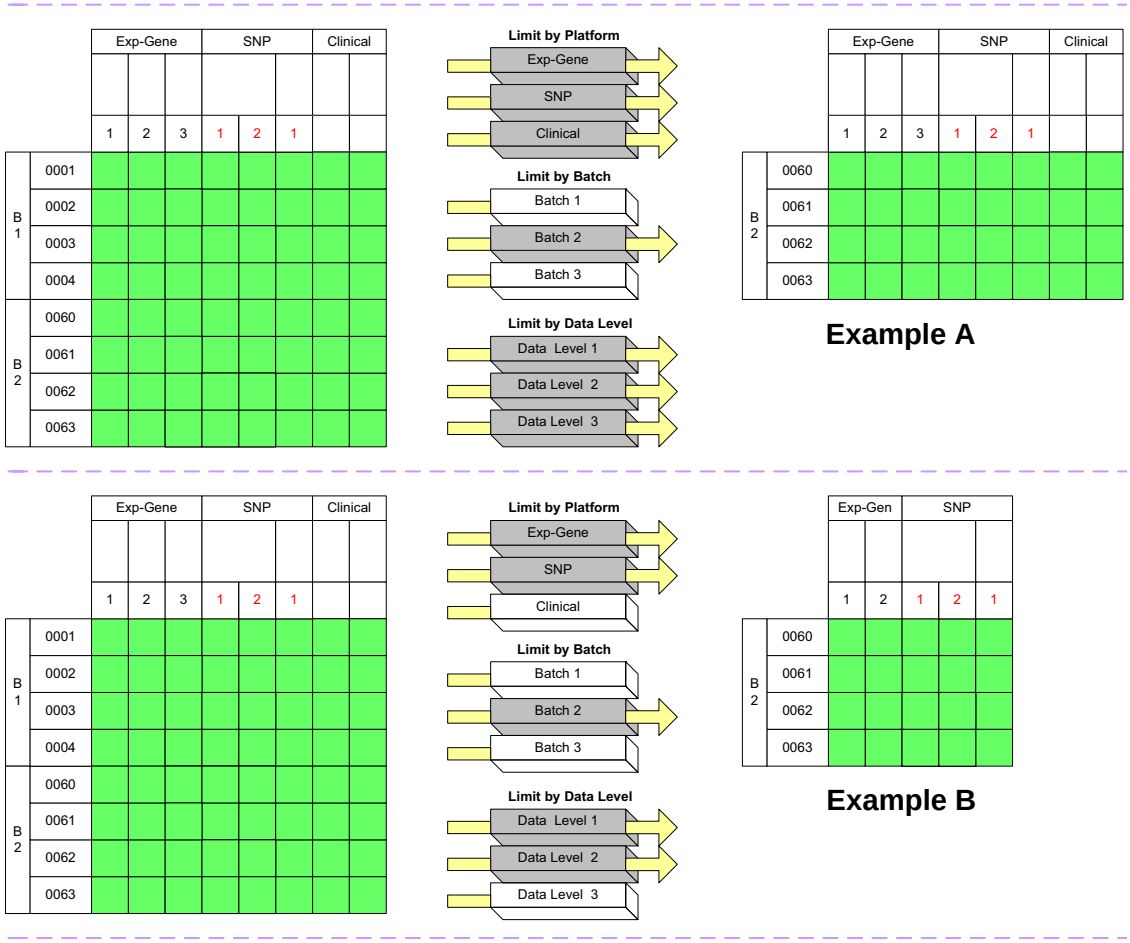

Figure 2.5 Filtering by Platforms, Batches, and Data Levels

Filtering Data by Sample Barcodes

The Biospecimen Core Repository collects tissue samples and clinical metadata from TCGA sites, aliquots them, codes them, and then sends them to the Cancer Genome Curation Centers (CGCCs) and Genome Sequencing Centers (GSCs) specified in the aliquot barcode. Matrix samples are coded in the format `TCGA-xx-yyyy-zz`

where,

`TCGA` is The Cancer Genome Atlas (project name)

`xx` is the site ID (See *Site ID Values* on page 24.)

`yyyy` is the patient ID (See *Patient ID Values* on page 24.)

`zz` is the sample type ID (See *Sample Type Values* on page 24.)

Figure 2.6 provides two examples of limiting the display of data to samples from the MD Anderson Cancer Center – Brain Bank as follows:

- Example A displays the following samples:
  - Sites: MD Anderson Cancer Center – Brain Bank (Site ID 02)

- Patients: All (The default Patient ID is a blank box, indicating that there are no limitations on patient IDs.)
- Sample type: Solid tumor (Sample Type ID 01)

- Example B displays the following samples:
  - Sites: MD Anderson Cancer Center – Brain Bank (Site ID 02)
  - Patients: All (The default Patient ID is a blank box, indicating that there are no limitations on patient IDs.)
  - Sample types: Solid tumor (sample type ID 01) *and* normal blood (sample type ID 10)

Each additional row of ID filters applied to the Matrix displays more data.

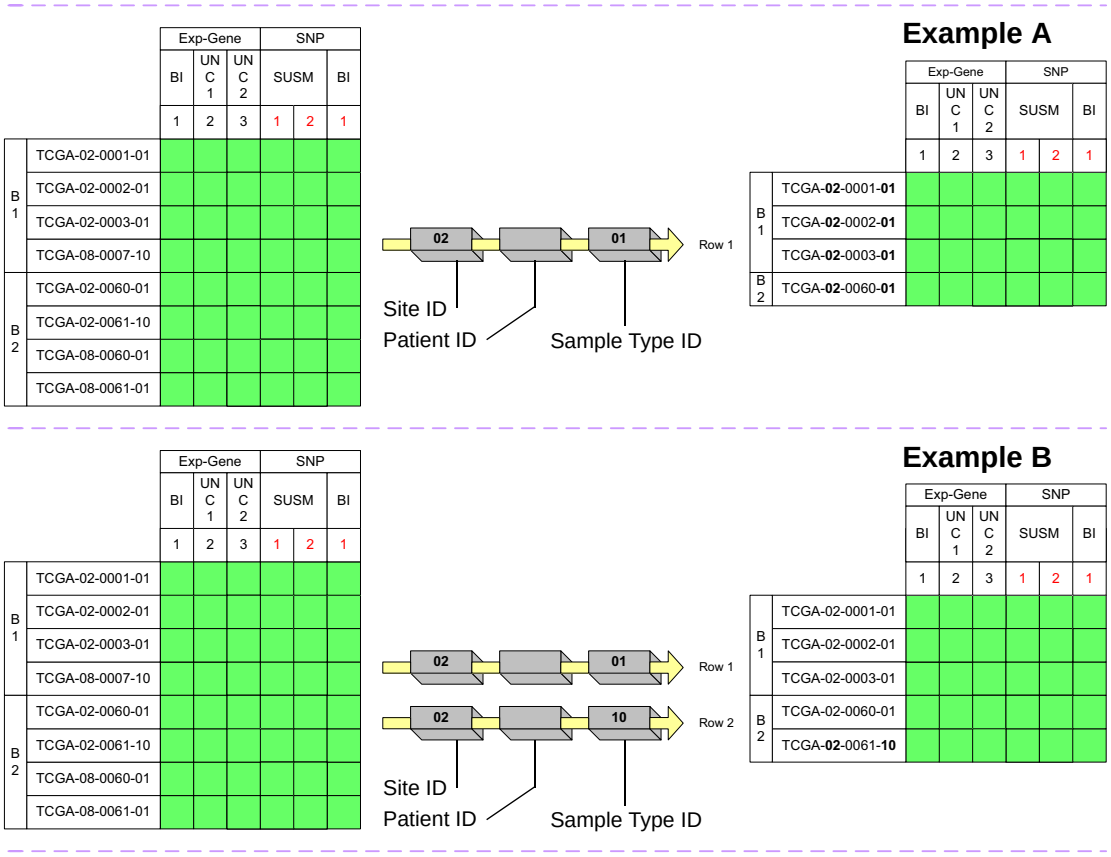

Figure 2.6 Filtering by Sample Barcodes

To produce the same samples as displayed in Example B, follow these steps:

1. Remove all Sample ID filter rows.
2. In the **Paste Sample List** text box, type ***TCGA-02-*-01, TCGA-02-*-10***.

   Note: The asterisk symbol (*) represents a wild card, which returns all values. Use a comma (,) to separate samples. Adding a comma is like adding another row of ID filters.

Site ID Values

Sites (collection sites), for example, the University of Texas MD Anderson Cancer Center, send tissue samples and clinical metadata to the Biospecimen Core Resource (BCR).

Table 2.3 provides values of all current TCGA collection sites.

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

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 Type Values

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

Table 2.4 provides values for sample types

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

Table 2.4 Sample type values

Color-Coding Data Sets

You can color-code data sets by data availability and matched and unmatched tumor/normal samples.

To color-code data sets, on the **Options** menu, next to **Color Cells By**, select a color scheme from the drop-down list, either **Availability** or **Tumor/Normal**.

Figure 2.7 illustrates the Options menu, with the color display in it's default state, **Color Cells By Availablity**. Other Options menu features are described in *Options Menu Features* on page 10.

Figure 2.7 Options Menu – Color-Coding

Figure 2.8 is an example of the Data Access Matrix as it appears when colored by Tumor/Normal.

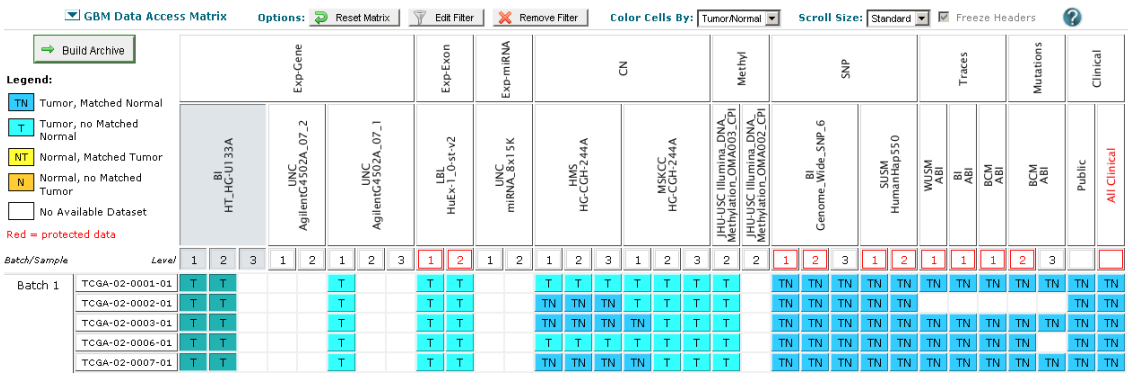

Figure 2.8 Data Access Matrix – Color-Coded by Tumor/Normal

The colors are defined in the legend on the left side of the Web page, and in *Figure 2.9*.

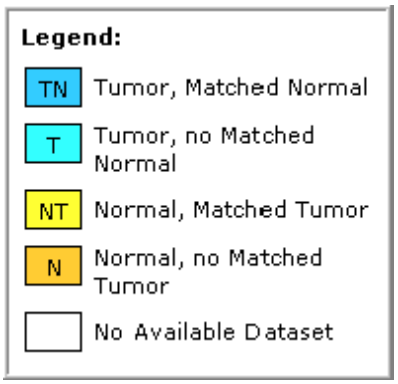

Figure 2.9 Data Access Matrix – Color-code Legend For Tumor/Normal

Data Selection Techniques

You can select data sets by cross-selecting a combination of platform type, center/platform, data level, batches, and patient samples. The result of your selections represents a subset of TCGA data files.

The Data Access Matrix enables you to create data sets for each sample or batch of samples as follows:

- Individual – Individual cell in the Matrix. One sample from one data level from a given center/platform for a single platform type.
- Sample – Row of cells in the Matrix. One patient sample from all data levels, all centers platforms, and all platform types.
- Batch – Set of biomaterials (patient samples) that the Biospecimen Core Repository (BCR) aliquots and sends out to all CGCCs and GCGs for analysis en masse. For further information about the batch process, refer to *TCGA Data Primer* at: http://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/other/TCGA_Data_Primer.pdf.zip
- Data level – Column of cells in the Matrix. All samples from a given data level.
- Center/Platform – Multiple (generally) columns on the Matrix. All samples from all data levels from a given center-platform pair.
- Data-type level data set – Multiple columns on the Matrix. All samples from all data levels, all platforms, and all centers for a single data type.

You can create more custom data sets by performing the following functions:

- Joining two or more data sets (*Figure 2.11*)
- Intersecting data sets (*Figure 2.12*)
- Subtracting data sets (*Figure 2.13*)

Note: You can create custom data sets more efficiently by using the Filter options. See *Data Filtering Techniques* on page 16.

Figure 2.10 provides a key to the data selection diagrams in this chapter. Each grid represents a section of the matrix. For detailed information about the matrix, see *Figure 1.2, Matrix Features,* on page 7.

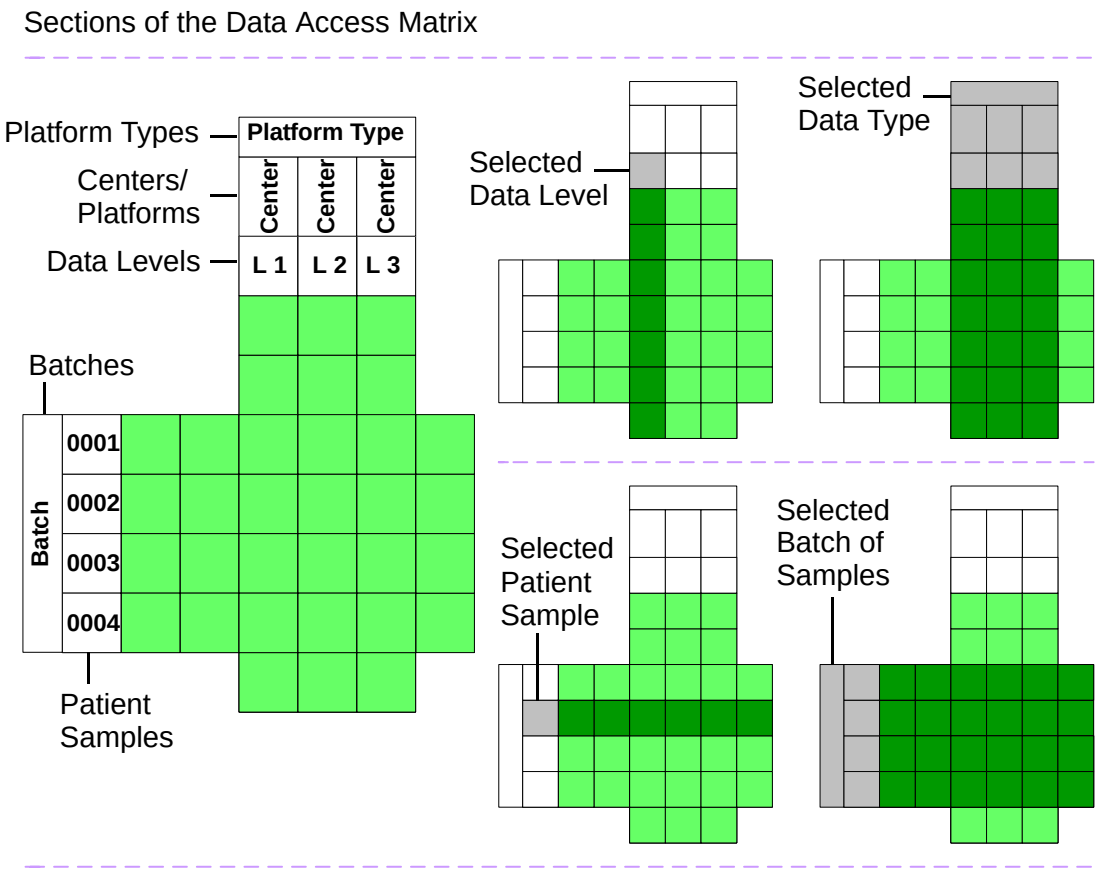

Figure 2.10 Key to Data Access Matrix Data Selection Diagrams

Creating a Union of Data Sets

You can join two or more data sets to create a group of all samples that are in a column, a row, or both. This is equivalent to an “OR” operation. See *Selecting a Union of Data Sets* on page 47.

Figure 2.11 provides two examples of a union of samples. The results of the union create a data set as follows:

- Data set 1 – All data from a given data type OR all data from a given batch, OR both, indicated by # 2 in Example A.
- Data set 2 – All data from a given data level OR all data in a given sample, OR both, indicated by # 2 in Example B.

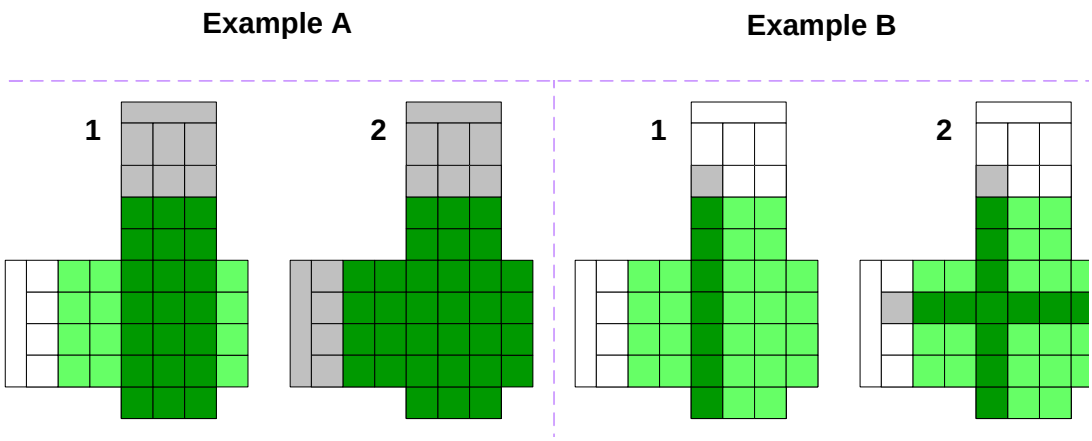

Figure 2.11 Union of Data Sets

To create a union of data, follow these steps:

1. Select a column (data level) or columns (platform/center) of sample data.
2. Click a row (sample) or rows (batches of samples).

Intersecting Data Sets

You can intersect data sets to create a data set that contains only samples that are common to both a column(s) and a row(s). This is equivalent to an "AND" operation.

Note: You can create an intersection of data sets more efficiently by using the Filter options. See *Data Filtering Techniques* on page 16.

See *Selecting All Data From a Single Center In a Single Batch* on page 49.

Figure 2.12 provides two examples of intersecting samples to create a data set as follows:

- Data set 1 – All data that is common to a platform type and a sample, indicated by # 2 in Example A.
- Data set 2 – All data that is common to a platform type and a batch of samples, indicated by # 2 in Example B.

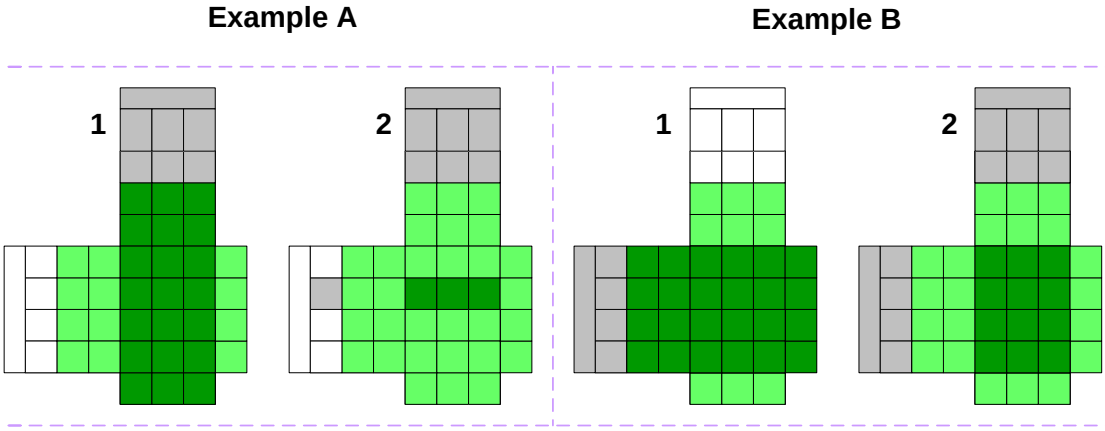

Figure 2.12 Intersecting Data Sets

To intersect data, follow these steps:

1. Select a column (data level) or columns (center/platform/data types) of sample data.
2. Press and hold the SHIFT key while you click a row (sample) or rows (batch of samples).

Subtracting Data Sets

You can subtract data sets to create a data set that excludes data from a selection. This is equivalent to a “NOT” operation. See *Selecting Data From a Platform Type* on page 45.

Figure 2.13 provides three examples of intersecting samples to create a data set as follows:

- Data set 1 – All samples from all data levels from a platform type except level 2 data, indicated by # 2 in Example A.
- Data set 2 – All samples in a batch except one patient sample, indicated by # 2 in Example B.
- Data set 3 – All samples in all data levels from a platform type except level 2 data, for all samples in a batch except one sample, indicated by # 3 in Example C.

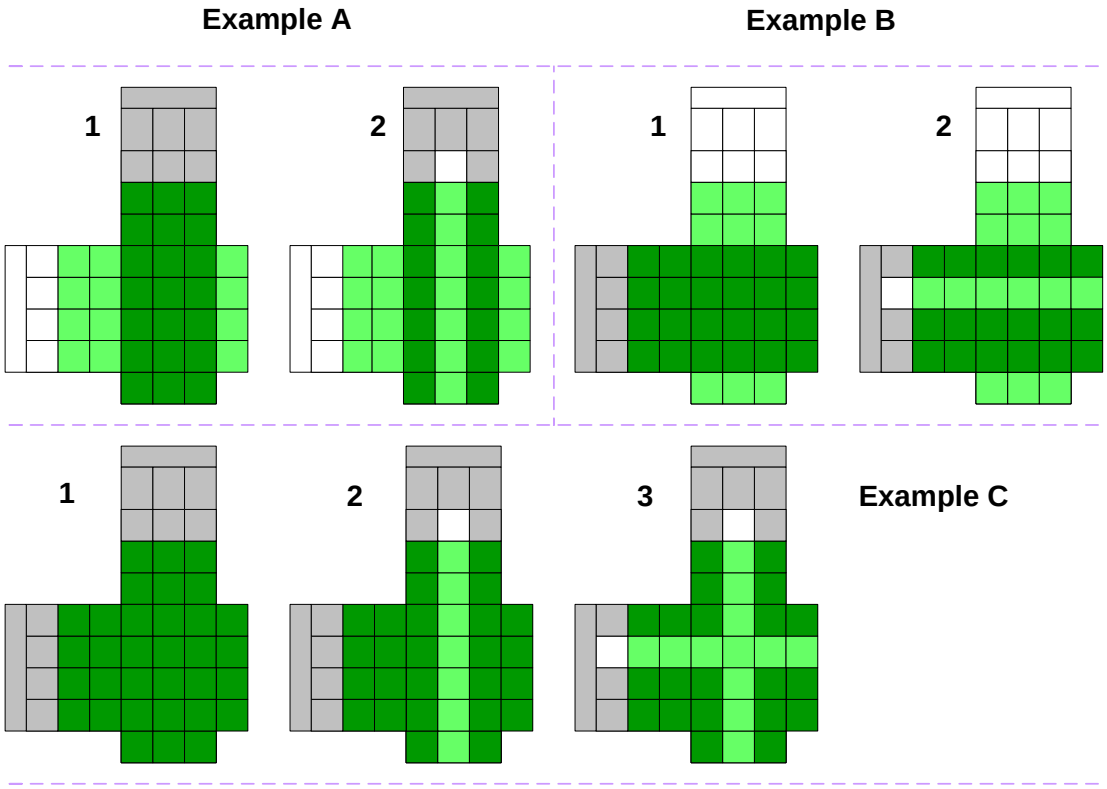

Figure 2.13 Subtracting Data Sets

To subtract data, follow these steps:

1. Select a column (data level) or columns (center/platform/data types) of sample data.
2. Click a row (sample) or rows (batch of samples).

Chapter 3

Downloading and Retrieving TCGA Data

This chapter provides instructions for selecting, downloading, and retrieving the files that contain your selected data.

Topics in this Chapter

- *Navigating Through the Data Access Download Page* on this page
- *Main Steps For Downloading Your Data Files* on page 35
- *Selecting All Data Files For Download* on page 37
- *Selecting a Subset of Data Files For Download* on page 37
- *Retrieving Your Data Archive* on page 39

Navigating Through the Data Access Download Page

The Data Access Matrix download page provides the means to review and select the files that contain the data sets you chose before you download archives. It appears automatically when you click the **Build Archive** button on the Data Access Matrix page.

See also *Main Steps For Selecting Data Sets* on page 15.

Understanding Data Access Download Components

Figure 3.1 illustrates components of the Data Access Download page. The elements are described in *Table 3.1* on page 32. See *Figure 3.5* on page 36 for descriptions of elements not identified here.

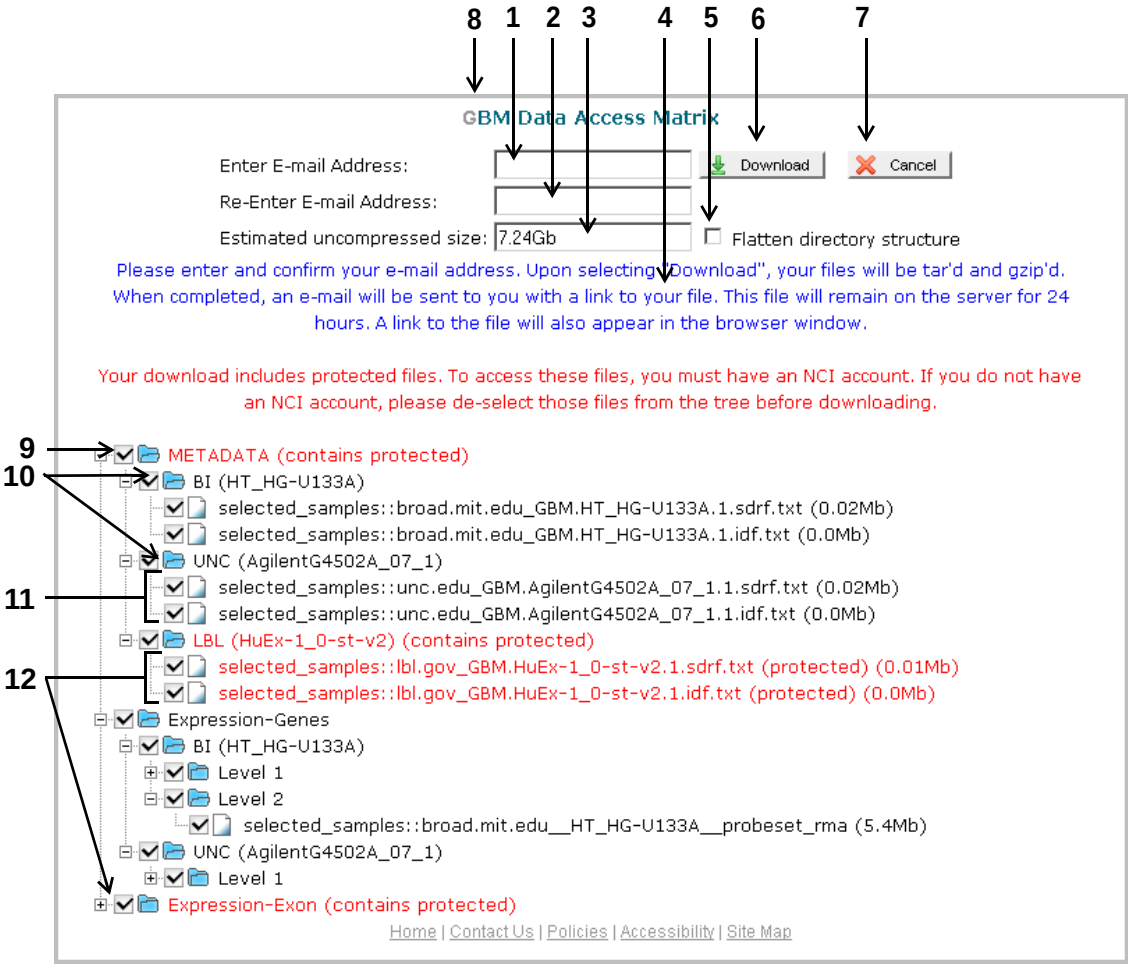

Figure 3.1 Data Access Download Page Components

Table 3.1 Describes each numbered component in *Figure 3.1*.

Component Number	***Description or Function***
1	E-mail address text box
2	E-mail address confirmation text box
3	Estimated size, in gigabytes (GB) of the data files selected.
4	Brief description of file processing. "tar'd" and "gzip'd" are archive-packaging and compression functions respectively.
5	Flatten Directory Structure check box – Allows you to choose the file structure to download as follows: • To store and download all files in a single directory, select the **Flatten Directory Structure** check box. • To maintain the directory structure of the file tree as it is displayed, clear the **Flatten Directory Structure** check box.

Table 3.1 Description of Data Access Download Page Components

Component Number	Description or Function
6	Download button – Starts the file processing and downloading stages. The final archive contains all the selected files. **Note:** This button will not be available to you if the size of your files exceeds 50 GB, and a warning will be displayed.
7	Cancel button – Cancels the data file process and displays the Data Access Matrix page.
8	Page title – Displays the targeted tumor type abbreviation. Currently data for the Glioblastoma Multiforme (GBM) is available.
9	Metadata – Folder that contains the following files: • Manifest – Lists the files included in the final archive • Investigation Description File (IDF) – 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. This file is available for CGCC data only. • Sample Data Relationship File (SDRF) – Tab-delimited file that describes the relationships between samples, array, data, and other objects used or produced in the experiment. This file is available for CGCC data only.
10	Nodes (subdirectories) of the file tree
11	Data files – Identified by the sample name followed by the file name and an estimated file size. The file name is the original file name as submitted by a center.
12	Protected data – To obtain access to protected TCGA data sets, please follow the instructions at http://cancergenome.nih.gov/data/access/closed/.

Table 3.1 Description of Data Access Download Page Components (Continued)

Understanding Directory Structure

The Matrix displays data files on the Download page according to the combination of data set selections you made. Each node, or folder, in the tree represents one of the platforms, centers, data levels, and/or batches, depending on the data sets you selected. You can expand and collapse these nodes.The file tree organizes the files by platform type, center/platform, and level. *Figure 3.2* Illustrates a typical tree-type directory structure with each node expanded.

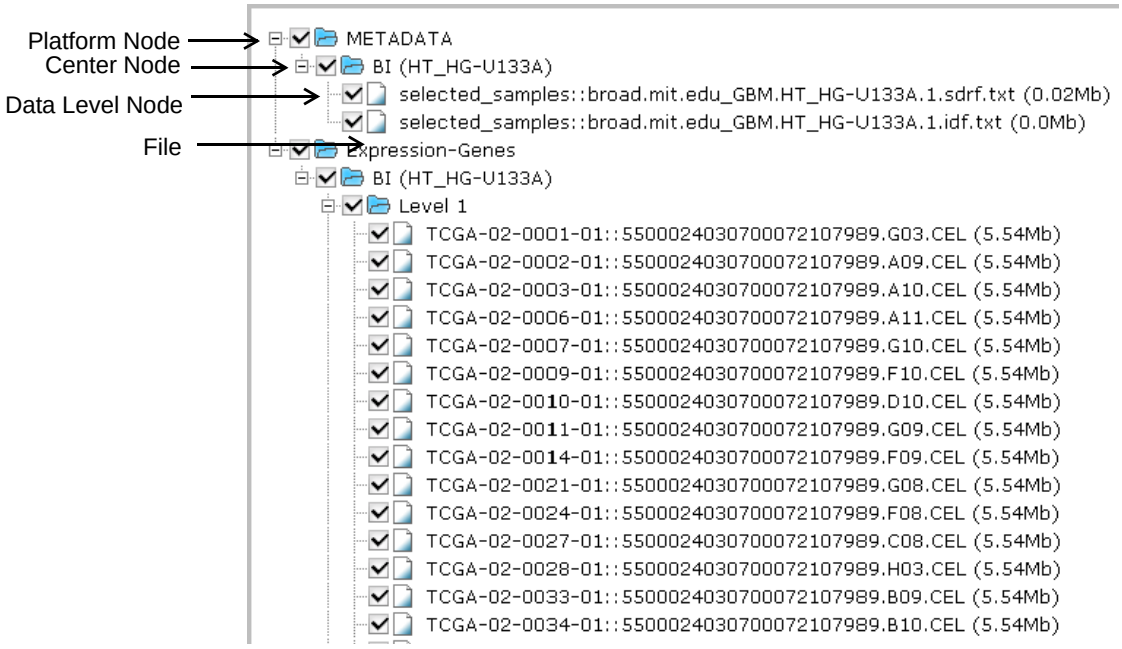

Figure 3.2 Simple File Tree – All Nodes Expanded

Figure 3.3 Illustrates the same tree as in *Figure 3.2*, but with the data level nodes collapsed. Although the files themselves are not visible, they remain selected for download. A check mark in a white check box beside a node indicates that all of its children are selected. See *Figure 3.4* on page 35 for more about parent/child relationships.

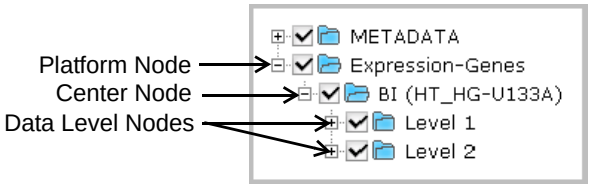

Figure 3.3 Simple File Tree – Collapsed Nodes

Figure 3.4 illustrates the same files and structure as *Figure 3.3* and *Figure 3.2*, but with several of the individual files removed from the selection. Check marks in gray check boxes, "semi-selected" boxes, indicate that one or more children have been deselected from the parent node. This visual cue is helpful in ascertaining file selection when you have collapsed all or part of the tree.

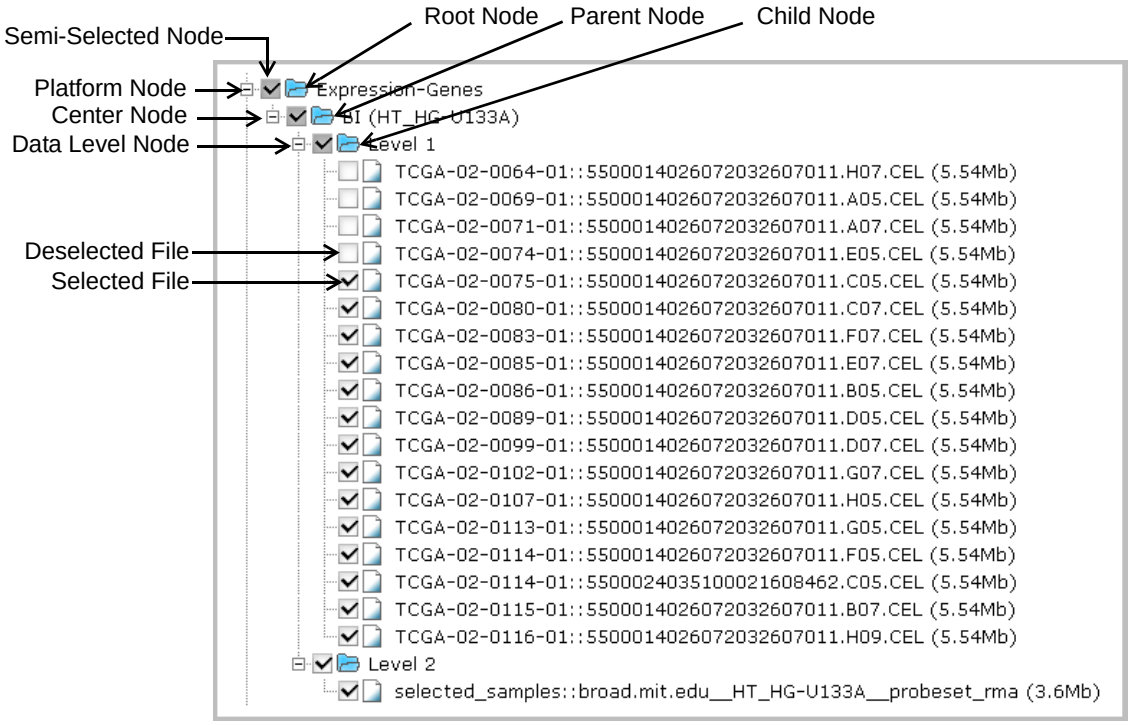

Figure 3.4 Simple File Tree – Subset of Files

Main Steps For Downloading Your Data Files

The Download page is displayed after you have selected your data sets and clicked the Build Batch button. For data set selection instructions, see *Chapter 2, Selecting Data Sets,* on page 15.

Figure 3.5 illustrates a typical set of data files on the Download page. All the files in all the directories are selected for download by default.

If the estimated uncompressed size of all the data files you selected exceeds the 22 GB limit, the **Download** button is disabled, and the page displays a warning, as illustrated in *Figure 3.5*. To reduce the total file size you can start over and limit your data set selection, or you can select a subset of the data files displayed on the Download page.

To limit your data set selection, click **Cancel**, and then follow the steps in *Data Selection Techniques* on page 25.

To select a subset of files, follow the instructions in *Selecting a Subset of Data Files For Download* on page 37.

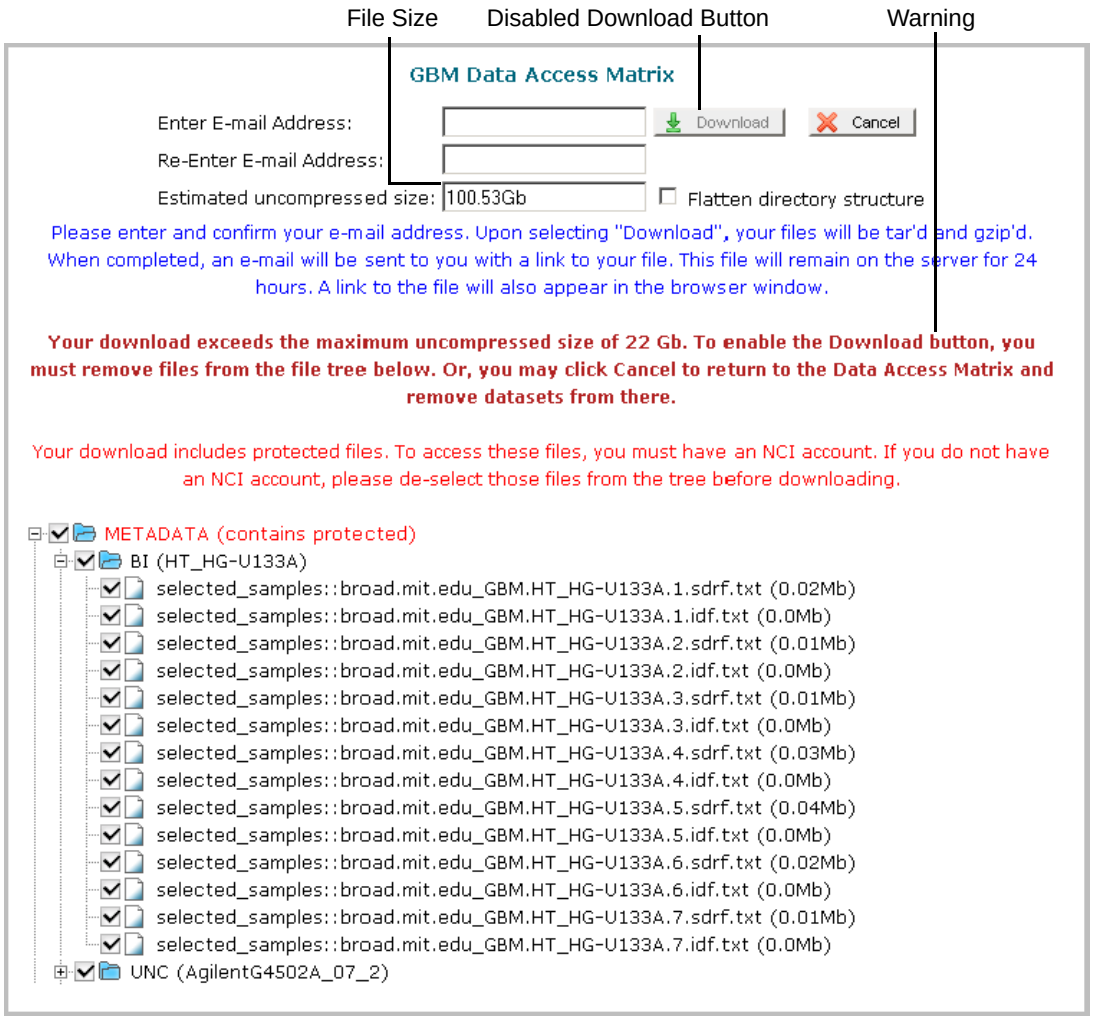

Figure 3.5 Download Page – Warning Message

Table 3.2 provides the main steps to follow to download your data files.

Step	Action
1	Enter and confirm your e-mail address so that the Matrix can alert you when the files have been processed.
2	Optionally, select the **Flatten directory structure** check box.
3	Optionally, select a subset of the data files you want to process.
4	Click **Download**.
5	Retrieve the data archive.

Table 3.2 Main Steps For Processing Data Files

Note: The Flatten directory structure option creates an archive of the data in a single directory, rather than maintaining the directories as displayed. All archives contain a `file_manifest.txt` file, a tab-delimited file that lists the Platform Type, Center, Platform, Level, Sample, and File Name for each data file. You can use this information to match each individual file with the Sample ID.

Selecting All Data Files For Download

By default the Matrix pre-selects all data files corresponding to the data sets you selected.

To download all data files, follow these steps:

1. In the **Enter E-mail address** text box, type your e-mail address.
2. In the **Re- Enter E-mail address** textbox, type the e-mail address again to verify that you have entered your e-mail address correctly.
3. Optionally, select the **Flatten data structure** check box.
4. Click **Download**.

   The Data Access Download – File Processing page appears.

Selecting a Subset of Data Files For Download

You may find that you need only a subset of the data you selected in the Matrix. In such cases you can indicate which files to download by selecting and deselecting any number of nodes in a file tree.

To select a subset of files for download, follow these steps:

1. Follow Steps 1and 2in *Selecting All Data Files For Download* on page 37.
2. Optionally, select the **Flatten DIrectory Structure** check box.
3. To collapse or expand a node, select the (**-**) or (**+**) signs respectively beside the node check boxes.
4. To exclude individual files from the download, select the check boxes beside the files.

   The check boxes beside the excluded files are cleared, and the parent nodes display a semi-selected state, indicated by a check mark in a gray square.

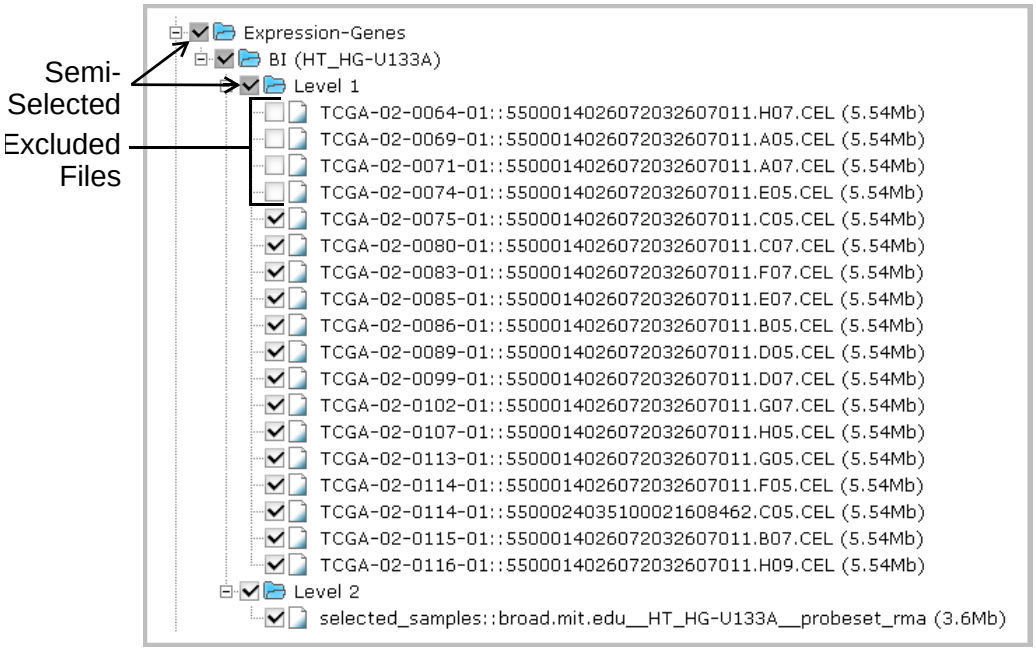

Figure 3.6 Download Page – Subset of Files Selected

5. Repeat the step above for each file you want to exclude from the archive.
6. To exclude all files in a node, select the check box beside the node that contains them.

   The node and the files it contains are excluded from the archive. In *Figure 3.7* all Level 1 files for UNC have been excluded.

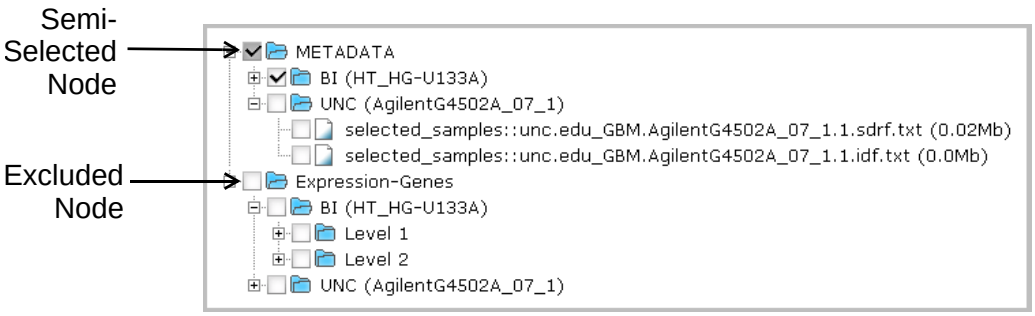

Figure 3.7 Download Page – Subset of Files Selected

7. When you are satisfied with the sub-selection, click **Download**.

   The Data Access Download page appears.

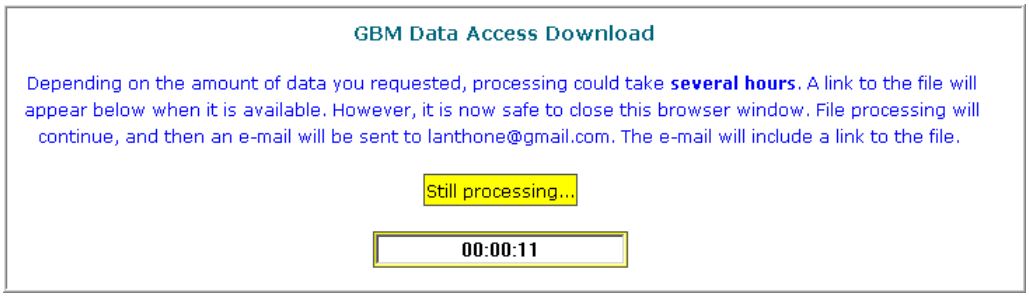

Figure 3.8 File Processing Message

During the processing phase, the Data Access Matrix system "`tars`" and "`gzips`" your files in a single archive, `{archive_name}.tar.gz`. The processing time depends on the amount of data you requested, but 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 TCGA server. Additionally, the system sends a link to the archive in a message to the e-mail address you provided in step 1 on page 37.

Retrieving Your Data Archive

You can retrieve your archive from the TCGA server when the system displays a link to the archive on the File Processing page, or when you receive a link to the TCGA server in an e-mail message.

Caution: Your archive file remains on the TCGA server for 24 hours. After that time the archive is deleted from the server

To retrieve your archive file for download, follow these steps:

1. Click the link to the archive that appears on the processing page, or in the e-mail message you receive from TCGA Data Portal.
2. If your archive contains protected data, continue with these instructions. Otherwise, skip to Step *7.* on page 40.

   The **Authentication Required** dialog box appears.

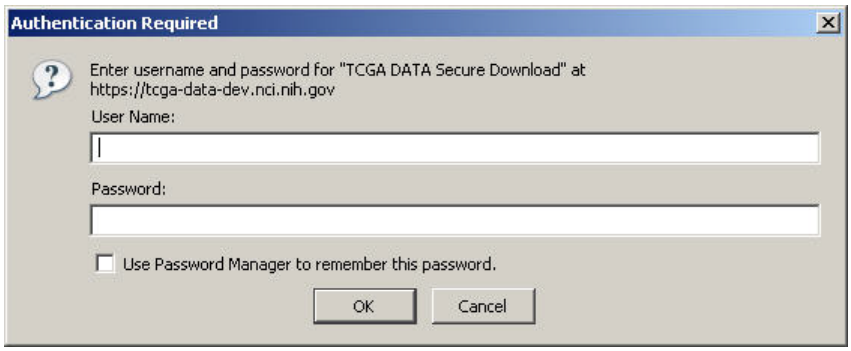

Figure 3.9 Authentication Required Dialog Box

3. In the **User Name** text box, type the username for your TCGA account. (If you do not have a username and password, please follow the instructions for protected data access at http://cancergenome.nih.gov/data/access/closed/.)
4. In the Password text box, type the password for this account.
5. Optionally, select the **Use Password Manager to remember this password** check box.
6. Click **OK**.

   The **File Download** dialog box appears. This dialog box may appear differently from the one here, depending on your browser and your system configuration.

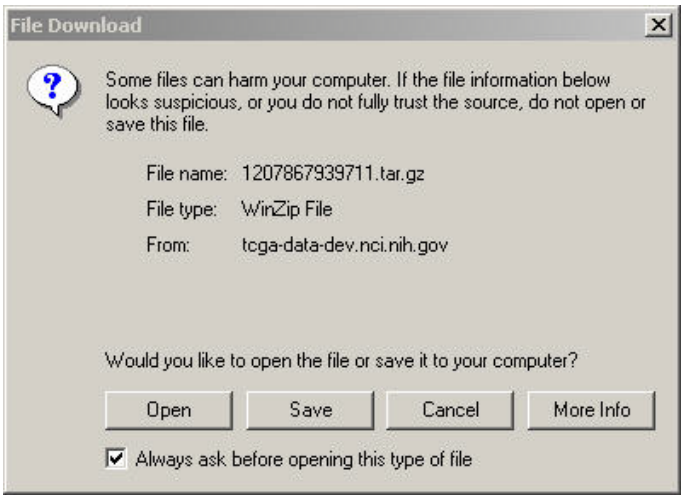

Figure 3.10 File Download Dialog Box

7. Do one of the following to access the compressed archive:
   - If you are running Windows, and your default compression application is WinZip, continue to the next step.

   - **or** -

   - If you are running another operating system, WinZip is not your default compression application, or you prefer to save the archive instead of opening it, skip to step 12 on page 41.
8. Click **Open**.

   WinZip recognizes that the `{archive_name}.tar.gz` contains a `.tar` file and displays an archive prompt.

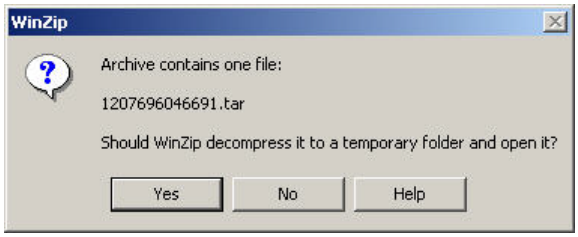

Figure 3.11 WinZip Prompt

9. Click **No**.

   The WinZip window displays the compressed `.tar` file.

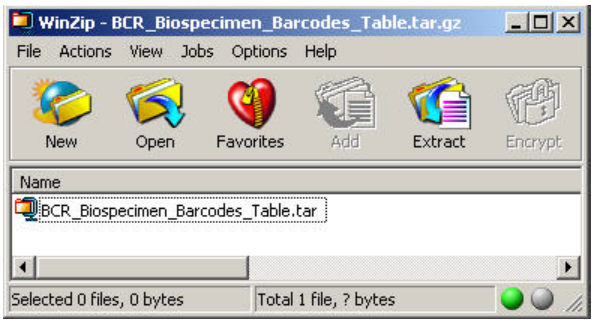

Figure 3.12 WinZip Window

10. Double-click the compressed file icon.

    The contents of the archive appear in the WinZip window.

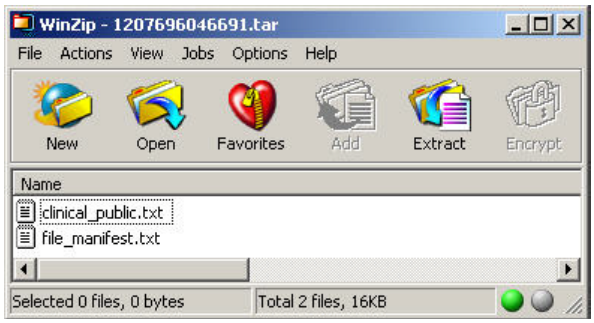

Figure 3.13 WinZip Window – Archive Content

Note: All archives contain a `file_manifest.txt` file, a tab-delimited file that lists the Platform Type, Center, Platform, Level, Sample, and File Name for each data file. You can use this information to match each individual file with the Sample ID.

11. Extract the files to a folder or open the files, according to the WinZip instructions. At this point you have completed the download process.

12. On the File Download dialog box (*Figure 3.10* on page 40), click **Save** and then browse to the location you want to store your archive in.

    The archive is transferred from the TCGA server to the location you indicated.

13. Once the transfer is complete, navigate to the archive on your computer and extract the compressed files according to the instructions provided with your archive application.

APPENDIX

A

DATA SET SELECTION – EXAMPLES

This chapter provides instructions for selecting specific data sets for your study. Refer to the *TCGA Data Primer* for further details about samples, centers, platform types, levels, and data availability. It is available for download at:
http://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/other/TCGA_Data_Primer.pdf.zip

Topics in this Chapter

- *Selecting Specific Data Sets* on page 43
- *Selecting All Data From a Batch* on page 44
- *Selecting Data From a Platform Type* on page 45
- *Selecting a Union of Data Sets* on page 47
- *Selecting All Data From a Specific Center* on page 48
- *Selecting All Data From a Single Center In a Single Batch* on page 49
- *Selecting All Data From a Data Level* on page 50

Selecting Specific Data Sets

To select specific data sets, follow these steps:

1. Click the button for each target data set.

   As illustrated in *Figure A.1*, only the individual cells you selected are activated as a result.

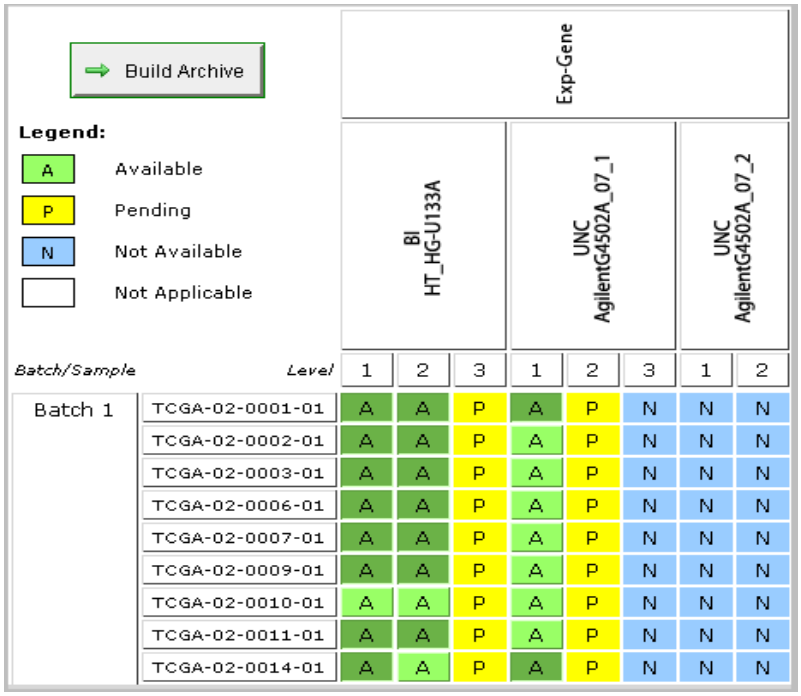

Figure A.1 Selection of Specific Data Sets

2. Click **Build Archive**.

Selecting All Data From a Batch

To select all data from a given batch, follow these steps:

1. Click the **Batch** button of interest.

   As illustrated in *Figure A.2*, the following elements are activated as a result:

   - The **Batch** button
   - All "available" samples in the batch
   - All cells under all column headers

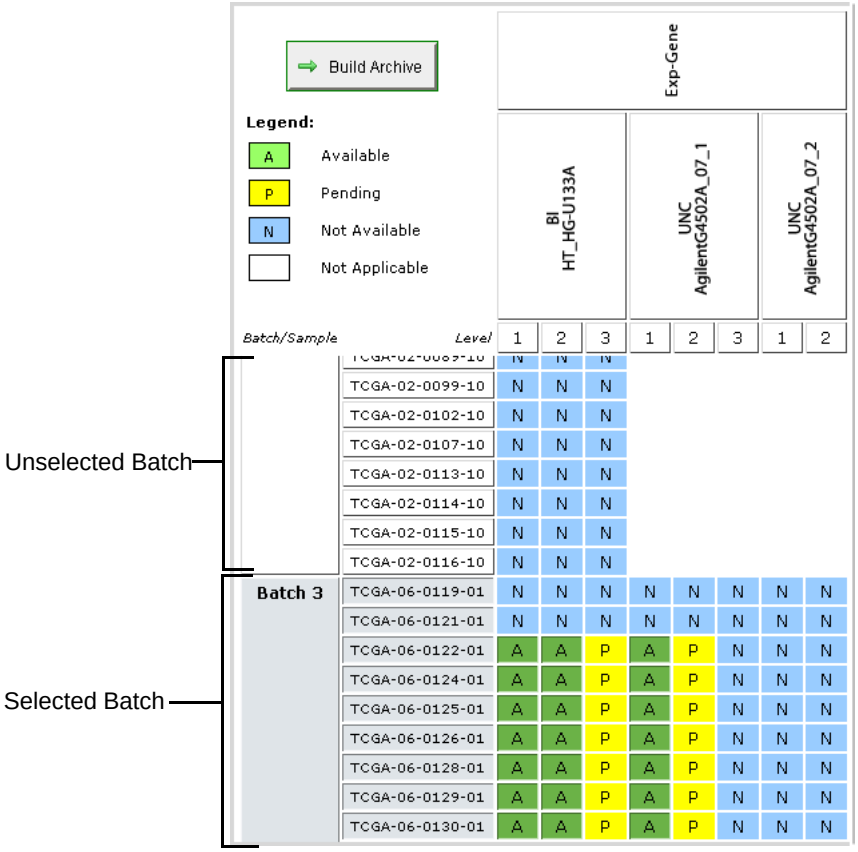

Figure A.2 Selection of All Data From a Batch

2. Click **Show Files for Download**.

Selecting Data From a Platform Type

To select all data from a data type, follow these steps:

1. Click the target **Platform Type** button. For example, click **Exp-Gene**.

   As illustrated in *Figure A.3*, the following elements are activated as a result:

   - **Exp-Gene** button (Data Type button)
   - CGCC buttons for the given data type
   - All data level buttons for each center in the data type
   - All “available” buttons for each sample under these columns of the Matrix

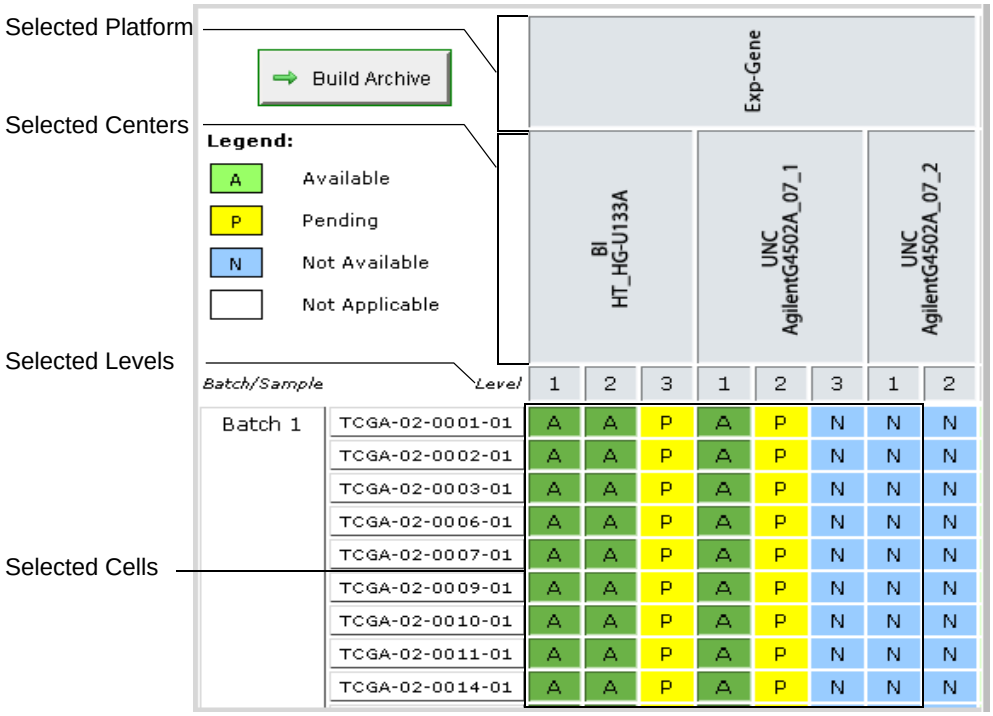

Figure A.3 All Data From a Platform Type and Data Level

2. You can exclude any cells you selected. For example to exclude all samples from the Broad Institute (**BI**), click the **Center** button labeled **BI**.

   As illustrated in *Figure A.4*, the following elements are deactivated as a result:

   - **BI** center button
   - All cells for each sample in the deactivated **BI** center column

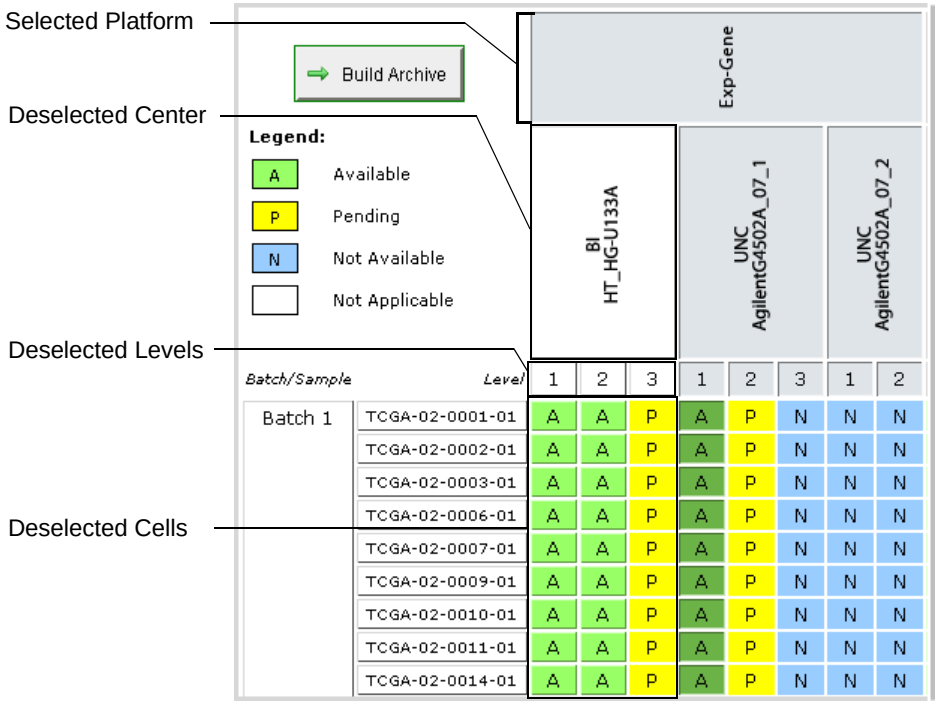

Figure A.4 Deactivated Sample Cells From a Platform Type

3. Click **Build Archive**.

Selecting a Union of Data Sets

You can join, or unite, data sets. For example, *Figure A.5* illustrates the union of all available data sets from a particular batch as well as from all SNP data from the Broad Institute (BI).

To perform a union of data sets, follow these steps:

1. Click the target **Batch** button. (See *Selecting All Data From a Batch* on page 44.).

   Note: Click the **Batch** button before you click the **Platform Type** button.

2. Click the target **Center**/**Platform Type** button. (See *Selecting Data From a Platform Type* on page 45.)

   As illustrated in *Figure A.5*, the following elements are activated as a result:

   - The **Batch** button
   - All “available” samples in the batch
   - The BI SNP **Center**/**Platform** button
   - All data sets under the BI SNP **Center**/**Platform** column header

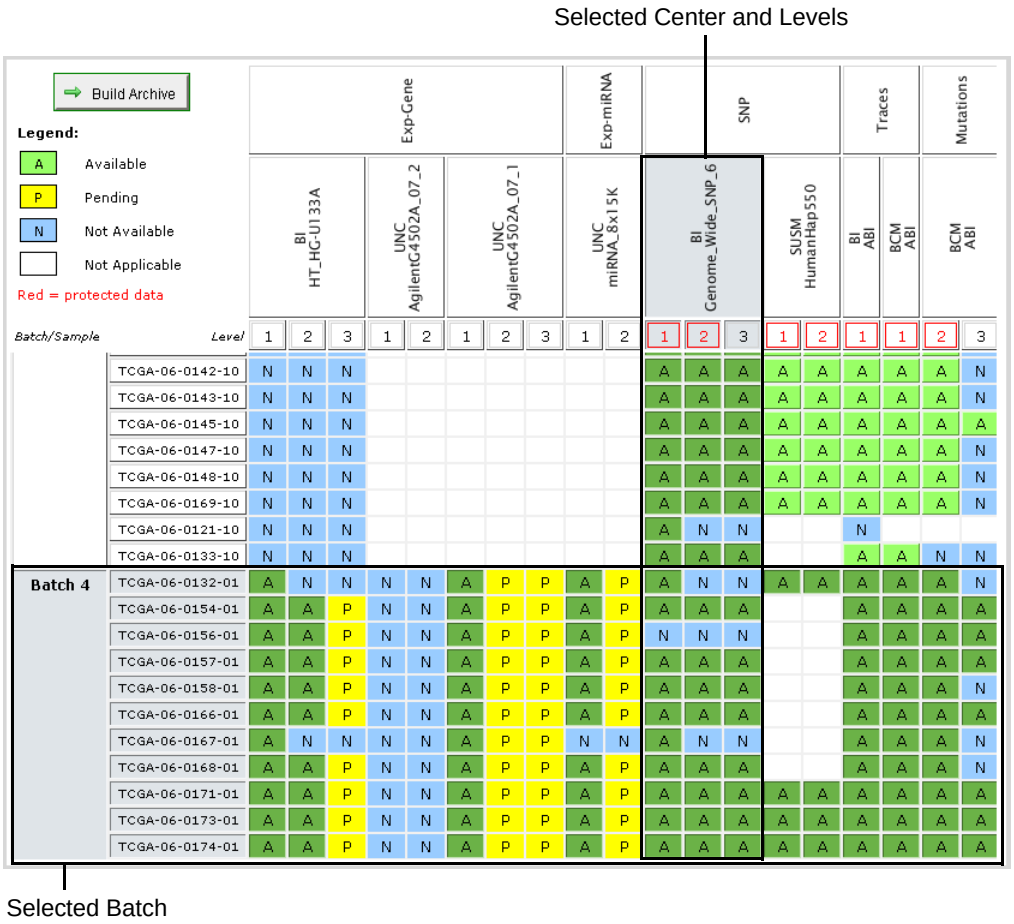

Figure A.5 Data Set Union

3. Click **Build Archive**.

Selecting All Data From a Specific Center

To select all data from a specific center, follow these steps:

1. Click a **Center** button from a given platform, for example the **Exp-Gene** platform for the Broad Institute (**BI**).

   As illustrated in *Figure A.6*, the following elements are activated as a result:

   - All Data Level buttons for BI, where applicable
   - All "available" data sets for each sample in the BI column

2. When applicable, click all remaining occurrences of the specific center's buttons, for example, the **SNP** platform and **Traces** for BI.

   Note: You can click the **Center** buttons in any order.

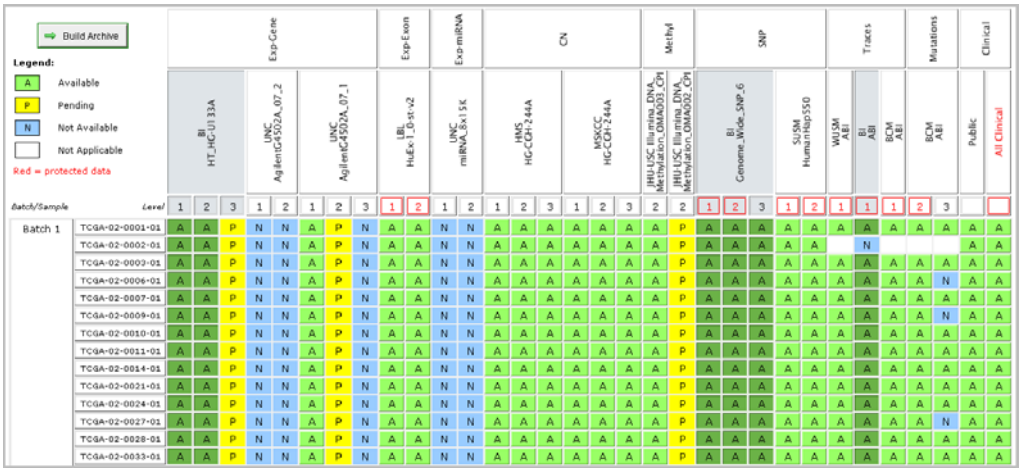

Figure A.6 All Data from a Specific Center

3. Click **Build Archive**.

Selecting All Data From a Single Center In a Single Batch

You can select all available data sets from a particular center, and then limit them to just a single batch. The resultant data sets are an intersection between the data sets for the center and the data sets for the batch.

Note: You can create an intersection of data sets more efficiently by using the Filter options. See *Data Filtering Techniques* on page 16.

To select all data for from a specific center in a given batch, follow these steps:

1. Click on the target **Center** button. (See *Selecting All Data From a Specific Center* on page 48.)
2. Press and hold the SHIFT key, and click the button for the target batch.

   Note: You can select the center and batch buttons in any order.

   As illustrated in *Figure A.7*, the following elements are activated as a result:

   - The target batch button
   - All samples in the batch

   The following elements are deactivated as a result:

   - All cells that are *not* in both the batch's row and the center's columns

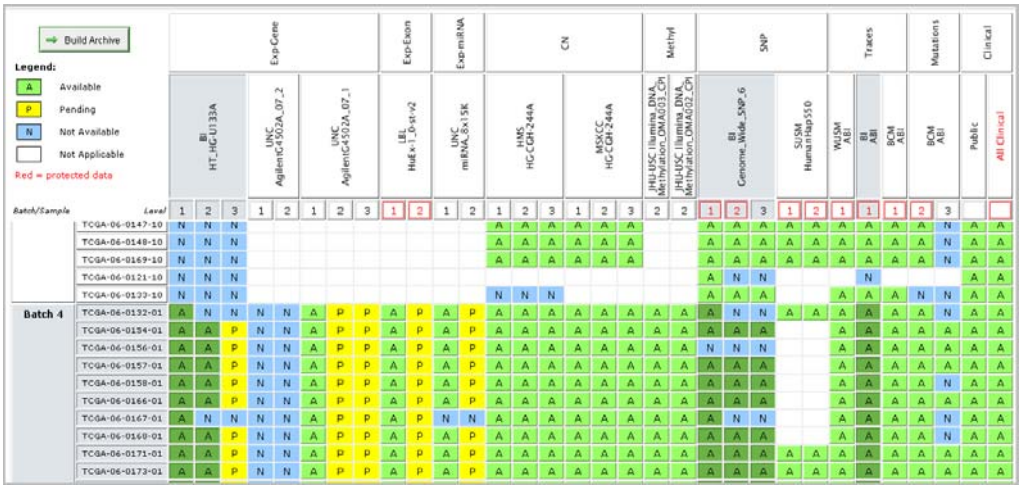

Figure A.7 All Data From a Specific Center in a Specific Batch

3. Click **Build Archive**.

Selecting All Data From a Data Level

To select all data from a data level, follow these steps:

1. Click the same **Data Level** button for each center of interest.

   As illustrated in *Figure A.8*, the following elements are activated as a result:

   - The selected **Data Level** buttons
   - All "available" data sets for each sample in the selected Data Level columns

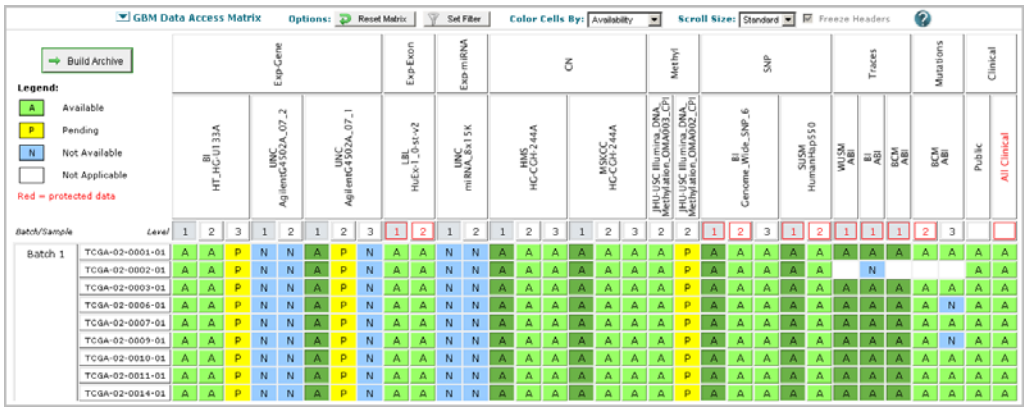

Figure A.8 All Data From a Data Level

2. Click **Build Archive**.

Appendix B

Accessing Data From an External Application or Website

This chapter provides instructions for using the Data Access Matrix to filter, archive, and retrieve data sets directly from external Web applications, without using the Matrix's data set selection features.

See *Downloading and Retrieving TCGA Data* on page 31 for instructions for selecting, downloading, and retrieving the files that contain your selected data.

Topics in this Chapter

- *Integrating TCGA Portal Data Access Service* on this page
- *Filtering Data From an External Application* on page 51
- *Valid Arguments For Data Set Filters* on page 56

Integrating TCGA Portal Data Access Service

If you have selected data sets via a Web application other than the Data Access Matrix, you can use the packaging, compression, and retrieval functions directly rather than using the data set selection techniques in the Matrix.

To integrate the data access service, invoke the following Web page in the Matrix:

```
http://tcga-data.nci.nih.gov/tcga/dataAccessExternalFilter.htm?
```

Filtering Data From an External Application

To create a data filter, identify the data criteria as a set of GET or POST arguments, as per *Example B.1*, *Example B.2*, and *Example B.3*.

Note that there is an optional parameter, "goto". The default is `goto=filetree`, so if you omit this parameter, the Matrix displays the file tree view (shown in *Figure B.1*,

Figure B.3 and *Figure B.4*). If you specify `goto=matrix`, the application goes directly to the matrix view, filtered according to the various parameters you specify (shown in *Figure B.2*). You would enter the goto= parameter in the position in the code shown in *Example B.1* in red text. To choose the default, leave out the text represented in red.

After applying the filter, select and download files as per the instructions in *Downloading and Retrieving TCGA Data* on page 31.

Example B.1 Criteria for selecting level 2 and 3 data for a specific set of patients

```
http://tcga-data-dev.nci.nih.gov/tcga/
   dataAccessExternalFilter.htm?goto=matrix&level=2,3&sampleList
   =tcga-02-0064-*,tcga-02-0069-*,tcga-02-0071-*,tcga-02-0074-
   *,tcga-02-0075-*,tcga-02-0080-*,tcga-02-0083-*,tcga-02-0085-*
```

The code in this example results in a file tree as illustrated *in Figure B.1.*

Figure B.1 Data Access Matrix Download Page – Selected Data Sets

The matrix view opens when you use the goto=matrix optional parameter or enter no goto= parameter (*Figure B.2*):

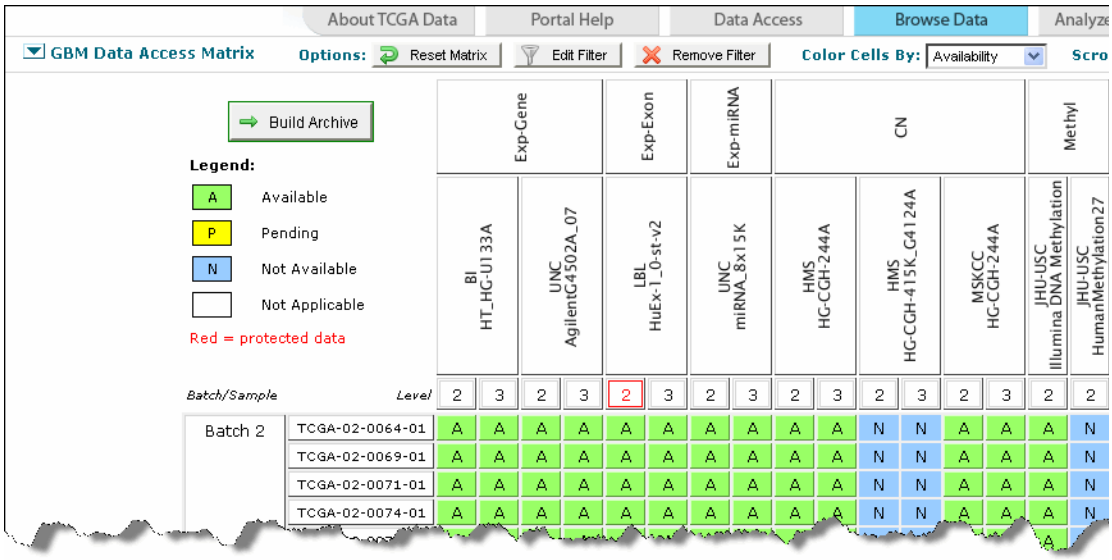

Figure B.2 Matrix view which opens with goto=matrix parameter

The criteria for the filter code in *Figure B.1* is described in *Table B.1*.

Filter Type	***Arguments***
Levels	`level=2,3` where, {`level=`}= data level parameter {`2`} = data level 2 {`,`} = attribute separator {`3`} = data level 3
Sample	`&sampleList=tcga-02-0064-*,tcga-02-0069-*` where, {`&`} = argument separator {`sampleList=`} = sample parameter {`tcga`} = project name {`-`} = Sample ID attribute separator {`02`} = Site ID (MD Anderson Cancer Center) {`0064`} = Patient ID {`*`} = any Sample Type ID (wildcard character) {`,`} = attribute separator {`tcga`} = project name {`02`} = Site ID (MD Anderson Cancer Center) {`0069`} = Patient ID {`*`} = any Sample Type ID (wildcard character)

Table B.1 Arguments for the filters in Figure B.1

Example B.2 Criteria for selecting Level 2 DNA Methylation Tumor samples only

```
http://tcga-data-dev.nci.nih.gov/tcga/
   dataAccessExternalFilter.htm?
   platformType=6&tumorNormal=T,TN&level=2
```

The code in this example results in a file tree as illustrated *in Figure B.3.*

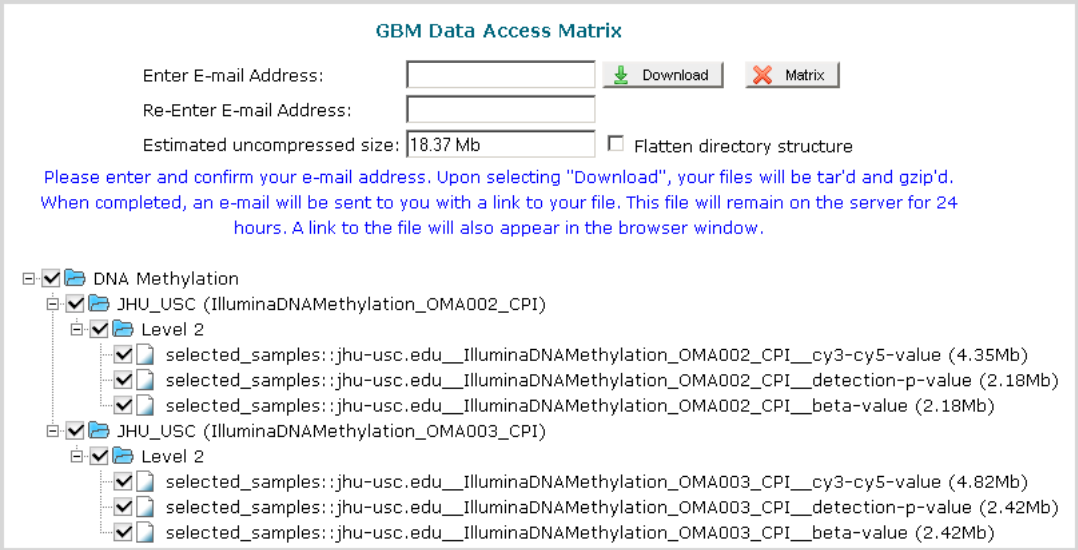

Figure B.3 Data Access Matrix Download Page – Selected Data Sets

The criteria for the filter code is described in *Table B.2*

Filter Type	***Arguments***
Platform type	`platformType=3,5` where, `{platformType=}` = platform type parameter `{6}` = Platform Type 6 (DNA Methylation) `{,}` = attribute separator `{3}` = Platform Type 5 (miRNA)
Tumor/Normal	`&tumorNormal=T,TN` where, `{&}` = argument separator `{tumorNormal=}` = tumor tissue/normal tissue parameter `{T}` = tumor tissue for which there is no matched normal tissue `{,}` = attribute separator `{TN}` = tumor tissue for which there is a matched normal tissue

Table B.2 Arguments for the filters in Figure B.3

Filter Type	***Arguments***
Level	`&level=2` where, `{&}` = argument separator `{level=}`= data level parameter `{2}` = data level 2

Table B.2 Arguments for the filters in Figure B.3

Example B.3 Criteria for selecting Clinical Batch 2 samples only

```
http://tcga-data-dev.nci.nih.gov/tcga/
   dataAccessExternalFilter.htm?platformType=-
   999&batch=Batch%202
```

The code in this example results in a file tree as illustrated *in Figure B.4.*

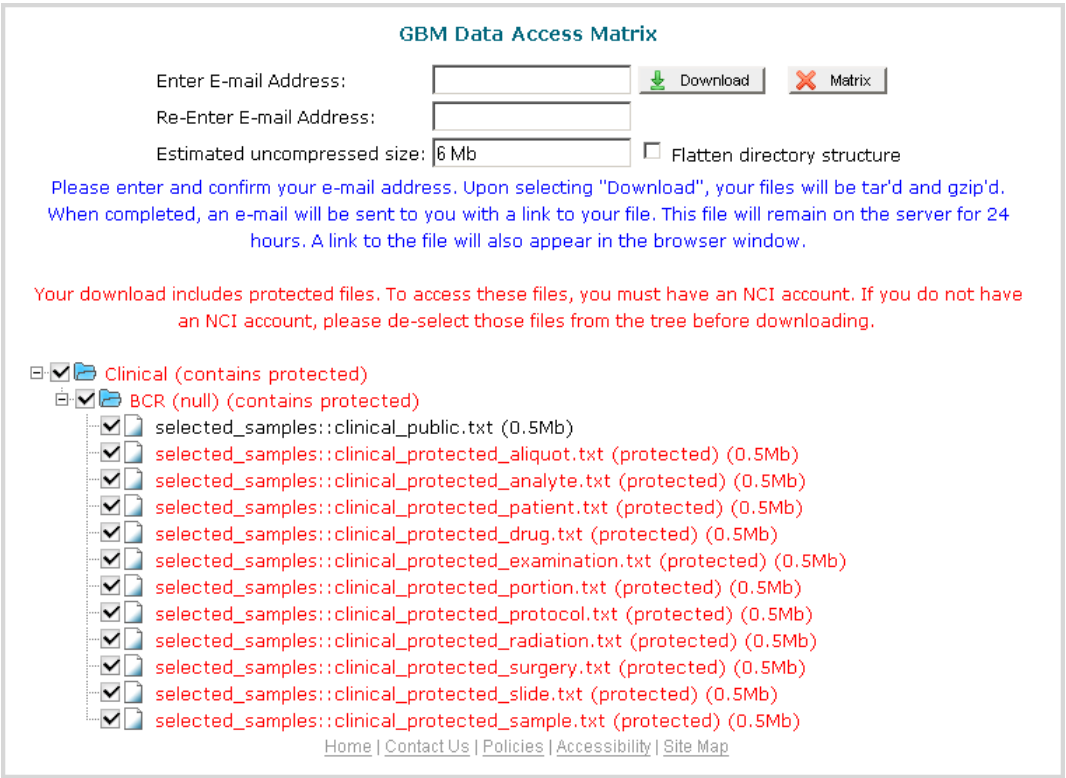

Figure B.4 Data Access Matrix Download Page – Clinical Data From Batch 2

The criteria for the filter code is described in *Table B.3*

Filter Type	***Arguments***
Platform type	`platformType=999` where, `{platformType=}` = platform argument `{999}` = Clinical data `{,}` = attribute separator `{3}` = Platform Type 5 (miRNA)
Batch	**`&batch=Batch%202`**`where,` `{&}` = argument separator `{batch=}` = batch number parameter `{%20}` = space character (URL code) `{2}` = batch number 2

Table B.3 Arguments for the filters in Figure B.4

Using the Matrix to Visualize and Modify Your Filter

The Matrix displays the data sets returned as per your filter criteria in a file tree on the Data Access Matrix Download page. (Refer to *Figure B.4* for an example of the data sets as they are displayed on the Download page.) You can display the same filter visually and edit or remove your filter using the Matrix filtering techniques. See *Modifying the Display of Data* on page 17.

How to Edit or Remove a Filter in the Data Access Matrix:

1. On the Download page, click **Matrix**, located beside the **Download** button.
2. Change or remove the filter criteria as described in *Data Filtering Techniques* on page 16.

Valid Arguments For Data Set Filters

Table B.4 lists the valid arguments for filtering data.

Filter	***Arguments***
diseaseType	if missing, the filter defaults to GBM
availability	A, P, N
batch	Batch 1, 2, 3, etc.
startDate	mm/dd/yyyy
endDate	mm/dd/yyyy
protectedStatus	P, N
tumorNormal	T, TN, NT, N

Table B.4 Valid arguments for data set filters

Filter	Arguments
sampleList	Barcode. Must start with TCGA (project name) followed by specific data sets as per *Example B.1* on page 52. Wildcard are accepted. See *Filtering Data by Sample Barcodes* on page 22.
platformType	Numeric constant. Constants are provided in *Table B.5* on page 57.
center	Numeric constant. You must select a center by its center ID and a platform ID, in the form c.p where, {c} = center ID {.} = separator {p}= platform ID For example, identify BI/HT_HG-U133A as 3.3 Constants are provided in *Table B.5* on page 57.
level	1,2,3, or C (for clinical data)

Table B.4 Valid arguments for data set filters (Continued)

Numeric Constants For Platforms, Platform Types, and Centers

Table B.5 lists the numeric constants for platform types and centers.

Criteria	Values	Constants
PlatformType	Clinical	999
PlatformType	Complete Clinical Set	1
PlatformType	Copy Number Results	8
PlatformType	DNA Methylation	6
PlatformType	Expression-Exon	4
PlatformType	Expression-Genes	3
PlatformType	Expression-miRNA	5
PlatformType	Minimal Clinical Set	2
PlatformType	SNP	7
PlatformType	Somatic Mutations	9
PlatformType	Trace-Gene-Sample Relationship	10
Platform	ABI	17
Platform	AgilentG4502A_07_1	25
Platform	AgilentG4502A_07_2	13639
Platform	Genome_Wide_SNP_6	19
Platform	H-miRNA_8x15K	21
Platform	HG-CGH-244A	10
Platform	HT_HG-U133A	3

Table B.5 Numeric constants for platforms, platform types and centers

Criteria	Values	Constants
Platform	HuEx-1_0-st-v2	4
Platform	HumanHap550	15
Platform	IlluminaDNAMethylation_OMA002_CPI	14
Platform	IlluminaDNAMethylation_OMA003_CPI	130
Platform	WHG-4x44K_G4112F	12
Center	BCM	9
Center	BCR	1
Center	BI	3
Center	HAIB	11
Center	HMS	2
Center	JHU_USC	6
Center	LBL	4
Center	MSKCC	7
Center	Protected and Public	888
Center	Public	777
Center	SUSM	8
Center	UNC	5
Center	WUSM	10

Table B.5 Numeric constants for platforms, platform types and centers (Continued)

APPENDIX

C

RULES FOR THE VISUAL DISPLAY OF DATA IN THE MATRIX

This chapter provides the rules that dictate the display of data in the Matrix. It explains the classifications of data as Available, Not Available, Not Applicable, Pending, and orphaned.

Topics in this Chapter

- *Rules For Displaying Sample Barcodes* on this page
- *Rules For Displaying Cells* on page 60
- *Rules For Displaying Data Sets* on page 62
- *Rules For Displaying Batches* on page 62
- *Origins of Orphaned Barcodes* on page 62
- *Rules For Displaying Orphaned Barcodes* on page 63

Rules For Displaying Sample Barcodes

The Matrix displays a sample's row as “Available,” if the following criteria apply:

1. The DCC received data for the sample from at least one center.

   – or –

   The DCC received clinical data for the sample that included at least one aliquot barcode.

– and –

2. At least one of the aliquot barcodes that are part of the sample is not marked “do not display.” See *BCR Barcode Exclusion* on page 62.

Note: The "do not display" feature will be available in a subsequent release of the Matrix.

Rules For Displaying Cells

The Matrix displays centers' cells differently from clinical cells, as explained in the following sections:

- *Rules For Center Cells* on page 60
- *Rules For Clinical Cells* on page 61

Rules For Center Cells

Clinical cells are displayed with any of the following indicators:

- Not applicable. See *Center Cells That Are Not Applicable* on page 60.
- Not available. See *Center Cells That Are Not Available* on page 60.
- Pending. See *Center Cells That Are Pending* on page 60.
- Available. See *Center Cells That Are Available* on page 61.

Center Cells That Are Not Applicable

The Matrix displays a center cell as blank, "Not Applicable," if the following criteria apply:

1. The DCC did not receive clinical data for any of the aliquot barcodes that the cell represents.

   – and –

   The DCC did not receive data for the sample from the center at the data level that the cell represents.

– or –

2. All aliquot barcodes associated with the cell are marked as "do not display" but the sample row is visible. See *BCR Barcode Exclusion* on page 62.

Center Cells That Are Not Available

The Matrix displays a center cell as "Not Available" if the following criteria apply:

1. The DCC received clinical data for the sample that included at least one of the aliquot barcodes that the cell represents.

– and –

2. DCC did not receive any data from the center for the sample at the data level that the cell represents.

Center Cells That Are Pending

The Matrix displays a center cell as "Pending" if the following criteria apply:

1. The DCC received data for the sample barcode from the center at the data level that the cell represents.

– and –

2. The data for the sample barcode at that level has not yet been processed by the portal team.

Center Cells That Are Available

The Matrix displays a center cell as “Available” if the following criteria apply:

1. The DCC received data for the sample from the center at the data level that the cell represents.

– and –

2. The data for the sample at that data level has been loaded/processed by the portal team

Rules For Clinical Cells

Clinical cells are displayed with any of the following indicators:

- Not applicable. See *Clinical Cells That Are Not Applicable* on page 61.
- Not available. See *Clinical Cells That Are Not Available* on page 61.
- Pending. See *Clinical Cells That Are Pending* on page 61.
- Available. See *Clinical Cells That Are Available* on page 61.

Clinical Cells That Are Not Applicable

The Matrix displays a clinical cell as blank, “Not Applicable” if the DCC did not receive any data for the sample barcode from the BCR.

Clinical Cells That Are Not Available

The Matrix displays a center cell as “Not Available” if the following criteria apply:

1. The DCC received clinical data for the sample that included at least one of the aliquot barcodes that the cell represents.

– and –

2. The DCC did not receive any data from the center for the sample at the data level that the cell represents.

Clinical Cells That Are Pending

The Matrix displays a clinical cell as “Pending” if the following criteria apply:

1. The DCC received full aliquot barcode data from the BCR for the sample that the cell represents.

– and –

2. TCGA portal team has not yet loaded the clinical data for the full barcodes.

Clinical Cells That Are Available

The Matrix displays a clinical cell as “Available” if the following criteria apply:

1. The DCC received full aliquot barcode data from the BCR for the sample that the cell represents.

– and –

2. TCGA portal team has loaded the clinical data.

Rules For Displaying Data Sets

The Matrix displays an available center cell if it contains a data set for each distinct full aliquot barcode found for the sample/center/platform level combination that the cell represents.

Note: If two or more data sets for a single full barcode exist in the data store, the Matrix displays the most recent data set received by the DCC.

The Matrix displays an available clinical cell if it contains a single data set containing all clinical information for all aliquots for that sample for which the aliquot data was received.

Rules For Displaying Batches

The Matrix displays a sample barcode's row within a batch if the BCR data for that sample is available, even if one or more aliquot barcodes for the sample are orphans (see *Origins of Orphaned Barcodes* on page 62). The BCR batch in which the sample was found will be used. Otherwise, the Matrix displays it in the "Unclassified" batch.

Origins of Orphaned Barcodes

An "orphaned" barcode is an aliquot barcode that contains data from a center but does not exist in any of the latest BCR archives.

The Matrix may consider a barcode "orphaned" for three reasons as follows:

- *CGCC File Generation Errors*
- *CGCC File Submission Error*
- *BCR Barcode Exclusion*

CGCC File Generation Errors

The CGCC may have made a mistake in generating their file. To correct the error, the CGCC submits a corrected file, and the Matrix displays the new data sets.

CGCC File Submission Error

A file submission error may occur if the CGCC submits a file before the BCR updates their archives with the latest barcodes. The Matrix automatically removes the orphan flag once the BCR submits the new archives.

BCR Barcode Exclusion

Barcode exclusion may occur if the BCR dropped a barcode from their archives.

- If the BCR dropped the barcode accidentally, the BCR submits corrected archives.
- If the BCR dropped the barcode intentionally, the BCR and/or CGCCs may request that the barcode be added to the “do not display.”

Rules For Displaying Orphaned Barcodes

The Matrix displays data differently for orphaned barcodes originating from CGCCs from the orphaned barcodes originating from the BCR.

A center may exclude barcodes from the Matrix intentionally by adding them to a “do not display” list. The Matrix does not display the designated barcodes regardless of whether associated data is available.

Rules for CGCC Orphans

If a CGCC archive contains an orphaned barcode, the Matrix displays the incorrect barcode in a Batch designated “Unclassified;” it does not display clinical data.

Rules for BCR Orphans

The Matrix displays BCR orphans as follows:

- If the BCR removes barcodes from their archives, the Matrix displays them only if one of the current center archives contains data for those barcodes. The Matrix displays these barcodes in a Batch designated “Unclassified.”
- If the BCR removes all aliquot barcodes from a sample but leaves the sample in their archives, the Matrix does not display clinical data. Clinical data is only available when aliquots have been specified.

INDEX

A

analyte barcode
- patient ID values 24
- sample type values 24

B

barcodes
- origins of orphaned 62
- rules for displaying orphaned 63

batches, rules for displaying 62

C

cells, rules for displaying 60
centers, numeric constants for 57

D

data
- accessing from external application 51
- download components 31
- downloading from the Matrix 31
- retrieving archive 39
- rules for displaying data sets 62
- selecting all files for download 37
- selecting all from a batch, example 44
- selecting all from a data level, example 50
- selecting all from single center, single batch, example 49
- selecting all from specific center, example 48
- selecting from platform type, example 45
- selecting specific, example 43
- selecting subset of files for download 37
- selecting union of data sets, example 47
- steps for accessing 5
- steps for downloading files 35
- TCGA compatible platforms 13
- valid arguments for filtering 56
- visualizing and modifying filter criteria 56

Data Access Matrix
- accessing 6
- additional references 2
- download page 31
- key to graphic elements 11
- navigating 7
- Options menu 10
- user interface 7
- visualizing and modifying filter criteria 56

download page, directory structure 33
download, understanding components 31
downloading
- data from the Matrix 31
- selecting all files for 37
- selecting subset of files for 37
- steps for data files 35

E

external application
- accessing data from 51
- filtering data from 51

F

filtering
- data from external application 51
- data, valid arguments for 56

M

Matrix See Data Access Matrix

O

orphaned barcodes
- origins 62
- rules for displaying 63

overview, TCGA project 5

P

platforms
- in TCGA data 13
- numeric constants for types 57

S

sample barcodes

rules for displaying 59
selecting
all data from a data level, example 50
all data from batch, example 44
all data from single center, single batch, example 49
all data from specific center, example 48
data from platform type, example 45
specific data sets, example 43
union of data sets, example 47

T

TCGA
data, compatible platforms 13
project overview 5
TCGA Portal
accessing Matrix from 6
data access service 51

U

user's guide
purpose 1
text conventions 2
topics in 2