CANCER MOLECULAR ANALYSIS PORTAL

User's Guide

NATIONAL® CANCER INSTITUTE

Center for Bioinformatics

July 12, 2008

TABLE OF CONTENTS

About This Guide

This section introduces you to the *Cancer Molecular Analysis Portal User's Guide*. It includes the following topics:

Purpose

This guide provides an overview of the Cancer Molecular Analysis Portal (CMA Portal) and instructions for using its tools and resources for querying and analyzing patient tissue *sample* data. This book is organized into chapters that parallel the CMA Portal's workflow.

Audience

This guide is designed for researchers who want to perform ad hoc querying and reporting across multiple domains, such as gene expression, chromosomal aberrations, and clinical data.

Topics Covered

If you are new to CMA Portal, read this brief overview, which explains what you will find in each chapter.

- *Chapter 1* provides information for launching, logging in to, and navigating the CMA Portal.
- *Chapter 2* describes how to use the CMA Portal to research gene expression, copy number, SNP, LOH, and pathway data.
- *Chapter 3* describes how to use the CMA Portal to review pathways and the anomalies associated with them.
- *Chapter 4* describes how to use the CMA Portal to investigate chromosomal regions of amplification, deletion, and over/under-expression.

- *Chapter 5* describes how to use the CMA Portal to study clinical data and explore the relationships between clinical and molecular study data.
- *Chapter 6* describes how to use the CMA Portal to perform principal component and gene pattern analyses, and to access associated research tools.
- *Chapter 7* describes how to select, create, and manage patient and/or gene lists.
- *Glossary* defines terms in this guide.

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 monospace 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.
{ }	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.

Cancer Molecular Analysis Portal Development and Management Teams		
Development	***Documentation***	***Project and Product Management***
Alex Jiang[4]	Lauren Anthone [2]	Anand Basu[1]
David Nassau[3]		Daniela Gerhard[1]
Erin Hickey[3]		Subha Madhavan[1]
Huaitien Liu[4]		Carl Schaefer[1]
Jianghin Zhang[4]		
Mark Lewis[4]		
Michael Harris[4]		
Mike Edmondson[4]		
Richard Finney[4]		
Robert Sfeir[3]		
Ryan Landy[4]		
Ying Long[4]		
1 National Cancer Institute Center for Bioinformatics (NCICB)	2 Lockheed Martin	3 SRA International, Inc.
4 Science Application International Corporation (SAIC)		

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 CANCER MOLECULAR ANALYSIS PORTAL

This chapter provides information for launching, logging in to, and navigating the Cancer Molecular Analysis Portal.

Topics in this chapter include:

About the Cancer Molecular Analysis Portal

The Cancer Molecular Analysis Portal (CMA Portal) provides tools and resources that enable you to query genomic characterization, sequencing, and clinical data. For more information about the available data, refer to The Cancer Genome Atlas' (TCGA's) website at:
http://cancergenome.nih.gov/components/dmbca.asp

Launching the Cancer Molecular Analysis Portal

To launch the Cancer Molecular Analysis Portal, follow these steps:

1. In Internet Explorer or other browser, type the following CMA Portal URL in the address field: https://cma.nci.nih.gov/cma/index.jsp

The CMA Portal home page appears *(Figure 1.1)* and displays the Gene View menu by default.

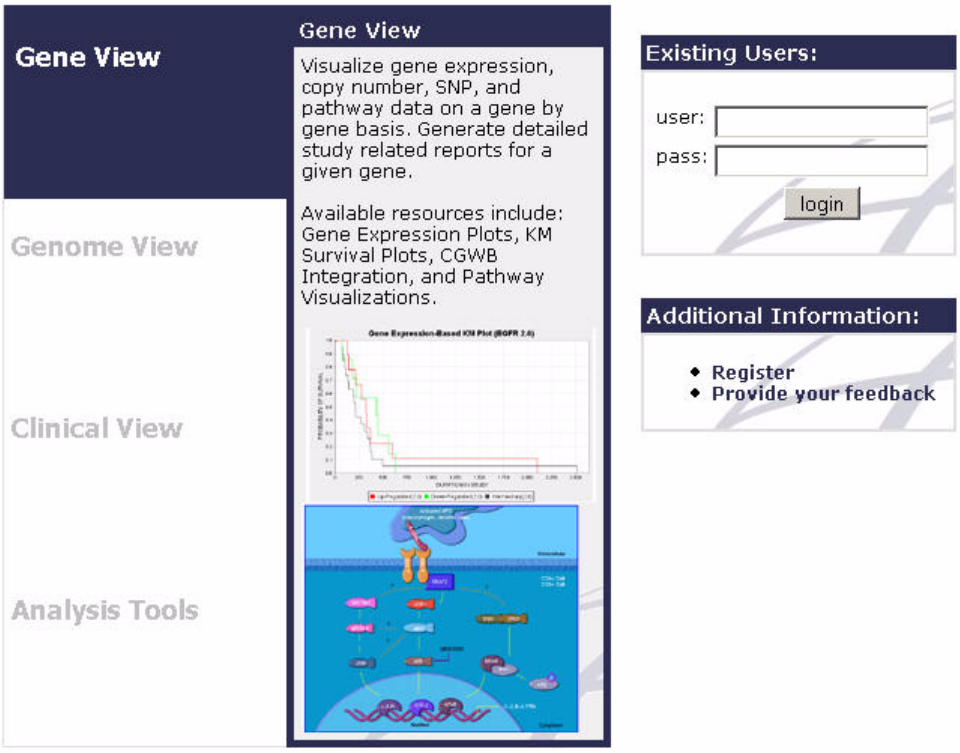


Figure 1.1 Cancer Molecular Analysis Portal – Home Page

Logging In To the Cancer Molecular Analysis Portal

If you intend to conduct your research using *open-access data*[1] only, you do not need a user ID to use CMA Portal resources. However, to take advantage of *controlled-access data*[2] in the Cancer Genome Atlas (TCGA) you must provide a username and password. See *Requesting a Username and Password* on page 5.

Logging In From the Home Page

If you have a CMA Portal account, you can log in from the Home page before you use any of the CMA Portal tools.

To log in to the CMA Portal, follow these steps:

1. In the **Existing Users** side bar, type your username and password in the text boxes provided.
2. Click **Log In**.

Logging In From Workspace Pages

If you do not have a CMA Portal account, you can use any of the CMA Portal tools with *open-access data*. Afterward, you may decide to register for access to controlled data.

To log in to the CMA Portal from workspace pages, follow these steps:

1. Open-access data – Data that is available to the general public.
2. Controlled-access data – Non-public, protected data.

1. Click any of the views or tools on the menu bar, for example, **Analysis Tools**.
2. Read and accept the provisions. See *Accepting the Cancer Molecular Analysis Portal Provisions* on page 5.

   In this example, the Analysis Tools workspace appears.

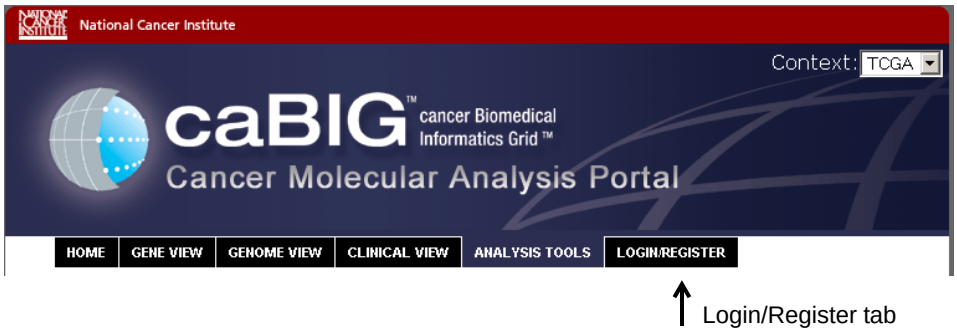


Figure 1.2 Analysis Tools Workspace – Tabs

3. Click the **Login/Register** tab.

   The Login page appears.

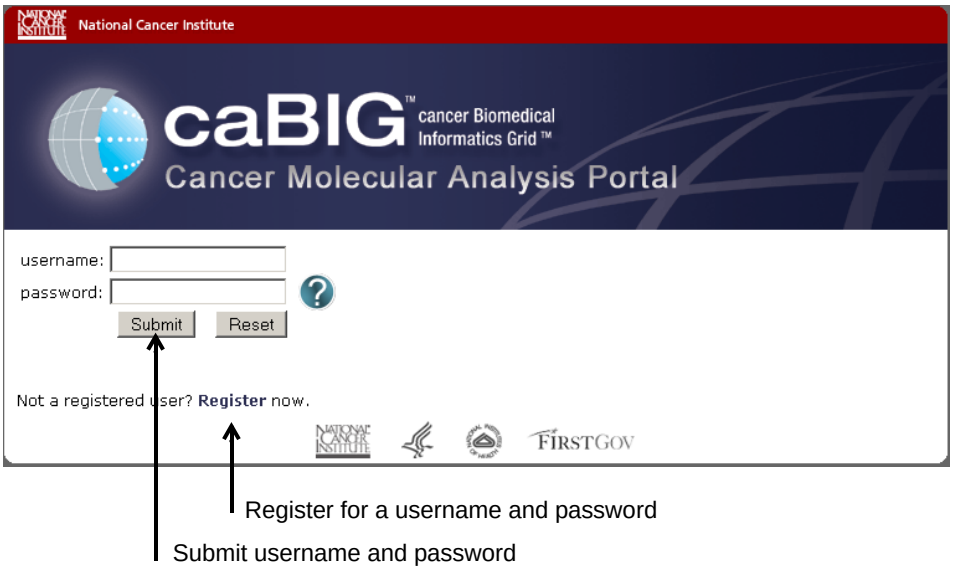


Figure 1.3 Login Page

4. Type your username and password in the text boxes provided, and then click **Submit**.If you do not have an account, click **Register** to register for a username and password, and then follow the instructions in *Requesting a Username and Password* on page 5.

Accepting the Cancer Molecular Analysis Portal Provisions

The Legal Rules of the Road page appears after you have logged in. This page also appears if you click any of the tabs on the left side of the page without having logged in first.

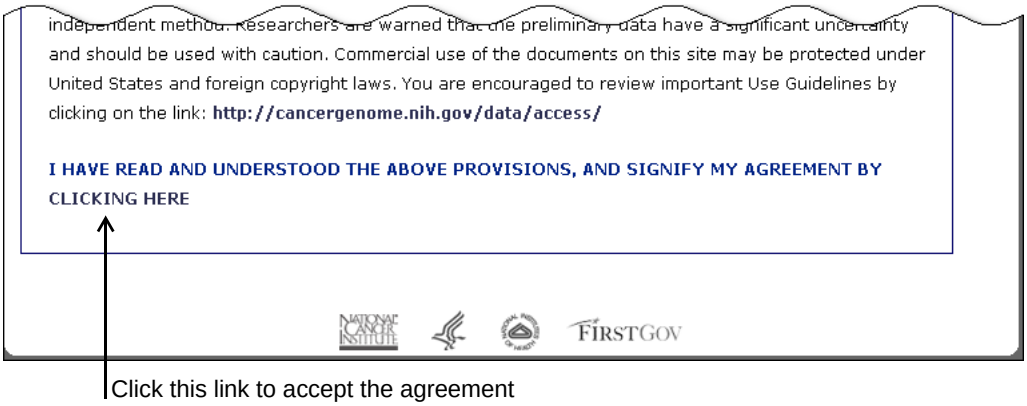


Figure 1.4 Legal Rules of the Road Page

Read the provisions, and then click **CLICKING HERE** (*Figure 1.4*).

Requesting a Username and Password

If you are a first-time CMA Portal user and want to access controlled TCGA datasets, click the **Register** link.

To register for a username and password, follow these steps:

1. In the **Additional Information** section of the side bar, click **Register**.

The New Users page appears.

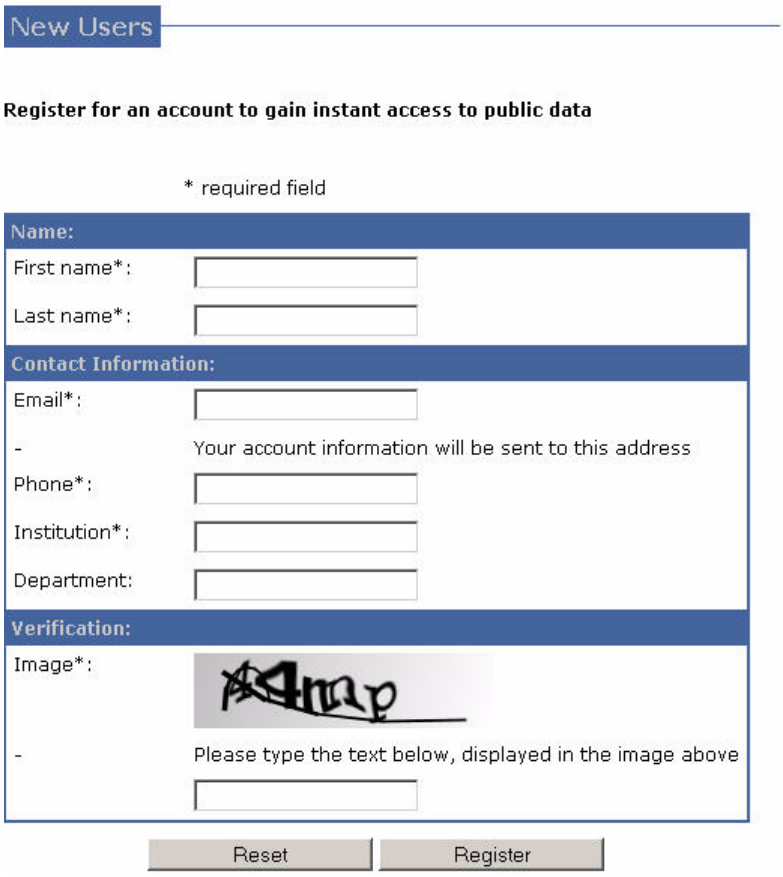


Figure 1.5 New Users Page

2. In the **Name** area, type your first and last names in the text boxes provided.
3. In the **Contact Information** area, type your email address and other contact information in the text boxes provided.

   Note: The system will send your account information to this email address.
4. In the **Verification** area, type the series of letters and numbers in the text box exactly as displayed in the verification image.
5. Click **Register**, or, to clear all text boxes, click **Reset**.

   Note: All registration information requested is required.
6. If you clicked **Register**, the system displays a message indicating that your account information will be sent to you by email.
7. When you receive your account information, follow the instructions in the email message to apply for a Data User Certificate (DUC). For further information about data access and DUCs, see http://cancergenome.nih.gov/dataportal/data/access/closed/duc/.

Navigating Through the Cancer Molecular Analysis Portal

The panes on the CMA Portal workspace enable quick access to all the workspaces. Each workspace has tools and resources that allow you to investigate genomic data.

Navigating the Cancer Molecular Analysis Portal Home Page

The Cancer Molecular Analysis Portal home page (*Figure 1.6*) appears after you have accepted the Legal Rules of the Road.

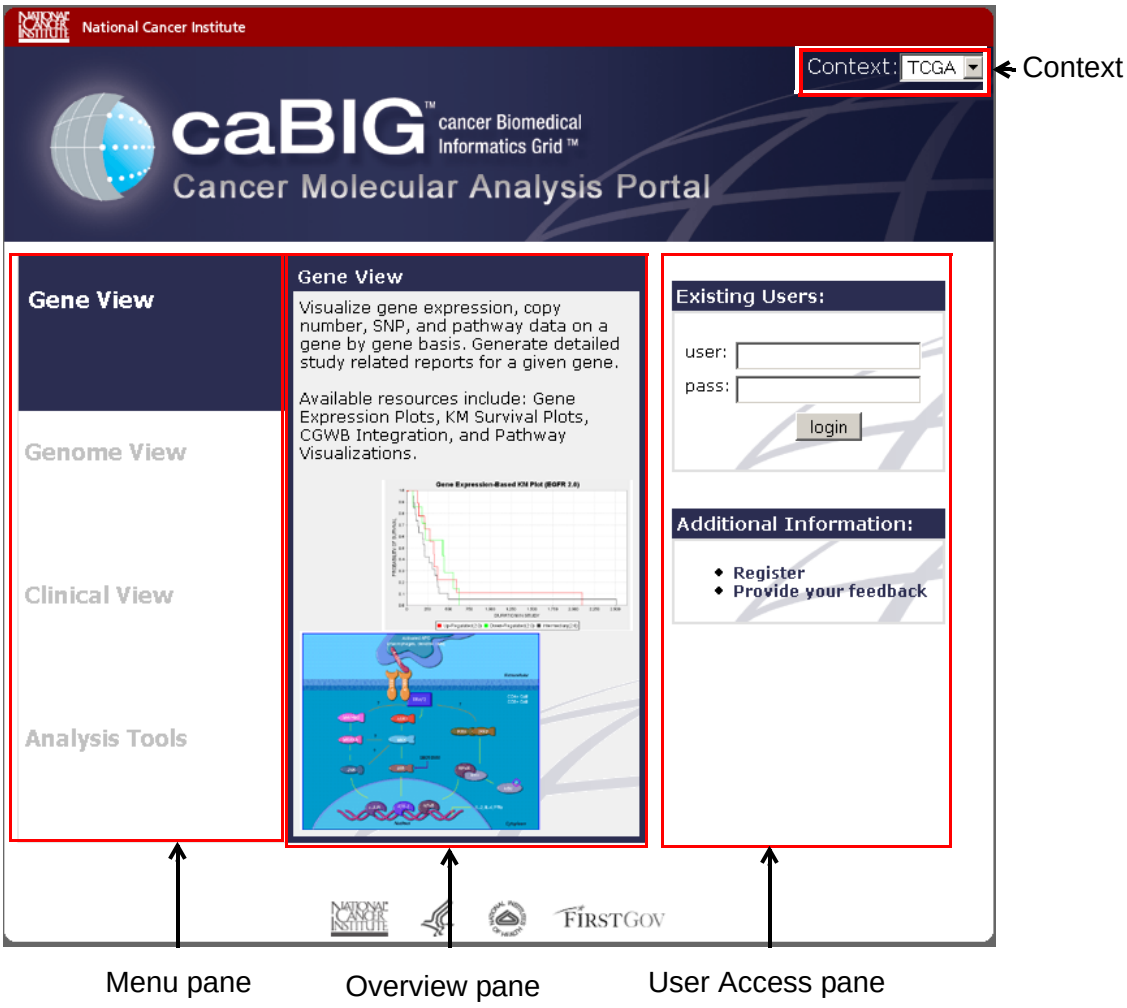


Figure 1.6 The Cancer Molecular Analysis Portal Home Page

Table 1.1 describes each of the panes on the Home page, and the Context drop-down list at the upper right side of the page.

Area	***Description/Function***
Menu pane	Provides links to query and analysis workspaces
Overview pane	Provides a brief overview of each of the workspaces
User Access pane	Login page for existing users. Provides links to register for an account and to provide feedback and request CMAP support.

Table 1.1 Cancer Molecular Analysis Portal – Home page elements

Area	Description/Function
Context	Indicates the current dataset used for your queries. Currently data from TCGA is available for research. Other datasets will become available in subsequent releases of the portal.

Table 1.1 Cancer Molecular Analysis Portal – Home page elements (Continued)

Menu Pane Links

Each section of the Menu pane acts as link to their respective workspaces. Click a link to access the tools and resources associated with each menu item. *Figure 1.7* illustrates the links to the workspace in the Menu pane.

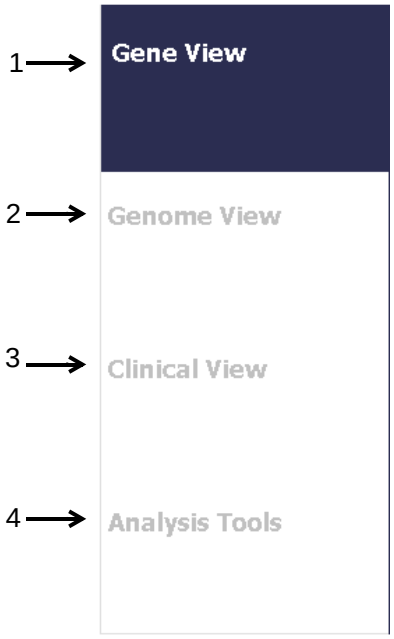


Figure 1.7 Menu Pane Links

Table 1.2 describes each item on the Menu pane.

Callout Number	Description/Function
1	**Gene View** link – Provides access to the gene analysis workspace
2	**Genome View** link – Provides access to the genome workspace
3	**Clinical View** link – Provides access to the clinical analysis workspace
4	**Analysis Tools** link – Provides access to the analysis tools workspace

Table 1.2 Cancer Molecular Analysis Portal – Menu pane elements

Overview Pane Elements

The Overview pane provides a brief overview of each workspace and the tools and resources available for research.

Figure 1.8 is an example of the overview and resource information that is provided for each workspace.

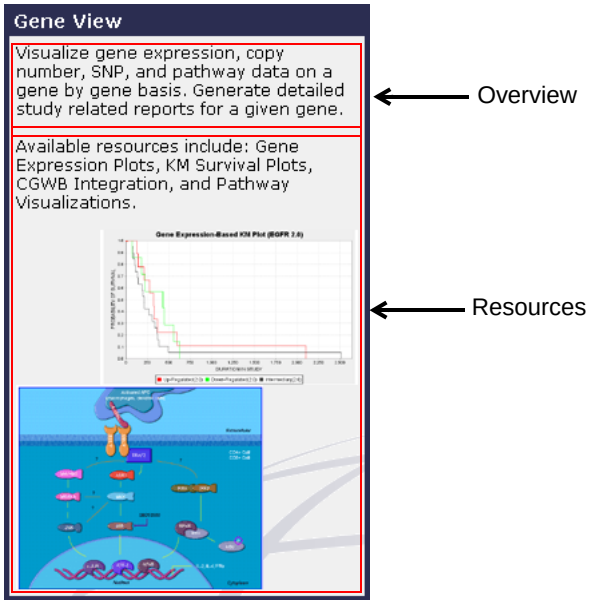


Figure 1.8 Overview Pane – Gene View

User Access Pane Elements

The Access pane provides access to *controlled-access data* for users who have registered for an account for TCGA (The Cancer Genome Analysis) data. New users can register for an account in the Additional Information section.

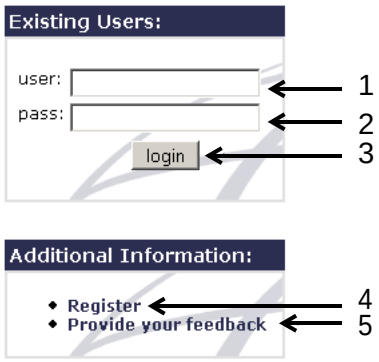


Figure 1.9 User Access Pane

Table 1.3 describes each item on the User Access Pane.

Callout Number	***Description/Function***
1	**User** – Type your username here to log in and obtain complete access to Cancer Molecular Analysis Portal
2	**Pass** – Type your password <<This
3	**login** – Click to log in to the portal

Table 1.3 CMA Portal home page – User Access side bar elements

Callout Number	Description/Function
4	**Register** – Click the link to register for access to the full dataset
5	**Provide Your Feedback** – Click the link to display the NCI Center for Bioinformatics Support page in a new browser window

Table 1.3 CMA Portal home page – User Access side bar elements (Continued) (Continued)

Navigating the Cancer Molecular Analysis Portal Workspaces

Each of the workspaces, Gene View, Genome View, Clinical View, and the Analysis Tools has its own workspace which is displayed whenever you select an item on the Home page menu bar. Each workspace contains tabbed pages, a side bar, and unique tools and resources. The tabs and sidebar are the same for all workspaces except the Login/Register workspace.

Tabs Elements

Tabs run across the top of each workspace except the Login/Register workspace. They allow you to access all CMAP workspaces.

Figure 1.10 Workspace Tabs

Side Bar Elements

The side bar appears on the right side of the of each workspace except the Login/ Register workspace.

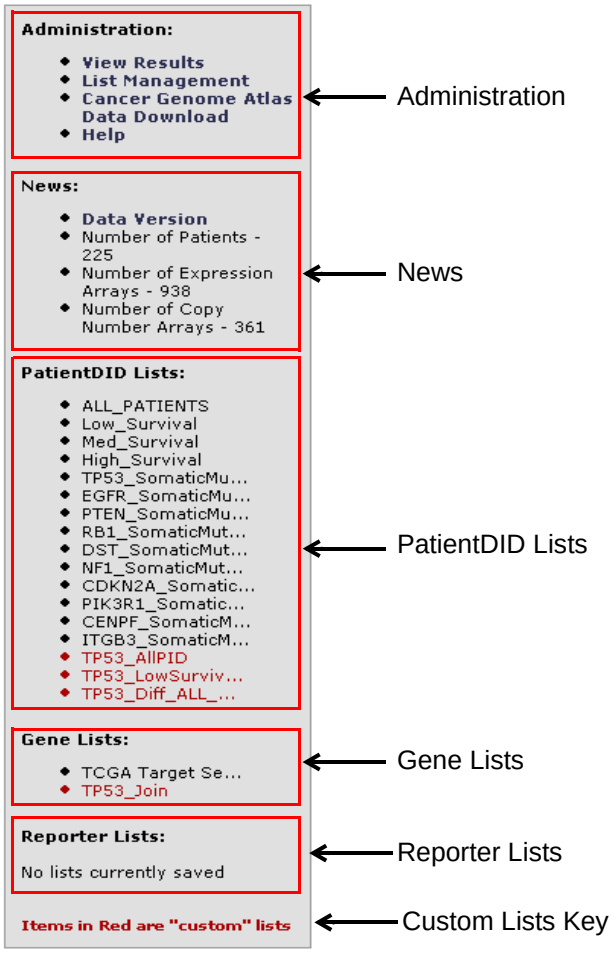


Figure 1.11 Side Bar

Administration Section Elements

Table 1.4 Describes each element in the Administration section.

Feature	***Description/Function***
View Results	Provides access to the workspace in which you have saved queries
List Management	Provides access to the default and any custom lists you may have created
Cancer Genome Atlas Data Download	Links to The Cancer Genome Atlas Data (TCGA) portal from which you can search for and download data from TCGA datastores. Also provides access to the Data Access Matrix. See *Accessing the Data Access Matrix* on page 12.

Table 1.4 Side Bar – Administration

Accessing the Data Access Matrix

The Data Access Matrix (the Matrix) application enables researchers to select data sets in TCGA (The Cancer Genome Atlas) servers through a user-friendly graphic-based selection system.

To access the Data Access Matrix follow these steps:

1. On the sidebar, click **the Cancer Genome Atlas Data Download** link.

   The Cancer Genome Atlas Data Portal appears in a new browser window.

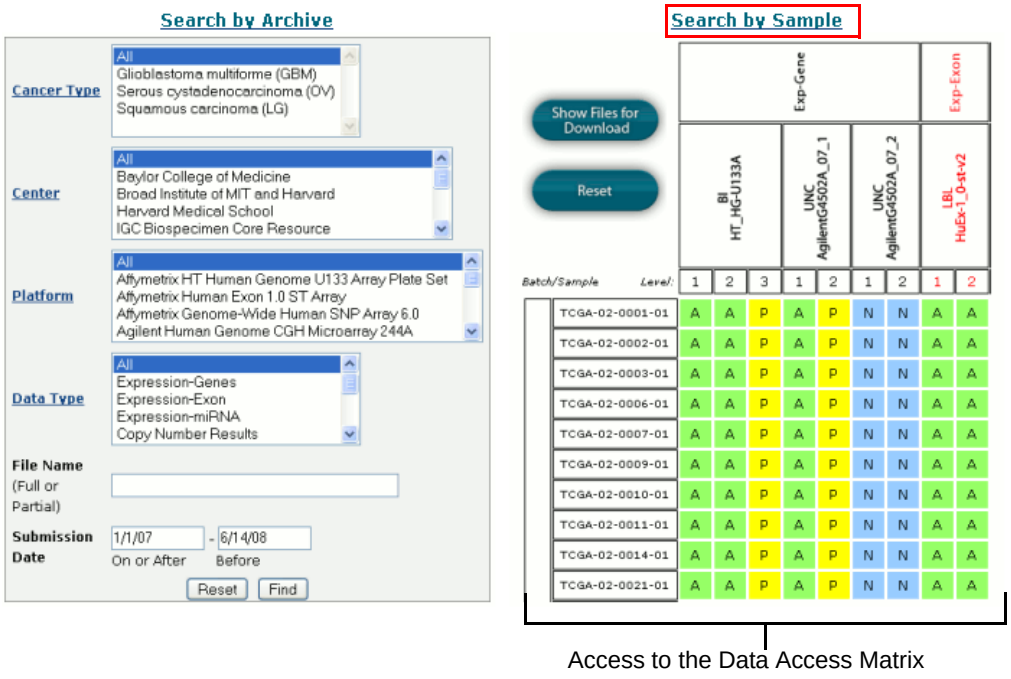


Figure 1.12 The Cancer Genome Atlas Portal – Search Data Page

2. Click the **Search Data** tab on the top of any Portal page, and then click the **Search by Sample** link. You can also click anywhere in the Data Access Matrix image.

Follow the instructions in the Data Access Matrix documentation for selecting and downloading data.

News Section

Table 1.5 describes each element in the News section.

Feature	***Description/Function***
Data Version	Provides information about the date and source of the data available for research and a link to a TXT file that contains a list of data source files.
Number of Patients	Number of patients sampled for the dataset
Number of Expression Arrays	Number of arrays used to obtain the available expression data

Table 1.5 Side Bar – News

Feature	Description/Function
Number of Copy Number Arrays	Number of arrays used to obtain the available copy number data
Custom Lists Key	Provides a key for any user-defined lists that appear in the side bar. For example, see the *PatientDID* section in *Figure 1.11*.

Table 1.5 Side Bar – News (Continued)

Note: The numbers of patients sampled and arrays used are updated continuously as new data sets from TCGA become available through the Data Coordination Center (DCC). For further information, see TCGA website at: http://cancergenome.nih.gov/components/dmbca.asp.

PatientDID List Section

De-identified ID (DID) information is data from tissue *sample* analysis from which personally identifiable information has been removed such that a researcher can no longer trace data back to an individual patient. For a description of PatientDIDs and the PatientDID lists, see *About PatientDID Lists* on page 68 and *Navigating the Sidebar Lists* on page 69.

Gene Lists Section

The Gene Lists section displays pre-defined sets of genes available for query (currently, the TCGA dataset) in black text and user-defined lists in red. See *About Gene Lists* on page 68 and *Creating Custom Lists* on page 71.

Reporter Lists section

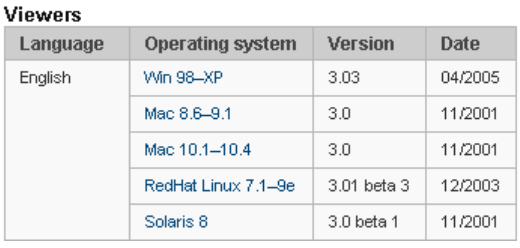
The Reporter Lists section displays user-defined sets of *gene reporters <<Subha*in red text. See *About Gene Reporter Lists* on page 69 and *Creating Custom Lists* on page 71.

Installing the SVG Plugin

you must install the Adobe *SVG Plugin* to display the pathway diagrams generated in CMA Portal. If you are an Internet Explorer user, Adobe's installer will add the plug-in automatically into Internet Explorer.

If you are a Mozilla Firefox user, follow these steps:

3. Navigate to http://www.adobe.com/svg/viewer/install/main.html and scroll down the page to the Installing Adobe SVG Viewer section.



Viewers

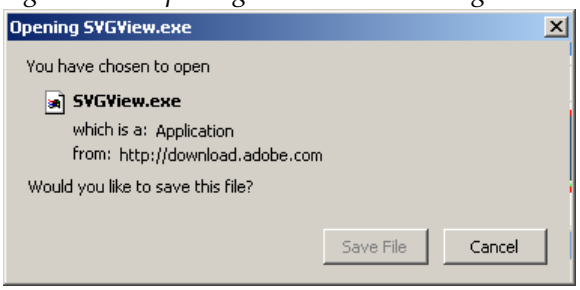
Language	Operating system	Version	Date
English	Win 98–XP	3.03	04/2005
	Mac 8.6–9.1	3.0	11/2001
	Mac 10.1–10.4	3.0	11/2001
	RedHat Linux 7.1–9e	3.01 beta 3	12/2003
	Solaris 8	3.0 beta 1	11/2001

Figure 1.13 Adobe SVG Viewer Download Links

4. Click the link that is appropriate to your language and operating system.

The Opening SVG Viewer dialog box appears.

Figure 1.14 Opening SVG Viewer Dialog Box



5. Click **Save File**.
6. Using Windows Explorer browse to the **`C:\Program Files\Common Files\Adobe`** folder.
7. Click the Plugins folder and copy the **`NPSVG3.dll`**and **`NPSVG3.zip`**. files.
8. Paste both files into your Firefox plugins folder (the default Firefox plugins folder is `C:\Program Files\Mozilla Firefox\plugins`).
9. In the Firefox browser address field, type *`about:config`*.

   A list of configurable files appears.

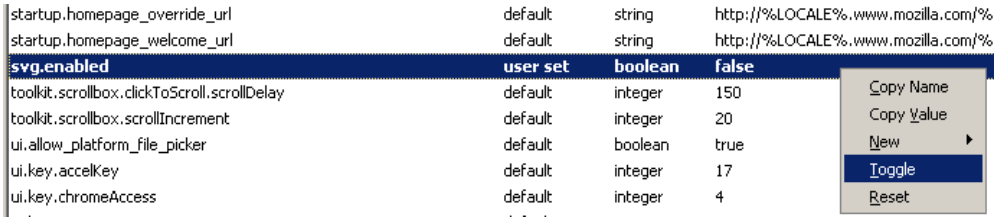


Figure 1.15 Plugins

10. Scroll down the list to **svg**.**enabled** to configure the setting.
11. To change the value from **true** to **false**, double click the row, or right-click the row, and select **Toggle**.
12. Close all browser windows and restart Firefox.

Logging Out

To log out of Cancer Molecular Analysis Portal, in the Cancer Molecular Analysis Portal workspace, click the **LOGOUT** tab in the upper right-hand corner.

You are logged out of Cancer Molecular Analysis Portal and returned to the Home page.

Caution: The system logs you out automatically after 30 minutes of inactivity. Because the CMA Portal does not save your custom lists from one session to the next, you will not be able to retrieve your lists after you have been logged out.

Application Support

For any general information about the application, application support, or to report a bug, contact *NCICB* Application Support.

Email: ncicb@pop.nci.nih.gov	When submitting support requests via email, please include: • Your contact information, including your telephone number • The name of the application/tool you are using • The URL of the application • A description of the problem and steps to recreate it • The text of any error messages you have received
Application Support URL	http://ncicb.nci.nih.gov/NCICB/support
Telephone: 301-451-4384 Toll free: 888-478-4423	Telephone support is available from: Monday to Friday, 8 am – 8 pm Eastern Time, excluding government holidays.

CHAPTER

2

WORKING WITH GENE-BASED VIEWS

This chapter describes how to use the Cancer Molecular Analysis Portal to research gene expression, copy number, SNP, and pathway data.

Topics in this chapter include:

Overview

The Gene View workspace in the CMA Portal enables researchers to analyze and visualize *gene expression*, *copy number*, *SNP*, *LOH*, and pathway data on a gene-by-gene basis, and to generate detailed study-related reports for a given gene.

This workspace provides access to gene expression plots, *Kaplan-Meier* survival plots, genome browser views, and pathway visualizations.

The CMA Portal analyzes data and categorizes gene anomalies (e.g., *overexpression*, *amplification*, etc.) by anomaly type. The following table illustrates the categories assigned to different data types.

Data Type	***Category***	
Gene expression	Overexpressed	Underexpressed
Copy number	Amplified	Deleted

Table 2.1 Categories of data analysis

Data Type	Category	
Mutation	Mutated	NA

Table 2.1 Categories of data analysis (Continued)

Note: CMA Portal also lists any agents (chemical compounds) that are associated with a genetic anomaly. The Cancer Gene Index Project is the source for gene-to-drug relationship.

Creating Gene Expression Plots

To create a gene expression plot, follow these steps:

1. Do one of the following to navigate to the Gene View workspace:

- From the Home page, click **Gene View**.

  - **or** -

- From any page in the portal, click the **Gene View** tab.

  The Gene View workspace appears.

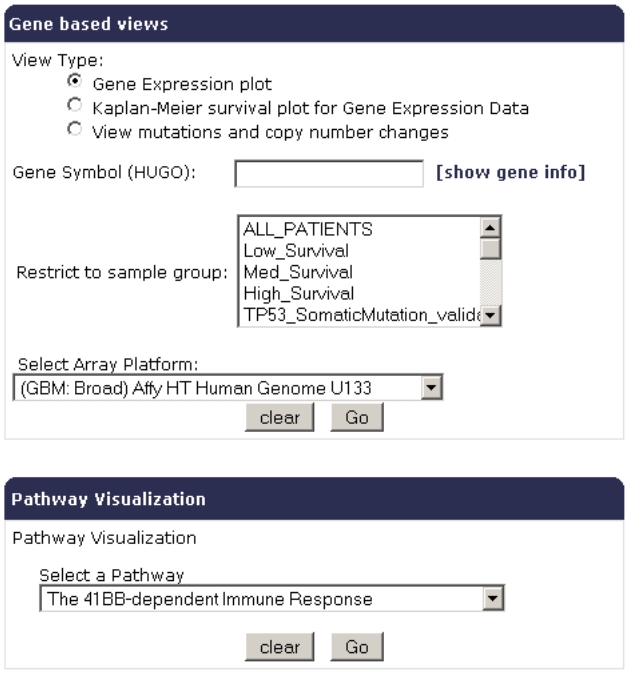


Figure 2.1 Gene View – Default Workspace

2. Under **View Type**, select the type of plot you want to generate.

Table 2.2 lists the view type options.

View Type	*Description*
Gene Expression Plot	Generates the following plots: • **Geometric Mean** – displays mean expression intensity (Geometric mean) versus Groups.For additional gene expression plot details, see *Understanding Geometric Mean Gene Expression Plots* on page 21. • **Log2 Intensity** – displays average expression intensities for the gene of interest. For additional graph details, see *Log2 Intensity Gene Expression Plot Details* on page 23. • **Box and Whisker Log2 Intensity** – displays a Box and Whisker plot or box plot. For additional plot details, see *Box and Whisker Log2 Intensity Gene Expression Plot Details* on page 24.
Kaplan-Meier Survival Plot	The Kaplan-Meier method is used for survival analysis. Kaplan-Meier curves are used to estimate survival probability for the user-defined set of criteria as a function of time and survival differences as analyzed by the log-rank test.
View mutations and copy number changes	Generates heatmaps and launches the Cancer Genome Workbench. For additional information, see *Viewing Mutation and Copy Number Changes* on page 33.
Pathway Visualization	Generates a graphic depiction of selected pathways and provides links to pathway information. For details, see *Visualizing Pathways* on page 34.

Table 2.2 Types of gene-based views

3. In the Gene Symbol (*HUGO*) box, type the gene symbol of interest, for example, *EGFR* or *WT1*, for which you want to generate an expression plot.
4. To view details about the gene of interest in the Cancer Genome Anatomy Project website, click **show gene info**, otherwise, skip to Step *6.*

   The Cancer Genome Anatomy Project website appears in a new browser window.

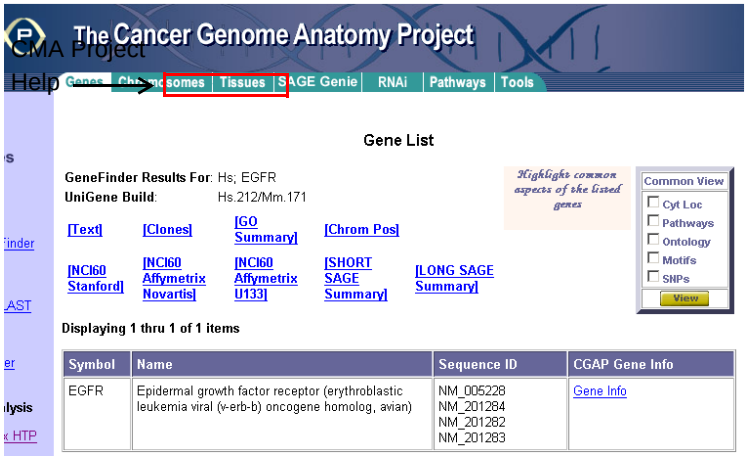


Figure 2.2 Cancer Genome Anatomy Project Web Page–EGFR

5. For help with using the Cancer Genome Anatomy Project, click **CGAP How To**.
6. To restrict your research to a specific *sample* group, from the **Restrict to sample group** drop-down list, select a sample group of interest. Custom lists appear in red text. To select multiple sample groups for comparison, press and hold the CTRL key and click each of the groups of interest.
7. From the **Select Array Platform** drop-down list, select an array platform.

   Table 2.3 provides a description of each of the available arrays.

Array Platform	***Description***
(GBM: Broad) Affy HT Human Genome U133	High-throughput expression profile of approximately 40,0000 transcripts and variants.
(GBM:UNC) Agilent Whole Human Genome	High-density profiling analysis tool that covers over 41,000 unique human genes and transcripts
(GBM:LBL) Affy Human Exon 1.0	Contains approximately one million predicted and confirmed exons.
Agilent 8 x 15K Human miRNA-specific Microarray	Contains probes for 530 human and 76 human viral microRNAs from the Sanger database v.10.1.

Table 2.3 Description of available array platforms

8. Click **Go**, or, to clear the entries on the page and start over, click **Clear**.

   The geometric mean gene expression plot appears.

   Note: The CMA Portal does not generate an expression plot if any of your criteria are incompatible, for example, if the gene you selected is not applicable for the array platform you selected. In such cases, a message at the top of the page is displayed to alert you to the incompatibility issue.

Understanding Geometric Mean Gene Expression Plots

The geometric mean gene expression plot (*Figure 2.3*) appears by default when you perform a gene expression search.

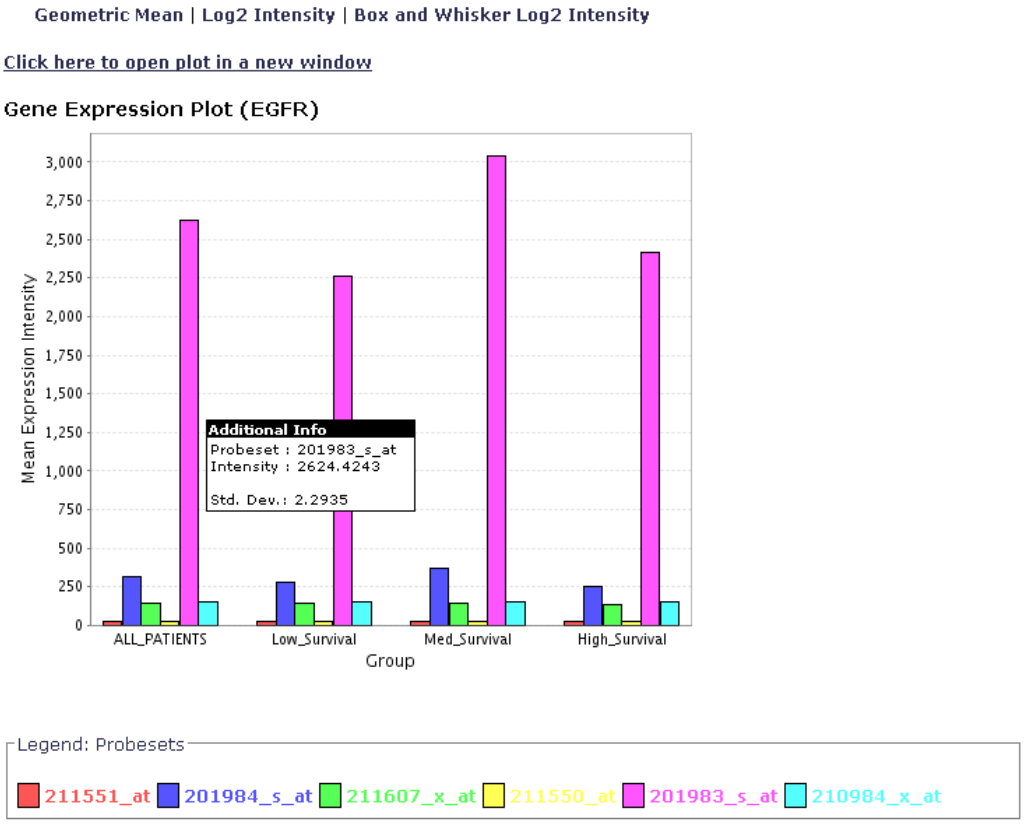


Figure 2.3 Geometric Mean Gene Expression Plot

Table 2.4 describes each area of the Gene Expression Plot page.

Item	*Special Instructions*
Plot Type Selection	Click a plot type name to display it. • **Geometric Mean** displays mean expression intensity (geometric mean) versus Groups.For additional graph details, see *Understanding Gene Expression-Based Kaplan-Meier Plots* on page 29. • **Log2 Intensity** displays average expression intensities (log2 values) for the gene reporter of interest. For additional graph details, see *Log2 Intensity Gene Expression Plot Details*. • **Box and Whisker Log2 Intensity** displays a Box and Whisker plot or box plot. For additional graph details, see *Box and Whisker Log2 Intensity Gene Expression Plot Details*.
Click here to open plot in a new window	Click the link to open the current plot in a new window and adjust the display.
Legend Probesets	**Gene Reporter information** – Indicates the color for each probeset appearing in the plot. To display gene reporter information, click a probeset number in the Legend.
Additional Info	To display additional information about each probeset, hover your mouse cursor over a bar in the graph. The pop-up window displays the following information: • **Probeset** – Each probeset contains multiple probe pairs. Each probe pair consists of two groups of probes—one called a perfect match (PM) and the other called a mismatch (MM). The perfect match is a set of oligonucleotides whose sequence exactly matches the gene of interest; the mismatch differs from the perfect match at one base position in the middle of the sequence. • **Intensity** – The geometric mean value calculated for each comparison group. • **Standard deviation** – The standard deviation value of a comparison group, such as GBM, for a particular probeset or gene. Standard deviation is a statistical measure of spread or variability.

Table 2.4 Understanding the Gene Expression Plot workspace

Visualizing Probesets In the Legend

Detailed information for each *probeset* in the gene expression plot legend is available via the CMA Reporter.

To view probeset details, click the probeset of interest.

The CMA Viewer displays the details.

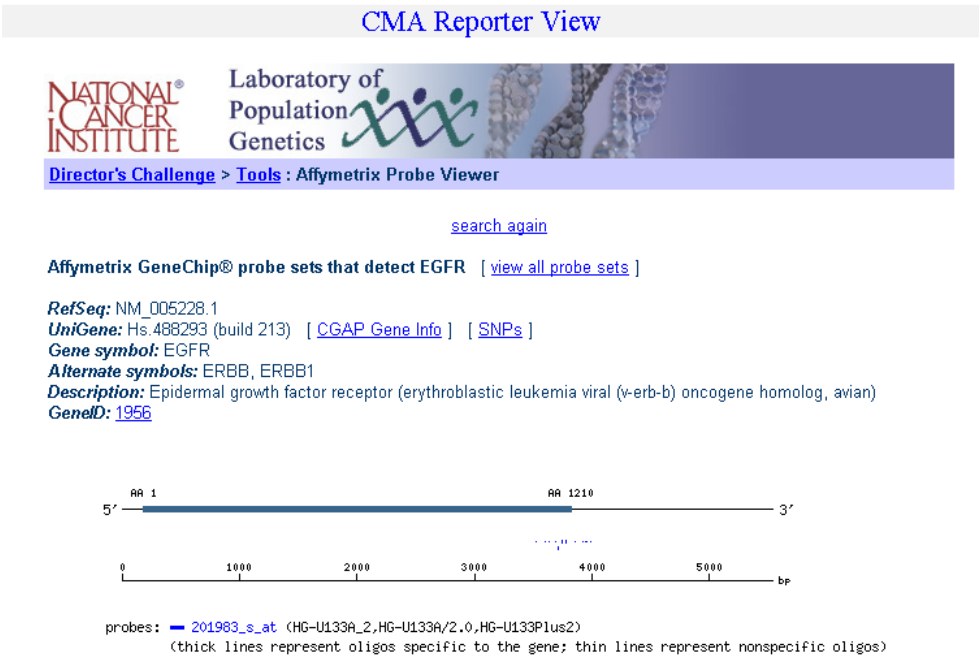


Figure 2.4 Section of a CMA Reporter View – Probesets That Detect EGFR

Log2 Intensity Gene Expression Plot Details

The Log2 Intensity Gene Expression Plot (*Figure 2.5*) displays average expression intensities for the gene of interest based on Affymetrix GeneChip arrays (U133A arrays). Multiple *probesets* (for some genes) are designed to measure the expression

of the gene of interest. For more information on the probeset design strategy for human genes, see the Affymetrix website at http://www.affymetrix.com.

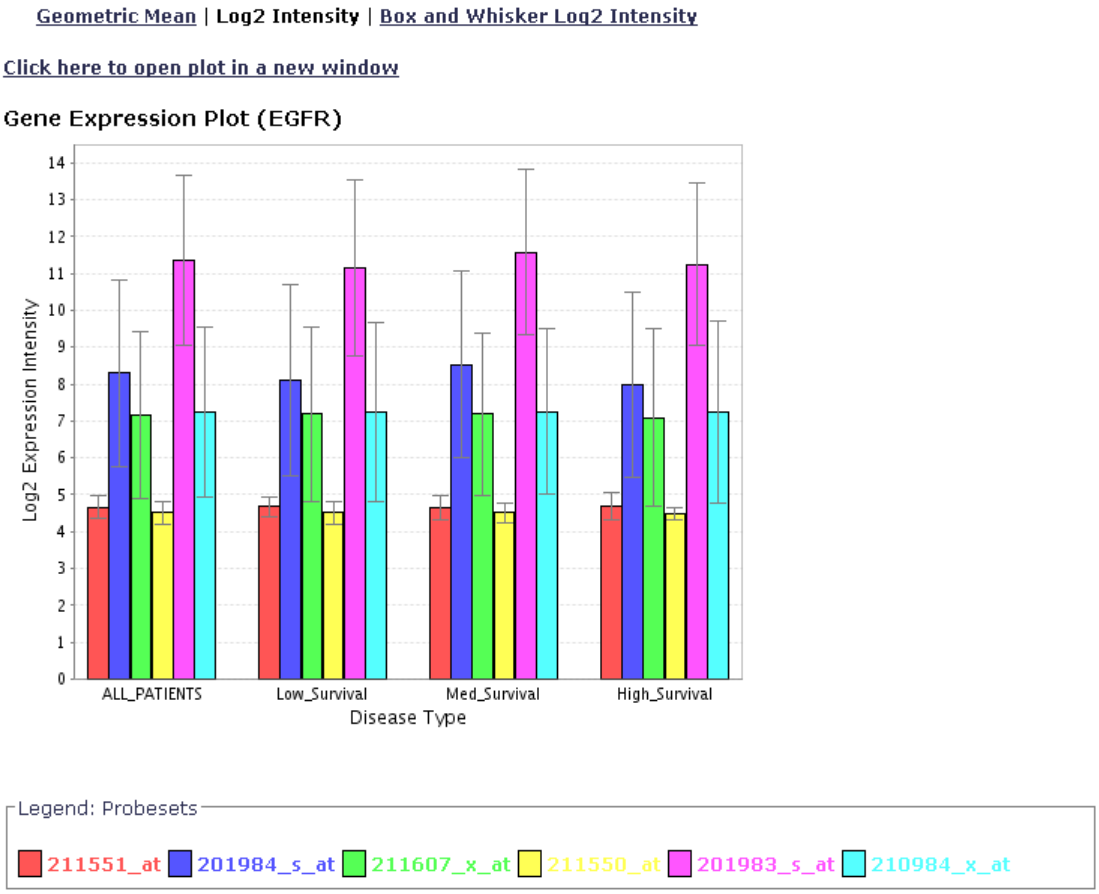


Figure 2.5 Log2 Intensity Gene Expression plot

Box and Whisker Log2 Intensity Gene Expression Plot Details

The Box and Whisker Log2 Intensity Gene Expression Plot (*Figure 2.6*) displays a box plot without all the individual data points for each *sample*. Examples of uses of box and whisker plots include the following:

- Indicate whether a distribution is skewed and whether there are potential unusual observations (outliers) in the dataset
- Compare two or more datasets

- Compare distributions; the center, spread, and overall range are immediately apparent

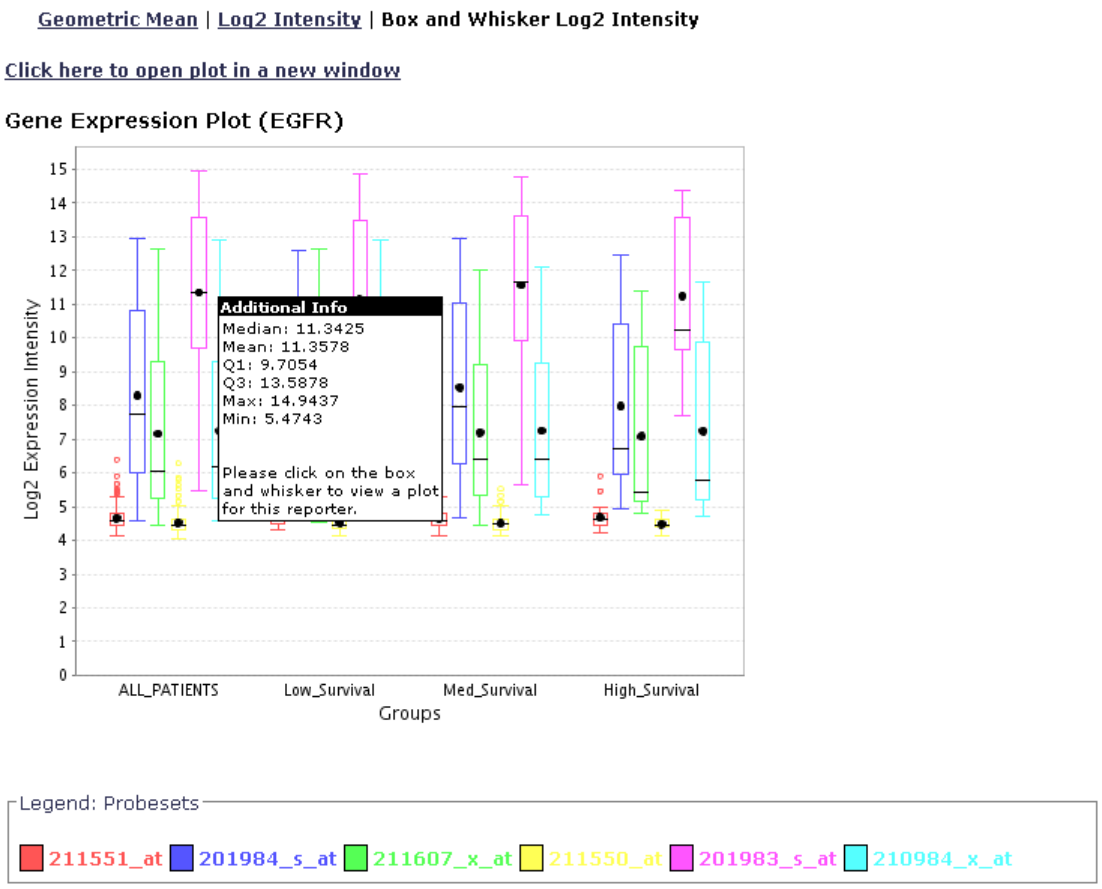


Figure 2.6 Box and Whisker Log2 Intensity Gene Expression plot

A box and whisker plot or box plot is a graph that presents information from a five-number summary. To display the summary about a probeset for one group, hover your cursor over the probeset on the plot to display the Additional Information.

Table 2.5 describes Additional Information details.

Item	***Description***
Median	Median value of log 2 (or ratio) gene expression values for a particular probeset or unified gene
Mean	Mean value of log 2 (or ratio) gene expression values for a particular probeset or unified gene
Q1	The bottom section of the box. The first quartile is the median of the lower part of the data
Q3	The top section of the box. The third quartile is the median of the upper part of the data
Min.	The minimum value
Max.	The maximum value

Table 2.5 Box and Whisker Log2 Intensity Gene Expression Plot information

Item	Description
...plot	Represents the probeset name

Table 2.5 Box and Whisker Log2 Intensity Gene Expression Plot information (Continued)

Note: To display a coin plot for the *gene reporter*, click anywhere in the box. A *coin plot* is a box-and-whisker plot with all the individual data points (see *Displaying a Coin Plot* on page 27).

In the box-and-whisker plot, the individual probeset summary is represented as follows (*Figure 2.7*): Horizontal lines (the "whiskers") extend to, at the most, 1.5 times the box length (the interquartile range) from either or both ends of the box. They end at an observed value, thus connecting all the values outside the box that are not more than 1.5 times the box width away from the box.

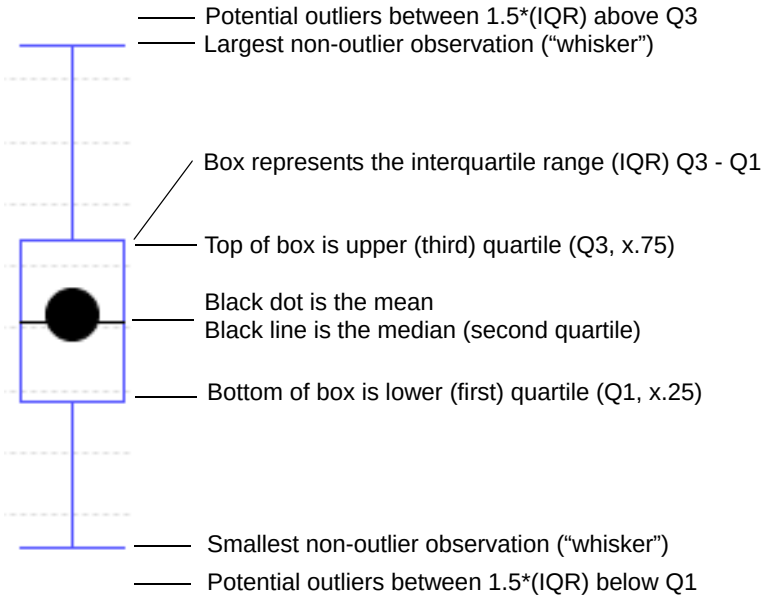


Figure 2.7 Box and Whisker Plot details

Displaying a Coin Plot

A coin plot is a box-and-whisker plot that includes all individual data points. This enables you to obtain a diagram representing a statistical summary of the data without the disadvantage of concealing the real data.

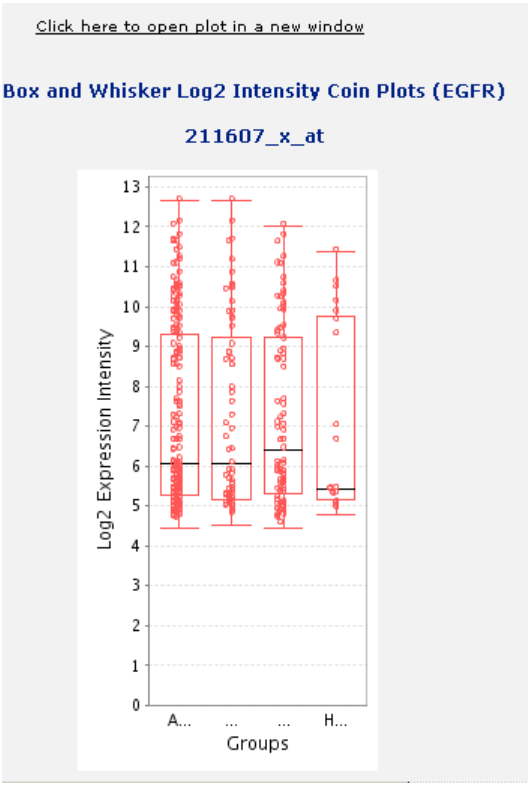


Figure 2.8 Coin Plot for a Probeset

In the coin plot, the individual *probeset* summary is represented as follows (*Figure 2.9*). Horizontal lines (the "whiskers") extend to a maximum of 1.5 times the box length (the interquartile range) from either or both ends of the box. They end at an observed value, thus connecting all the values outside the box that are not more than 1.5 times the box width away from the box.

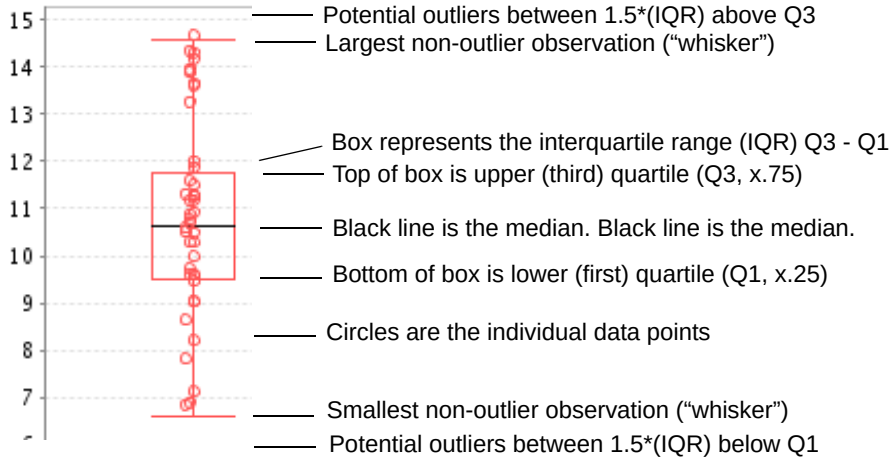


Figure 2.9 Coin Plot Details

Creating Gene Expression-Based Kaplan-Meier Plots

To create a gene expression-based Kaplan-Meier (KM) plot, follow these steps:

1. Do one of the following to navigate to the Gene View workspace:

- From the Home page, click **Gene View**.

  - **or** -

- From any page in the portal, click the **Gene View** tab.

  The Gene View workspace appears.

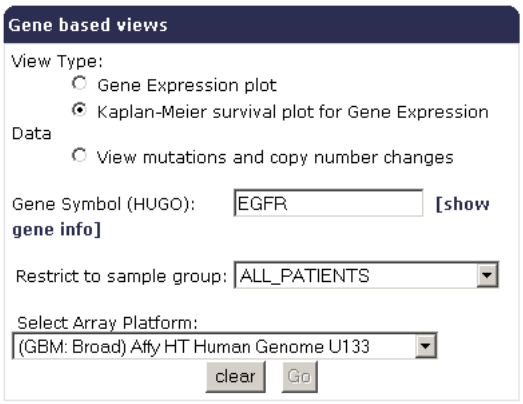


Figure 2.10 Gene View Workspace – Gene-based Views Section

2. In the Gene Symbol (*HUGO*) box, type the gene symbol of interest, for example, *EGFR* or *WT1*, for which you want to generate an expression plot.
3. To view details about the gene of interest, click **show gene info**.

4. From the **Restrict to sample group** drop-down list, select a sample group. Custom patient lists you have saved appear in red text.
5. From the **Select Array Platform** drop-down list, select an array platform.

   Table 2.3 on page 20 provides a description of each of the available arrays.
6. Click **Go**, or, to clear the entries on the page and start over, click **Clear**.

   The Gene Expression-based *Kaplan-Meier Survival* Plot appears.

   If the system does not generate an expression plot, check for messages above the Gene-based View section.

Understanding Gene Expression-Based Kaplan-Meier Plots

A Gene Expression-Based *Kaplan-Meier* plot (*Figure 2.11*) displays the survival rate at each time point for samples with certain expression characteristics (e.g., EGFR expression levels in tumor samples greater than those in the non-tumor samples by 3 fold or higher). Kaplan-Meier estimates are calculated based on the last follow-up (FU) time and the censor status (0=alive, 1=dead) from the samples of interest.The Kaplan-Meier estimates are then plotted against the survival time. The points that correspond to the events with censor status of 0 are indicated on the graph. You can dynamically modify the fold change (up and down regulation) thresholds and redraw the plot.

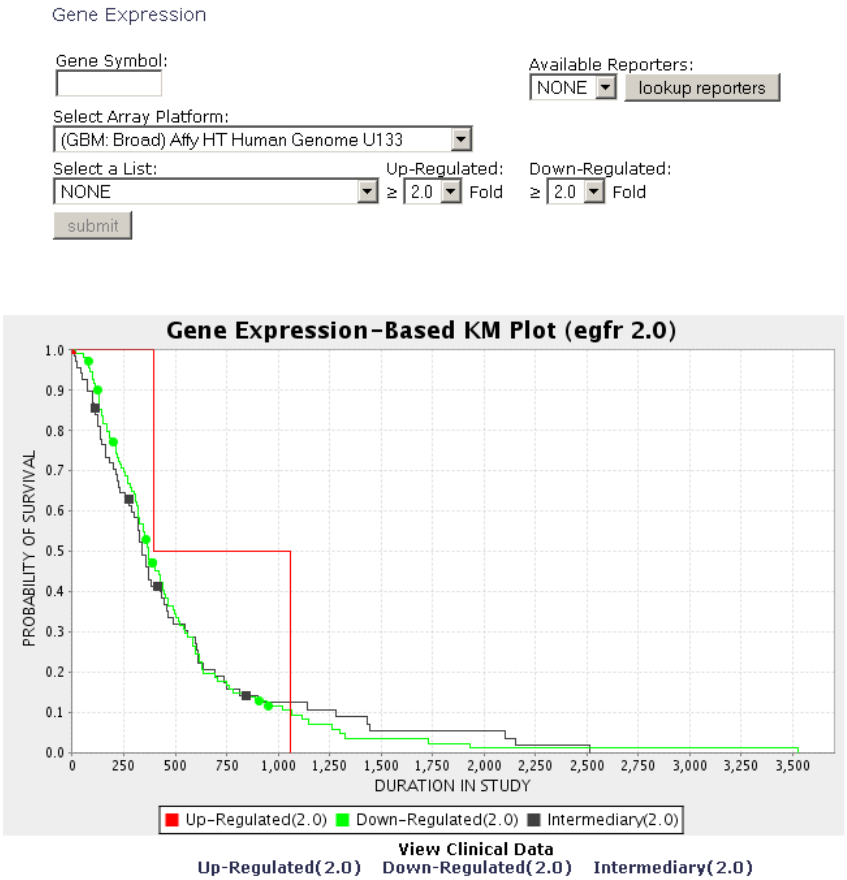


Figure 2.11 Gene Expression-Based KM Plot

Table 2.6 describes areas on the *Kaplan-Meier Survival Plot* for Gene Expression data page.

Item	***Function/Description***
Gene Expression/Copy Number Filter	When you apply a copy number filter, the CMA Portal provides links to display the copy number data for samples
View Clinical Reports	When you apply a gene expression filter, the CMA Portal provides links to display the gene expression for **Up-regulated 2.0**, **Down-regulated 2.0**, and **Intermediary2.0** samples. To generate a clinical report, click the appropriate link. For more information, see *Viewing Clinical Reports* on page 32.

Table 2.6 Gene Expression-Based KM Plot page

Item	Function/Description
Statistical Report	• **Number of Samples** specifies the number of Up-Regulated2.0, Intermediary2.0, Down-Regulated2.0 samples, if any. • **Log-rank *p*-Value** indicates the significance of the difference in survival between any two groups of samples segregated based on gene expression of the gene of interest. The log rank *p*-value is calculated using Mantel-Haenszel method. The *p*-values are recalculated every time a new threshold is selected.

Table 2.6 Gene Expression-Based KM Plot page (Continued)

Creating and Filtering Gene Expression Queries

The criteria selection boxes on the top of the Gene Expression-based *Kaplan-Meier plot* allows you to create and filter queries, and to modify fold change thresholds dynamically.

Figure 2.12 shows query criteria selection boxes for gene expression plots.

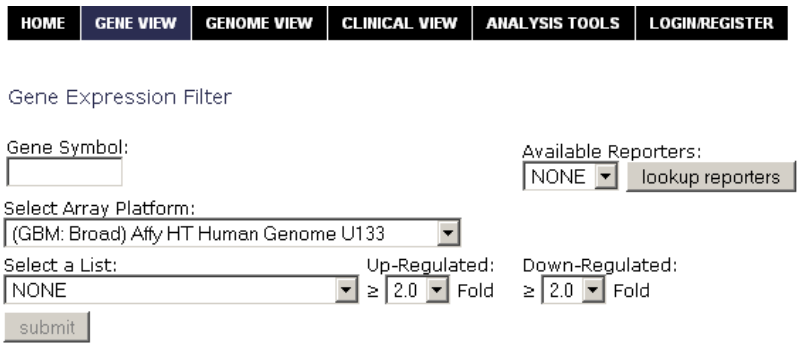


Figure 2.12 Gene Expression Filter

To create the gene expression plot, follow these steps:

1. In the **Gene Symbol** box, type the gene symbol of interest, for example, *EGFR*, for which you want to generate the plot.
2. From the **Select Array Platform** drop-down list, select an array platform.

   Table 2.3 on page 20 provides a description of each of the available arrays.
3. From the **Select a List** drop-down list, select a *sample* group. Patient lists you have saved appear in red.
4. To generate a list of gene reporters, click **look up reporters**.

The list of available gene reporters is displayed in the Available Reporters box.

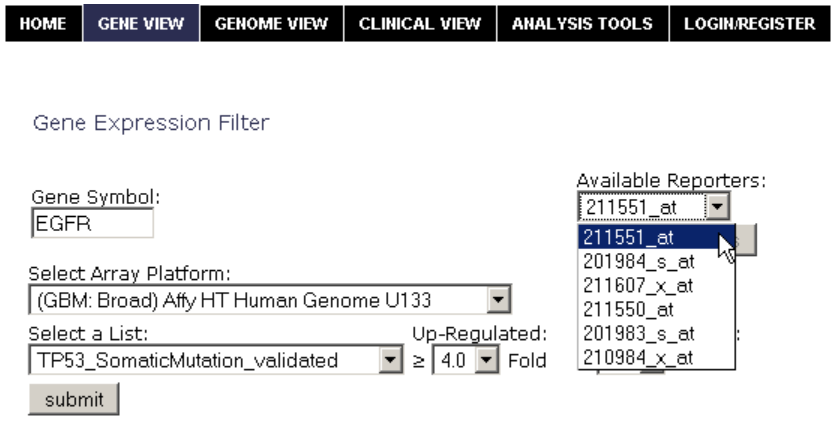


Figure 2.13 Gene Expression Filter – Reporters List

5. Select a gene reporter from the drop-down list.
6. In the **Up-Regulated** criterion box, select the "greater than or equal to" fold change value, and then do the same in the **Down-Regulated** criterion box.
7. Click **Submit**.

   The KM plot is redrawn with the criteria you selected.

Viewing Clinical Reports

Once you have generated a KM survival plot, you can view and download clinical data items for the selected patient list (*Figure 2.15*). On the Clinical Report page, you can save your *sample* group selection to a custom *PatientDID list*.

To view a clinical report, follow these steps:

1. Generate a gene expression-based Kaplain-Meier plot. For detailed instructions, see *Creating Gene Expression-Based Kaplan-Meier Plots* on page 28.

   The KM plot appears, with the clinical report links at the bottom of the graph.

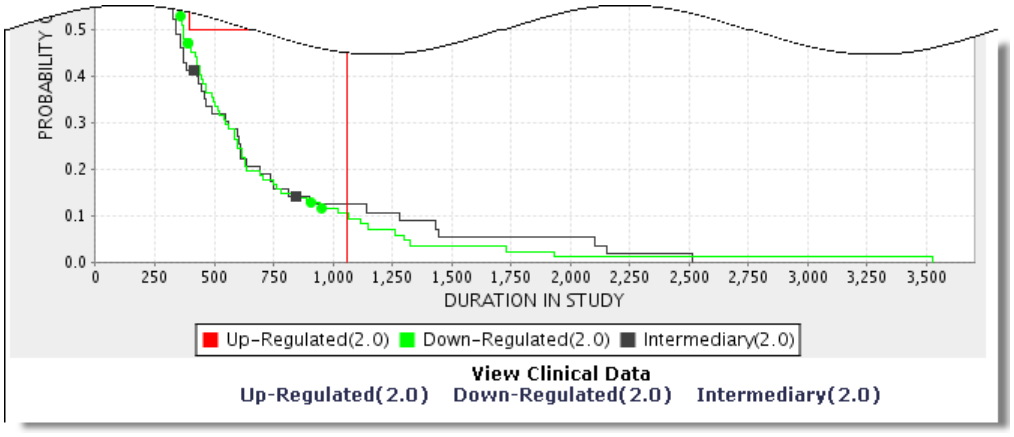


Figure 2.14 Gene Expression-based KM Plot – View Clinical Data

2. Under **View Clinical Data**, select the link for the sample group data of interest, for example, **Down-Regulated(2.0)**.

   The clinical report appears.

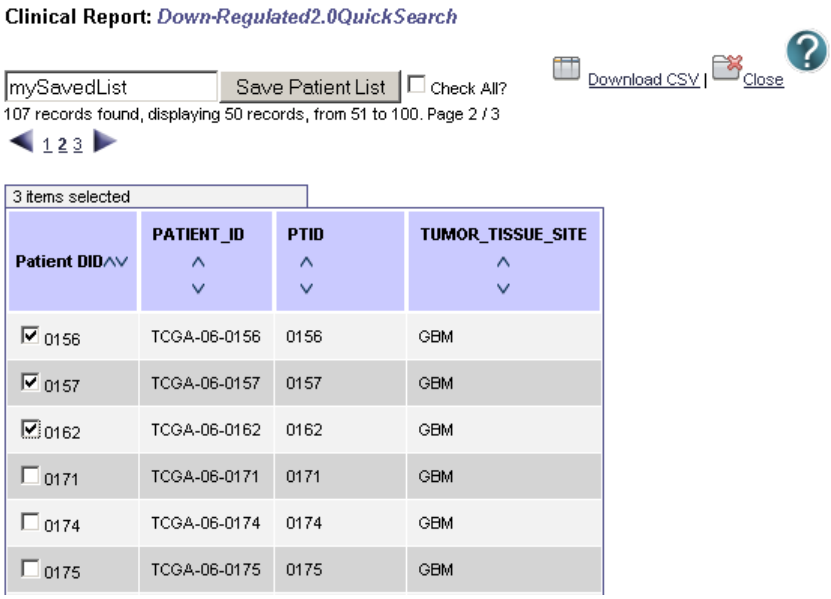


Figure 2.15 Clinical Report Page

3. To sort the data numerically, click the up/down carets (**^**) in the heading row for the column you want to sort. See *Navigating Report Table Results* on page 52 for details.
4. Do one of the following to select patient data for download:
   - To select individual patients, click the check box beside each patient of interest.

   -or-

   - To select all patients, select the **Check All** check box.
   - To clear all selected patients, select the **Check All** check box again.
   - To save the list of patients that you selected, type a unique name in the **mySavedList box**, and click **Save Patient List**.
5. To download a comma-separated value file containing the selected patient data, click **Download CSV**.
6. To close the report, click **Close**.

Viewing Mutation and Copy Number Changes

The Gene View workspace provides a link to the Cancer Genome Workbench (CGWB), an application that allows you to view and drill down into copy number and mutation data of a particular sequence. The CGWB integrates clinical tumor mutation profiles with the reference human genome.

To view mutation and copy number changes, follow these steps:

1. In the **Gene Symbol** box, type the gene symbol of interest, for example, *EGFR*, for which you want to view mutation and copy number data.
2. Click **Go**, or, to clear the entries on the page and start over, click **Clear**.

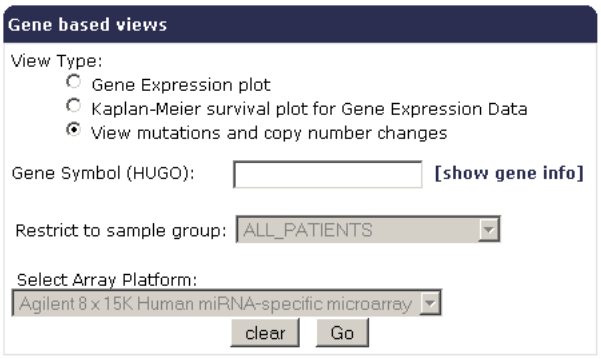


Figure 2.16 Gene-based Views – View Mutation and Copy Number Changes

For instructions on using the CGWB features, refer to the online Help in that application.

Visualizing Pathways

The pathway visualization section of the Gene View workspace provides a graphic depiction of the pathway of interest and provides detailed information about associated genes, agents, and anomalies. See *Working With Pathways and Associated Anomalies* on page 35.

CHAPTER

3

WORKING WITH PATHWAYS AND ASSOCIATED ANOMALIES

This chapter describes how to use the Cancer Molecular Analysis Portal to review pathways and the anomalies associated with them.

Topics in this chapter include:

OverviewThe pathway visualization section of the Gene View workspace provides access to graphic depictions of pathways of interest and provides detailed information about associated genes, agents, and anomalies.

Generating Pathway Diagrams

Pathway diagrams enable you to investigate genetic anomalies at a cellular level.

To generate a pathway diagram, follow these steps:

1. Do one of the following to navigate to the Gene View workspace:

- From the Home page, click **Gene View**.

  - **or** -

- From any page in the portal, click the **Gene View** tab.

The Gene View workspace appears.

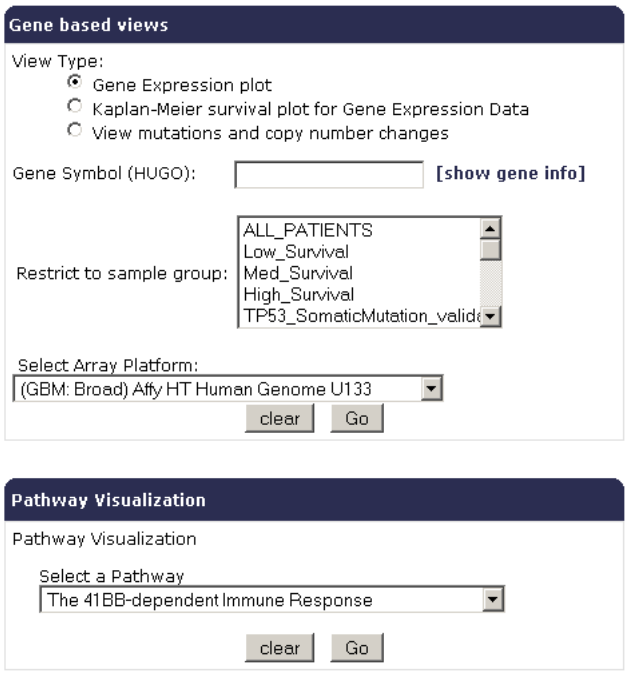


Figure 3.1 Gene View – Default Workspace

2. Under **Pathway Visualization**, select the pathway of interest from the **Select Pathway** drop-down list. *Table A.2* on page 74 lists the pathways that are available for investigation.
3. Click **Go**. Or, to return to the top of the pathway list, click **Clear**.

The Pathway Visualization and Pathway Gene Anomalies page appears in a new browser.

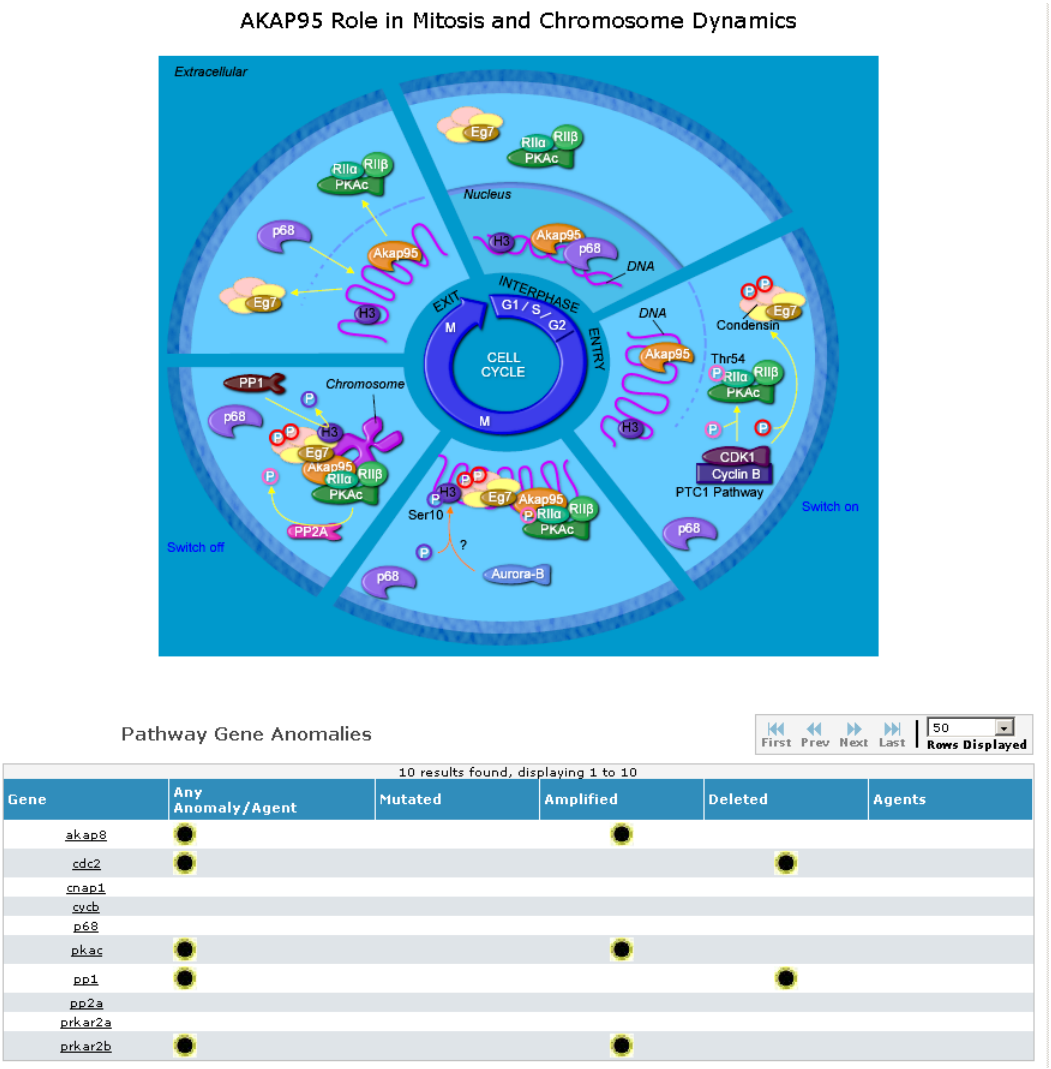


Figure 3.2 Pathway Visualization and Pathway Gene Anomalies Page

Note: If the pathway diagram is not displayed, install the *SVG plugin* on your computer. For more information about installing the SVG plugin, see *Installing the SVG Plugin* on page 13.

The Pathway Visualization and Pathway Gene Anomalies page is divided into 2 sections as follows:

- An interactive diagram of the pathway of interest appears at the top of the page. The basis for the visualization is a BioCarta pathway diagram. The selected pathway is depicted as if projected onto a morphological illustration of a cell. It depicts gene-to-gene and other molecular interactions in a graphical interface.
- A Pathway Genes Anomalies table appears at the bottom of the page. It lists the genes for the selected pathway and the anomalies associated with each of them.

Investigating Genes Via Pathway Diagrams

The pathway diagram contains hyperlinks that allow you to investigate each gene associated with the pathway. The graphic representations of the genes link to, and launch, the UCSC Genome Bioinformatics website's Genome Browser, which provides detailed information about the genes of interest.

Additionally, the pathway diagram interacts with the Gene Pathways Anomalies table at the bottom of the page so that you can isolate genes that are of interest to you.

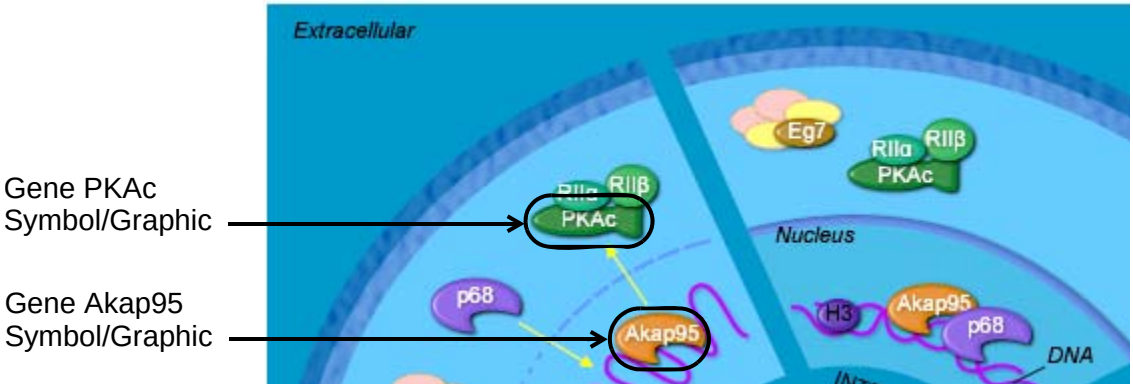


Figure 3.3 Pathway Diagram – Close-up View of Gene Symbol PKAc

To display information about a given gene, follow these steps:

1. To investigate a specific gene via the UCSC Genome Browser, click its symbol/graphic on the diagram. For example, to investigate the PKAc gene, click any of the graphic symbols labeled **PKAc**.

   The UCSC Genome browser opens and displays information about the selected gene. For complete documentation, click the Help button on the UCSC Genome Bioinformatics website.

2. To highlight the genes associated with a given column of data in the Gene Pathways Anomalies table, scroll down to the table and click the column heading. For example, to isolate the *amplified* genes on the Biocarta diagram, click the **Amplified** column heading.

All amplified genes are highlighted in the diagram. Compare *Figure 3.4* with *Figure 3.2* on page 37.

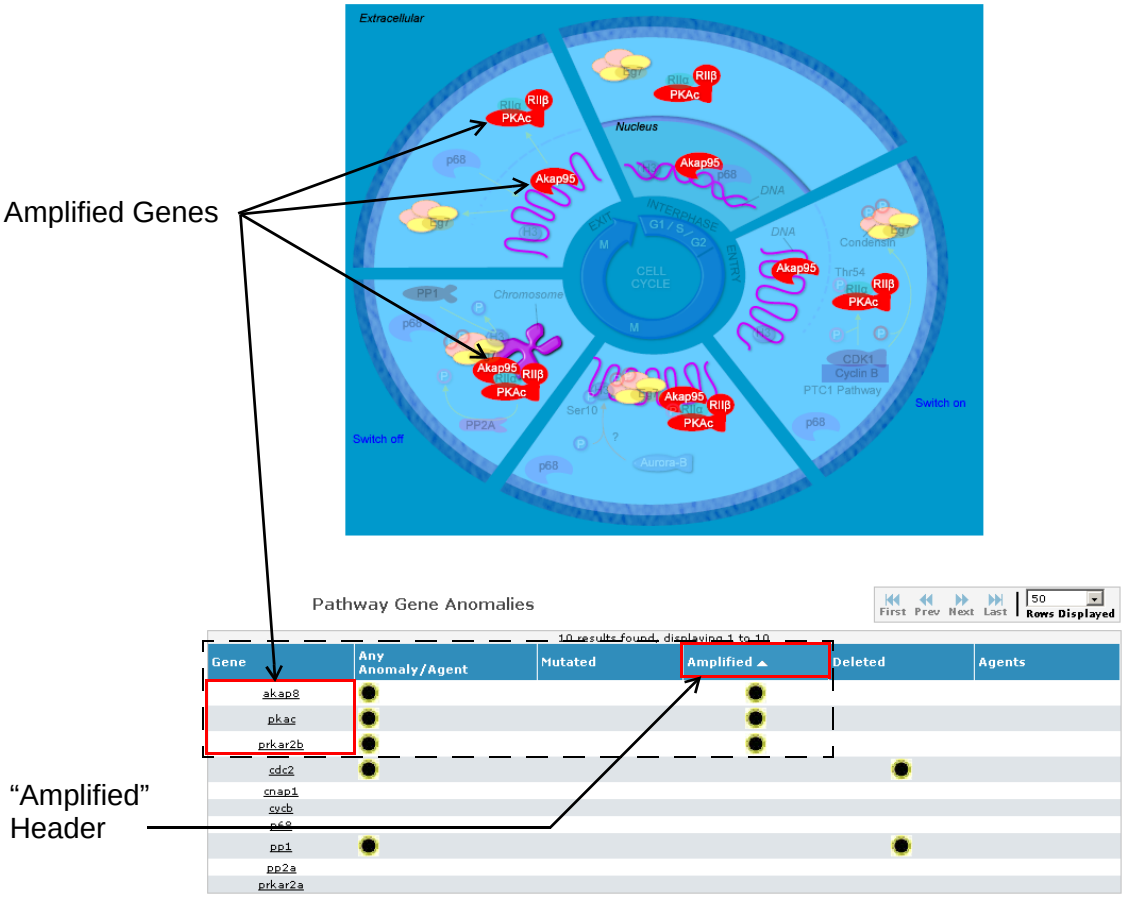


Figure 3.4 Pathway Diagram – Isolated Amplified Genes

3. To review genes for the selected pathway and the anomalies, mutations, and agents associated with them, scroll down to the Pathway Gene and Anomalies table, and follow the instructions in *Investigating Genes Via Pathway Gene Anomalies Tables* on page 39.

Investigating Genes Via Pathway Gene Anomalies Tables

The Pathway Gene Anomalies table (*Figure 3.5*) lists the genes for the selected pathway alphabetically, and indicates the anomalies associated with each gene with a

black dot. Any agents associated with a gene are listed in the Agents column. They are hyperlinked to the NCI Thesaurus.

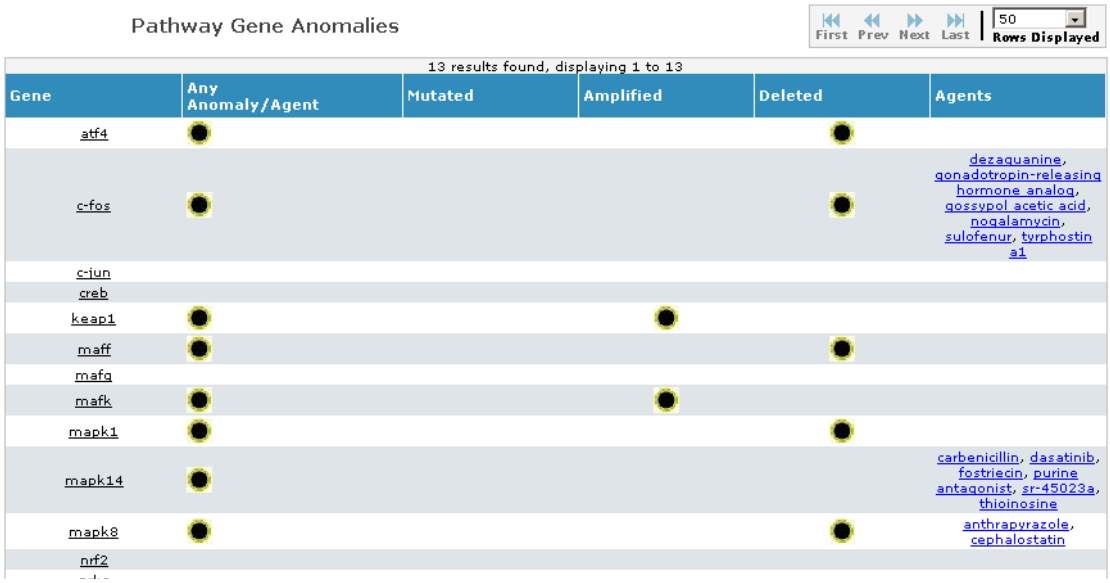


Figure 3.5 Pathway Gene Anomalies table

Table 3.1 describes each column on the Pathways and Associated Anomalies table.

Column Name	***Description***
Gene	Genes that are associated with a given pathway
Any Anomaly/Agent	A black dot appears when a gene within the pathway has an anomaly and/or drug agent associated with it.
Mutated	A black dot appears when genes within the pathway are categorized as *mutated*.
Amplified	A black dot appears when the computed *copy number* of the gene is greater than, or equal to, 2.5.
Deleted	A black dot appears when genes within the pathway have been *deleted*.
Overexpressed	A black dot appears when genes within the pathway are categorized as *overexpressed*.
Underexpressed	A black dot appears when genes within the pathway are categorized as *underexpressed*.
Agents	A black dot appears when genes within the pathway are associated with a drug agent.

Table 3.1 Column descriptions for Pathway Genes and Anomalies

To display detailed information about genes and agents, follow these steps:

1. Scroll to the bottom of the **Pathway Visualization and Pathway Gene Anomalies** page.
2. For detailed gene information, click the gene name hyperlink.

A new window displays the NCBI Entrez Gene page with detailed information about the selected gene.

3. For detailed agent information, click the agent name hyperlink.

   A new window displays the NCI Thesaurus page with detailed information about the selected agent.

Navigating the Gene Anomalies Table

The Gene Anomalies table provides several mechanisms for viewing and sorting the information provided.

By default, the table is sorted alphabetically by gene. Black dots in a Gene row indicate the anomaly(ies) with which a given gene is associated.

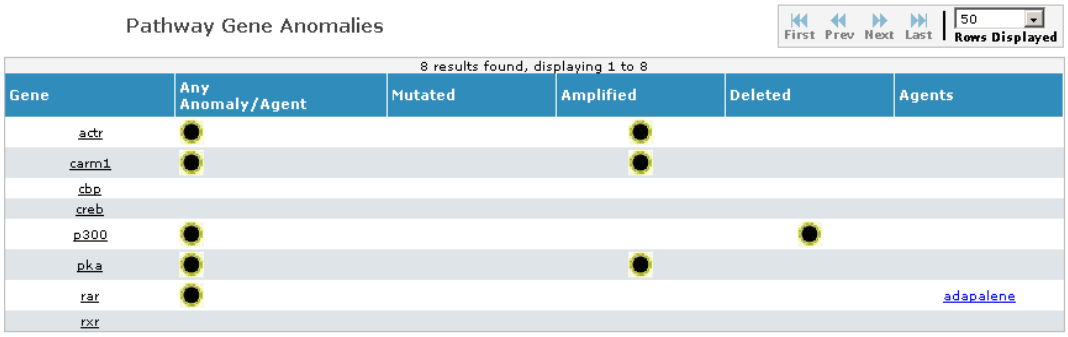


Figure 3.6 Gene Anomalies Table – Sorted Alphabetically

To change the sort order and your view of the table, follow these steps:

1. To sort the table by anomaly, click the column heading for the anomaly of interest. For example, click the **Amplified** column heading.

   An “up” arrow appears next to the column name and the table is re-sorted such that the genes that have been amplified appear at the top of the Gene column.

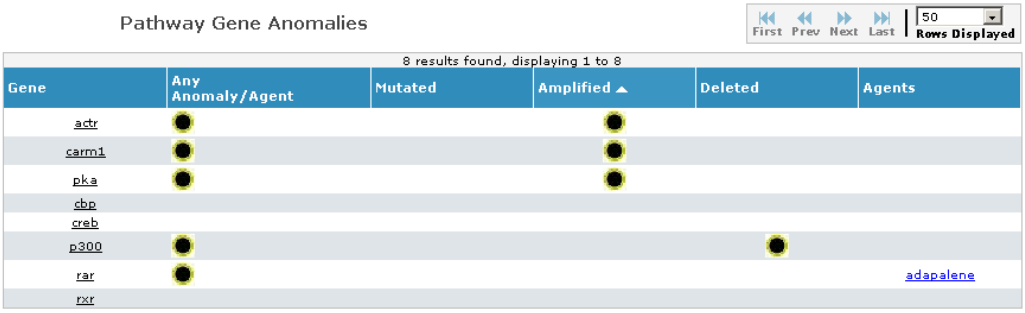


Figure 3.7 Gene Anomalies Table – Sorted by Amplification

2. To move all genes with anomalies to the top of the column, click the **Any Anomaly/Agent** column name.
3. To move all genes without anomalies to the top of the column, click the **Any Anomaly/Agent** column name again.

To display a longer list of items on one page, click the **Rows Displayed** list box () just above the table on the right side. Select the number of items to display at once, either **50** or **100**.

Figure 3.8 Pathway Gene Anomalies Table Navigation Toolbar

The number of genes associated with a given pathway can exceed the limit that can be displayed on one page. The navigation arrows become active when the gene list exceeds one page, and allow you to move from page to page.

4. To move to another page of the list, click one of the arrow buttons next to the **Rows Displayed** list. You can move directly to the first or last page, or move to a previous page or subsequent page.

CHAPTER

4

WORKING WITH GENOME VIEWS

This chapter describes how to use the Cancer Molecular Analysis Portal to investigate chromosomal regions of amplification, deletion, and over-expression.

Topics in this chapter include:

About Genomic Views of Data

The CMA Portal enables you to explore data in one genome-level visualization and to investigate chromosomal regions of *amplification*, *deletion*, and *over-expression* via the integrated Heatmap Viewer.

Characterizations include the following types of data:

- Copy number
- Gene expression
- Methylation
- Mutation

Note: A Java plugin is required for viewing heatmaps.

Do one of the following to navigate to the Genome View workspace:

- From the Home workspace, click **Genome View**.

- **or** -

- From any workspace in the portal, click the **Genome View** tab.

The Genome View workspace appears.

Viewing Copy Number Data

Differences in the number of copies of certain genes contributes to genetic variability. Variability can be caused by deletions of some genes on only one chromosome, or multiple copies of some genes.

Copy number data in CMA Portal is provided by the following sources:

- Broad Institute (Broad)
- Harvard Medical School (HMS)
- Memorial Sloan-Kettering Cancer Center (MSKCC)
- Stanford University School of Medicine (SUSM)

Note: Copy numbers are also known as copy number variants (CNVs) and copy number polymorphisms (CNPs)

For more info on characterization sources, refer to TCGA site at: http://cancergenome.nih.gov/data/types/genomic/

*Table 4.1*Lists the CGCCs and the platforms used to derive their genome copy number data.

Broad	***HMS***	***MSKCC***	***SUSM***	***Combined Data***
Affymetrix	Agilent	Agilent	Illumina	Combined
Paired Affymetrix	Paired Agilent	Paired Agilent	Paired Illumina	Combined Paired
Unpaired Affymetrix	Unpaired Agilent	Unpaired Agilent	Unpaired Illumina	Combined Unpaired
Normal Affymetrix	Normal Agilent	Normal Agilent		

Table 4.1 Genome copy number data sources and platforms

*Table 4.1*Lists the CGCCs and the platforms used to derive their gene copy number data

Broad	***HMS***	***MSKCC***	***SUSM***
Affymetrix	Agilent	Agilent	Illumina
Paired Affymetrix	Paired Agilent	Paired Agilent	Paired Illumina
Unpaired Affymetrix	Unpaired Agilent	Unpaired Agilent	Unpaired Illumina

Table 4.2 Gene copy number data sources and platforms

To view copy number data, in the Genome workspace, click a CGCC/platform link.

The Heatmap Viewer displays a heatmap of the selected genes.

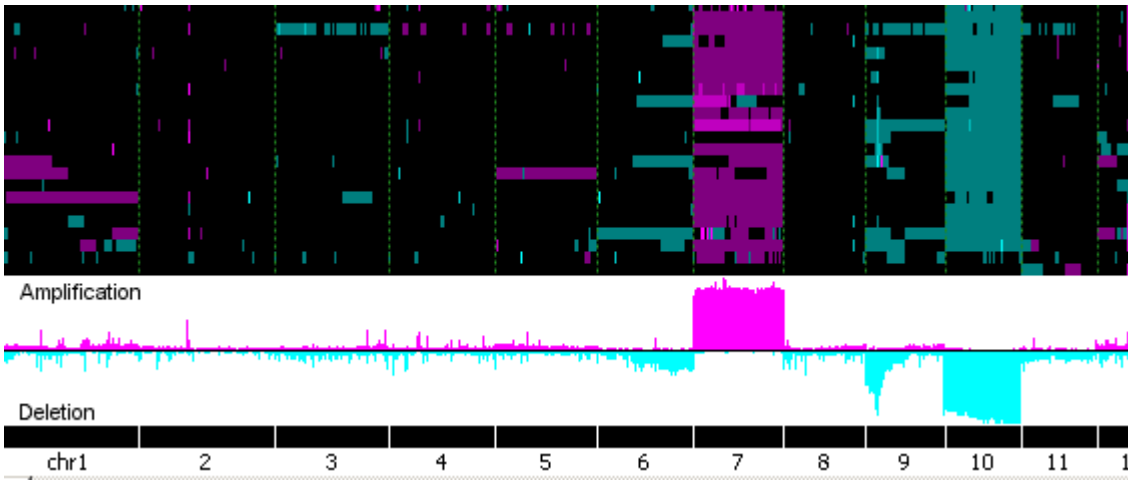


Figure 4.1 Heatmap Viewer – Section of Copy Number Heatmap

For Heatmap Viewer instructions, see *Navigating the Heatmap Viewer* on page 46.

Viewing Gene Expression Data

Gene expression data in CMA Portal is derived from the following sources:

- Gene expression of 188 GBM samples using Affymetrix U133. Submitted by the Broad Institute
- Gene expression of 188 GBM samples using Affymetrix U133 organized by 1000 genes with variable expression

To view gene expression data, click the a CGCC/platform link.

The Heatmap Viewer displays a heatmap of the selected genes.

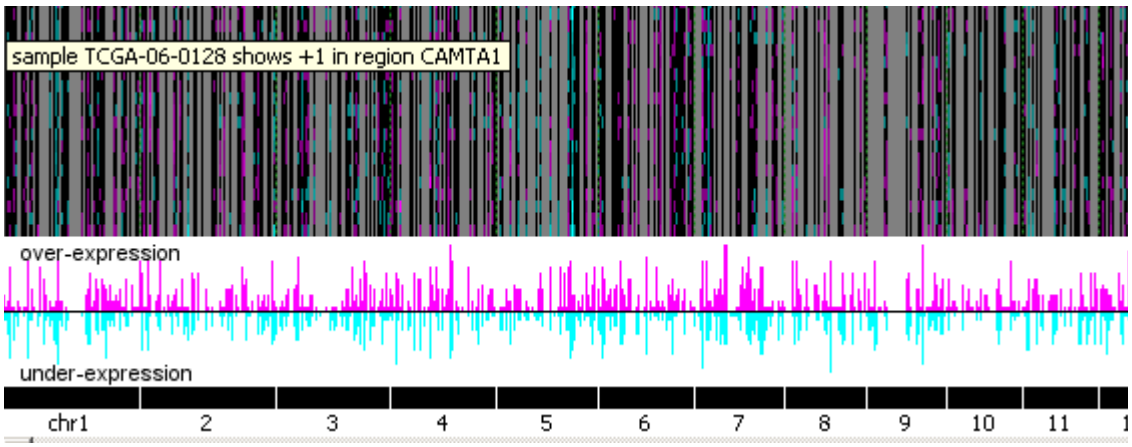


Figure 4.2 Heatmap Viewer – Section of Gene Expression Heatmap

For Heatmap Viewer instructions, see *Navigating the Heatmap Viewer* on page 46.

Viewing Methylation Data

Methylation data in CMA Portal is derived from the following sources:

- DNA methylation of 169 GBM samples in 1961 genes using Illumina. Submitted by the Sidney Kimmel Comprehensive Cancer Center At Johns Hopkins University (JHU-USC)
- DNA methylation of 169 GBM samples in 2178 genes using Illumina. Submitted by the Sidney Kimmel Comprehensive Cancer Center At Johns Hopkins University (JHU-USC)

To view gene methylation data, click one of the links under **Genomic View of Methylation Data**.

The Heatmap Viewer displays a heatmap of the selected genes.

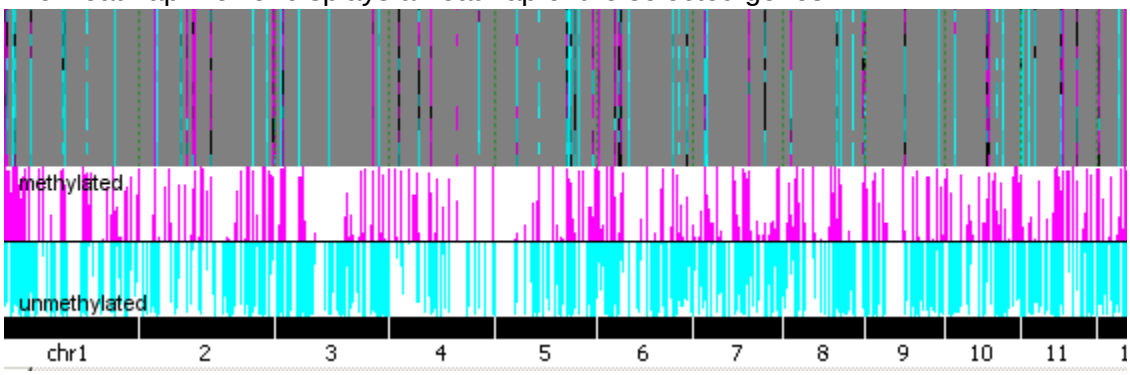


Figure 4.3 Heatmap Viewer – Section of Methylation Heatmap

For Heatmap Viewer instructions, see *Navigating the Heatmap Viewer* on page 46.

Viewing Mutation Data

The link to the Cancer Genome Workbench launches a genome view of data. For assistance with using the Cancer Genome Workbench, refer to the documentation provided in the Genome Workbench browser.

Navigating the Heatmap Viewer

Controls in the Heatmap Viewer allow you to zoom in to any location, determine the heatmap's color and contrast, and to research your data further through links to the Cancer Genome Workbench.

To access the Heatmap Viewer documentation, click the **Help** button at the top of the Heatmap Viewer, and select **Documentation**.

CHAPTER

5

WORKING WITH CLINICAL VIEWS

This chapter describes how to use the Cancer Molecular Analysis Portal to study clinical data and explore the relationships between clinical and molecular study data.

Topics in this chapter include:

Specifying Clinical Search Criteria

CMAP provides several tools for querying, visualizing, and downloading patient and *sample* data. You can search for data directly from the Clinical View workspace or via the Data Access Matrix.

Specifying Criteria in the Clinical View Workspace

To search for patient data in the Clinical View workspace, follow these steps:

1. Do one of the following to navigate to the Clinical View workspace:

- From the Home workspace, click **Clinical View**.

  - **or** -

- From any workspace in the portal, click the **Clinical View** tab.

  The Clinical View workspace appears. Preselected *open-access data* appears in the Selected Groups Box.

  Note: *Open-access data* groups are available to all CMA Portal users. Other search criteria are available to registered viewers only. This serves to protect patient privacy. For more information about patient privacy and

data access policy, see TCGA Program Components page at: http://cancergenome.nih.gov/components/hsp.asp

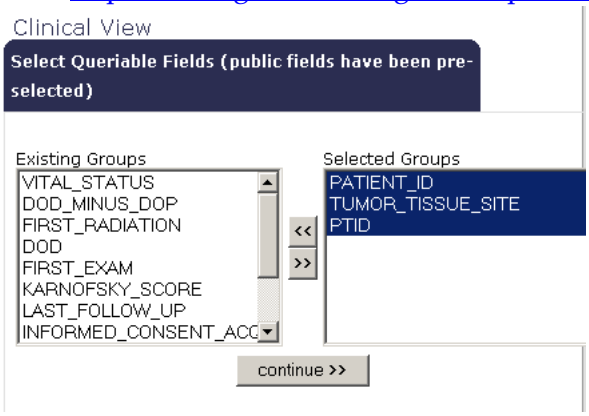


Figure 5.1 Clinical View – Group Selection Criteria

2. Select the groups to query from the **Existing Groups** list by doing any of the following:
   - To select a single group, click the group name.
   - To select several contiguous groups, click the first group, press and hold the SHIFT key, and select the last group in the series. Or, drag your cursor from the first group to the last group in the series.
   - To select several discontinuous groups, click one group and CTRL + click additional groups.
3. Click **>>** to move the selected groups to the **Selected Groups** list.
4. To remove a group from the **Selected Groups** list, select the name and click **<<** to return it to the **Existing Groups** list.
5. Click **Continue**.

The Clinical View criteria filter page appears. The number and types of criteria available depend on the groups you selected for your query. Your set of filters may look different than the ones displayed in *Figure 5.2*.

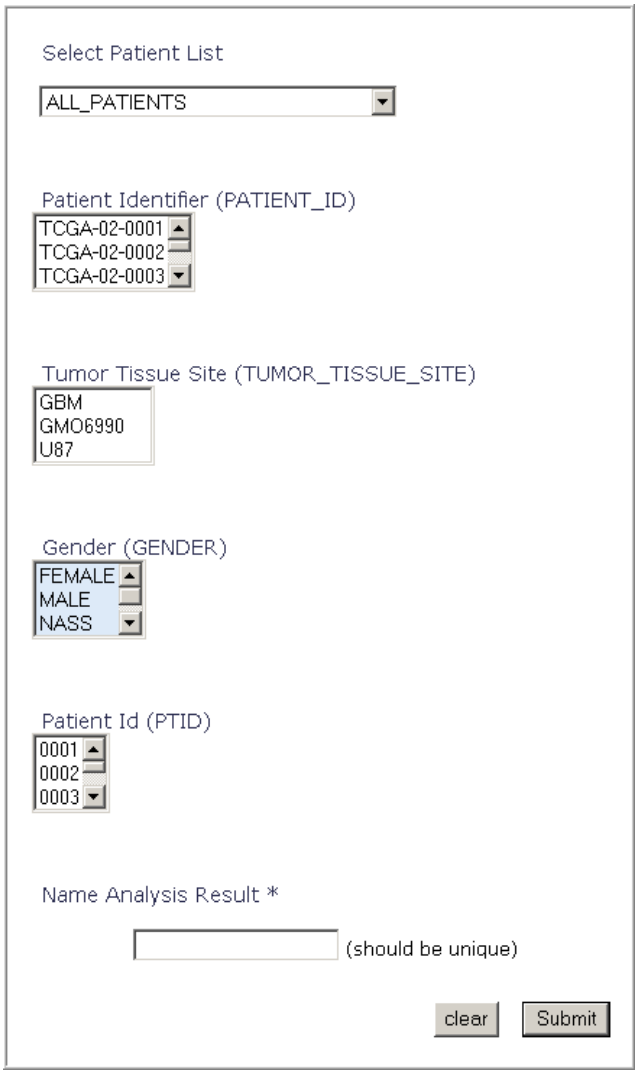


Figure 5.2 Clinical View – Criteria Filter Page

Filtering Clinical Search Criteria

The Clinical Query Filter page enables you to refine your query.

To specify search criteria, follow these steps:

1. Follow the instructions in *Specifying Clinical Search Criteria* on page 47 to select the groups of data you want to query in order to access the list of filters.
2. On the Criteria Filter page (*Figure 5.2* on page 49), in the **Select Patient List**, select a patient group to query. Available groups of patients with certain characteristics are the same as those listed in the *PatientDID* Lists shown on the sidebar. If you created and added a custom list, the list name appears in red text in the list of patients. See *About PatientDID Lists* on page 68.

3. Continue to select criteria from the options displayed. The number/type of criteria vary according to the groups you selected in your initial query criteria selection. See *Specifying Clinical Search Criteria* on page 47.

4. In the **Name Analysis Result** box, type a unique name for the query.

   Caution: You will lose prior query data if you enter the same name as a previous query during the current session.

   To clear the values entered on the page and enter new values, click **Clear**.

5. To submit the query, click **Submit**.

   CMA Portal processes your query and displays **completed** when the result has been returned.

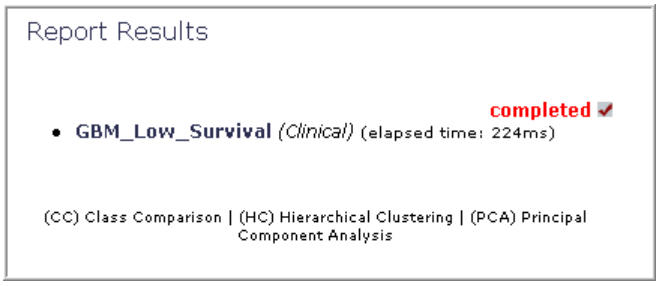


Figure 5.3 Clinical View – Report Results Page

   The CMA Portal stores the query for the duration of your current CMA Portal session.

6. To view the results of a query, click the query name (link).

   The results appear in a new browser window.

For information about the results, see *Viewing Clinical Reports* on page 32.

Navigating To the Report Results Page.

Query results are generated on the View Results page, and appear there immediately after you have submitted your query. You can return to the Report Results page at any time during your session.

To view the Report Results page, on the sidebar, click **View Results**. See *Side Bar Elements* on page 11.

Working With Clinical Report Results

Results of your sample-based query appear in table format on the Clinical Report page.

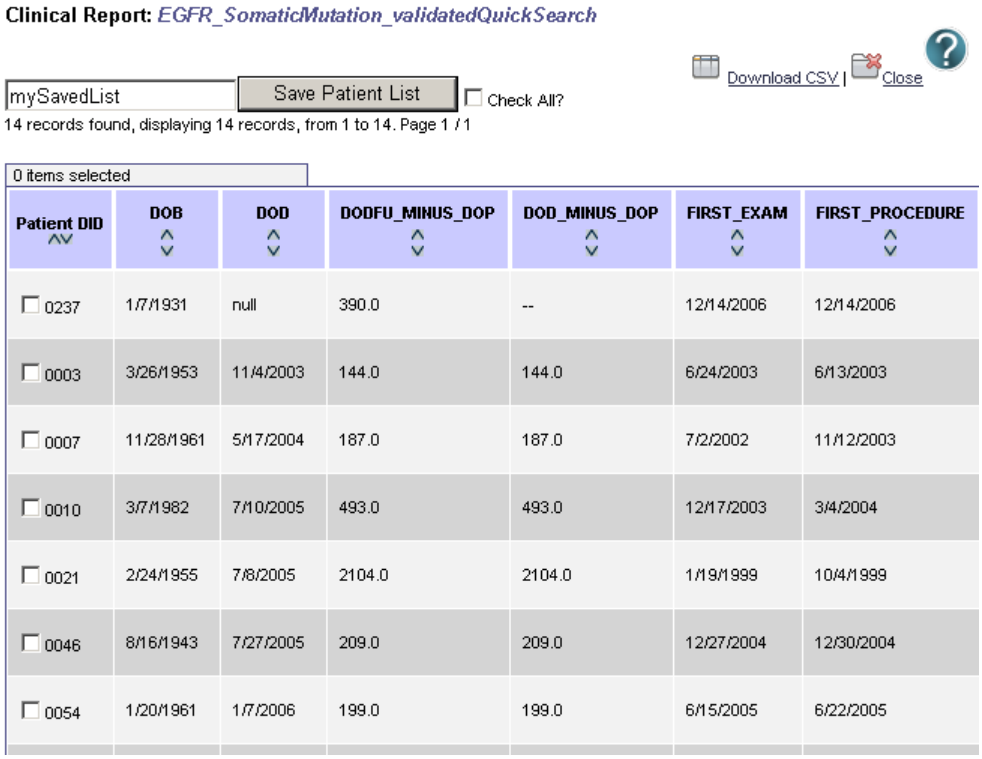


Figure 5.4 Clinical Report Page

Table 5.1 lists the types of patient information available for each *sample* in your query. The type of data displayed in the report table reflects the criteria and filters you selected for your query.

Column Header	***Definition/Description***
PatientDID	Patient de-identified ID – Data that has been disassociated from a patient's personally identifiable information (PII). See *About PatientDID Lists* on page 68.
DOB	Date of birth – Calendar date of a patient's birth
DOD	Date of death – Calendar date of a patient's death
DODFU Minus DOP	Date of death or last followup minus the date of the procedure – Number of days between the date of the patient's first procedure and date of death
DOD Minus DOP	Date of death minus the date of the procedure – Number of days between the date of the patient's first procedure and date of death
First Exam	Date of the patient's first examination
First Procedure	Date of the patient's first procedure
First Radiation	Date of the patient's first radiation treatment

Table 5.1 Clinical Report patient information

Column Header	Definition/Description
Gender	Patient's gender
Informed Consent Acquired?	Indicates whether or not the patient's consent to provide samples was acquired
Karnofsky Score	Karnofsky Performance status scale – represents the patient's functional capabilities
Last Followup	Date of the patient's last followup examination
Patient_ID	Patient identifier – Number associated with a patient as per TCGA code standards
PTID	Patient identifier – Number associated with a patient without the TCGA prefix
Tumor Tissue Site	Location of the tumor
Vital Status	Indicates whether the patient is alive or dead

Table 5.1 Clinical Report patient information (Continued)

Navigating Report Table Results

To assist you in selecting result records and analyzing the data, you can do the following:

- View multiple pages of records
- Sort the records by any column
- Select records and download result data
- Create a *PatientDID* list to use in other queries

displays the tools for accomplishing these tasks.

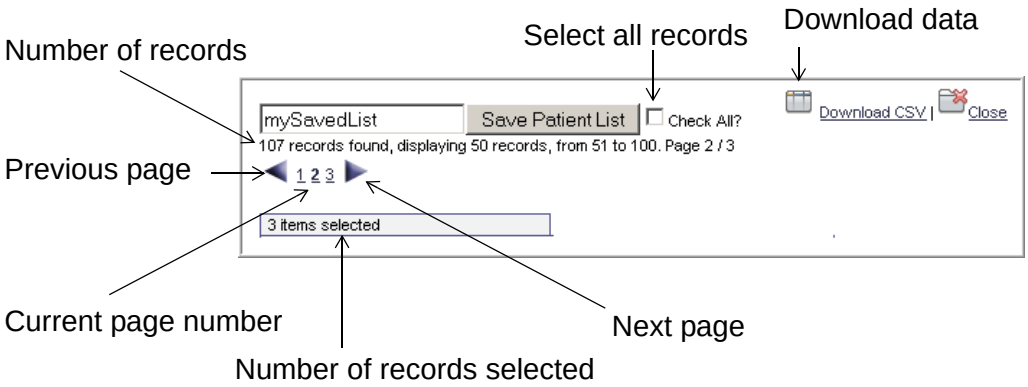


Figure 5.5 Clinical Report Table – Tools

Viewing Multiple Pages of Records

To view a subsequent page of the results table, click the **Next** arrow.

To view a previous page of the results table, click the **Previous** arrow.

Sorting Records

You can sort the results table by any of the columns that display up/down carets (^) in the headers.

To sort a column in a report table, follow these steps:

1. To sort a column in ascending order, click the caret pointing up.
2. To sort a column in descending order, click the caret pointing down.

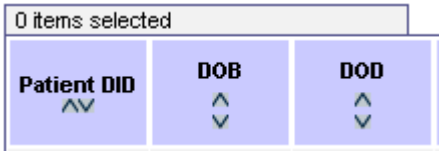


Figure 5.6 Sort Results by Column

Selecting and Downloading Sample-based Clinical Report Records

To select and download report records, follow these steps:

1. Do one of the following to select sample records for download:
   - To select individual records, click the check box beside each desired record.

   -**or**-

   - To select all records, select the **Check All** check box. This check box acts like a toggle switch. To clear all selected records, select the **Check All** check box again.
2. To save the list of patient records that you selected, type a unique name in the **mySavedList** box, and click **Save Patient List**.
3. To download a comma-separated value file containing the selected records, click **Download CSV**. When prompted, open or save the file to your computer.
4. To close the report, click **Close**.

Creating a PatientDID List From a Sample-based Clinical Report

From the Clinical Report page, you can select patients and save them to a customized *PatientDID* list to use as criteria in future queries.

To create a PatientDID List, follow these steps:

1. Select one or more patients to save to the list. See *step 1.* on page 53.
2. To save the selected records to a *PatientDID* list, type a unique name in the **mySavedList** box, and click **Save Patient List**.

   The new *sample* set is listed in red text in the *PatientDID* List in the side bar.

Creating Sample-Based KM Plots

To create a sample-based *Kaplan-Meier* plot, follow these steps:

1. Do one of the following to navigate to the Clinical View page:

- From the Home page, click **Clinical View**.

  - **or** -

- From any page in the portal, click the **Clinical View** tab.

  The Clinical View page displays the Kaplan-Meier parameters at the bottom of the page.

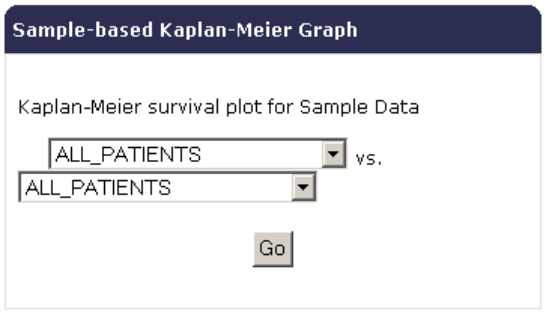


Figure 5.7 Clinical View – Sample-based Kaplan-Meier Graph Criteria

2. From the drop-down lists, select *sample* patient groups for comparison purposes.
3. Click **Go**.

The Sample-Based Kaplan-Meier survival plot appears.

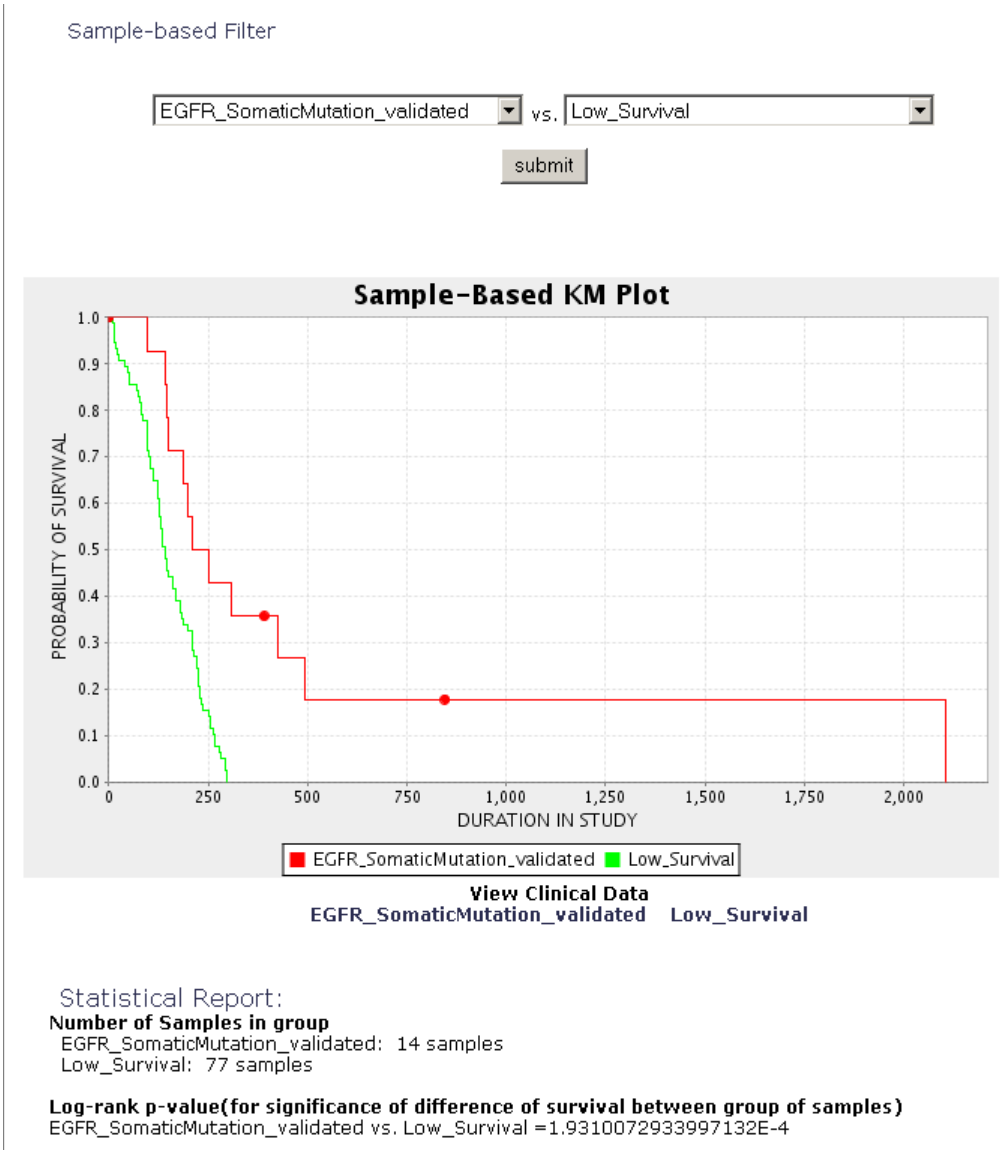


Figure 5.8 Sample-Based KM PLot

4. To compare other *sample* groups, select the two groups of interest from the Sample-based Filter drop-down lists at the top of the page, and then click **Submit**.

   A new Sample-Based Kaplan-Meier survival plot appears.

5. To view clinical data associated with the sample groups you selected, directly below the plot, click a sample group link, for example, EGFR_Somatic_validated.

   The Clinical Report appears in a new browser window. See *Creating a PatientDID List From a Sample-based Clinical Report* on page 53.

Understanding Sample-Based KM Plots

A Sample-Based *Kaplan-Meier* plot *(Figure 5.9)* shows the survival rate at each time point for *samples* with certain genome characterization characteristics (e.g., EGFR mutation levels in tumor samples are greater than those in the non-tumor samples by 3 fold or higher). Kaplan-Meier estimates are calculated based on the last follow-up time and the censor status (0=alive, 1=dead) from the samples of interest. The Kaplan-Meier estimates are then plotted against the survival time. The points that correspond to the events with censor status of 0 are indicated on the graph.

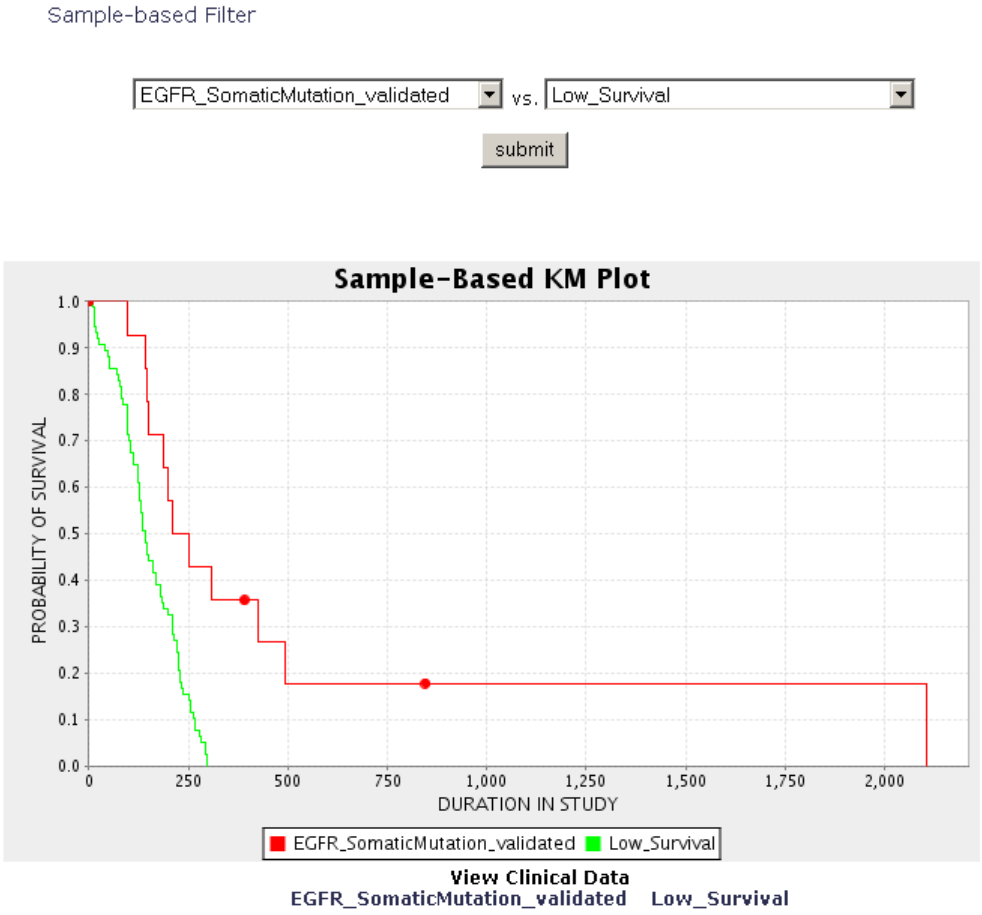


Figure 5.9 Sample-Based KM Plot

Table describes areas on the Sample-based Kaplan-Meier Survival Plot.

Item	***Special Instructions***
Sample-based Filter	To filter the plot, select new criteria from the drop-down lists at the top of the page, and then click **Submit**.

Table 5.2 Sample-Based KM Plot page description

Item	*Special Instructions*
View Clinical Data	To display clinical data for the selected *sample* groups, click the group link. For more information, see *Working With Clinical Report Results* on page 51
Statistical Report	• **Number of Samples in the group** specifies the number of samples, per group in the plot. • **Log-rank *p*-value** indicates the significance of the difference in survival between any two groups of samples segregated based on gene expression of the gene of interest. The log rank *p*-value is calculated using the Mantel-Haenszel procedure. The *p*-values are recalculated every time a new threshold is selected.

Table 5.2 Sample-Based KM Plot page description (Continued)

CHAPTER 6
ANALYSIS TOOLS

This chapter describes how to use the Cancer Molecular Analysis Portal to perform principal component and gene pattern analyses, and to access associated analysis applications.

Topics in this chapter include:

Overview

The Analysis Tools page provides access to the following tools and applications:

- Principal Component Analysis (PCA)
- Gene Pattern Analysis
- Gene Pattern Home
- Integrated Heatmap Viewer
- Cancer Genome Workbench (CGWB)

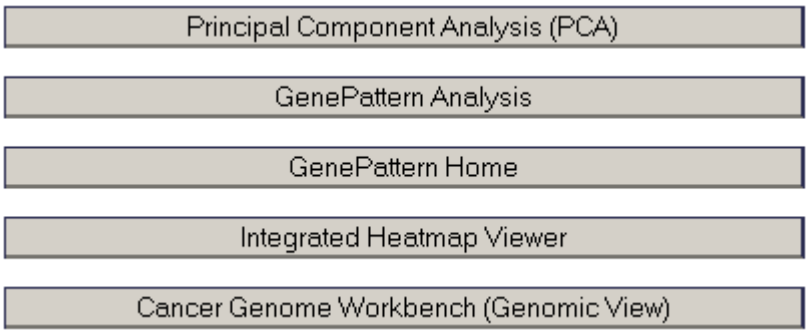


Figure 6.1 Analysis Tools Page

Principal Component Analysis

Principal component analysis (PCA) is method of identifying and highlighting patterns in data for the purpose of finding similarities and differences. PCA algorithms compress the number of dimensions of data to make the visual display more meaningful in terms of pattern recognition.

Selecting Criteria

To select criteria for analysis, follow these steps:

1. On the Analysis Home page, click **Principal Component Analysis**.

   The Principal Component Analysis page appears

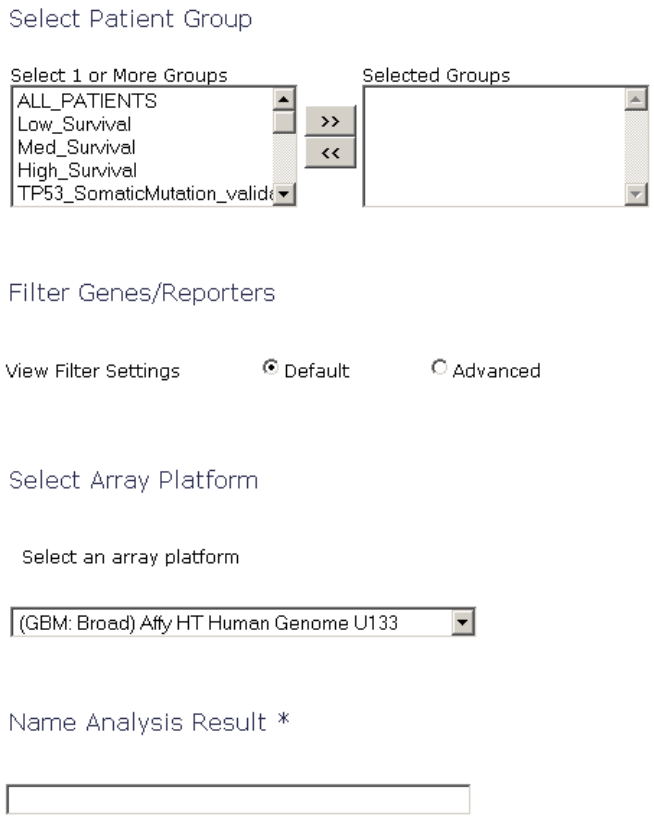


Figure 6.2 Analysis Tools – Principal Component Analysis

2. Select the *patient groups* to query from the patient group list by doing any of the following:
   - To select a single group, click the group name.
   - To select several contiguous groups, click the first group, press and hold the SHIFT key, and select the last group in the series. Or, drag your cursor from the first group to the last group in the series.
   - To select several discontinuous groups, click one group and CTRL + click additional groups.

   Note: You must select a minimum of two lists.
3. Click **>>** to move the selected groups to the **Selected Groups** list.
4. To remove a group from the **Selected Groups** list, select the name and click **<<** to return it to the patient groups list.
5. To view results with the default filter, select the **Default** option.
6. To constrain your query by variance (Gene Vector) percentile, click the **Advanced** filter option.
7. The Advanced options are dislplayed.

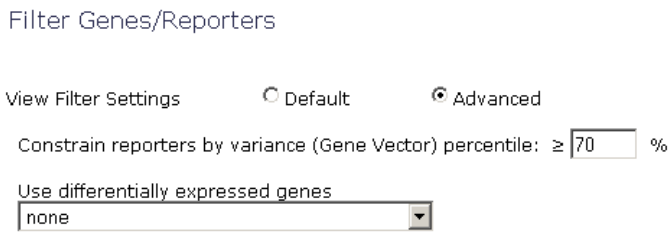


Figure 6.3 Advanced Filter Settings

8. Type a percentage in the text box provided to select the gene reporters whose variances of the log ratio (or log2 signals) across all experiments were among the top percentile of variance of all gene reporters identified. For example, 70% selects gene reporters with the top 30 (100 - 70) percentile of variance.
9. To filter by differentially expressed genes, select TCGA from the drop-down list.

   Note: Additional targets will be available as data is uploaded to the system.
10. Select an array platform from the drop-down list.
11. Type a unique name for the query.
12. Click **Submit**.

    The system processes your query and presents a link to the results.

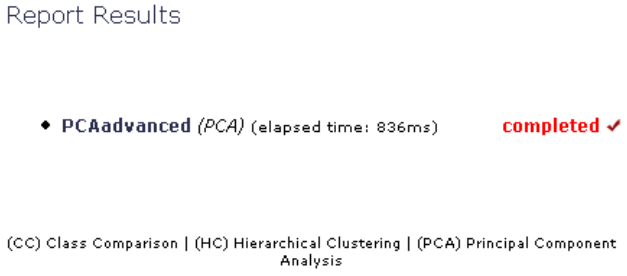


Figure 6.4 Query Process Completed

13. When the **completed** message is displayed, click the link to access the PCA plot.

    The PCA plot appears. (See *Figure 6.5*.)

Working With PCA Plots

The PCA plot displays the results of your query. Each point on the graph represents a *sample*, and each sample group is color-coded.

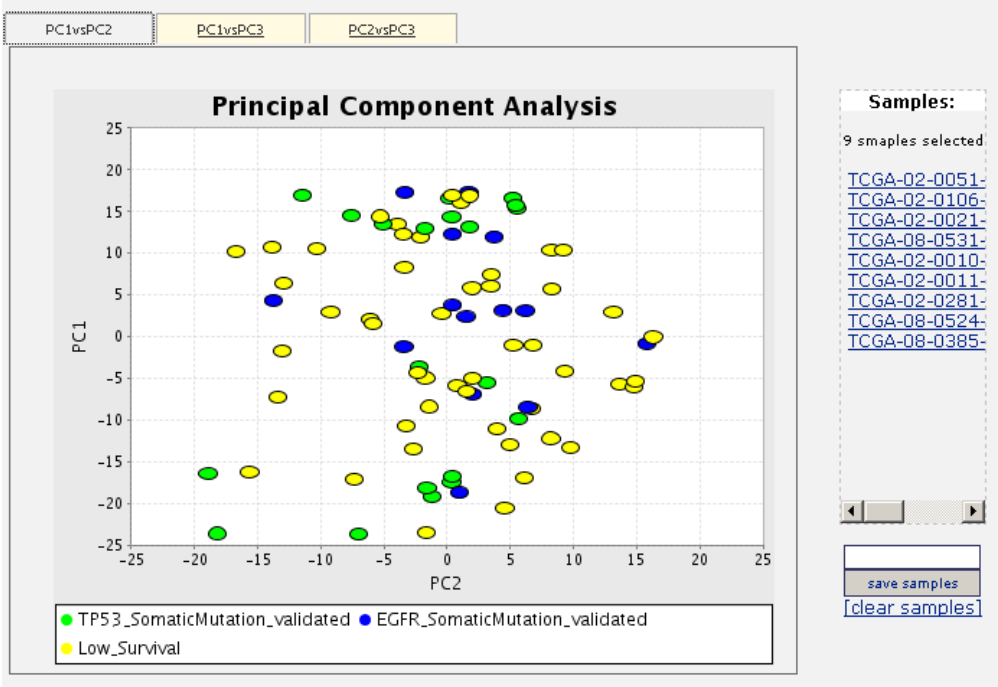


Figure 6.5 PCA Plot

Changing the Display

To display **PC1 versus PC2**, **PC1 versus PC3**, or **PC2 versus PC3**, click the appropriate tab above the plot.

Selecting Samples in PCA Plots

The Samples area enables you to select, review, and save *samples* in the plot.

To select individual samples, drag your cursor over the sample of interest. A red outline appears around your sample, and the sample name is added to the Samples list to the right of the plot.

Tip: Hover your cursor over a sample name to view the entire ID.

To select a sample group, drag your cursor across two or more samples.

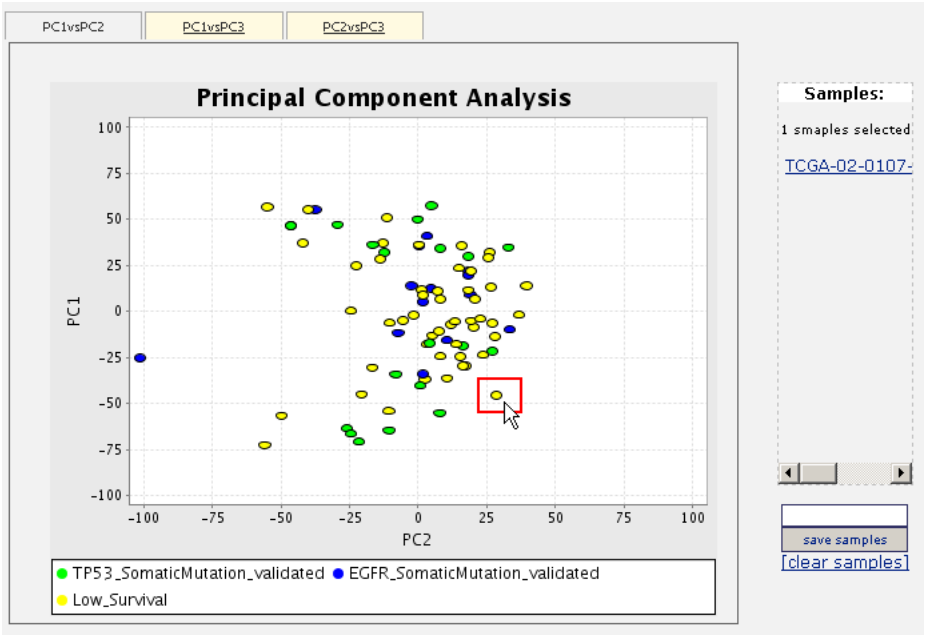


Figure 6.6 PCA Plot – Selected Sample

Gene Pattern Analysis

The Gene Pattern Analysis tool enables you to select groups of patients, genes, and array platform data to use as criteria for analysis in the GenePattern application.

Selecting Criteria

To select criteria for gene pattern analysis, follow these steps:

1. On the Analysis Tools page, click **Gene Pattern Analysis**.

The Analysis Module page appears

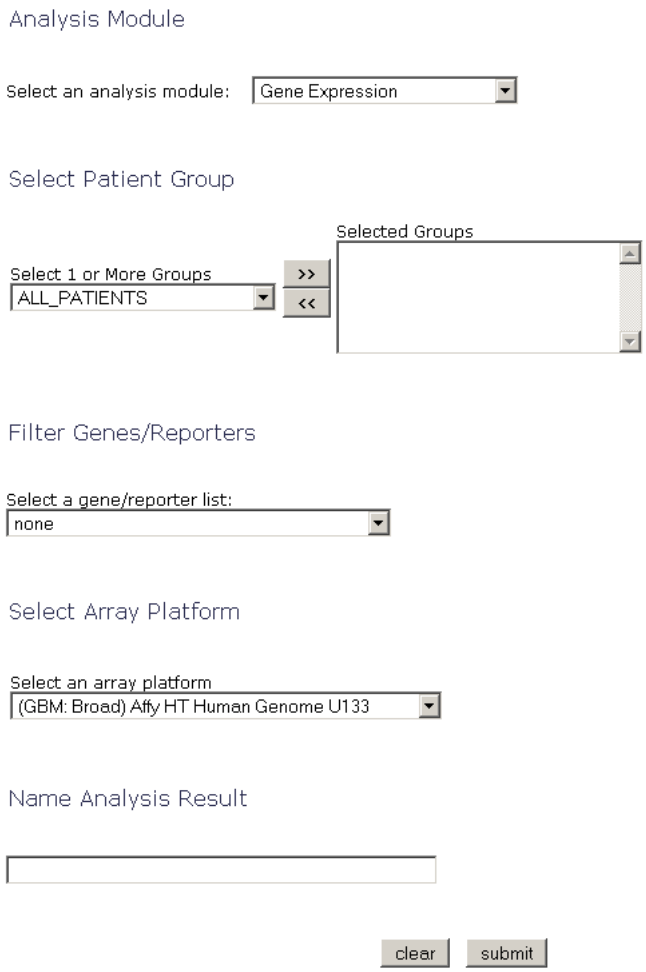


Figure 6.7 Analysis Tools – Gene Pattern Analysis

2. Select either **Gene Expression** or **Copy Number** from the module drop-down list.
3. Select the patient groups to query from the patient group list.
4. To filter by differentially expressed genes, select TCGA from the drop-down list.

   Note: Additional targets will be available as data is uploaded to the system.
5. Select an array platform from the drop-down list.
6. Type a unique name for the query.
7. Click **Submit**.

The system processes your query and presents a link to the results.

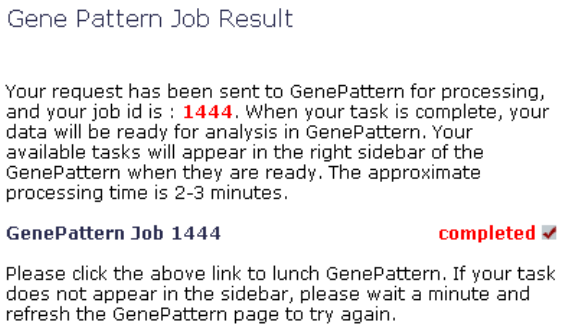


Figure 6.8 Query Process Completed

8. When the **completed** message is displayed, click the link to access the GenePattern application.
9. Follow the instructions provided in GenePattern to continue your analysis.

GenePattern Home

The GenePattern Home tool provides direct access to the GenePattern application. Follow the instructions provided on the GenePattern website to conduct your study.

Integrated Heatmap Viewer

The Heatmap Viewer tool provides direct access the CMA Portal Genome View.For details, see *Chapter 4, Working With Genome Views,* on page 43.

Cancer Genome Workbench

The Cancer Genome Workbench tool provides direct access to the Cancer Genome Workbench application. Follow the instructions provided in the application Help on the Cancer Genome Workbench website to conduct your study.

CHAPTER

7

MANAGING LISTS

This chapter describes how to manage patient and/or gene lists by editing application-defined lists and creating new custom lists.

Topics in this chapter include:

List Management Overview

The Manage Lists function centralizes all activities pertaining to the creation and management of user-defined, as well as study-defined, *PatientDID* lists, gene lists, and *gene reporter lists*. With these lists, you can further refine queries to facilitate analysis.

See *Side Bar Elements* on page 11 for a description of the sidebar content.

List Types

The CMA Portal sidebar contains a number of lists designed to facilitate analysis by providing sets of predefined data elements (patient ID, gene names, etc.).
These sidebar lists include:

- *PatientDID* Lists
- Gene Lists
- *Gene Reporter* Lists

About PatientDID Lists

DID information is data that has been disassociated from a patient's personally identifiable information (PII). The CMA Portal system provides DIDs rather than complete IDs in order to protect patient privacy.

The PatientDID list section of the sidebar contains somatic mutations (data type)/ somatic variants compiled from genomic sequencing centers. Each list in the section comprises a set of patients with a given set of unique characteristics. You can use default or custom lists to filter your queries.

Tip: To display the patientDIDs in any list, hover your cursor over the list name in the sidebar. A popup displays the data items.

Biorepositories that contain high quality cancer collections were chosen to provide clinical data and biospecimens (tumor and normal tissues) for analysis. Clinical datapoints collected include, but are not limited to:

- Clinical diagnosis
- Treatment history
- Histologic diagnosis
- Pathologic status
- Tissue anatomic site
- Surgery

Caution: CMA Portal saves your custom *PatientDID* lists for your current session only. Once you log out of your session, you can not retrieve them.

Table 7.1 describes the system-defined patient survival groups.

TCGA PatientDID List	***Description***
ALL_Patients	List of all patients available for investigation.
Low_Survival	List of patients who survived up to, and including, 300 days.
Med_Survival	List of patients who survived longer than 300 days but fewer than 900 days.
High_Survival	List of patients who survived longer than 900 days.

Table 7.1 PatientDID List descriptions

About Gene Lists

Each list in the Genes Lists section of the sidebar contains a set of genes of interest. Currently you can use default, TCGA genes, or custom lists to filter your queries.

Note: Currently the available data displayed is provided by TCGA Centers as of a given date. Other gene targets will be available in subsequent releases of this portal.

You can download TCGA gene list at:
http://gforge.nci.nih.gov/docman/view.php/259/7051/TCGA%20target%20lists.xls

About Gene Reporter Lists

You can create and use custom *gene reporter* lists for your queries.

Note: No predefined gene reporter lists are available currently. However, you can define custom lists.

Navigating the Sidebar Lists

Each of the lists in the sidebar contain predefined and custom (if you have defined them) lists. For example, *Figure 7.1* illustrates the *PatientDID* List section of the sidebar. Default CMA Portal lists are displayed in black text; custom lists you may have created in your current session are displayed in red text. See *Creating Custom Lists* on page 71.

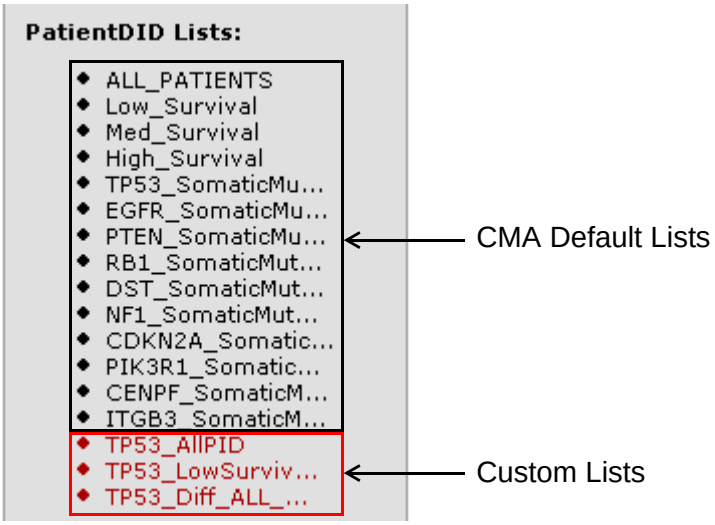


Figure 7.1 Sidebar Section – *PatientDID* List

To display the data included in analysis, hover your cursor over the list name of interest.

A popup lists the data items.

For example, *Figure 7.2* displays a partial list of patientDIDs for the EFGR Somatic Mutation (validated) list.

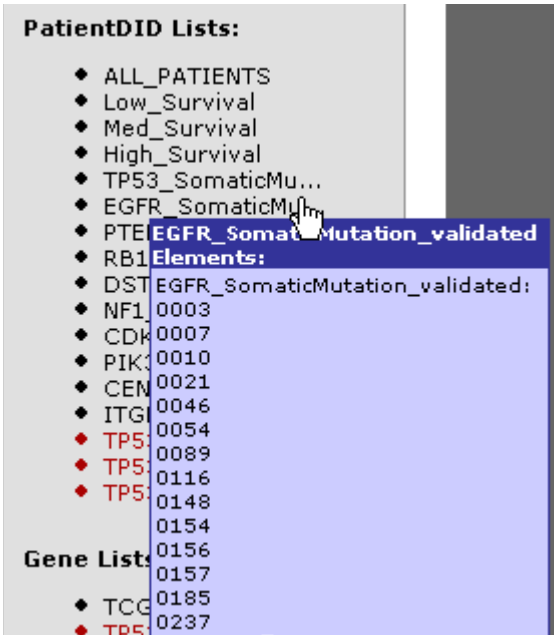


Figure 7.2 PatientDID Lists Section – List of EGFR patientDIDs

To export a list to a spreadsheet, double-click the list name. The CMA Portal system launches a spreadsheet application and displays the data in a worksheet.

Creating Custom Lists

You can define custom lists to facilitate your analysis in the following ways:

- Combine two or more lists. See *Creating New Lists From Existing Lists* on page 72.
- Upload your own lists automatically. See *Creating New Lists From Your Files* on page 74.
- Type lists manually. See *Creating New Lists Manually* on page 74.

Figure 7.3 illustrates the PatientDID, Gene, and Reporter lists as they appear with their respective content when managing lists. In addition to the lists displayed in the illustration, registered CMA Portal users can create custom clinical data lists.

Note: You may have to scroll down the page to see all lists.

Figure 7.3 List Management Page

Tip: To hide a group of lists, click the list header. Click the list header again to reveal the content.

Creating New Lists From Existing Lists

You can create new lists from two or more existing lists by joining or intersecting two or more lists; or by subtracting one list from another. *Figure 7.4* illustrates the end result of each of these operations. In the illustration, List A is represented by a circle and List B by a hexagon. Each list contains multiple data items, for example, gene symbols, some of which they have in common. As a result of commingling the lists, new lists are created as follows:

- List C (Join) – Data elements of Lists A and B combined.
- List D (Intersect) – Data elements that are common to both List A and List B
- List E (Difference) – Data elements that are *not* common to both List A and List B, where common data elements are subtracted from List A
- List F (Difference) – Data elements that are *not* common to both List A and List B, where common data elements are subtracted from List B

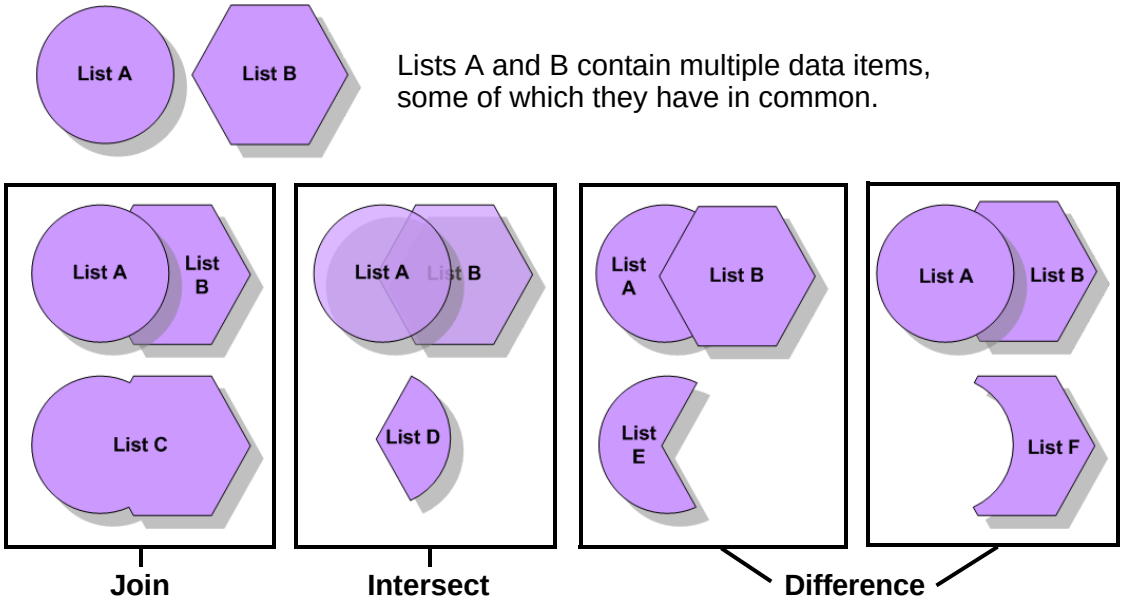


Figure 7.4 List Functions

To create a custom list from existing lists, follow these steps:

1. At the top of the sidebar, click **List Management**.
2. To navigate to the group of lists from which you want to create a custom list, click the appropriate link at the top of the List Management page. For example, to add to the list of genes, click Gene List.

Figure 7.5 List Management Menu

3. Select the check box next to the list name(s) that you want to include in a given operation to create a new list (*Figure 7.3*).

   Note: You cannot select more than two lists to use the **Difference** operator.

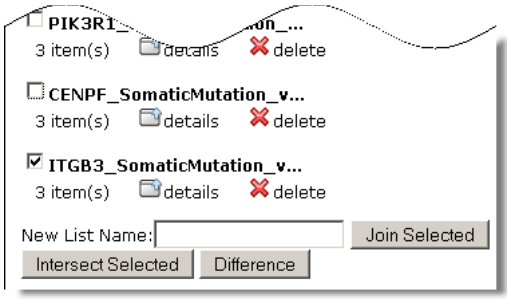


Figure 7.6 Patient DID List

4. In the New List Name box, type a unique name for the new list and then do one of the following operations:

   - To create a new list that contains all data items from two or more lists, click **Join**.
   - To create a new list that contains only data items that the selected lists have in common, click **Intersect**. If the selected lists have no items in common, the list is created but contains no data items.*Figure 7.7* provides an example of a list that is created as a result of the Intersect operation.
   - To create new lists that contain only data items that are unique to each of two lists, click **Difference**. *Figure 7.7* provides examples of the two lists that are created as a result of the Difference operation.

     For example, if you select **All Patients** and those with **TP53** mutations, two new lists are created as follows:

     — **Difference_High_Survival** – Contains items unique to the High Survival group. (High Survival DIDs minus TP53 DIDs)

     — **Difference_TP53** – Contains items unique to the TP53 group. (TP53 DIDs minus High Survival DIDs)

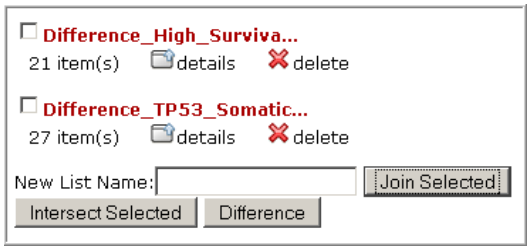


Figure 7.7 New Difference lists

   New lists appear on the Manage Lists page and in the side bar in red text.

5. To view the items in new lists, hover your cursor over the name of the list in the sidebar.

   The list's data items appear in a popup window. See on page 69.

6. To view the date and time that the list was created, hover your cursor over the name of the list in the **List Management** page.

   The list's creation date and time appear in a popup window.

Creating New Lists From Your Files

You can add a new list by uploading a file from your computer. To upload a list, follow these steps:

1. At the top of the Manage List page, click **Add List**, or scroll down the page until you reach the **Add List** section.

   The **Upload List or Manually Type List** section appears.

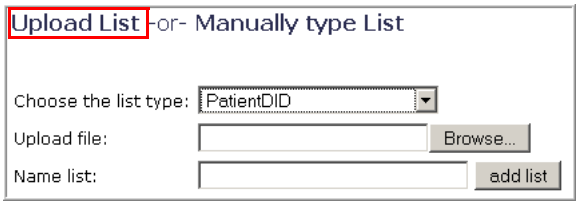


Figure 7.8 Add List Section – Upload List

2. Click **Upload List**.
3. From the **Choose the list type** drop-down list box, select the type of list you want to upload.
4. Next to the **Upload File** box, click **Browse** and then navigate to and select the file on your computer that you want to upload.
5. In the **Name List** box, type a unique name for the list, and then click **Add List**.

   The name of the list appears on the Manage Lists page and in the sidebar under the appropriate list type.

6. To view the items in new lists, do one of the following:

   - Hover your cursor over the name of the list in the sidebar. The list's content appears in a popup window. See on page 69.

   - **or** -

   - Next to the list name, click **Details**.

Creating New Lists Manually

You can create a new list type by manually typing data in the text box provided. To enter a list manually, follow these steps:

1. At the top of the Manage List page, click **Add List**, or scroll down the page until you reach the **Add List** section.

The **Upload List or Manually Type List** section appears.

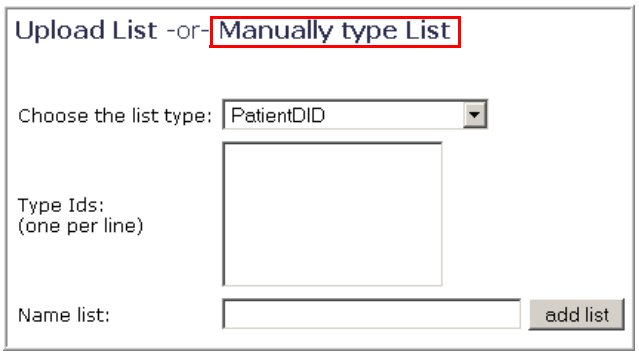


Figure 7.9 Add List Section – Manually Type List

2. Click **Manually Type List**.
3. From the **Choose the list type** drop-down list box, select the type of list you want to enter.

   Table 7.2 lists examples of correctly formatted codes for each list type.

List Type	***Correctly Formatted Examples***
PatientDID	CB160831 K03193
Gene-GENBANK_ACCESSION_NUMBER	AF125253 S75264
Gene-GENESYMBOL	BPIL2 IVL
Gene-LOCUS_LINK	10 100 10017
Reporter-AFFY_GHU133A_PROBE_SET	1007_s_at 1053_at
Reporter-IMAGE_CLONE	IMAGE:1407831 IMAGE:143995
Reporter-DBSNP	rs1000015 rs1000025
Reporter-AFFY_100K_SNP_PROBE_SET	SNP_A-1708471 SNP_A-1655302

Table 7.2 List type code formats

4. In the **Type Ids** box, type the data in the text box. Type only one item per line.
5. In the **Name List** box, type a unique name for the list, and then click **Add List**.

   The name of the list appears on the Manage Lists page and in the sidebar under the appropriate list type.

6. To view the items in new lists, do one of the following:
   - Hover your cursor over the name of the list in the sidebar. The list's content appears in a popup window. See on page 69.

   - or -

   - Next to the list name, click **Details**. Click **Details** again to hide the details.

   Note: If the format of the values entered in the **Type Ids** box was not correct, you must **Delete** the list and start again.

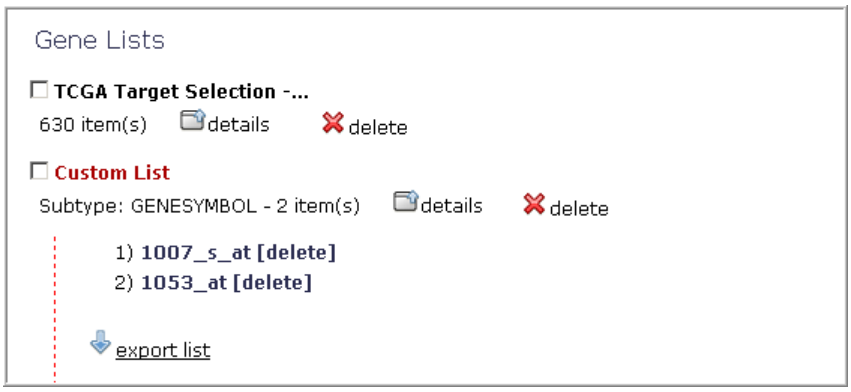


Figure 7.10 New List Content

Viewing and Managing List Content

CMA Portal enables you to manage the content of predefined and custom lists as follows:

- View details of a list's content. See *Viewing List Details* on page 76.
  - View a list's creation date and notes. See *Viewing List Creation Data and Notes* on page 77.
- Delete a list. See *Deleting Lists* on page 77.
  - Delete an item from a list. See *Deleting Items From a List* on page 78.
- Export a list. See *Exporting Lists* on page 78.

Viewing List Details

To view the individual data items in a list, follow these steps:

1. At the top of the **List Management** page, click the type of list you want to view (***PatientDID*** **Lists**, **Gene Lists**, or **Reporter Lists**).
2. To view the items in a list, do one of the following:
   - Hover your cursor over the name of the list in the sidebar. The list's content appears in a popup window. See on page 69.

   - or -

   - Next to the list name, click **Details**. Click **Details** again to hide the details.

The list details appear below the list name.

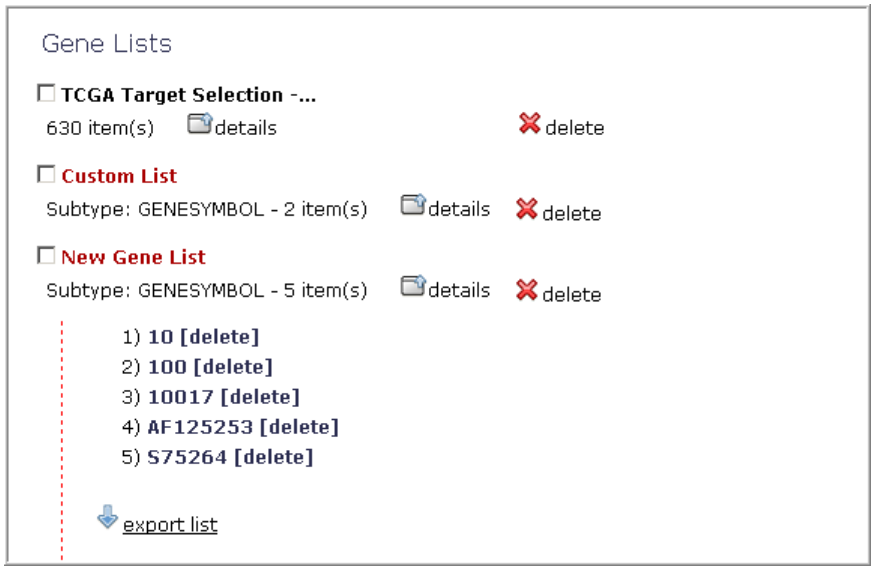


Figure 7.11 List Details

Viewing List Creation Data and Notes

To view the date and time that the list was created, and any notes about the list that may have been included, follow these steps:

1. At the top of the **List Management** page, click the type of list you want to view (***PatientDID*** **Lists**, **Gene Lists**, or **Reporter Lists**).
2. Hover your cursor over the name of the list in the **List Management** page. The list's creation date and notes appear in a popup window.

Deleting Lists

To delete a list, follow these steps:

1. At the top of the **List Management** page, click the type of lists you want to delete (***PatientDID*** **Lists**, **Gene Lists**, or **Reporter Lists**).

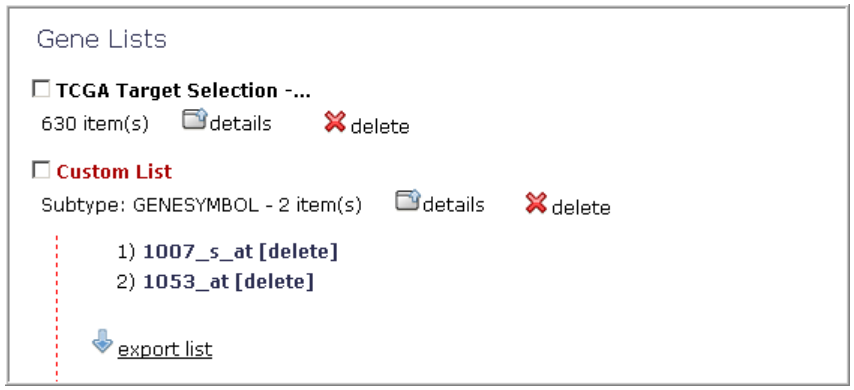


Figure 7.12 Gene Lists – Custom List Details

Caution: You can not retrieve a list once you have deleted it.

2. Click the **delete** icon (**x) next** to the name of the list you want to delete.

The list is deleted.

Deleting Items From a List

To delete one or more list items, follow these steps:

1. At the top of the **List Management** page, click the type of list (***PatientDID*** **Lists**, **Gene Lists**, or **Reporter Lists**) that contains the data you want to delete.
2. Click the **details** icon next to a list to display its content.
3. To the right of the item you want to delete, click **[delete]**.

   The item is deleted from the list.

Exporting Lists

To export a list, follow these steps:

4. At the top of the **List Management** page, click the type of lists you want to export (***PatientDID*** **Lists**, **Gene Lists**, or **Reporter Lists**).
5. Next to the list you want to export, click the **Details** icon to display all of the items in the list.
6. Scroll to the bottom of the list of data items, and click the **export list** link.

   The file download dialog box appears.

7. Open the list in Notepad or save the list as a file on your computer.

Tip: To export a list from the side bar, double-click the list name. Open the file in Notepad or save the patient identifiers to a spreadsheet file.

GLOSSARY

Acronyms, objects, tools and other terms referred to in the chapters or appendixes of this Cancer Molecular Analysis Portal User's Guide are described in this glossary.

Term	***Definition***
Affy HT Human Genome U133 (GBM: Broad)	High-throughput expression profile of approximately 40,0000 transcripts and variants.
Affy Human Exon 1.0 (GBM:LBL)	Contains approximately one million predicted and confirmed exons.
Agilent 8 x 15K Human miRNA-specific Microarray	Contains probes for 530 human and 76 human viral microRNAs from the Sanger database v.10.1.
Agilent Whole Human Genome (GBM:UNC)	High-density profiling analysis tool that covers over 41,000 unique human genes and transcripts
amplified	Condition when the number of copies of a gene is equal to or greater than 2.5 times the original.
ANOVA	Analysis of Variance which simplifies the F-test, where F-test is the mean square for each main effect and the interaction effect divided by the *within* variance. A one-way ANOVA or single factor ANOVA tests differences between the groups classified only on one independent variable.
BCM	Baylor College of Medicine
controlled data	Non-public, protected data. Ensures patient privacy.
copy number	Differences in the number of copies of a particular gene in a single *sample*. Also known as copy number variation (CNV).
deleted	Condition when one or more genes have been deleted.
gene expression	Process by which proteins are made from the instructions encoded in DNA.
gene reporter	Gene that codes for a product that can readily be measured, such as a fluorescing protein. Often used for expression studies of heterologous promoters.
intensity	Geometric mean value calculated for each comparison group.

Table .1 Key terms

Term	Definition
Kaplan-Meier survival plot	Survival probability for the user-defined set of criteria as a function of time and survival differences as analyzed by the log-rank test.
LOH	Loss of heterozygosity
methylation	Enzymatic addition of methyl (CH3) group to DNA which causes inactivation of that region.
mismatch (MM)	Condition in which DNA bases from one strand are not complementary to the bases from the other strand.
mutated	Alteration in DNA sequence that is either induced by a mutagen or is spontaneous
NCI	National Cancer Institute
NCICB	National Cancer Institute Center for Bioinformatics
normalization	Used to designing relational database tables and minimizing duplicated data.
open data	Data that is available to the general public.
overexpressed	TheCancer Molecular Analysis Portal when the fold ratio is twice the control value or average.
patient group	Pre-defined or user-defined patientDID list comprising patient identifiers with certain characteristics
patientDID	Data from tissue *sample* analysis from which personally identifiable information has been removed such that a researcher can no longer trace data back to an individual patient.
perfect match (PM)	Set of oligonucleotides whose sequence exactly matches the gene of interest
probe	Labeled segment of DNA that is used to bind to and identify a gene or mRNA transcript.
probeset	Multiple probe pairs. Each probe pair consists of two groups of probes—a *perfect match* (PM) and a *mismatch* (MM).
sample	Biological tissue sample
sample threshold	A final percentage threshold applied to the *samples* used to determine whether the gene is an anomaly.
SNP	Single-nucleotide polymorphism
standard deviation	Statistical measure of spread or variability.
SVG plugin	Integrates with your Web browser as a plug-in and enables you to display SVG images like the pathway diagram.
TCGA	The Cancer Genome Atlas
tumor mutation samples	The subset of tumor *samples* where a mutation has been found in that particular gene.
underexpressed	TheCancer Molecular Analysis Portal when the fold ratio is less than twice the control value or average.
value threshold	The initial threshold applied to data to determine an anomaly.

Table .1 Key terms (Continued)

INDEX

N

O

P

Q

R

S