TCGA Portal Data Browser

User's Guide

Center for Biomedical Informatics
and Information Technology

October 27, 2009

Credits and Resources

TCGA Data Portal Prototype Development and Management Teams		
Development	**Documentation**	**Product and Program Management**
David Nassau	Jill Hadfield	Carl Schaefer
Silpa Nanan		
Jessica Chen		
Robert Sfeir		
Ari Kahn		
Kavitha Thulasiraman		
Namrata Rane		
Shelley Alonso		

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 if it is a Web-based 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: Monday to Friday, 8 am – 8 pm Eastern Time, excluding government holidays.

TABLE OF CONTENTS

ABOUT THIS GUIDE

This section introduces you to the *TCGA Portal Data Browser User's Guide*. It includes the following topics:

- *Purpose* on this page
- *Audience* on this page
- *Topics Covered* on this page
- *Text Conventions Used in the User's Guide* on page 2

Purpose

This guide provides an overview of the TCGA Portal Data Browser, also called the Data Browser. This book is organized into chapters that parallel the TCGA Portal Data Browser workflow.

Audience

This guide is designed for the biomedical research community. The Cancer Genome Atlas (TCGA) is a comprehensive and coordinated effort to accelerate our understanding of the molecular basis of cancer.

Topics Covered

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

- *Chapter 1* provides instructions to start using TCGA Portal Data Browser, including configuring data browser search parameters and launching a search.
- *Chapter 2* describes how to view and interpret search results.
- *Glossary* defines certain terms used in this guide.

Text Conventions Used in the User's Guide

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.

CHAPTER
1

GETTING STARTED WITH TCGA PORTAL DATA BROWSER

This chapter introduces you to TCGA Portal Data Browser Cancer Gene Abnormalities page. It describes how you can perform a search for cancer genes submitted to TCGA and for patients and biological pathways associated with those genes.

Topics in this chapter include:

- *About TCGA Portal Data Browser* on this page
- *Search Modes* on this page
- *Launching a Data Browser Search* on page 5

About TCGA Portal Data Browser

On the Data Browser Cancer Gene Abnormalities page, you can configure queries of data that have been submitted to TCGA. You can set search criteria relating to genes, patients and abnormality types. The system performs the search, summarizes the data and displays results as gene lists, patient lists and biological pathways.

Using TCGA Portal Data Browser Online Help

The TCGA Portal Data Browser online help explains how to use all of the features of the TCGA Portal Data Browser portal.

Note: You can open online help without being logged into TCGA Portal Data Browser.

To access online help in TCGA Portal Data Browser, use either of these options:

- Click the **Help** icon/ available in the upper right corner of the user interface (?). This opens online help at the Welcome page.

- With the mouse, hover over any underlined field title in the interface.This opens a tool tip for the field. Click the **For More Info** link in the tool tip which opens online help positioned in the Glossary at the term definition. From there you can navigate to any area in online help, as described below.

Online help organization:

- The left panel displays the Table of Contents (TOC), and also offers access to the Index and Search features of online help. The TOC can be expanded. All topics listed in the TOC and index are hypertext links to the referenced topics.
- The right panel displays the Welcome to TCGA Portal Data Browser Online Help page and other topic contents.

Navigating online help:

- The bread crumb trail at the top of the page shows the relative location of the current help topic relative to neighboring topics. Click a breadwinner link to display that help topic.
- Click the **Back** or **Forward** links at the top of the page to display help topics you have previously viewed.
- Follow hypertext links or the **Related Topics** buttons in the help topics to open other closely related topics. If the current help page has related topics associated with it, you can also view them by clicking the **Related Topics** button (▤) at the top right of the help page.
- Locate topics using the table of contents that displays in the left pane of the online help project or the **Index** tab that displays at the top of the Table of Contents pane.
- Perform word searches of Help by entering query text in the search text box.
- Print the current topic by clicking the **Print** button (🖶) at the top right of the help page.

Search Modes

The search panel on the left of the Data Browser page described below displays three color-coded tabs: **Genes**, **Patients** and **Pathways**, or search “modes”. Search criteria that you define on these tabs filter the search results. You can click a button that carries search criteria defined on one tab to all three tabs.

- **Genes** tab – When you select this tab, all search criteria are enabled. Queries launched from the Genes mode search for genes that meet the defined criteria. Search results are summarized by gene symbol.
- **Patients** tab – When you select this tab, the **Average across patients** and **Correlations** filters are disabled in the search criteria. Queries launched from the Patients tab search for patients whose genes meet the search criteria. Search results are summarized by patient ID.
- **Pathways** tab – When you select this tab, all search criteria are enabled. Queries launched from the Pathways mode search for pathways whose gene products stem from genes that meet the search criteria. Search results consist of a list of pathways in which any of the defined genes found in the search are involved.

Any results/tabs that are open for one mode are saved when you switch to another mode, but are hidden until you switch back to that original mode. A search that is proceeding in one mode continues in the background while you are working with another mode.

Some results are expressed as ratios or percentages. When listing by genes, many calculations give the ratio of patients affected over patients tested for that abnormality, for each gene. When listing by patients, it is the ratio of genes affected over genes tested, for each patient

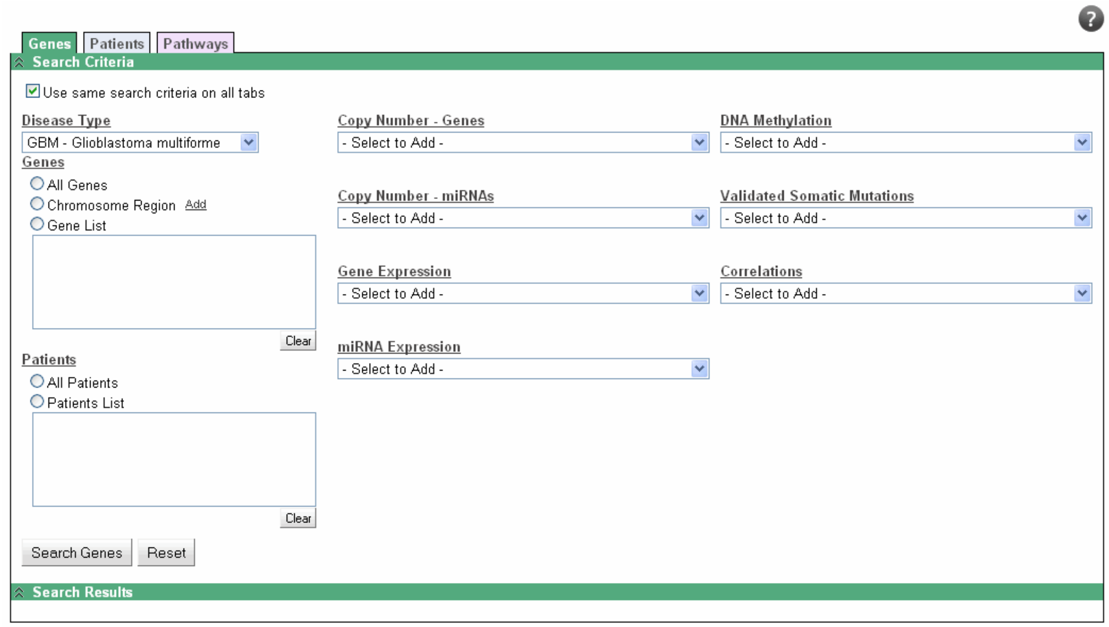
Copy Number results may alternately be expressed as an Average Across Patients which is an average value across patients for each gene.

For more information, see *Chapter 2 Viewing Search Results*

Launching a Data Browser Search

To configure a search in TCGA Data Browser, follow these steps.

1. Open the browser. The Data Portal displays the Cancer Gene Abnormalities page with fields for configuring search parameters (*Figure 1.1*).



☑ Use same search criteria on all tabs

Disease Type: GBM - Glioblastoma multiforme

Figure 1.1 TCGA Data Portal cancer genes abnormalities search page

Tip: The page for configuring search parameters is the same for all three tabs, although availability of fields may vary. The display of results depends on the mode you select. Distinctions are described in the following steps.

2. Select or enter criteria for the fields as described.

   Note: An implicit AND is used between different criteria.

   a. **Disease Type** – Select from the drop-down list a cancer type for the search.

Tip: After you have run a search, if you modify the disease type, existing results are deleted from the results panel. The browser also resets the types of filters that you can select based on the data available for that disease.

b. **Genes** – Choose either **All Genes**, **Chromosome Region(s)**, or **Gene List**. For a detailed description of these options, see *Genes Options* on page 7.

c. **Patients** – Select **All Patients** or specify a **Patient List** of IDs. If you type any text in the text box, the **Patients List** button is automatically selected. For a detailed description of these options, see *Patients Options* on page 7.

d. **Copy Number** – The drop-down lists all the center/platform combinations available for the *copy number* abnormality type. Click **Select to Add** to add one to the display. Once you make a selection, additional query fields appear. For a detailed explanation of the fields and their interpretation, see *Copy Number Options* on page 8.

e. **Copy Number miRNAs** – The drop-down lists all the center/platform combinations available for the copy number miRNAs abnormality type. Click **Select to Add** to add one to the display. Once you make a selection, additional query fields appear. For a detailed explanation of the fields and their interpretation, see *Copy Number miRNA Options* on page 9.

f. **Gene Expression** – The drop-down lists all the center/platform combinations available for the gene expression abnormality type. Click **Select to Add** to add one to the display. Once you make a selection, additional query fields appear. For a detailed explanation of these fields and their interpretation, see *Gene Expression Options* on page 10.

g. **miRNA Expression** – The drop-down lists the center/platform combinations available for this abnormality type. Click **Select to Add** to add one to the display. Once you make a selection, additional query fields appear. For a detailed explanation of the fields and their interpretation, see *miRNA Expression Options* on page 11.

h. **DNA Methylation** – The drop-down lists the center/platform combinations available for this abnormality type. Click **Select to Add** to add one to the display. Once you make a selection, additional query fields appear. For a detailed explanation of the fields and their interpretation, see *DNA Methylation Options* on page 12.

i. **Validated Somatic Mutations –** The drop-down lists all the center/platform combinations available for the somatic mutations abnormality type. Click **Select to Add** to add one to the display. Once you make a selection, additional query fields appear. For a detailed explanation of these fields, and their interpretation, see *Validated Somatic Mutations Options* on page 12.

j. **Correlations** – This option is available only when you select the **Genes** or **Pathway** tab. This mode lists results by gene.

Select and define correlations to search for genes or pathways with significant correlations between two data sets, as defined by the search criteria. Correlations are available for specific combinations of abnormalities/platforms, for example, a *copy number* platform vs. a *gene expression* platform. For a detailed explanation of these fields, and their interpretation, see *Correlations Options* on page 13.

Note the following regarding tab selection: If you make all of your selections on the Genes tab, the results display as a gene list. If you select the Patients tab before you launch the search, results display as a patients list. If you select the Pathways tab before launching the search, the results panel displays a list of pathways that match the parameters you set in relation to genes and patients and abnormality type(s).

Note: If you select no abnormality types, you will get a list of all genes (in gene mode), a list of all patients (in patient mode), or a list of all pathways (in pathway mode).

3. To launch the query, click **Search**.
4. To clear the fields on the search panel, click **Reset**. To clear results displaying in the right panel, click the **Refresh** button for your browser.

Genes Options

- **All Genes** – Includes all genes that are in TCGA database, a total that increases daily. The search considers all of these genes, but presumably does not return all of them because of the various filters.
- **Chromosome Location** – Defined as a chromosome number, a start location and a stop location. X and Y chromosomes are excluded. If a Start location is omitted, the Portal assumes that the Start is the beginning of the chromosome. If the Stop location is omitted, the application assumes that the Stop if the end of the chromosome. If both are omitted, the region is defined as the whole chromosome.

  A Chromosome Region is equivalent to a Gene List containing all the genes that are contained within, or overlap, that region. If any part of a gene overlaps a region, the gene is included.
- **Gene List** – A gene list executes as a “pre-filter” before abnormality filters are applied, limiting the genes considered for the search. Only genes listed can appear in search results.

When Patient mode is selected, a gene list defines how the abnormality data is generated. When expressed as a ratio, the number of genes in the numerator gives the number of genes affected over the denominator which is all of the genes tested for that abnormality for each patient.

Tip: Genes on a gene list can be separated by spaces, new lines, commas, or semicolons in this text box.

After you have defined the genes criteria, return to *Launching a Data Browser Search* on page 5.

Patients Options

Patient List – A patient list executes as a “pre-filter” before gene abnormality filters are applied, limiting the patients considered for the search. Only patients listed can appear in search results.

When Gene mode is selected, a patient list defines how the abnormality data is generated. For example, when abnormalities are expressed as a ratio, the number of patients in the numerator gives the number of patients affected with abnormal genes over the denominator which is all of the patients who had this gene tested for the given abnormality. Note that if the patient list includes patients that were not tested, they are not included in the denominator. So the denominator shows how many of your patients in the list actually were tested.

Tip: Patients can be separated by spaces, new lines, commas, or semicolons in this text box.

After you have defined the patients criteria, return to *Launching a Data Browser Search* on page 5.

Copy Number Options

By default, the limit is set at "**<=-0.5 or >=0.5**" but you can change this. For instance, you could omit one or the other side of this limit. For example, "**>=0.5**" finds amplifications but not deletions. You cannot leave both sides empty.

A copy number abnormality is counted when the copy number (log2 ratio) is qualified by this limit. For a typical, non-sex chromosome, not-CNV case, a stored value of 0 means 2 copies (2^{0+1}). A value of 1 means 4 copies (2^{1+1}), and -1 means 1 copy (2^{-1+1}).

By setting the limit statement, you are specifying what is to be considered abnormal. The first number must always follow a "<" or "<=" operator, and the second number must always follow a ">" or ">=" operator.

Frequency – By defining this threshold, you limit the returned rows to only those that show a certain minimum percent abnormality. In Gene or Pathway mode, this means the percentage of patients for each gene showing an abnormality (patient ratio). In Patient mode, it means the percentage of genes showing an abnormality (gene ratio). You can change the default value to any number between zero and 100. A frequency of zero reports all results for the column without filtering.

Note: If the Frequency text box is left blank, that also means zero. Decimal numbers are allowed, for example, 0.01.

This filter applies to most abnormality types, but it does not apply to copy number Averaged Across Patients (the checkbox selected), which does not display as a patient or gene ratio. It also does not apply to correlations.

The **Average across Patients** checkbox is available only when you select the Genes tab. With the Average Across Patients option, a different method is used to calculate results. Rather than considering each patient, the search program obtains the average across all patients for each gene. In this case, the limit statement is used to filter results, but not to define what counts as abnormal.

When listing by patient, the search program counts each gene that shows an abnormality. The resulting ratio gives the number of genes affected over all genes.

The search is influenced by lists you specify in the following way:

- **Gene List**: Limits search to specified genes.

- **Patient List**: Limits search to specified patients.

The search is influenced by the tab you select in the following way:

- **Gene mode**: The program, for each gene, finds the number of patients with copy number variations, as defined by the search criteria. Results are summarized by gene symbol, one row per gene; values are calculated base on the total number of patients for whom that gene was characterized.
- **Patient mode**: Add copy number data sets to search for patients with copy number variations, as defined by the search criteria. Results are summarized by patient ID, one row per patient; values are the percent (or ratio) of genes that meet the search criteria for each patient.
- **Pathway mode**: The program searches for pathways containing genes with frequent copy number variations, as defined by the search criteria. Results are displayed as a list of pathways whose genes meet the search criteria.

After you have defined the copy number criteria, return to *Launching a Data Browser Search* on page 5.

Copy Number miRNA Options

The Copy Number miRNAs search is similar to *Copy Number Options* except the search is for miRNAs instead of for genes.

By default, the limit is set at "**<=-0.5 or >=0.5**" but you can change this. For instance, you could omit one or the other side of this limit. For example, "**>=0.5**" finds amplifications but not deletions. You cannot leave both sides empty.

A copy number miRNA abnormality is counted when the copy number (log2 ratio) is qualified by this limit. For a typical case, a stored value of 0 means 2 copies (2^{0+1}). A value of 1 means 4 copies (2^{1+1}), and -1 means 1 copy (2^{-1+1}).

By setting the limit statement, you are specifying what is to be considered abnormal. The first number must always follow a "<" or "<=" operator, and the second number must always follow a ">" or ">=" operator.

Frequency – By defining this threshold, you limit the returned rows to only those that show a certain minimum percent abnormality. In Gene or Pathway mode, this means the percentage of patients for each miRNA showing an abnormality (patient ratio). In Patient mode, it means the percentage of miRNAs showing an abnormality (miRNA ratio). You can change the default value to any number between zero and 100. A frequency of zero reports all results for the column without filtering.

Note: If the Frequency text box is left blank, that also means zero. Decimal numbers are allowed, for example, 0.01.

This filter applies to most abnormality types, but it does not apply to copy number Averaged Across Patients (the checkbox selected), which does not display as a patient or miRNA ratio. It also does not apply to correlations.

The **Average across Patients** checkbox is available only when you select the Genes tab. With the Average Across Patients option, a different method is used to calculate results. Rather than considering each patient, the search program obtains the average across all patients for each miRNA. In this case, the limit statement is used to filter results, but not to define what counts as abnormal.

When listing by patient, the search program counts each miRNA that shows an abnormality. The resulting ratio gives the number of miRNAs affected over all miRNAs.

The search is influenced by lists you specify in the following way:

- **Gene List**: Limits search to miRNAs associated with specified genes.
- **Patient List**: Limits search to specified patients.

The search is influenced by the tab you select in the following way:

- **Gene mode**: The program, for each miRNA, finds the number of patients with copy number variations, as defined by the search criteria. Results are summarized by miRNA symbol, one row per miRNA and grouped by gene; values are calculated base on the total number of patients for whom that miRNA was characterized.
- **Patient mode**: Add copy number miRNA data sets to search for patients with copy number miRNA variations, as defined by the search criteria. Results are summarized by patient ID, one row per patient; values are the percent (or ratio) of miRNAs that meet the search criteria for each patient.
- **Pathway mode**: The program searches for pathways containing miRNAs with frequent copy number variations, as defined by the search criteria. Results are displayed as a list of pathways whose miRNAs meet the search criteria.

After you have defined the copy number miRNA criteria, return to *Launching a Data Browser Search* on page 5.

Gene Expression Options

Gene expression values stored in the database represent the ratio of tumor expression over normal expression. The expression value for a given gene for a given patient is the log2 ratio of the tumor expression of the gene in the patient to the mean expression of the gene in a pool of normal samples. Consequently, a value of 1.0 would mean that the tumor expression is twice as great in the patient’s tumor as in the normal pool (two-fold over-expression in tumor samples), while a value of -1.0 would mean that the tumor expression is half as great in the patient’s tumor as in the normal pool (two-fold under-expression in tumor samples). A value of 0 would mean no difference in gene expression between tumor samples and the normal pool.

For glioblastoma (GBM), a synthetic normal has been derived based on the limited set of available normal tissue samples. For other cancer types, an actual paired normal might be used, depending on availability of those tissues.

Frequency – By defining this threshold, you limit the returned rows to only those that show a certain minimum percent abnormality. In Gene or Pathway mode, this means the percentage of patients for each gene showing an abnormality (patient ratio). In Patient mode, it means the percentage of genes showing an abnormality (gene ratio). You can change the default value to any number between zero and 100. A frequency of zero reports all results for the column without filtering.

The search is influenced by lists you specify in the following way:

- **Gene List**: Limits search to specified genes.
- **Patient List**: Limits search to specified patients.

The search is influenced by the tab you select in the following way:

- **Gene mode:** Searches for genes that are over- or under-expressed, as defined by search criteria. Results are summarized by gene symbol, one row per gene; values are the percent (or ratio) of patients who have genes that meet the search criteria.
- **Patient mode:** Searches for patients with a certain frequency of over- or under-expressed genes, as defined by search criteria. Results are listed by patient ID, one row per patient; values are the percent (or ratio) of genes that meet the criteria for that patient.
- **Pathway mode:** Searches for pathways containing genes that are over- or under-expressed as defined by search criteria. Results display as a list of pathways that meet the criteria for the specified gene(s).

After you have defined the gene expression criteria, return to *Launching a Data Browser Search* on page 5.

miRNA Expression Options

The values for this parameter are fold ratios expressed as log2, as in the *Gene Expression Options*. The default filter looks for extremes of both under- and over-expression. You can define parameters to look for only underexpression or overexpression.

Frequency – By defining this threshold, you limit the returned rows to only those that show a certain minimum percent abnormality. In Gene or Pathway mode, this means the percentage of patients for each miRNA showing an abnormality (patient ratio). In Patient mode, it means the percentage of miRNAs showing an abnormality. You can change the default value to any number between zero and 100. A frequency of zero reports all results for the column without filtering.

The search is influenced by lists you specify in the following way:

- **Gene List**: Limits search to miRNAs associated with specified genes.
- **Patient List**: Limits search to specified patients.

The search is influenced by the tab you select in the following way:

- **Gene mode:** Searches for miRNAs that are over- or under-expressed, as defined by search criteria. Results are summarized by miRNA symbol, one row per miRNA and grouped by gene; values are the percent (or ratio) of patients who have miRNAs that meet the search criteria.
- **Patient mode:** Searches for patients with a certain frequency of over- or under-expressed miRNAs, as defined by search criteria. Results are listed by patient ID, one row per patient; values are the percent (or ratio) of miRNAs that meet the criteria for that patient.
- **Pathway mode:** Searches for pathways containing miRNAs that are over- or under-expressed as defined by search criteria. Results display as a list of pathways that meet the criteria.

After you have defined the miRNA expression criteria, return to *Launching a Data Browser Search* on page 5.

DNA Methylation Options

The values for this parameeter represent beta values in the range 0 to 1. The default filter, >=0.5, looks for instances of methylation. By flipping the operator (to < or <=) and changing the filter value, the system can look for instances of non-methylation.

Frequency – By defining this threshold, you limit the returned rows to only those that show a certain minimum percent abnormality. In Gene or Pathway mode, this means the percentage of patients for each methylation region showing an abnormality (patient ratio). In Patient mode, it means the percentage of methylation region showing an abnormality (gene ratio). You can change the default value to any number between zero and 100. A frequency of zero reports all results for the column without filtering.

The search is influenced by lists you specify in the following way:

- **Gene List**: Limits search to methylation regions associated with specified genes.
- **Patient List**: Limits search to specified patients.

The search is influenced by the tab you select in the following way:

- **Gene mode:** Searches for methylation regions, as defined by search criteria. Results are summarized by methylation region, one row per methylation region, grouped by gene. Values are the percent (or ratio) of patients who have genes that meet the search criteria.
- **Patient mode:** Searches for patients with a certain frequency of over- or under-expressed genes, as defined by search criteria. Results are listed by patient ID, one row per patient; values are the percent (or ratio) of genes that meet the criteria for that patient.
- **Pathway mode:** Searches for pathways containing methylation regions that are over- or under-expressed as defined by search criteria. Results display as a list of pathways that meet the criteria for the specified methylation region[s])..

After you have defined the DNA methylation criteria, return to *Launching a Data Browser Search* on page 5.

Validated Somatic Mutations Options

You can add any of the following types of mutations to the query: **any non-silent**, **missense**, **nonsense**, **silent**, and **splice site**.

No limit statement is applied to mutations. The decision is binary: if a mutation has been reported within a gene in a patient, that patient is considered to have the abnormality. When listing by gene, the search program counts patients with the specified type of mutation in each gene. The resulting ratio shows the number of patients affected over the total patients tested for the mutations.

Frequency – By defining this threshold, you limit the returned rows to only those that show a certain minimum percent abnormality. In Gene or Pathway mode, this means the percentage of patients for each gene showing an abnormality (patient ratio). In Patient mode, it means the percentage of genes showing an abnormality (gene ratio).

You can change the default value to any number between zero and 100. A frequency of zero reports all results for the column without filtering.

The search is influenced by lists you specify in the following way:

- **Gene List**: Limits search to specified genes.
- **Patient List**: Limits search to specified patients.

The search is influenced by the tab you select in the following way:

- **Gene mode:** Searches for genes with frequent mutations of the type defined in the search criteria. Results are summarized by gene symbol, one row per gene; values are the percent (or ratio) of patients that meet the criteria for that gene.
- **Patient mode:** Searches for patients with a high number of genes with mutations of the type defined in the search criteria. Results are listed by patient ID, one row per patient; values are the percent (or ratio) of genes that meet the criteria for that patient.
- **Pathway mode:** Searches for pathways containing genes with frequent mutations of the type defined in the search criteria. Results display as a list of pathways that meet the criteria for the specified gene(s).

After you have defined the validated somatic mutations criteria, return to *Launching a Data Browser Search* on page 5.

Correlations Options

A correlation appears in the results as a column of real numbers between -1 and 1, with 1 being perfect correlation, -1 being perfect inverse correlation, and 0 being noncorrelation.

Correlations are calculated using the Pearson algorithm, which looks at each patient. For example, if patients who have amplification of gene ABC also tend to have overexpression of ABC, this would result in a high correlation value. A P-value is also calculated, indicating the probability that the correlation is significant and not occurring by chance.

You can specify a limit statement, filtering by p-value and or r-value (the correlation itself).

Note: If correlation value is left blank on both sides, it means “show correlations, but don’t filter by r-value”. If P-value is left blank, it means “don’t filter by P-value”. If all are left blank, it means “show all correlations”.

For more information including the interpretation of p-values, see *Calculating P-values* on page 20.

After you have defined the correlations criteria, return to *Launching a Data Browser Search* on page 5.

Modifying a Search

Once you set query parameters, launch a search and view the results, you can modify the search parameters and rerun the search.

For example, you could copy gene names or patients names from the genes, patients or pathway search results summary to expand the corresponding lists for a new search.

You could use different combinations of anomalies/correlations, or redefine the limit statements. You could enter a patient list to show abnormality results for a specific patient or group of patients. For more information about specific fields, see *Launching a Data Browser Search* on page 5.

To modify the search, follow these steps:

1. On the search results page, click **Modify Search Criteria** link.

   This allows you to return to the query page to modify parameters.

2. Redefine the search parameters, all of which serve to filter the data when you re-run the search.
3. Click **Search** to rerun the search.

Note: When displaying a single pathway and clicking the Modify Search Criteria link, the text on this button changes to say **Update Pathway**. It will only refresh the currently displayed pathway, using the current filter criteria. The accompanying data in table format is reconfigured to reflect modified search parameters.

CHAPTER 2

VIEWING SEARCH RESULTS

This chapter describes how TCGA Portal Data Browser displays search results and how you can best view and interpret the results.

Topics in this chapter include:

- *Search Results Overview* on this page
- *Search Results Examples* on page 21
- *Downloading Source Data Files* on page 30
- *Tutorial: Gene Expression to Pathway* on page 32

Search Results Overview

Once you have launched a search, the Data Browser displays search results below a search criteria summary.

Paging

Initially, a summary of search results displays on a new page, color coded to the tab from which the search was launched. The page counter at the bottom of the page says "1 of 1+". A progress bar shows pages being gathered on the server. As it does so, the page counter increases, for example, "1 of 5+". When all pages have been gathered, the progress bar disappears and the page counter no longer shows a plus sign.

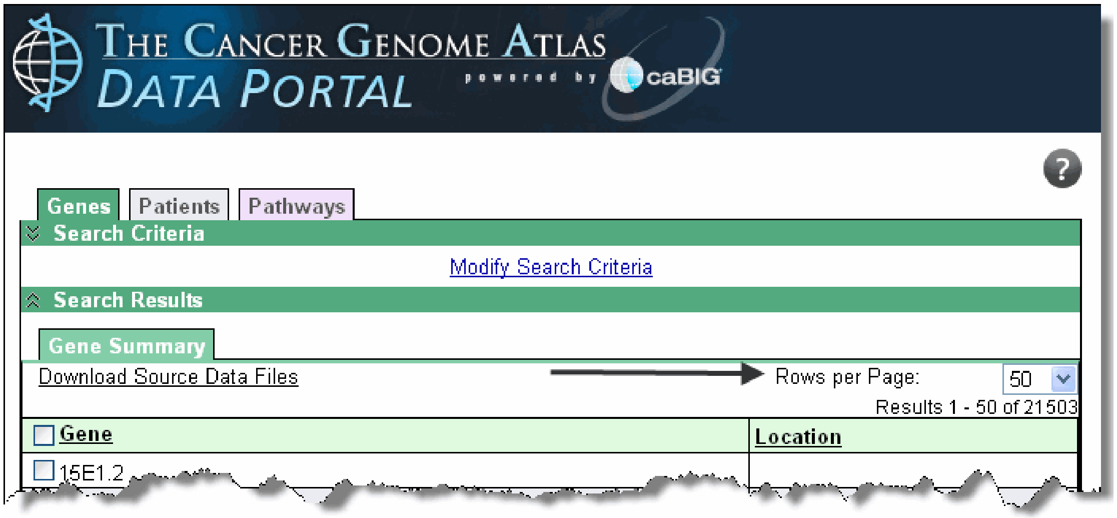
Pages are gathered on the server and only served to the client as you request them. The client does not keep pages in memory other than the currently displayed page.

If you try to navigate to a page that has not yet been gathered on the server, there may be a delay while the server catches up with the request.

You can page through results using the forward (>) or back (<) arrows or enter a page number to go directly to that page.

Note: Search criteria defined for the search always display at the top of the page, above the search results.

You can specify the number of rows of results to display on a page. Click the drop-down arrow in the upper right, above the results columns (*Figure 2.1*). (



Rows per Page: 50

Results 1 - 50 of 21503

Figure 2.1 The drop-down arrow in the upper right governs the number of rows displaying on a page

Gene and Patients Results

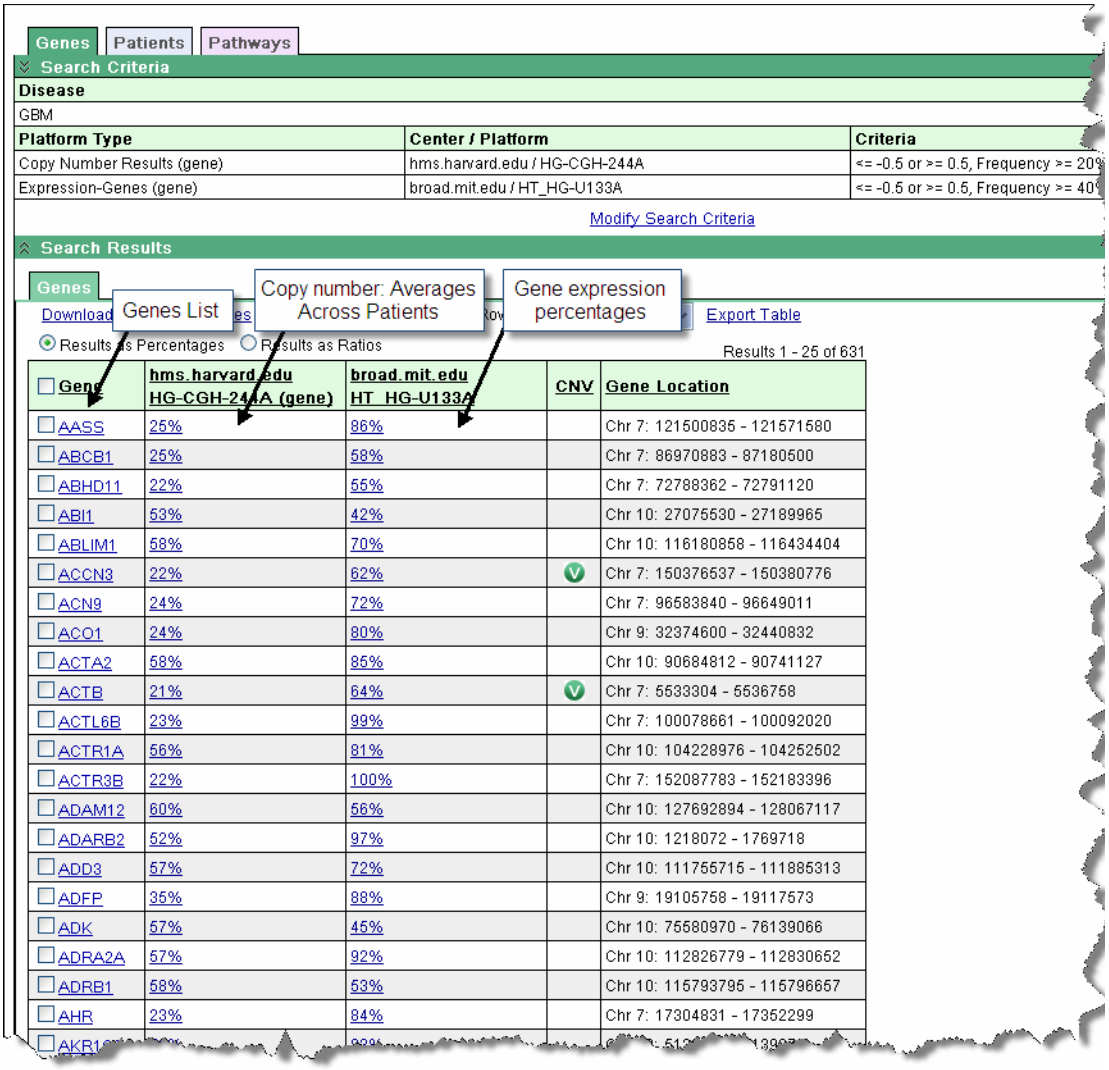
Queries launched in Genes and Patients modes generate tables that display the search results.

In Genes mode, the left column in the results table is always a list of gene symbols (*Figure 2.2*). In Patients mode, the left column in the table is always a list of patient IDs. In Pathway mode, it is always a list of pathway descriptions.

In Patient and Gene modes, a checkbox by each ID allows you to select one or more items to copy them *en masse* into the Patient List or Gene List textbox in the query panel. The header for this column also contains a checkbox that selects all the genes or patients only for the current page. Once items are selected, click the **Copy Genes** (or **Patients**) **to Search** button to copy them to the lists. This allows you to rerun the search with modified criteria. For more information, see *Modifying a Search* on page 13.

Numeric Columns

If you have selected the **Average Across Patients** option in the Copy Number abnormality type, that column appears just to the right of the Gene column (*Figure 2.2*). Other abnormality columns display initially as percentages, with the **Results as Percentages** button selected (*Figure 2.2*). Select the **Results as Ratio** button above the table to display a ratio showing patients/genes tested over total patients/genes. You can toggle between percentages and ratios using these buttons.



Search Criteria

Disease		
GBM		
Platform Type	**Center / Platform**	**Criteria**
Copy Number Results (gene)	hms.harvard.edu / HG-CGH-244A	<= -0.5 or >= 0.5, Frequency >= 20
Expression-Genes (gene)	broad.mit.edu / HT_HG-U133A	<= -0.5 or >= 0.5, Frequency >= 40

Modify Search Criteria

Search Results

Genes List

Copy number: Averages Across Patients

Gene expression percentages

● Results as Percentages ○ Results as Ratios

Results 1 - 25 of 631

□ Gene	hms.harvard.edu HG-CGH-244A (gene)	broad.mit.edu HT_HG-U133A	CNV	Gene Location
□ AASS	25%	86%		Chr 7: 121500835 - 121571580
□ ABCB1	25%	58%		Chr 7: 86970883 - 87180500
□ ABHD11	22%	55%		Chr 7: 72788362 - 72791120
□ ABI1	53%	42%		Chr 10: 27075530 - 27189965
□ ABLIM1	58%	70%		Chr 10: 116180858 - 116434404
□ ACCN3	22%	62%	✓	Chr 7: 150376537 - 150380776
□ ACN9	24%	72%		Chr 7: 96583840 - 96649011
□ ACO1	24%	80%		Chr 9: 32374600 - 32440832
□ ACTA2	58%	85%		Chr 10: 90684812 - 90741127
□ ACTB	21%	64%	✓	Chr 7: 5533304 - 5536758
□ ACTL6B	23%	99%		Chr 7: 100078661 - 100092020
□ ACTR1A	56%	81%		Chr 10: 104228976 - 104252502
□ ACTR3B	22%	100%		Chr 7: 152087783 - 152183396
□ ADAM12	60%	56%		Chr 10: 127692894 - 128067117
□ ADARB2	52%	97%		Chr 10: 1218072 - 1769718
□ ADD3	57%	72%		Chr 10: 111755715 - 111885313
□ ADFP	35%	88%		Chr 9: 19105758 - 19117573
□ ADK	57%	45%		Chr 10: 75580970 - 76139066
□ ADRA2A	57%	92%		Chr 10: 112826779 - 112830652
□ ADRB1	58%	53%		Chr 10: 115793795 - 115796657
□ AHR	23%	84%		Chr 7: 17304831 - 17352299
□ AKR1				

Figure 2.2 Example search results displaying corresponding search criteria in the search panel. The third column calculations could convert to ratio display by clicking the Results as Ratio button.

The application filters out rows with blank cells. If a calculation results in less than 0.5%, the result displays as <1%.

If you selected the copy number **Average across Patients**, the resulting column contains real numbers. If you selected a correlation, the column contains real numbers plus a p-value in scientific notation.

Correlations are found using the Pearson algorithm. For more information on the calculation, see *Correlations Options* on page 13.

Gene-Patient Pivot Tables

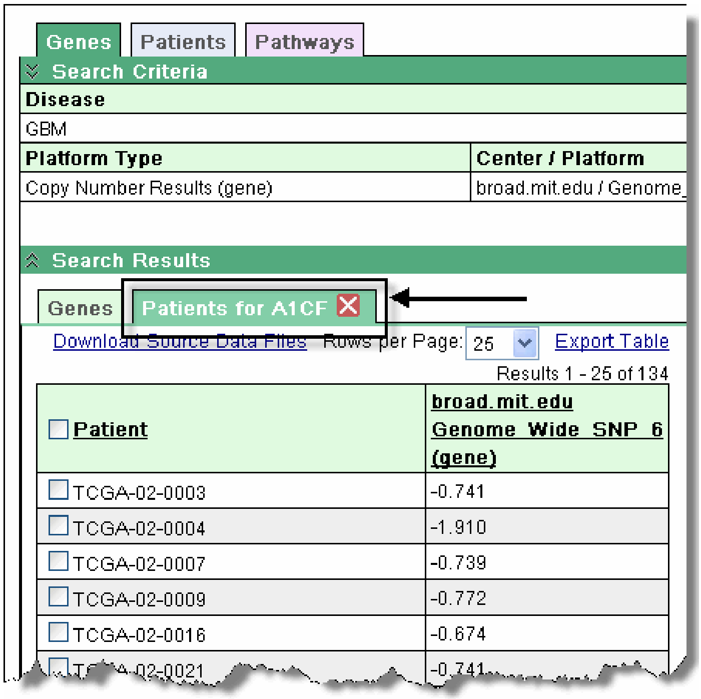
After you have performed a gene or patient search, in Gene or Patient modes you can “pivot” the information on a particular gene or genetic element to show data points for each patient or on a particular patient to show data points for each gene or genetic element.

The data that initially displays in the data browser represents aggregate numbers per gene or per patient. The pivot feature displays raw values for each patient or genetic element.

You can generate a pivot table from any column in query results except correlation: copy number, gene expression, mutation, methylation. Click a selected value which is hypertext linked. A link is present only in percent values. Only the column selected will be present in the pivoted table.

Gene Pivot Table

The browser now displays a new tab to the right of the Gene Summary tab (*Figure 2.3*). The new tab shows the gene name you selected. The table reveals raw values corresponding to that gene for each of the patients used to calculate the aggregate figure in the previous tab.



Disease	
GBM	
Platform Type	**Center / Platform**
Copy Number Results (gene)	broad.mit.edu / Genome_

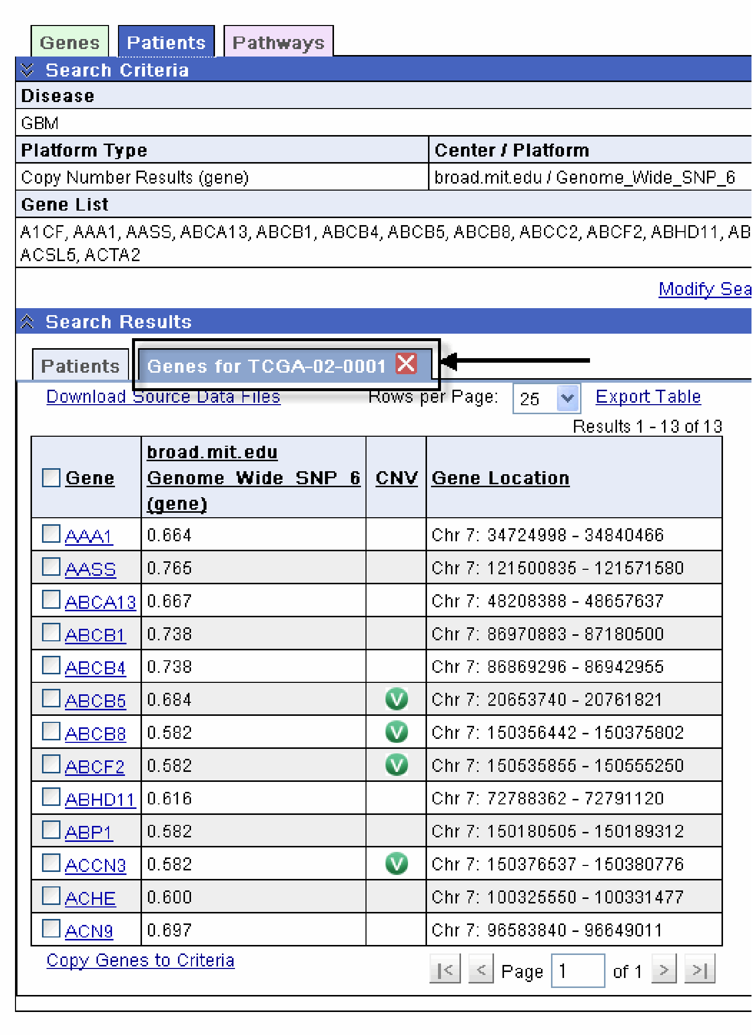
Results 1 - 25 of 134

□ Patient	broad.mit.edu Genome_Wide_SNP_6 (gene)
□ TCGA-02-0003	-0.741
□ TCGA-02-0004	-1.910
□ TCGA-02-0007	-0.739
□ TCGA-02-0009	-0.772
□ TCGA-02-0016	-0.674
□ TCGA-02-0021	-0.741

Figure 2.3 Gene pivot table displaying raw values for all patients in the aggregate figure for the given gene

Patient Pivot Table

The browser now displays a new tab to the right of the Patient Summary tab (*Figure 2.3*). The new tab name shows the patient you selected. The table reveals raw values corresponding to that patient for each of the genes used to calculate the aggregate figure in the previous tab. .



Disease	
GBM	
Platform Type	**Center / Platform**
Copy Number Results (gene)	broad.mit.edu / Genome_Wide_SNP_6
Gene List	
A1CF, AAA1, AASS, ABCA13, ABCB1, ABCB4, ABCB5, ABCB8, ABCC2, ABCF2, ABHD11, AB ACSL5, ACTA2	

Results 1 - 13 of 13

□ Gene	broad.mit.edu Genome_Wide_SNP_6 (gene)	CNV	Gene Location
□ AAA1	0.664		Chr 7: 34724998 - 34840466
□ AASS	0.765		Chr 7: 121500835 - 121571580
□ ABCA13	0.667		Chr 7: 48208388 - 48657637
□ ABCB1	0.738		Chr 7: 86970883 - 87180500
□ ABCB4	0.738		Chr 7: 86869296 - 86942955
□ ABCB5	0.684	V	Chr 7: 20653740 - 20761821
□ ABCB8	0.582	V	Chr 7: 150356442 - 150375802
□ ABCF2	0.582	V	Chr 7: 150535855 - 150555250
□ ABHD11	0.616		Chr 7: 72788362 - 72791120
□ ABP1	0.582		Chr 7: 150180505 - 150189312
□ ACCN3	0.582	V	Chr 7: 150376537 - 150380776
□ ACHE	0.600		Chr 7: 100325550 - 100331477
□ ACN9	0.697		Chr 7: 96583840 - 96649011

Figure 2.4 Patient pivot table displaying raw values for all genes in the aggregate figure for the given patient

Annotations

The data browser can show two types of annotations in the results table.

A “row annotation” does not apply to any particular column, but to each row as a whole. These include:

- *CNV:* In Gene mode, an icon (V) appears for each gene that is in a known region of *copy number* variation.
- *Location*: In Gene mode, the chromosome number, start and stop codons are listed for each gene.

A “value annotation” applies to a particular cell.

- *Paired*: in Patient mode, an icon (P) appears for each copy number value where the value is derived by comparing tumor against normal for the same patient.
- *Correlation p-value*: in Gene mode, each correlation has a corresponding P-number value indicating the probable significance of the correlation. For more information, see *Calculating P-values* on page 20.

Note: The colors of the "V" and "P" icons correspond to the mode in which you are working.

Pathways Results

In Pathway mode, when you click a pathway name link, on the Pathway Summary page, a new tab containing information about that pathway appears:

- A *Biocarta* pathway diagram. Three main cellular components (nucleus, cytoplasm and cellular membrane) are shown along with graphical elements that represent individual proteins. Gene products whose genes match the filter are highlighted on the illustration.

  Note: Gene symbol(s) on the figure may not match the symbol(s) used on other Data Portal pages. The notation may follow the Biocarta naming convention. These are indicated in the table underneath the diagram.
- By default, the **Show Genes Matching Search** radio button is selected and a table of genes that match the filter (the same genes whose products are highlighted in the diagram) displays. The browser also shows abnormalities defined in the search panel, as well as correlation values.
- Click **Show All Pathway Genes** to display a table that shows all genes represented in the pathway, along with their abnormality values. In this case, the columns are defined by the filter, but the rows are not filtered. Some duplicate gene names may appear, if any of the data sources identified a gene as having different start/stop codons.
- In Pathway mode, a p-value indicating the probable significance of the pathway appears in the list. For more information about the p-value calculations, see *Calculating P-values* on page 20.

Calculating P-values

Two types of p-value can be calculated: one for correlations, the other for pathways. Both are expressed using scientific notation.

- The correlation p-value tells how likely it is that the accompanying correlation was found by chance. A low p-value signals the probability that the correlation is significant and not occurring by chance. A high p-value indicates the opposite, the probability that the correlation is not significant and is occurring by chance.

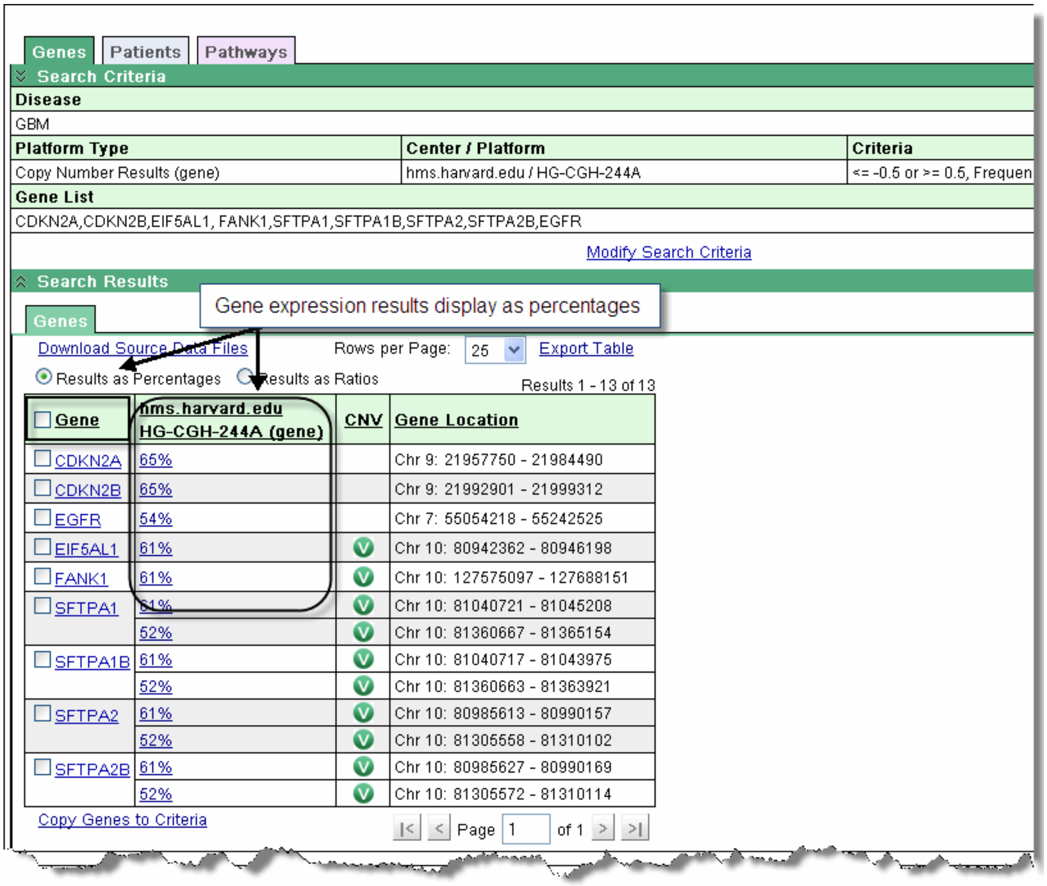
  While the p-value may suggest whether or not the results occur by chance, it is its relationship with the accompanying correlation value that makes the calculation meaningful.
- The pathway p-value tells how good a "fit" the pathway is for the filter. A lower number means a better fit, and higher probable significance. The Fischer's Exact algorithm is used for the calculation.

Search Results Examples

Once you have launched a search, the Data Browser displays search results below a search criteria summary. The results are color coded according to the tab on which you launched the search. This section displays examples of search results.

Genes Mode

When you selected the **Genes** tab in setting search criteria, all areas of the filter panel were enabled. Results are then summarized by gene symbol. The results table's leftmost column shows gene symbols; the rightmost column displays the gene location on a chromosome (*Figure 2.6*). ..



Disease		
GBM		
Platform Type	Center / Platform	Criteria
Copy Number Results (gene)	hms.harvard.edu / HG-CGH-244A	<= -0.5 or >= 0.5, Frequen
Gene List		
CDKN2A,CDKN2B,EIF5AL1, FANK1,SFTPA1,SFTPA1B,SFTPA2,SFTPA2B,EGFR		

Gene expression results display as percentages

Rows per Page: 25

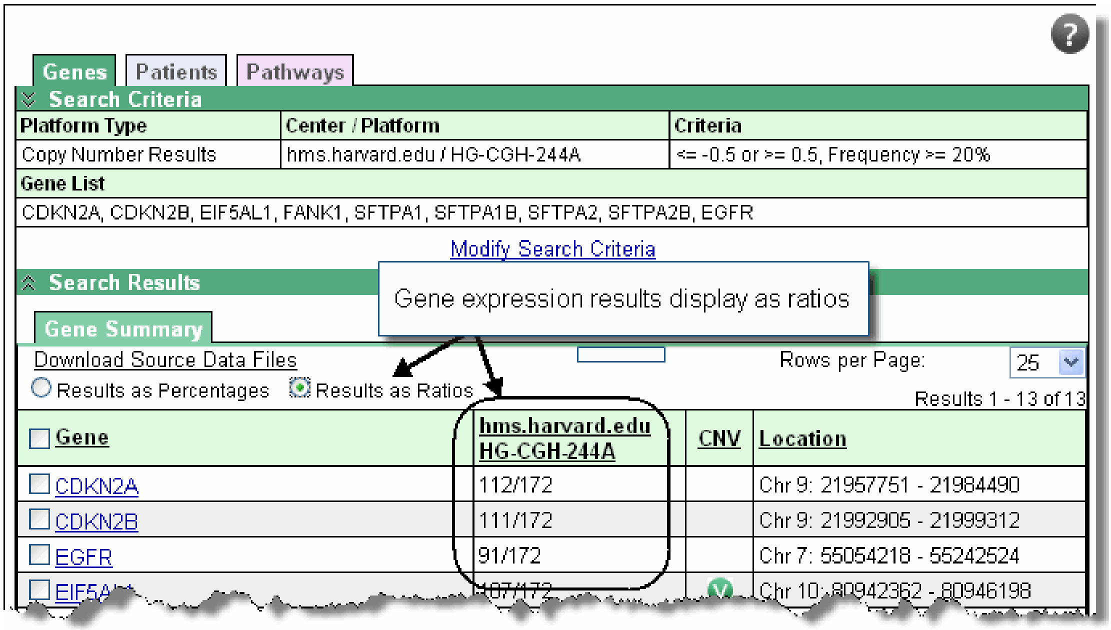
● Results as Percentages ○ Results as Ratios

Results 1 - 13 of 13

□ Gene	hms.harvard.edu HG-CGH-244A (gene)	CNV	Gene Location
□ CDKN2A	65%		Chr 9: 21957750 - 21984490
□ CDKN2B	65%		Chr 9: 21992901 - 21999312
□ EGFR	54%		Chr 7: 55054218 - 55242525
□ EIF5AL1	61%	✓	Chr 10: 80942362 - 80946198
□ FANK1	61%	✓	Chr 10: 127575097 - 127688151
□ SFTPA1	61%	✓	Chr 10: 81040721 - 81045208
	52%	✓	Chr 10: 81360667 - 81365154
□ SFTPA1B	61%	✓	Chr 10: 81040717 - 81043975
	52%	✓	Chr 10: 81360663 - 81363921
□ SFTPA2	61%	✓	Chr 10: 80985613 - 80990157
	52%	✓	Chr 10: 81305558 - 81310102
□ SFTPA2B	61%	✓	Chr 10: 80985627 - 80990169
	52%	✓	Chr 10: 81305572 - 81310114

Figure 2.5 Gene summary displayed as percentages

By default *gene expression* results display as percentages (*Figure 2.5*), but you can change the display to ratio (*Figure 2.6*).



Search Criteria

Platform Type	Center / Platform	Criteria
Copy Number Results	hms.harvard.edu / HG-CGH-244A	<= -0.5 or >= 0.5, Frequency >= 20%

Gene List

CDKN2A, CDKN2B, EIF5AL1, FANK1, SFTPA1, SFTPA1B, SFTPA2, SFTPA2B, EGFR

Modify Search Criteria

Search Results

Gene expression results display as ratios

Gene Summary

Download Source Data Files

○ Results as Percentages ● Results as Ratios

Rows per Page: 25

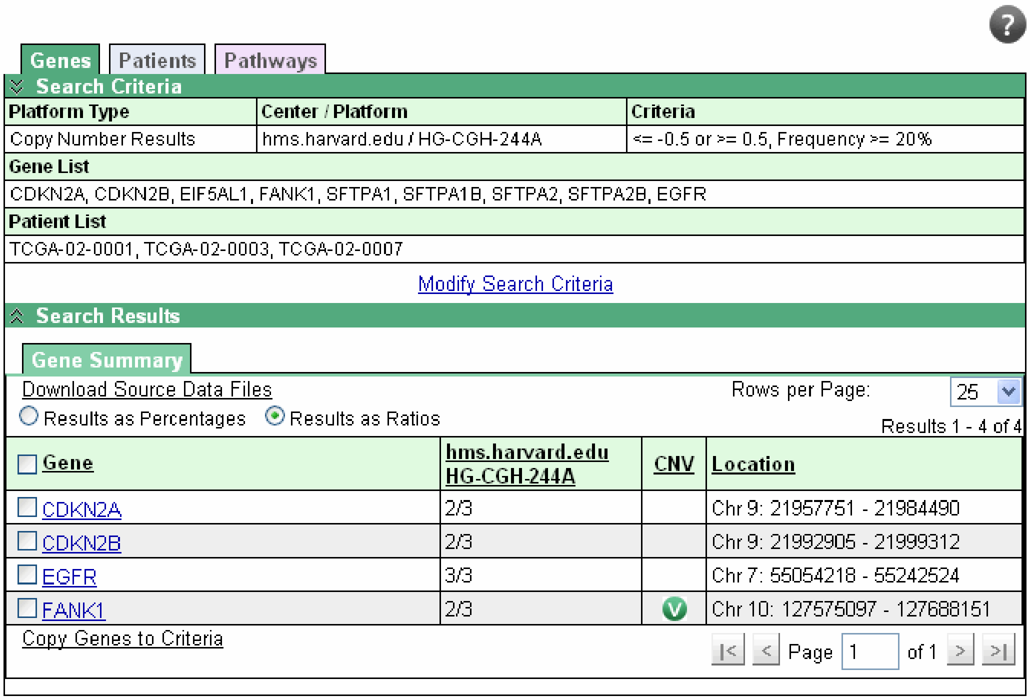
Results 1 - 13 of 13

□ Gene	hms.harvard.edu HG-CGH-244A	CNV	Location
□ CDKN2A	112/172		Chr 9: 21957751 - 21984490
□ CDKN2B	111/172		Chr 9: 21992905 - 21999312
□ EGFR	91/172		Chr 7: 55054218 - 55242524
□ EIF5AL1	107/172	V	Chr 10: 80942362 - 80946198

Figure 2.6 Gene summary results displayed as ratios

If you specified one or more genes by name as in the example used in *Figure 2.6*, the results display one row per gene, with the gene(s) listed in the first column. Because **All Patients** was selected in the left panel, the second column of the results table displays initially the percentage calculated of patients who tested positive for the gene versus all patients who were tested using that platform. Select **View as Ratio**. The second column now displays the ratio results for the abnormality you defined. For example, in *Figure 2.6*, 172 patients were tested using the HG-CGH-244A platform. The ratio results shows a range of 87 - 112 out of 172 patients were positive for the *copy number* abnormality represented by the genes on the first page of results.

If you restrict the query to a patient list and re-run the query, the search results could display as shown in *Figure 2.7*: all four patients have at least one copy number abnormality defined in the search parameters.



Platform Type	Center / Platform	Criteria
Copy Number Results	hms.harvard.edu / HG-CGH-244A	<= -0.5 or >= 0.5, Frequency >= 20%

Gene List

CDKN2A, CDKN2B, EIF5AL1, FANK1, SFTPA1, SFTPA1B, SFTPA2, SFTPA2B, EGFR

Patient List

TCGA-02-0001, TCGA-02-0003, TCGA-02-0007

Rows per Page: 25

○ Results as Percentages ● Results as Ratios

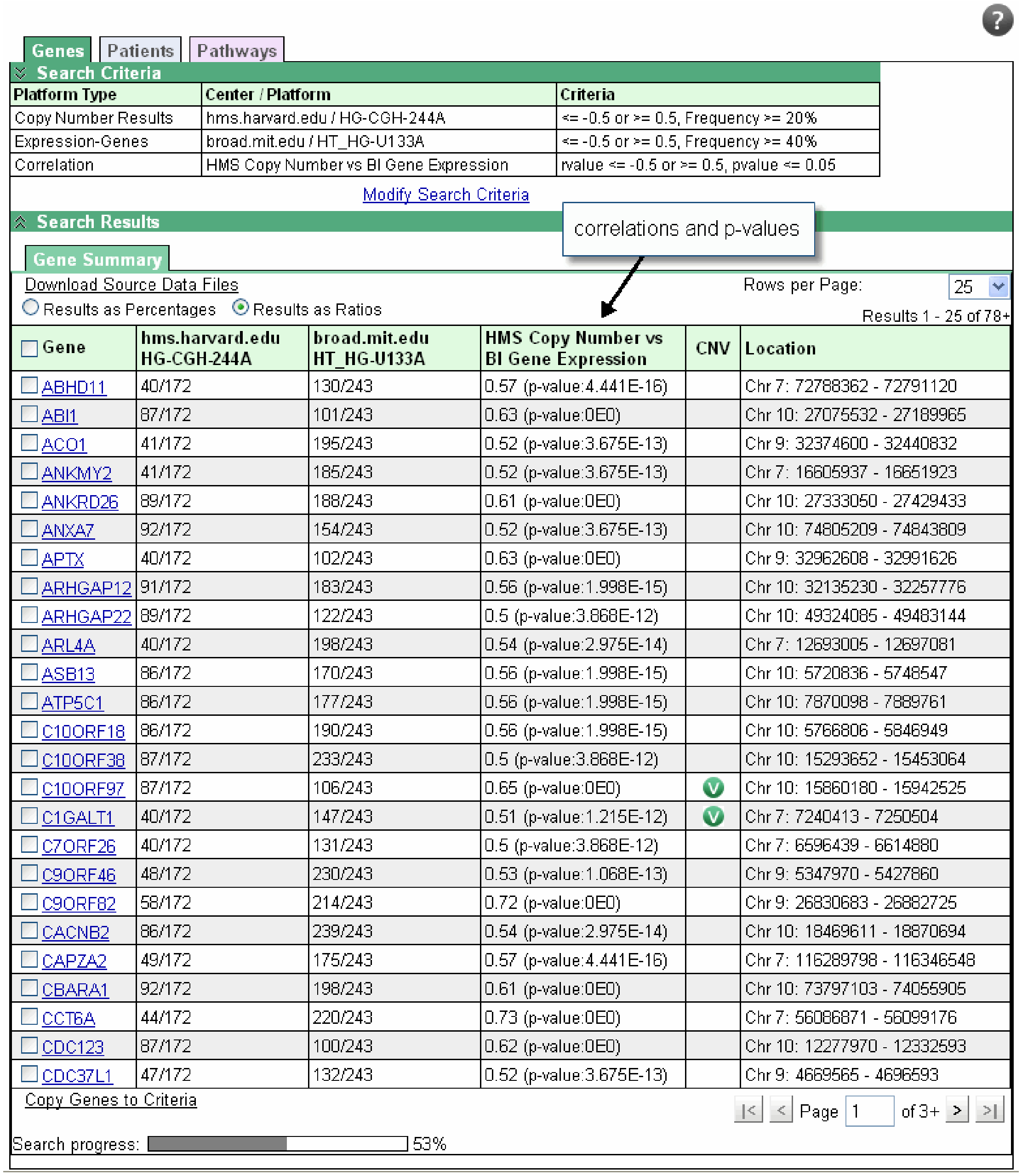
Results 1 - 4 of 4

□ Gene	hms.harvard.edu HG-CGH-244A	CNV	Location
□ CDKN2A	2/3		Chr 9: 21957751 - 21984490
□ CDKN2B	2/3		Chr 9: 21992905 - 21999312
□ EGFR	3/3		Chr 7: 55054218 - 55242524
□ FANK1	2/3	V	Chr 10: 127575097 - 127688151

|< < Page 1 of 1 > >|

Figure 2.7 Query parameters restricted by a patients list display results calculated accordingly

If you select the **Correlations** option on the query panel, the results display a column that shows both the correlation number followed by a p-value (*Figure 2.8*). For more information about these calculations, see *Correlations Options* on page 13 and *Calculating P-values* on page 20.



Platform Type	Center / Platform	Criteria
Copy Number Results	hms.harvard.edu / HG-CGH-244A	<= -0.5 or >= 0.5, Frequency >= 20%
Expression-Genes	broad.mit.edu / HT_HG-U133A	<= -0.5 or >= 0.5, Frequency >= 40%
Correlation	HMS Copy Number vs BI Gene Expression	rvalue <= -0.5 or >= 0.5, pvalue <= 0.05

○ Results as Percentages ● Results as Ratios

Results 1 - 25 of 78+

□ Gene	hms.harvard.edu HG-CGH-244A	broad.mit.edu HT_HG-U133A	HMS Copy Number vs BI Gene Expression	CNV	Location
□ ABHD11	40/172	130/243	0.57 (p-value:4.441E-16)		Chr 7: 72788362 - 72791120
□ ABI1	87/172	101/243	0.63 (p-value:0E0)		Chr 10: 27075532 - 27189965
□ ACO1	41/172	195/243	0.52 (p-value:3.675E-13)		Chr 9: 32374600 - 32440832
□ ANKMY2	41/172	185/243	0.52 (p-value:3.675E-13)		Chr 7: 16605937 - 16651923
□ ANKRD26	89/172	188/243	0.61 (p-value:0E0)		Chr 10: 27333050 - 27429433
□ ANXA7	92/172	154/243	0.52 (p-value:3.675E-13)		Chr 10: 74805209 - 74843809
□ APTX	40/172	102/243	0.63 (p-value:0E0)		Chr 9: 32962608 - 32991626
□ ARHGAP12	91/172	183/243	0.56 (p-value:1.998E-15)		Chr 10: 32135230 - 32257776
□ ARHGAP22	89/172	122/243	0.5 (p-value:3.868E-12)		Chr 10: 49324085 - 49483144
□ ARL4A	40/172	198/243	0.54 (p-value:2.975E-14)		Chr 7: 12693005 - 12697081
□ ASB13	86/172	170/243	0.56 (p-value:1.998E-15)		Chr 10: 5720836 - 5748547
□ ATP5C1	86/172	177/243	0.56 (p-value:1.998E-15)		Chr 10: 7870098 - 7889761
□ C10ORF18	86/172	190/243	0.56 (p-value:1.998E-15)		Chr 10: 5766806 - 5846949
□ C10ORF38	87/172	233/243	0.5 (p-value:3.868E-12)		Chr 10: 15293652 - 15453064
□ C10ORF97	87/172	106/243	0.65 (p-value:0E0)	V	Chr 10: 15860180 - 15942525
□ C1GALT1	40/172	147/243	0.51 (p-value:1.215E-12)	V	Chr 7: 7240413 - 7250504
□ C7ORF26	40/172	131/243	0.5 (p-value:3.868E-12)		Chr 7: 6596439 - 6614880
□ C9ORF46	48/172	230/243	0.53 (p-value:1.068E-13)		Chr 9: 5347970 - 5427860
□ C9ORF82	58/172	214/243	0.72 (p-value:0E0)		Chr 9: 26830683 - 26882725
□ CACNB2	86/172	239/243	0.54 (p-value:2.975E-14)		Chr 10: 18469611 - 18870694
□ CAPZA2	49/172	175/243	0.57 (p-value:4.441E-16)		Chr 7: 116289798 - 116346548
□ CBARA1	92/172	198/243	0.61 (p-value:0E0)		Chr 10: 73797103 - 74055905
□ CCT6A	44/172	220/243	0.73 (p-value:0E0)		Chr 7: 56086871 - 56099176
□ CDC123	87/172	100/243	0.62 (p-value:0E0)		Chr 10: 12277970 - 12332593
□ CDC37L1	47/172	132/243	0.52 (p-value:3.675E-13)		Chr 9: 4669565 - 4696593

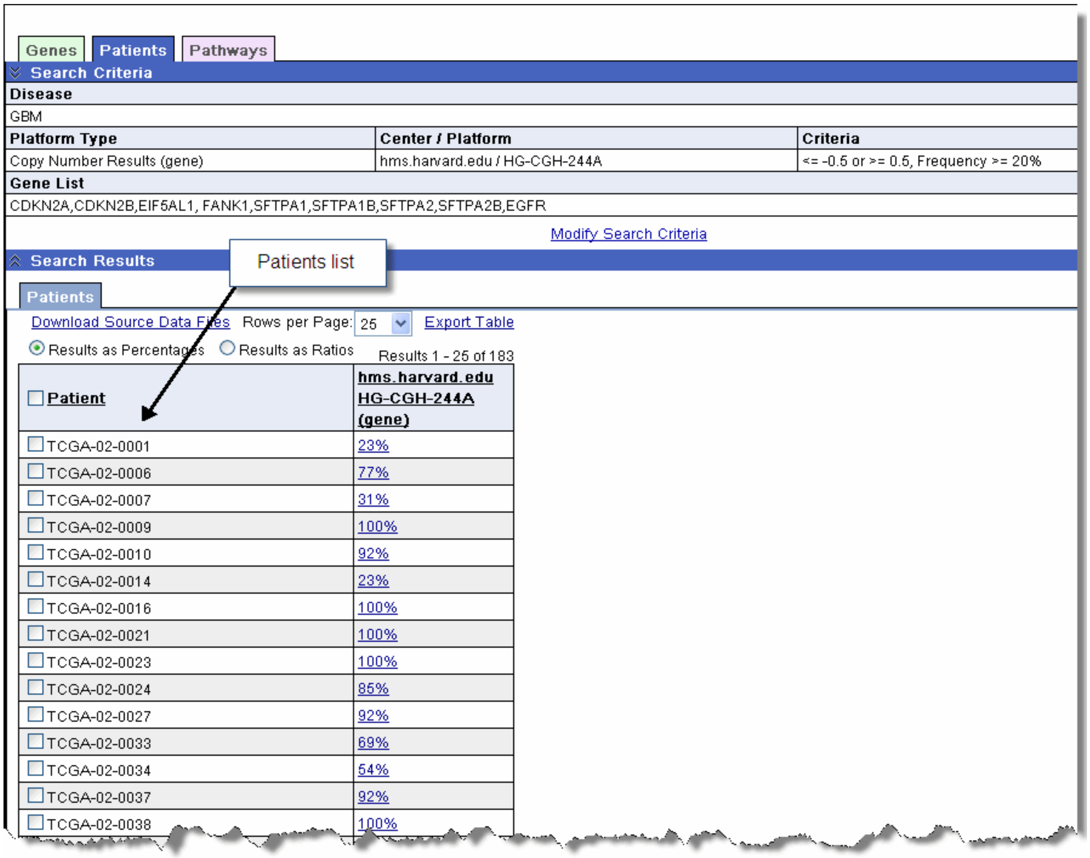
Search progress: 53%

Figure 2.8 A Gene search filtered by copy number and gene expression. Column 4 shows correlation and p-value calculations, as defined in the search parameters

To copy gene names to the Gene List text box on the query panel, check the box(es) corresponding to the name in the Gene list and click the **Copy Genes to Search** button. Click the Modify Search Criteria to return to the search page. At this point, you can rerun the search with the specified gene names added to your criteria.

Patients Mode

*I*f you select the **Patients** tab before launching a search, the search results display patient IDs in the left column, one patient per row (*Figure 2.9*).



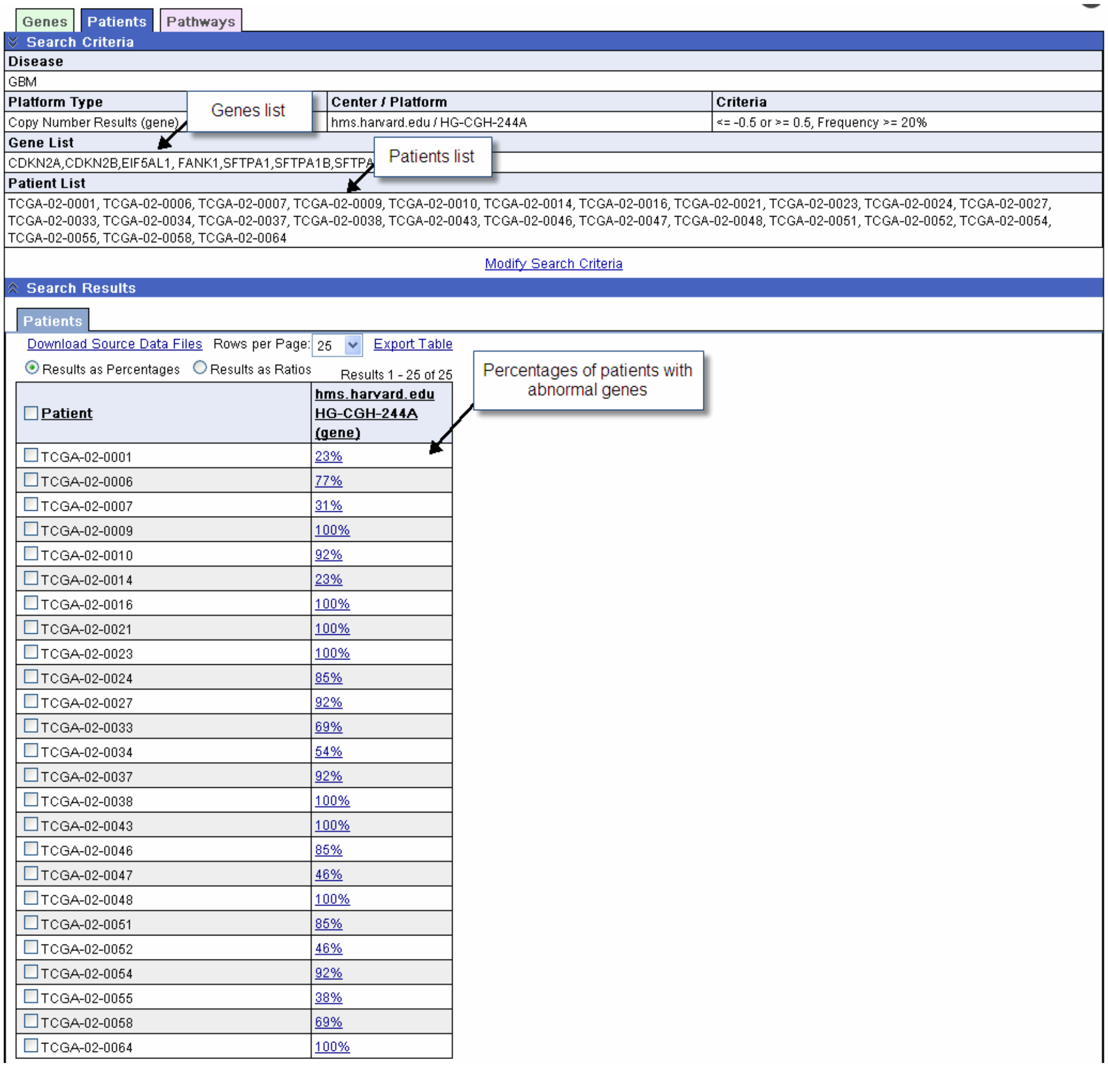
Disease		
GBM		
Platform Type	**Center / Platform**	**Criteria**
Copy Number Results (gene)	hms.harvard.edu / HG-CGH-244A	<= -0.5 or >= 0.5, Frequency >= 20%
Gene List		
CDKN2A,CDKN2B,EIF5AL1, FANK1,SFTPA1,SFTPA1B,SFTPA2,SFTPA2B,EGFR		

Results 1 - 25 of 183

□ Patient	hms.harvard.edu HG-CGH-244A (gene)
□ TCGA-02-0001	23%
□ TCGA-02-0006	77%
□ TCGA-02-0007	31%
□ TCGA-02-0009	100%
□ TCGA-02-0010	92%
□ TCGA-02-0014	23%
□ TCGA-02-0016	100%
□ TCGA-02-0021	100%
□ TCGA-02-0023	100%
□ TCGA-02-0024	85%
□ TCGA-02-0027	92%
□ TCGA-02-0033	69%
□ TCGA-02-0034	54%
□ TCGA-02-0037	92%
□ TCGA-02-0038	100%

Figure 2.9 In Patients mode, this search was run with a list of abnormal genes against all patients.

If you run the search with a specified group of genes AND a specified list of patients in the search panel, the results display the proportion of abnormal genes for each of these patients out of the genes tested using this platform. In other words, the results display the percentage of patients that have the abnormal genes (*Figure 2.10*).



Search Criteria

Disease		
GBM		
Platform Type	**Center / Platform**	**Criteria**
Copy Number Results (gene)	hms.harvard.edu / HG-CGH-244A	<= -0.5 or >= 0.5, Frequency >= 20%
Gene List		
CDKN2A,CDKN2B,EIF5AL1, FANK1,SFTPA1,SFTPA1B,SFTPA		
Patient List		
TCGA-02-0001, TCGA-02-0006, TCGA-02-0007, TCGA-02-0009, TCGA-02-0010, TCGA-02-0014, TCGA-02-0016, TCGA-02-0021, TCGA-02-0023, TCGA-02-0024, TCGA-02-0027, TCGA-02-0033, TCGA-02-0034, TCGA-02-0037, TCGA-02-0038, TCGA-02-0043, TCGA-02-0046, TCGA-02-0047, TCGA-02-0048, TCGA-02-0051, TCGA-02-0052, TCGA-02-0054, TCGA-02-0055, TCGA-02-0058, TCGA-02-0064		

Modify Search Criteria

Search Results

Patients

Download Source Data Files Rows per Page: 25 Export Table

● Results as Percentages ○ Results as Ratios Results 1 - 25 of 25

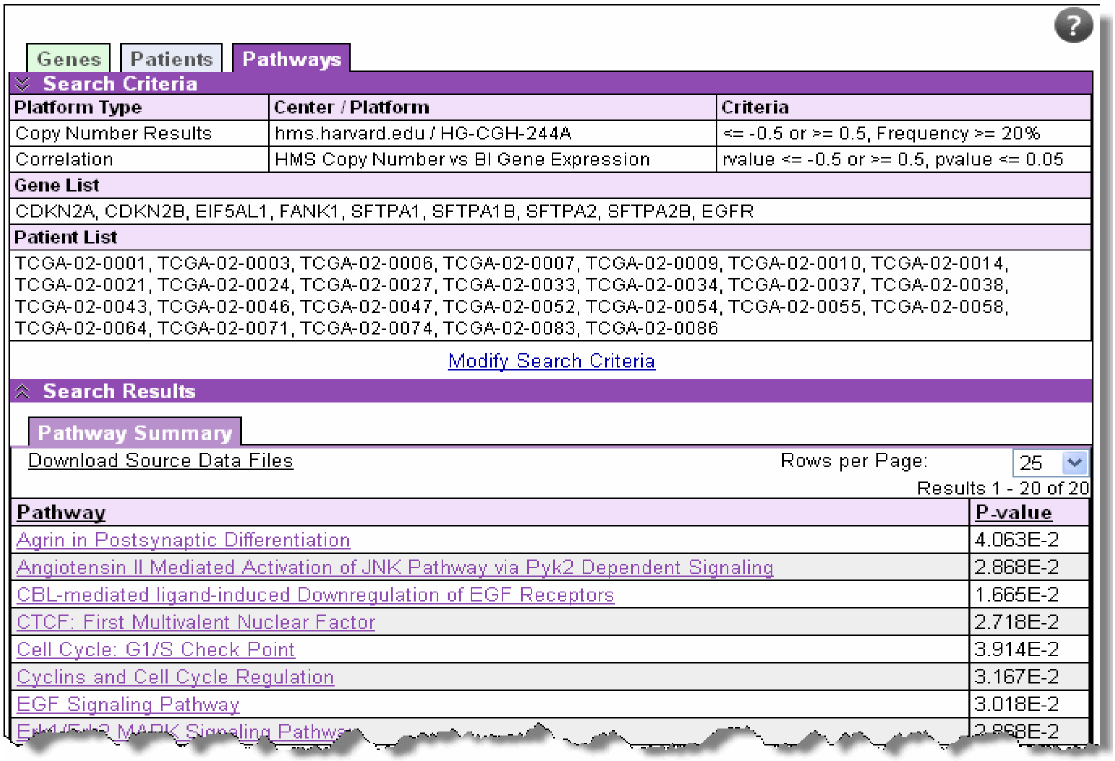
□ Patient	hms.harvard.edu HG-CGH-244A (gene)
□ TCGA-02-0001	23%
□ TCGA-02-0006	77%
□ TCGA-02-0007	31%
□ TCGA-02-0009	100%
□ TCGA-02-0010	92%
□ TCGA-02-0014	23%
□ TCGA-02-0016	100%
□ TCGA-02-0021	100%
□ TCGA-02-0023	100%
□ TCGA-02-0024	85%
□ TCGA-02-0027	92%
□ TCGA-02-0033	69%
□ TCGA-02-0034	54%
□ TCGA-02-0037	92%
□ TCGA-02-0038	100%
□ TCGA-02-0043	100%
□ TCGA-02-0046	85%
□ TCGA-02-0047	46%
□ TCGA-02-0048	100%
□ TCGA-02-0051	85%
□ TCGA-02-0052	46%
□ TCGA-02-0054	92%
□ TCGA-02-0055	38%
□ TCGA-02-0058	69%
□ TCGA-02-0064	100%

Figure 2.10 A specified list of patients tested against a specified list of abnormal genes. The second column displays the percentage of patients with abnormal genes against the total patients in the list

To copy patient names to the Patient List text box on the query panel, check the box(es) corresponding to the name(s) in the patient list and click the **Copy Patients to Search** button. Click Modify Search Criteria to return to the search page. At this point, you can rerun the search with the specified patient names added to your criteria.

Pathways Mode

Pathway results display alphabetically on a summary tab (*Figure 2.11*).



Platform Type	Center / Platform	Criteria
Copy Number Results	hms.harvard.edu / HG-CGH-244A	<= -0.5 or >= 0.5, Frequency >= 20%
Correlation	HMS Copy Number vs BI Gene Expression	rvalue <= -0.5 or >= 0.5, pvalue <= 0.05

Gene List

CDKN2A, CDKN2B, EIF5AL1, FANK1, SFTPA1, SFTPA1B, SFTPA2, SFTPA2B, EGFR

Patient List

TCGA-02-0001, TCGA-02-0003, TCGA-02-0006, TCGA-02-0007, TCGA-02-0009, TCGA-02-0010, TCGA-02-0014, TCGA-02-0021, TCGA-02-0024, TCGA-02-0027, TCGA-02-0033, TCGA-02-0034, TCGA-02-0037, TCGA-02-0038, TCGA-02-0043, TCGA-02-0046, TCGA-02-0047, TCGA-02-0052, TCGA-02-0054, TCGA-02-0055, TCGA-02-0058, TCGA-02-0064, TCGA-02-0071, TCGA-02-0074, TCGA-02-0083, TCGA-02-0086

Modify Search Criteria

Rows per Page: 25

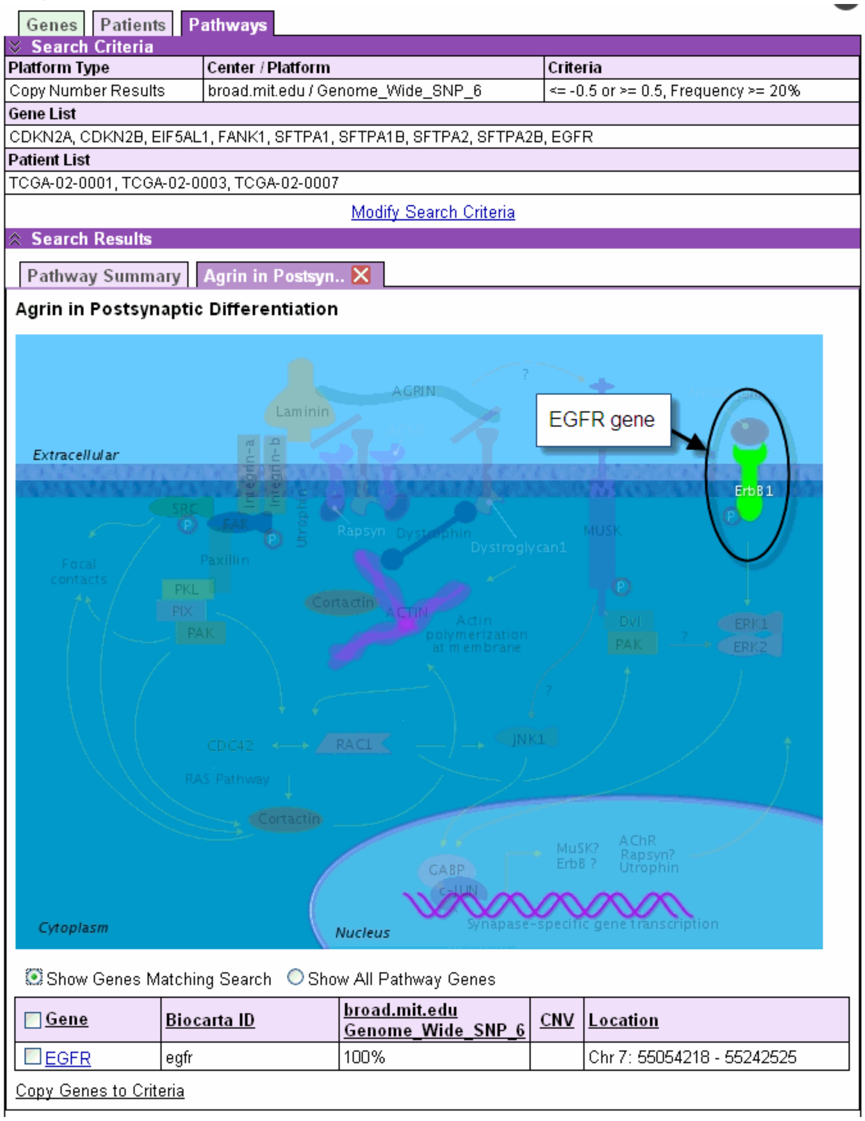
Results 1 - 20 of 20

Pathway	P-value
Agrin in Postsynaptic Differentiation	4.063E-2
Angiotensin II Mediated Activation of JNK Pathway via Pyk2 Dependent Signaling	2.868E-2
CBL-mediated ligand-induced Downregulation of EGF Receptors	1.665E-2
CTCF: First Multivalent Nuclear Factor	2.718E-2
Cell Cycle: G1/S Check Point	3.914E-2
Cyclins and Cell Cycle Regulation	3.167E-2
EGF Signaling Pathway	3.018E-2
Erk1/Erk2 MAPK Signaling Pathwa	2.868E-2

Figure 2.11 A search launched in Pathway mode displays results in a Pathway Summary tab. Each pathway displayed on the list uses at least one of the genes in the genes list on the search panel.

Pathways that are listed include at least one of the gene products stemming from genes in the genes list on the search panel. The Significance column indicates p-values. For more information about pathway p-values, see *Calculating P-values* on page 20.

Click on a pathway to open a new tab that is specific to the selected pathway (*Figure 2.12*).



Genes | Patients | Pathways

Search Criteria

Platform Type	Center / Platform	Criteria
Copy Number Results	broad.mit.edu / Genome_Wide_SNP_6	<= -0.5 or >= 0.5, Frequency >= 20%

Gene List

CDKN2A, CDKN2B, EIF5AL1, FANK1, SFTPA1, SFTPA1B, SFTPA2, SFTPA2B, EGFR

Patient List

TCGA-02-0001, TCGA-02-0003, TCGA-02-0007

Modify Search Criteria

Search Results

Pathway Summary | Agrin in Postsyn..

Agrin in Postsynaptic Differentiation

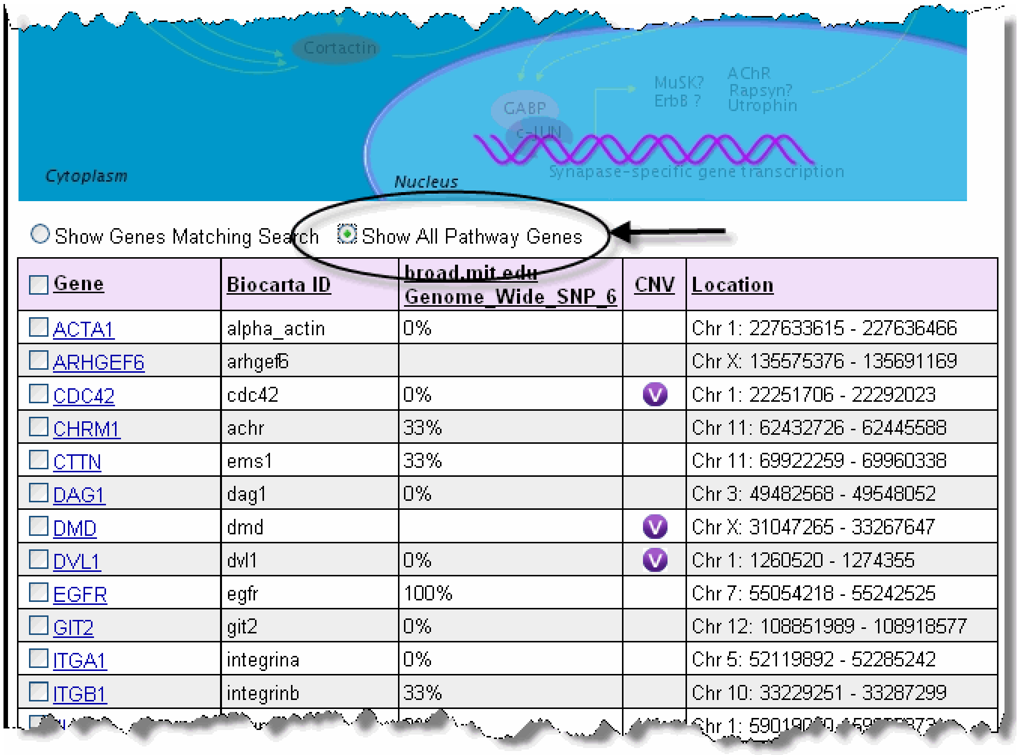
● Show Genes Matching Search ○ Show All Pathway Genes

□ Gene	Biocarta ID	broad.mit.edu Genome_Wide_SNP_6	CNV	Location
□ EGFR	egfr	100%		Chr 7: 55054218 - 55242525

Copy Genes to Criteria

Figure 2.12 Agrin pathway diagram highlighting the EGFR gene, identified as erbB1 on the diagram.

This tab displays a pathway diagram highlighting the specified gene(s) whose product is a component of the pathway. A gene summary table below the figure displays data for any of the abnormality criteria you set in the query panel. Initially the table displays only the gene(s) you specified in the search. Click the **Show All Pathway Genes** button to display all genes found in the pathway (*Figure 2.13*).



□ Gene	Biocarta ID	broad.mit.edu Genome_Wide_SNP_6	CNV	Location
□ ACTA1	alpha_actin	0%		Chr 1: 227633615 - 227636466
□ ARHGEF6	arhgef6			Chr X: 135575376 - 135691169
□ CDC42	cdc42	0%	V	Chr 1: 22251706 - 22292023
□ CHRM1	achr	33%		Chr 11: 62432726 - 62445588
□ CTTN	ems1	33%		Chr 11: 69922259 - 69960338
□ DAG1	dag1	0%		Chr 3: 49482568 - 49548052
□ DMD	dmd		V	Chr X: 31047265 - 33267647
□ DVL1	dvl1	0%	V	Chr 1: 1260520 - 1274355
□ EGFR	egfr	100%		Chr 7: 55054218 - 55242525
□ GIT2	git2	0%		Chr 12: 108851989 - 108918577
□ ITGA1	integrina	0%		Chr 5: 52119892 - 52285242
□ ITGB1	integrinb	33%		Chr 10: 33229251 - 33287299

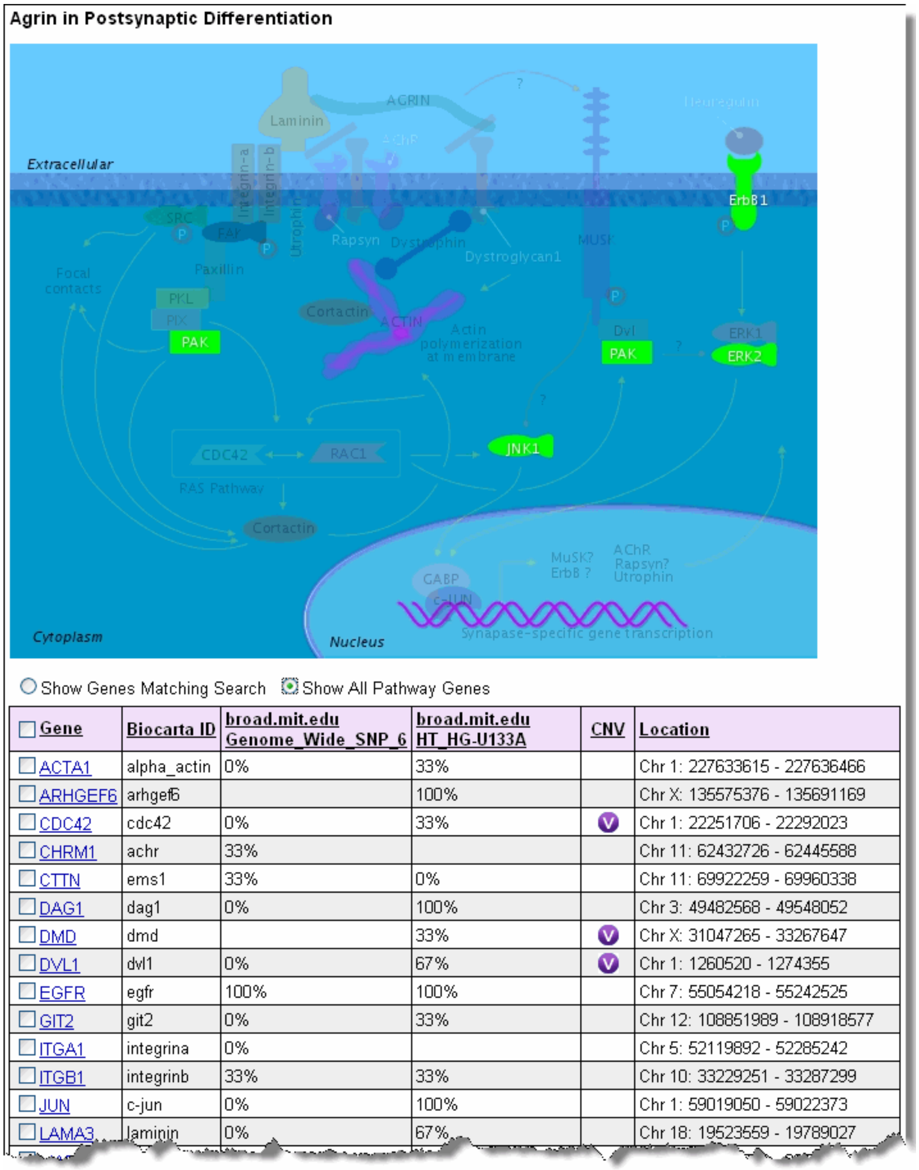
Figure 2.13 All genes represented in a pathway display in the pathway table if you so indicate by selecting the ***Show All Pathway Genes*** *button below the pathway diagram.*

Modifying a Pathway

With one pathway displaying, you can modify the criteria and update the pathway view. Note that after you do that, the **Search** button on the query page changes to **Update Pathway**.

After you modify the filter criteria, clicking the button redisplays the pathway, originally shown in *Figure 2.12*, and the table below the pathway is reconfigured, adding new columns, as appropriate, to display the information based on modified criteria (*Figure 2.14*).

Additionally, the Pathway Summary section shows all pathways that have the additional genes in common with the current pathway.



Agrin in Postsynaptic Differentiation

○ Show Genes Matching Search ● Show All Pathway Genes

☐ Gene	Biocarta ID	broad.mit.edu Genome_Wide_SNP_6	broad.mit.edu HT_HG-U133A	CNV	Location
☐ ACTA1	alpha_actin	0%	33%		Chr 1: 227633615 - 227636466
☐ ARHGEF6	arhgef6		100%		Chr X: 135575376 - 135691169
☐ CDC42	cdc42	0%	33%	V	Chr 1: 22251706 - 22292023
☐ CHRM1	achr	33%			Chr 11: 62432726 - 62445588
☐ CTTN	ems1	33%	0%		Chr 11: 69922259 - 69960338
☐ DAG1	dag1	0%	100%		Chr 3: 49482568 - 49548052
☐ DMD	dmd		33%	V	Chr X: 31047265 - 33267647
☐ DVL1	dvl1	0%	67%	V	Chr 1: 1260520 - 1274355
☐ EGFR	egfr	100%	100%		Chr 7: 55054218 - 55242525
☐ GIT2	git2	0%	33%		Chr 12: 108851989 - 108918577
☐ ITGA1	integrina	0%			Chr 5: 52119892 - 52285242
☐ ITGB1	integrinb	33%	33%		Chr 10: 33229251 - 33287299
☐ JUN	c-jun	0%	100%		Chr 1: 59019050 - 59022373
☐ LAMA3	laminin	0%	67%		Chr 18: 19523559 - 19789027

Figure 2.14 A pathway updated from Figure 2.13 shows in the results table additional abnormality data. and additional genes in the pathway.

These examples illustrate that the results that display depend on the way you define the gene and patient lists and the mode you select for the query.

See also *Modifying a Search* on page 13.

Downloading Source Data Files

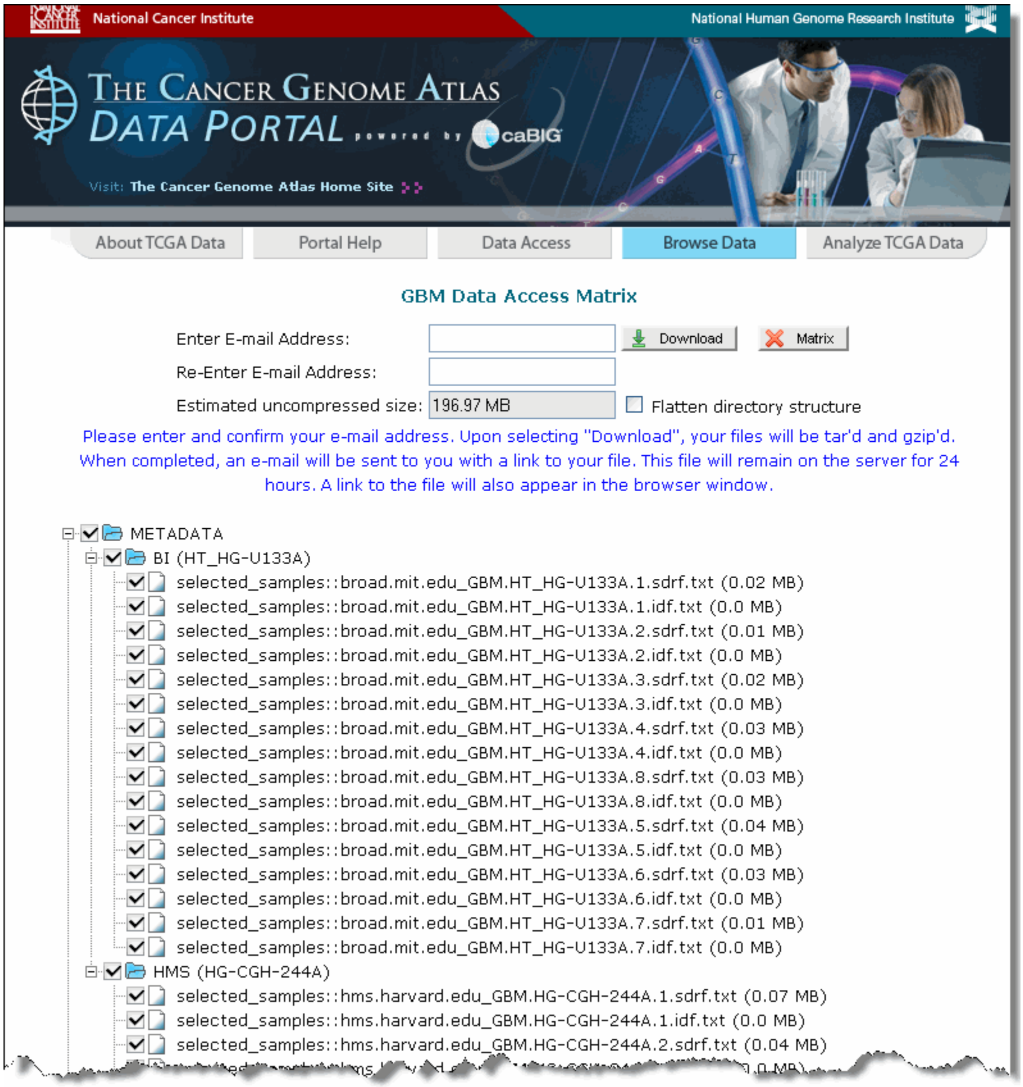
From the Genes tab or the Patients tab search results, you can download data source files. To do so, follow these steps:

1. In the Search Results sections of the page, click the **Download Source Data Files** hypertext link under the Genes or Patients tab.

The application filters the Data Access Matrix based on the current query criteria. For example, if the current query has HMS copy number and B1 expression, it filters the DAM to those two data types, level 3 files only.

If you launch the download with a patient list, the Matrix is filtered according to that list, returning sample IDs corresponding to those patients. A gene list has no effect on filtering in the Matrix, which does not "understand" genes. It only knows data types, levels, samples, etc.

The browser displays the "file tree" page of the Matrix (*Figure 2.15*).


National Cancer Institute
National Human Genome Research Institute
The Cancer Genome Atlas
Data Portal powered by caBIG
Visit: The Cancer Genome Atlas Home Site
About TCGA Data
Portal Help
Data Access
Browse Data
Analyze TCGA Data
Download
Matrix

GBM Data Access Matrix

Enter E-mail Address:

Re-Enter E-mail Address:

Estimated uncompressed size: 196.97 MB ☐ Flatten directory structure

Please enter and confirm your e-mail address. Upon selecting "Download", your files will be tar'd and gzip'd. When completed, an e-mail will be sent to you with a link to your file. This file will remain on the server for 24 hours. A link to the file will also appear in the browser window.

```
⊟ ☑ METADATA
  ⊟ ☑ BI (HT_HG-U133A)
      ☑ selected_samples::broad.mit.edu_GBM.HT_HG-U133A.1.sdrf.txt (0.02 MB)
      ☑ selected_samples::broad.mit.edu_GBM.HT_HG-U133A.1.idf.txt (0.0 MB)
      ☑ selected_samples::broad.mit.edu_GBM.HT_HG-U133A.2.sdrf.txt (0.01 MB)
      ☑ selected_samples::broad.mit.edu_GBM.HT_HG-U133A.2.idf.txt (0.0 MB)
      ☑ selected_samples::broad.mit.edu_GBM.HT_HG-U133A.3.sdrf.txt (0.02 MB)
      ☑ selected_samples::broad.mit.edu_GBM.HT_HG-U133A.3.idf.txt (0.0 MB)
      ☑ selected_samples::broad.mit.edu_GBM.HT_HG-U133A.4.sdrf.txt (0.03 MB)
      ☑ selected_samples::broad.mit.edu_GBM.HT_HG-U133A.4.idf.txt (0.0 MB)
      ☑ selected_samples::broad.mit.edu_GBM.HT_HG-U133A.8.sdrf.txt (0.03 MB)
      ☑ selected_samples::broad.mit.edu_GBM.HT_HG-U133A.8.idf.txt (0.0 MB)
      ☑ selected_samples::broad.mit.edu_GBM.HT_HG-U133A.5.sdrf.txt (0.04 MB)
      ☑ selected_samples::broad.mit.edu_GBM.HT_HG-U133A.5.idf.txt (0.0 MB)
      ☑ selected_samples::broad.mit.edu_GBM.HT_HG-U133A.6.sdrf.txt (0.03 MB)
      ☑ selected_samples::broad.mit.edu_GBM.HT_HG-U133A.6.idf.txt (0.0 MB)
      ☑ selected_samples::broad.mit.edu_GBM.HT_HG-U133A.7.sdrf.txt (0.01 MB)
      ☑ selected_samples::broad.mit.edu_GBM.HT_HG-U133A.7.idf.txt (0.0 MB)
  ⊟ ☑ HMS (HG-CGH-244A)
      ☑ selected_samples::hms.harvard.edu_GBM.HG-CGH-244A.1.sdrf.txt (0.07 MB)
      ☑ selected_samples::hms.harvard.edu_GBM.HG-CGH-244A.1.idf.txt (0.0 MB)
      ☑ selected_samples::hms.harvard.edu_GBM.HG-CGH-244A.2.sdrf.txt (0.04 MB)
```

Figure 2.15 Data Access Matrix page from which you can download source data files

2. From the Matrix page that opens, click **Download** to build the archive with all selected files, or make your selections of what files to download. For example, you could choose to include level 2 or level 1 files.

Note: When selecting files to download, be careful not to exceed the maximum download size.

For more information about working within the Data Access Matrix, see http://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/docs/tcga_DataAccessMatrix_UserGuide.pdf.

Tutorial: Gene Expression to Pathway

You can follow this example scenario as a tutorial to learn how to integrate the search features in the data browser to find meaningful data.

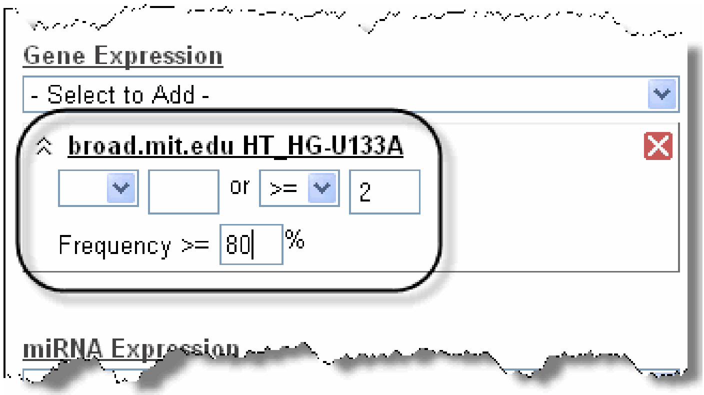
1. To begin, search for genes that are severely *overexpressed* in GBM patients. Make sure the **Disease Type** selected is **GBM** (the default), then select **All Genes** and **All Patients**.
2. Define the *gene expression* settings as described below and shown in *Figure 2.16*.

   a. In the drop-down list, select **broad.mit.edu HT_HG-U133A**.

   b. Leave the [Limit] **<=** box blank and set **>=** at **2**. The **Limit of >= 2** represents a *fold ratio* of 4 (expression is 4 times normal) because the values are fold ratio expressed as log(2).

   c. Define the **Frequency >= at 80%**. This represents the percentage of patients for each gene showing an abnormality (patient ratio).

   In summary, these settings represent "Show all the genes which had 4 times normal expression in at least 80% of patients."



Gene Expression

- Select to Add -

broad.mit.edu HT_HG-U133A

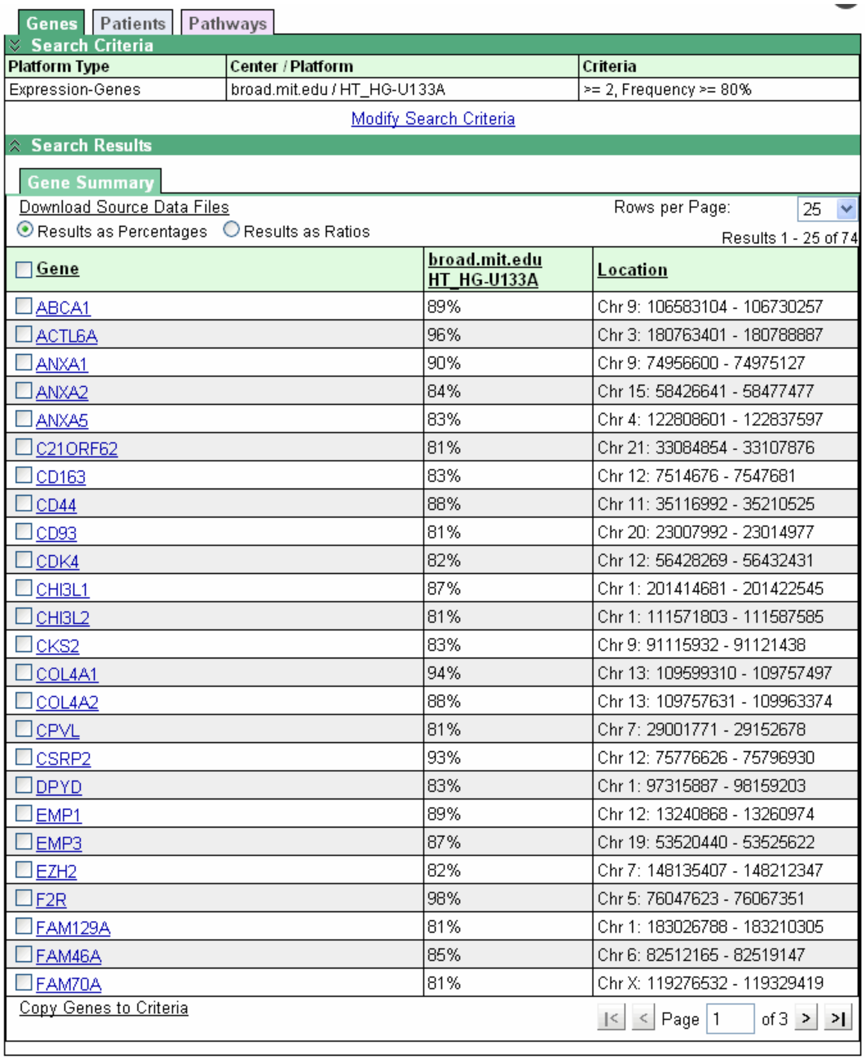
or >= 2

Frequency >= 80 %

miRNA Expression

Figure 2.16 Gene expression search criteria for use case

3. Click **Search Genes** to launch the search. The search returns 74 genes (*Figure 2.17*).



Platform Type	Center / Platform	Criteria
Expression-Genes	broad.mit.edu / HT_HG-U133A	>= 2, Frequency >= 80%

Modify Search Criteria

Download Source Data Files

Rows per Page: 25

● Results as Percentages ○ Results as Ratios

Results 1 - 25 of 74

□ Gene	broad.mit.edu HT_HG-U133A	Location
□ ABCA1	89%	Chr 9: 106583104 - 106730257
□ ACTL6A	96%	Chr 3: 180763401 - 180788887
□ ANXA1	90%	Chr 9: 74956600 - 74975127
□ ANXA2	84%	Chr 15: 58426641 - 58477477
□ ANXA5	83%	Chr 4: 122808601 - 122837597
□ C21ORF62	81%	Chr 21: 33084854 - 33107876
□ CD163	83%	Chr 12: 7514676 - 7547681
□ CD44	88%	Chr 11: 35116992 - 35210525
□ CD93	81%	Chr 20: 23007992 - 23014977
□ CDK4	82%	Chr 12: 56428269 - 56432431
□ CHI3L1	87%	Chr 1: 201414681 - 201422545
□ CHI3L2	81%	Chr 1: 111571803 - 111587585
□ CKS2	83%	Chr 9: 91115932 - 91121438
□ COL4A1	94%	Chr 13: 109599310 - 109757497
□ COL4A2	88%	Chr 13: 109757631 - 109963374
□ CPVL	81%	Chr 7: 29001771 - 29152678
□ CSRP2	93%	Chr 12: 75776626 - 75796930
□ DPYD	83%	Chr 1: 97315887 - 98159203
□ EMP1	89%	Chr 12: 13240868 - 13260974
□ EMP3	87%	Chr 19: 53520440 - 53525622
□ EZH2	82%	Chr 7: 148135407 - 148212347
□ F2R	98%	Chr 5: 76047623 - 76067351
□ FAM129A	81%	Chr 1: 183026788 - 183210305
□ FAM46A	85%	Chr 6: 82512165 - 82519147
□ FAM70A	81%	Chr X: 119276532 - 119329419

Copy Genes to Criteria

Figure 2.17 First page of genes returned in the example search

4. To see which of these genes are highly correlated with amplification, click the **Modify Search Criteria** link.

5. Leaving the gene expression settings as they are, add Correlations criteria described below and shown in *Figure 2.18*,.

   a. In the drop down list, select **HMS Copy Number vs B1 Gene Expression**.

b. To restrict to positive correlations, leave **<=** value blank and enter a minimum **>=** value of **0.4** and a high **p-value** of **0.05** as shown in *Figure 2.18*.

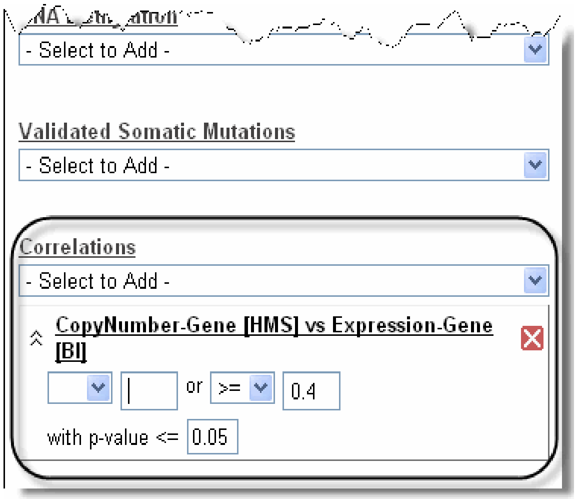
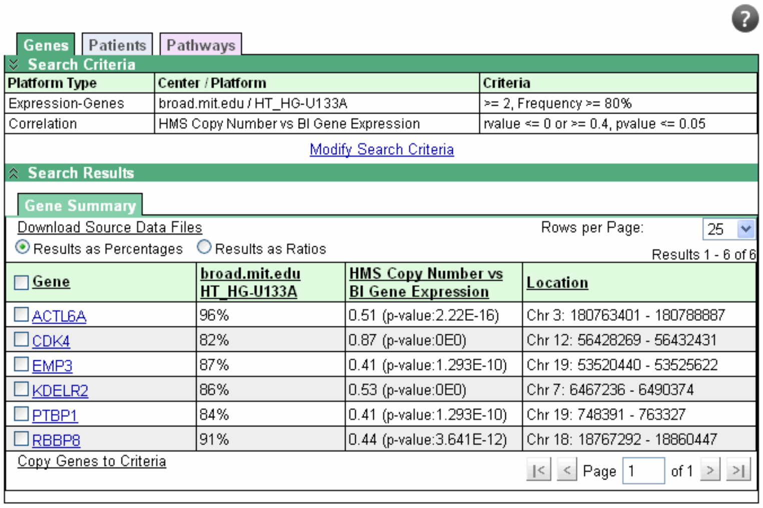


Figure 2.18 Correlation configurations for narrowing the search

After re-running the search, the gene list is filtered by the additional criteria. The new filters return a list of six genes (*Figure 2.19*). These are genes with very high overexpression as well as high correlation with *copy number*.



Platform Type	Center / Platform	Criteria
Expression-Genes	broad.mit.edu / HT_HG-U133A	>= 2, Frequency >= 80%
Correlation	HMS Copy Number vs BI Gene Expression	rvalue <= 0 or >= 0.4, pvalue <= 0.05

Modify Search Criteria
Search Results
Gene Summary
Download Source Data Files
Rows per Page: 25
● Results as Percentages ○ Results as Ratios

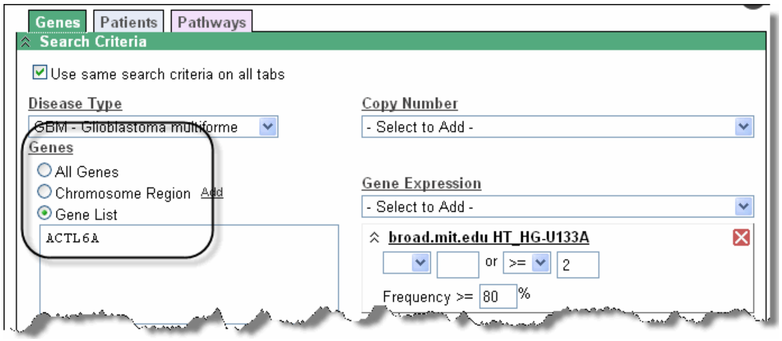
Results 1 - 6 of 6

□ Gene	broad.mit.edu HT_HG-U133A	HMS Copy Number vs BI Gene Expression	Location
□ ACTL6A	96%	0.51 (p-value:2.22E-16)	Chr 3: 180763401 - 180788887
□ CDK4	82%	0.87 (p-value:0E0)	Chr 12: 56428269 - 56432431
□ EMP3	87%	0.41 (p-value:1.293E-10)	Chr 19: 53520440 - 53525622
□ KDELR2	86%	0.53 (p-value:0E0)	Chr 7: 6467236 - 6490374
□ PTBP1	84%	0.41 (p-value:1.293E-10)	Chr 19: 748391 - 763327
□ RBBP8	91%	0.44 (p-value:3.641E-12)	Chr 18: 18767292 - 18860447

Figure 2.19 List of six genes after filtering the search with additional correlation and p-value criteria

6. Based on reviewing the low p-value associated with its moderate correlation value, the gene ACTL6A might be a candidate for further study. Select the gene

and click the **Copy Genes to Search** button, which inserts the gene ID in the Genes List text box of the search pane (*Figure 2.20*).



Genes | Patients | Pathways

Search Criteria

☑ Use same search criteria on all tabs

Disease Type: GBM - Glioblastoma multiforme

Genes

○ All Genes
○ Chromosome Region Add
● Gene List

ACTL6A

Copy Number: - Select to Add -

Gene Expression: - Select to Add -

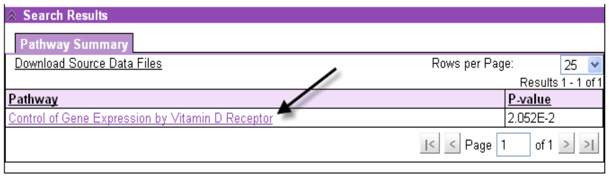
broad.mit.edu HT_HG-U133A

or >= 2

Frequency >= 80 %

Figure 2.20 The gene of choice is added to the Gene List, ready to launch a new search

7. Click **Modify Search Criteria** again. Make sure **Use same search criteria in all tabs** is selected.
8. Select the **Pathways** tab and launch a new search. The search returns just one pathway, the control of gene expression by the vitamin D receptor (*Figure 2.21*).



Search Results

Pathway Summary

Download Source Data Files

Rows per Page: 25

Results 1 - 1 of 1

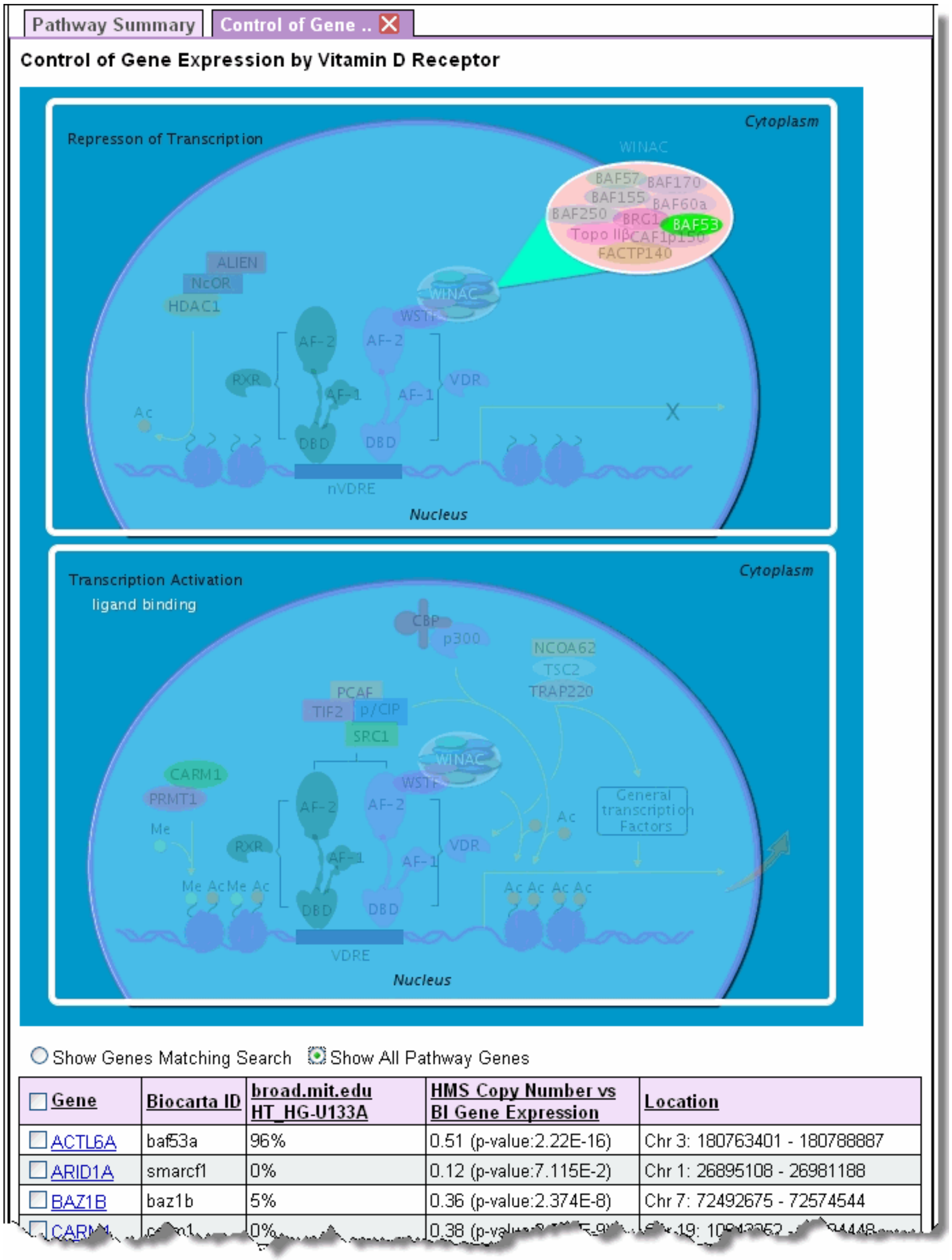
Pathway	P-value
Control of Gene Expression by Vitamin D Receptor	2.052E-2

Page 1 of 1

Figure 2.21 One pathway is listed in the search results

9. Click the pathway hypertext link which opens the pathway diagram. Select the button **Show All Pathway Genes**. The browser displays the new table showing

all of the genes involved in the control of gene expression by the vitamin D receptor (*Figure 2.22*).



○ Show Genes Matching Search ● Show All Pathway Genes

☐ Gene	Biocarta ID	broad.mit.edu HT_HG-U133A	HMS Copy Number vs BI Gene Expression	Location
☐ ACTL6A	baf53a	96%	0.51 (p-value:2.22E-16)	Chr 3: 180763401 - 180788887
☐ ARID1A	smarcf1	0%	0.12 (p-value:7.115E-2)	Chr 1: 26895108 - 26981188
☐ BAZ1B	baz1b	5%	0.36 (p-value:2.374E-8)	Chr 7: 72492675 - 72574544
☐ CARM1	[illegible]	0%	0.38 (p-value:[illegible])	Chr 19: [illegible]

Figure 2.22 Pathway diagrams representing vitamin D transcription repression and activation. The table represents a segment of all genes involved in the selected pathway

The gene of choice, ACTL6A, is highlighted in the first figure of the *Biocarta* diagram, the repression of transcription. Note the difference in the Biocarta ID of the gene, baf53a, from the gene ID in the TCGA database, ACTL6A. This illustrates the value of the data browser showing both IDs.

The results located in this example might leave you wondering whether the overexpression of ACTL6A in the vitamin D receptor has something to do with GBM. This could lead to further hypotheses to be tested against available data or in the laboratory.

APPENDIX A

GLOSSARY

Acronyms, objects, tools and other terms referred to in chapters or appendixes of this TCGA Portal Data Browser*TCGA Portal Data Browser User’s Guide* are described in this glossary.

Term	***Definition***
anomaly	In the context of TCGA, an anomaly is a gene abnormality
BCM	Baylor College of Medicine
Biocarta	Source for biological pathways, displayed in a graphical format, mapping known genomic and proteomic relationships
copy number abnormality type	Variances in the number of copies of a genes in a cell from the number found in normal cell samples
deleted	The when there are 2.5 or greater copies of the gene.
fold ratio	The fold ratio (also called fold change) is the ratio of the measured gene expression value for an experimental sample to the expression value for the control sample.
gene expression values	Values in TCGA database represent the ratio of tumor expression over normal expression.
mutated	TheTCGA Portal Data Browser when . . . ..
NCI	National Cancer Institute
NCICB	National Cancer Institute Center for Bioinformatics
normalization	Used to designing relational database tables and minimizing duplicated data.
overexpressed	Gene expression when the fold ratio is twice the control value or average.
sample threshold	A final percentage threshold applied to the samples used to determine whether the gene is an abnormality.

Table A.1 Glossary of TCGA Portal Data Browser terms

Term	***Definition***
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	Gene expression when the fold ratio is less than twice the control value or average.
value threshold	The initial threshold applied to data to determine an abnormality.
WIBR	Whitehead Institute for Biomedical Research

Table A.1 Glossary of TCGA Portal Data Browser terms

INDEX