REMBRANDT 1.0

User's Guide

NATIONAL® CANCER INSTITUTE

Center for Bioinformatics

Credits and Resources

REMBRANDT Development and Management Teams		
Development	***User's Guide***	***Program Management***
Subha Madhavan	Subha Madhavan	Subha Madhavan
Alex Jiang	Neysa Narena	
Kevin Rosso	Ying Long	
Ryan Landy	Huaitian Liu	
Himanso Sahni		
David Bauer		
Huaitian Liu		
Michael Harris		
Ram Bhattaru		
David Hall		
Ye Wu		
Ying Long		
Don Swan		
Dana Zhang		
Vesselina Bakalov		
Nick Xiao		
Jean-Claude Zenklusen		
Yuri Kotliarov		
Gregg Silk		

Application Support
If you have a question about REMBRANDT that is not answered in the *REMBRANDT User's Guide*, contact Application Support at ncicb@pop.nci.nih.gov.
When submitting support requests via email, please include: - Your contact information - The REMBRANDT application URL - A description of the problem and steps to re-create it - The text of any error messages you have received

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER

1

USING THE REMBRANDT USER'S GUIDE

Welcome to the *REMBRANDT User's Guide.* This book introduces you to REMBRANDT and shows you how to work with the REMBRANDT repository to conduct searches and to create graphs from search results.

Topics in this chapter include:

About REMBRANDT

REMBRANDT (REpository for Molecular BRAin Neoplasia DaTa), a joint initiative of NIH's National Cancer Institute (NCI) and the National Institute of Neurological Disorders and Stroke (NINDS), provides a bioinformatics knowledge base framework that leverages data warehousing technology to host and integrate clinical and functional genomics data from clinical trials involving patients suffering from gliomas (tumors).

REMBRANDT works in conjunction with tools and infrastructure components developed by the NCI Center for Bioinformatics (NCICB) to support brain tumor studies such as the Glioma Molecular Diagnostic Initiative (GMDI) led by the NCI's Center for Cancer Research (CCR) Neuro-Oncology Branch. GMDI's primary goal is to develop a molecular classification schema that is both clinically and biologically meaningful, based on gene expression and genomic data from gliomas of patients followed through the natural history and treatment phases of their illness. GMDI's secondary objective is to explore gene expression profiles to determine patient responsiveness and to correlate profiles with discrete chromosomal abnormalities.

REMBRANDT's knowledge framework provides researchers with the ability to perform *ad hoc* querying and reporting across multiple data domains, such as gene expression, chromosomal aberrations and clinical data.

Researchers can use REMBRANDT to answer questions related to a patient or patient population and to view integrated data sets in a variety of contexts. Tools that link data to other annotations such as cellular pathways, gene ontology terms and genomic information are embedded. The system also allows users to perform various higher-order analyses on clinical and genomic data sets.

Organization of this Guide

This book is organized into chapters that parallel REMBRANDT's workflow.

Chapter 1. Using the REMBRANDT User's Guide, describes REMBRANDT, how this guide is organized and how text conventions are represented.

Chapter 2. Getting Started, explains how to launch and log in to REMBRANDT, get additional help if needed, navigate through the application and use REMBRANDT's basic features to get results.

Chapter 3. Conducting Simple Searches, describes how to search by gene keyword and Reporter ID and to create gene expression plots, Kaplan-Meier surival plots, and copy number-based graphs from those search results.

Chapter 4. Conducting Advanced Searches, tells you how to create gene expression, copy number and clinical queries. This chapter further describes how to view and save your results.

Chapter 5. High Order Analysis, extends the basic knowledge of the previous chapters and shows you how to work with class comparisons, hierarchical clustering and principal component analysis.

In addition, there is a glossary that provides terms used throughout this guide with their definitions.

Font Styles and Their Meaning

Table *1.1* illustrates how text conventions are represented in this guide. The various typefaces differentiate between regular text and menu commands, keyboard keys, toolbar buttons, dialog box options and text that you type.

Convention	*Description*	*Example*
Boldface type	Options that you select in dialog boxes or drop-down lists. Buttons or icons that you click.	In the Patterns Search dialog box, click the **Add** button.
Italics	Used to reference other documents, sections, figures and tables.	*caCORE Software Development Kit 1.0 Programmer's Guide*

Table 1.1 REMBRANDT User's Guide Text Conventions

Convention	*Description*	*Example*
Italic boldface type	Text that you type	In the New Subset text box, enter ***GBM_EGFR_query.***
TEXT IN SMALL CAPS	Keyboard key(s) that you press	Press ENTER
"**Note:**"	Highlights a concept of particular interest	**Note:** This concept is used throughout the installation manual.

Table 1.1 REMBRANDT User's Guide Text Conventions (Continued)

CHAPTER

2

GETTING STARTED

This chapter introduces the REMBRANDT interfaces, navigation and common features used on REMBRANDT pages. How to use REMBRANDT to conduct queries as well as how to find additional resources is also described.

Topics in this chapter include:

Starting REMBRANDT

Start REMBRANDT by following these steps:

1. Go to the REMBRANDT portal on the NCICB website: http://rembrandt.nci.nih.gov/.
2. Click the **REMBRANDT Application** button located in the lower left-hand column as indicated in *Figure 2.1*.

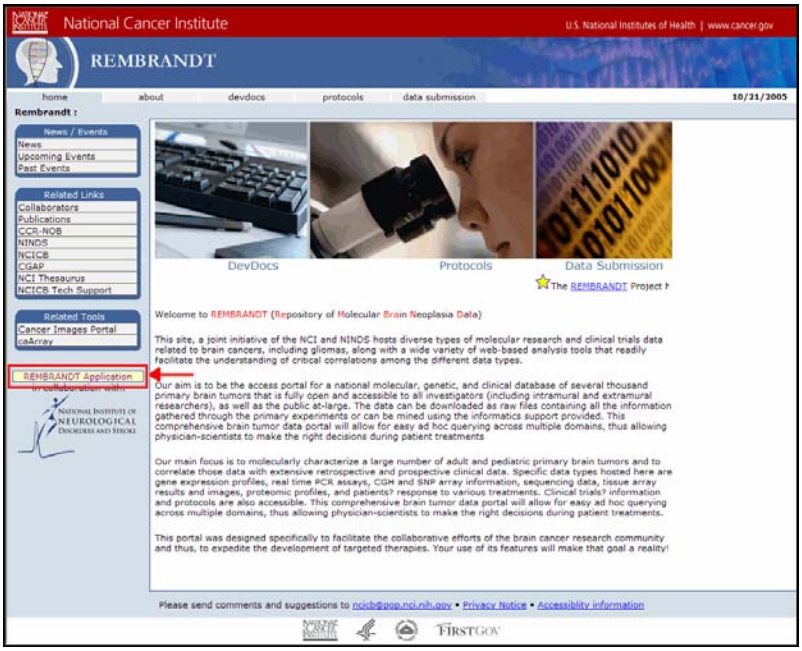


Figure 2.1 The REMBRANDT portal on the NCICB website

The REMBRANDT login page appears *(Figure 2.2)*.

Figure 2.2 REMBRANDT login page

If you already have a REMBRANDT user account, enter the login information and click **Submit.**

Creating a User Account

Each REMBRANDT user is given a unique userID and password. The userID and password you are assigned determines your access rights for the software. To set up a user account, you must

- Contact NCICB Application Support:
  - Email: NCICB@pop.nci.nih.gov
  - Phone: 888-478-4423 (toll-free) or 301-451-4384 (local)

OR

- Go to the NCICB REMBRANDT login page http://rembrandt-db.nci.nih.gov/Rembrandt/login.jsp and click the **request username/password** link to send an email requesting a userID and password to NCICB Application Support.

Note: If you would like to offer feedback via email to the REMBRANDT development team, click the **feedback** link.

Figure 2.3 Request and feedback links

Online Help and Tutorials

Information about how to use REMBRANDT is easily accessed via:

- Online help
- Context-sensitive help
- Tutorials

Note: You can also contact Application Support. See *Credits and Resources* for more details.

Online Help

If you need help, REMBRANDT has an online help system that you can access by clicking the **Help** button that displays throughout REMBRANDT.

Context-Sensitive Help

Clicking the question mark *(Figure 2.4)* that displays adjacent to a field will open a pop-up window that provides helpful information about that field.

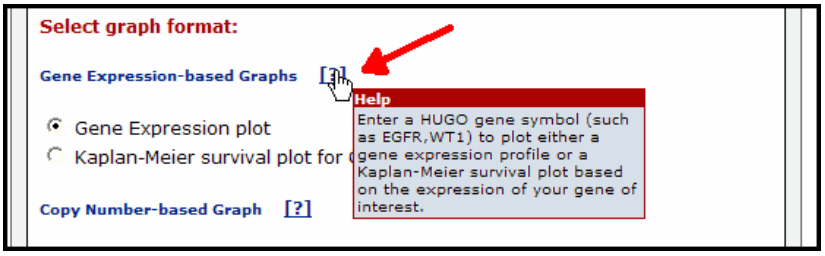


Figure 2.4 Context-sensitive Help button

You can access more information about a report by clicking the question mark button *(Figure 2.5)* on the report.

Figure 2.5 Finding more information about a report

Tutorials

To access REMBRANDT tutorials, click the **tutorials** link *(Figure 2.6)* located on the top of the REMBRANDT application's opening page.

Note: You don't have to log in to access the tutorials.

Figure 2.6 Tutorials link

Related links and related tools are provided on the left-hand menu *(Figure 2.7)* of the REMBRANDT website, located at http://rembrandt.nci.nih.gov.

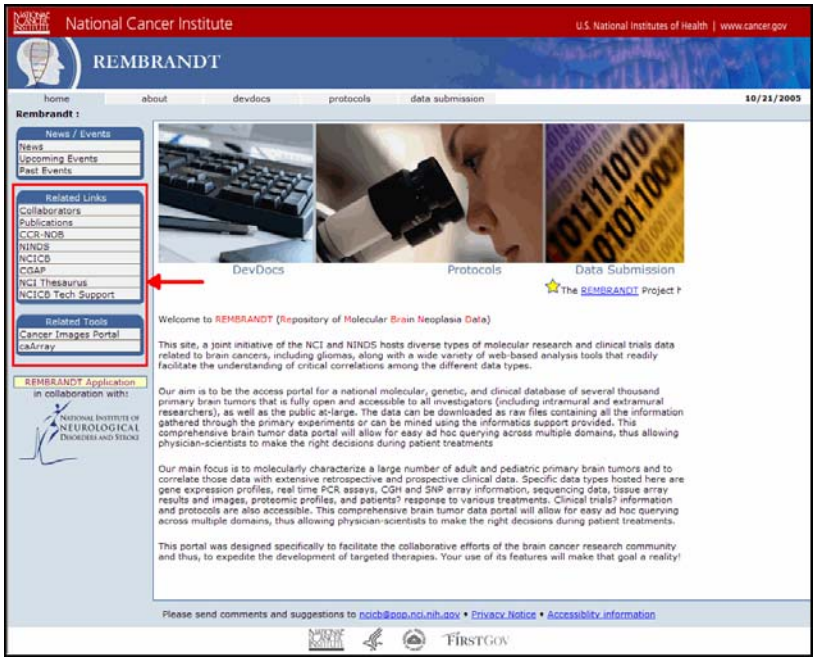


Figure 2.7 Related links and tools

Working with REMBRANDT

If you are a first-time user, probably one of the most helpful things to do is to open two browser windows, one for your work and one for the tutorial, so that you can work in conjunction with the tutorial. (See *Figure 2.6 Tutorials link*.)

Once you have logged in, the Legal Rules of the Road page appears. After you have read the provisions, click the **Clicking Here** link *(Figure 2.8)* in the lower right-hand corner.

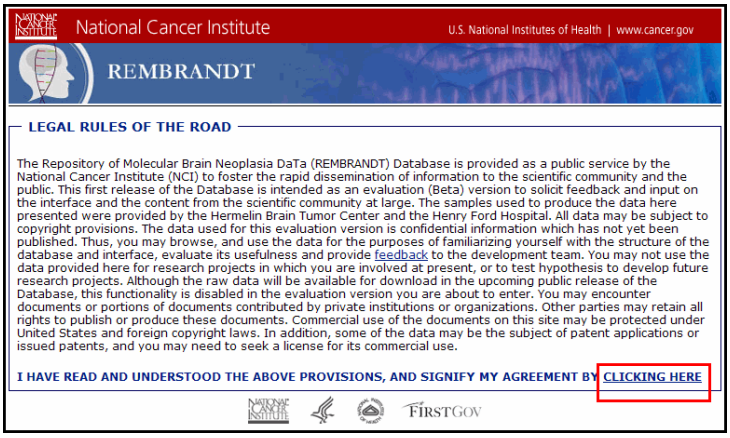


Figure 2.8 Legal Rules of the Road page

The next page that appears shows the REMBRANDT workspace *(Figure 2.9)*.

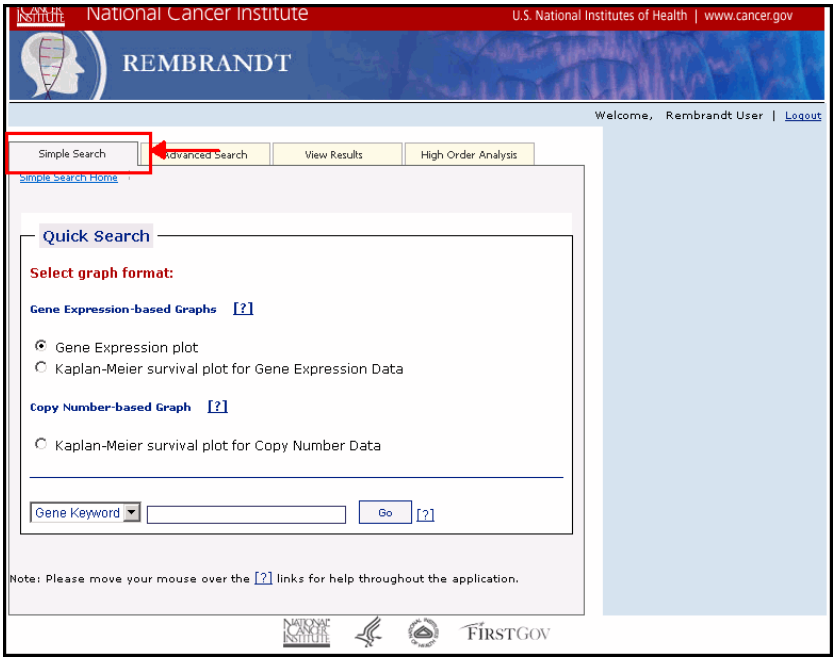


Figure 2.9 REMBRANDT workspace

Available Tabs and Queries

The REMBRANDT workspace is organized by the following tabs:

- **Simple Search**
- **Advanced Search**
- **View Results**
- **High Order Analysis**

You can begin your search from either the **Simple Search** or **Advanced Search** tab. Each of these two tabs displays a menu of buttons and each button creates a different type of query.

The **Simple Search** tab provides three search categories:

- **Gene Expression plot** (see page 19)
- **Survival plot for Gene Expression Data** (see page 21)
- **Kaplan-Meier survival plot for Copy Number Data** (see page 21)

The **Advanced Search** tab provides the ability to add:

- **Gene Expression Analysis** (see page 28)
- **Copy Number Data Analysis** (see page 46)
- **Clinical Study Analysis queries** (see page 50)

Search results can be viewed from the **View Results** tab (*Figure 2.10*). This page shows a collection of reports previously viewed in a particular user session, which allows you to compare different reports (for example, clinical and Gene expression reports for a set of patient samples) by opening them in different windows. You can also view the higher-order analysis results from this page.

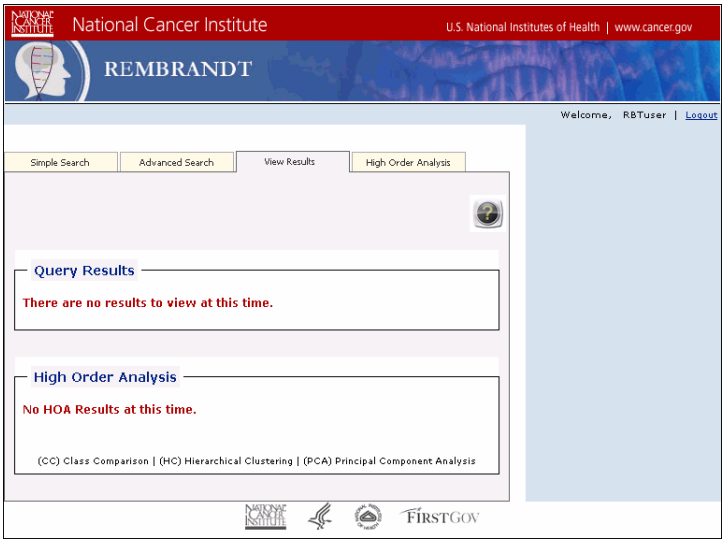


Figure 2.10 View Results tab

Higher order analyses can be conducted from the **Higher Order Analysis** tab. This tab provides the ability to perform Class Comparison (see page 54), Hierarchical Clustering (see page 59) and Principal Component Analysis (see page 57).

Clicking one of the menu buttons from either the **Simple Search** or **Advanced Search** tabs opens subsequent pages with required and optional search criteria fields. After you have submitted your query, that query can be refined by filtering, selecting samples, and so on, as needed. Once you are finished with your query, you can run your report. From this report, you may print, download for Excel or view query details.

Common Buttons and Other Features

Each page displays certain navigational features and common buttons. In addition, the **Queries** column (shown in blue in *Figure 2.11*) shows the status of your queries. This is particularly helpful when you have completed one query and then are entering search criteria on another tab for another type of query.

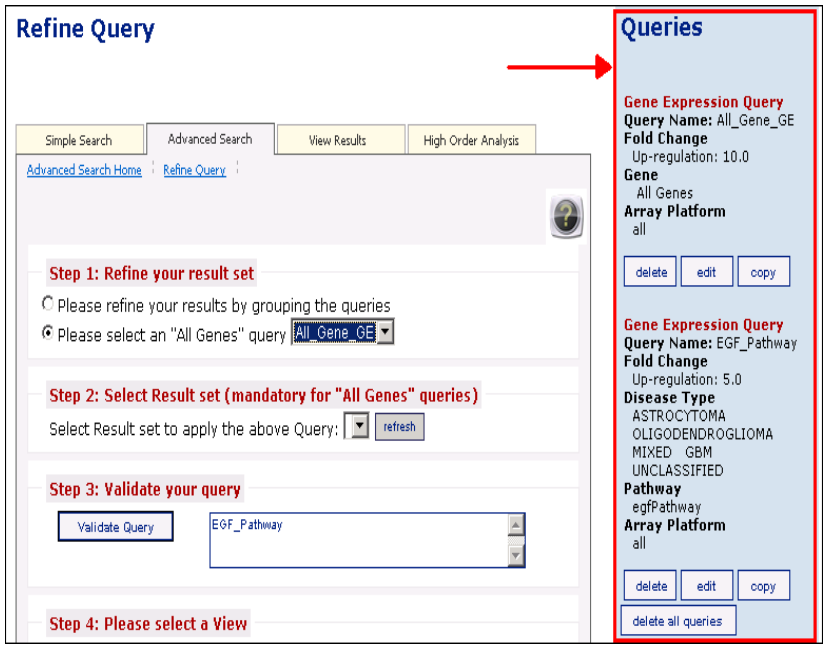


Figure 2.11 Queries column

Common buttons displayed throughout REMBRANDT are listed in Table *2.1*.

Common Buttons	*Purpose*
Cancel button	Clicking the **Cancel** button eliminates all data currently entered in a form; a query will not be generated.
Check boxes	Small boxes that can be turned on or off by clicking with the mouse. When selected, an **X** displays in a square, when off, the square is blank.
Clear button	Clicking the **Clear** button restores a report to its original state and clears any highlighting.
Copy button	Located in the **Queries** column. Clicking the **Copy** button creates a copy of a query.
Delete button	Located in the **Queries** column. Clicking the **Delete** button deletes a query.
Delete All Queries button	Located in the **Queries** column. Clicking the **Delete All Queries** button deletes all queries displayed in the **Queries** column.
Edit button	Located in the **Queries** column. Clicking the **Edit** button returns you to the original query page for editing purposes.
Find button	Use the **Find** button to open a screen from which you can search.
Preview button	Clicking the **Preview** button displays a preview of your report.
Refresh button	Clicking the **Refresh** button brings up previously saved result sets in a drop-down list.
Submit button	Clicking the **Submit** button returns you to the Build Query page where you can either finalize your query or add another query.

Table 2.1 Common buttons

Selecting Reports

Several types of reports are available in REMBRANDT. Each type is described below.

Gene Expression Plot (Simple Search)

The Gene expression plot displays average expression intensities for the gene of interest based on Affymetrix GeneChip arrays (U133 Plus 2.0 arrays).

Multiple probe sets (for some genes) are designed to measure the expression of the gene of interest.

Note: See http://www.affymetrix.com for more information on their probe set design strategy for human genes.

Group average (samples are averaged based on tumor subtypes into six categories: Glioblastoma Multiforme, Oligodendroglioma, Astrocytoma, Mixed, Unclassified and Unknown tumors) is calculated for each probe set and is plotted on the *Y*-axis for each tumor type. As an indication of probabilities of obtaining the differences in expression values between tumor (or a subtype of tumor) and non-tumor samples, a *p*-value is displayed with mouse-overs.

Note: For detailed information on data processing, go to http://rembrandt.nci.nih.gov (the information portal of this application). Click the **Protocols** tab and then click the **Data Processing** link.

Kaplan-Meier Survival Plot

Kaplan-Meier (KM) survival plots show survival. There are two types of KP survival plots that are available:

- Gene Expression KM Plot
- Copy Number KM Plot

Gene Expression KM Plot

Users can query gene expression and graph changes in survival rate at each time point on the study.

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.

You can dynamically modify the fold change (up and down regulation) thresholds and redraw the plot.

A log-rank *p*-value is provided as an indication of 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 procedure[1]. The *p*-values are recalculated every time a new threshold is selected.

You can toggle to a unified gene expression view with lesser reporters to get a gene-based view of the expression data. To obtain the unified gene expression values, the probe-level data is processed with custom CDF (Chip Definition Files) that rearranges Affymetrix probes into gene-based probe sets. Probes mapped to alternatively spliced exons are grouped into a distinct probe sets. The most 3` probes are selected for processing. Non-specific probes are masked before processing.

Copy Number KM Plot

Users can query amplication or deletions and graph changes on survival rate at each time point on the study.

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

1. Kaplan, E.L. and Meier, P. Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* 53, 457-481 (1958).

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.

A log-rank *p*-value is provided as indication of significance of the difference in survival between any two groups of samples segregated based on amplification/deletion of the gene or SNP 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.

Clinical Report (Sample Report)

Clinical information including patient demographics, therapy and outcome data (either in just a single domain such as Gene Expression or by including a combination of queries from multiple domains such as Gene expression, Chromosomal aberrations and clinical areas) is displayed in this report.

When either a gene expression filter and/or a copy number filter are applied, hyperlinks are provided in this report to display the gene expression and/or copy number data for a particular sample.

Gene Expression Data Per Sample View

The Gene Expression Data Per Sample report displays gene expression ratios (between the tumor sample and the average of non-tumor samples) for each probe set (or IMAGE clone) for the genes selected in the queries.

Each column represents a sample and the samples are grouped by tumor sub-type.

For Affymetrix probe sets, the ratio of the absolute expression values of the tumor sample and the average expression value of the non-tumor samples.

For each Image Clone, missing values are handled and ratio of expression values between the tumor and average of non-tumor group is calculated for each sample.

Note: See http://rembrandt.nci.nih.gov for more information on data processing.

Gene Expression Data Per Disease Group View

The Gene Expression Data Per Disease Group report displays mean gene expression ratios (between the tumor group and the average of non-tumor samples) for each probe set (or IMAGE clone) for the genes selected in the queries.

Each column represents a sample group (tumor sub-type).

Group average samples were based on tumor subtype categories:

- All
- Glioblastoma Multiforme
- Oligodendroglioma
- Astrocytoma
- Mixed tumors
- Unknown

Group average samples were also calculated for each probe set (or IMAGE clone). As an indication of probabilities of obtaining the differences in expression values between

tumor (or a sub-type of tumor) and non-tumor samples, a *p*-value is displayed within the parenthesis for each ratio.

Note: See http://rembrandt.nci.nih.gov for more information on data processing.

Copy Number Data Per Sample View

The Copy Number Data Per Sample report displays the copy number data from Affymetrix 100K SNP arrays.

The CHP files from the Affymetrix Gene Chip Operating System were processed using the Affymetrix Copy Number Analysis Tool (CNAT) and copy number data was collected for each mapping SNP reporter on the Chip, for all the tumor samples.

Each column represents a sample and the samples are grouped based on the tumor sub-type.

Where To Go From Here

For more information on using REMBRANDT, see:

- *Chapter 3*, *Conducting Simple Searches*, to learn how to do simple searches.
- *Chapter 4*, *Conducting Advanced Searches*, to understand how to do advanced searches.
- *Chapter 5*, *High Order Analysis*, to learn how to conduct high order analyses.

CHAPTER

3

CONDUCTING SIMPLE SEARCHES

This chapter describes how to use REMBRANDT to conduct simple searches of the REMBRANDT repository and to create graphs from the results obtained.

Topics in this chapter include:

Simple Searches Overview

You can use REMBRANDT to conduct simple searches from specified gene keywords or SNP probe set ID's. Search results can then be used to generate either a gene expression profile or a Kaplan-Meier survival plot based on the expression of your gene of interest.

Alternatively, you can generate a Kaplan-Meier survival plot based on a specified gene copy number or SNP reporter.

Simple searches are initiated on the **Simple Search** tab *(Figure 3.1).* The **Simple Search** tab provides three search categories:

- Gene Expression plot (by gene keyword)
- Kaplan-Meier survival plot for Gene Expression Data (by gene keyword)
- Kaplan-Meier survival plot for Copy Number Data (by gene keyword or SNP probe set ID)

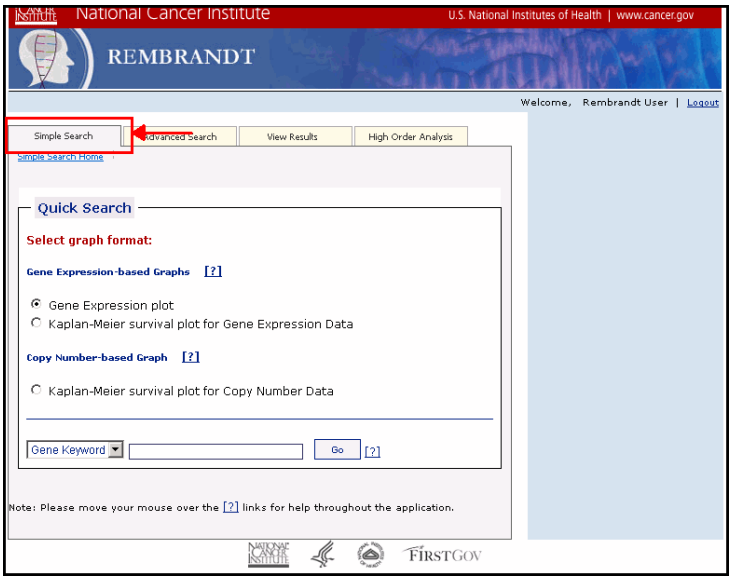


Figure 3.1 Simple Search tab

Simple Searches Workflow

When conducting a simple search, complete the following steps in this suggested order:

1. After logging in, click the **Simple Search** tab, then select a graph.
2. Select either the **Gene Keyword** or **SNP Probe set ID** option, as applicable, from the drop-down list and enter the keyword or ID in the text box.

   Note: The **SNP Probe set ID** option is available only for the Copy Number-based Graph format.
3. Click **Go** to generate a graph from the search results.
4. If an intermediate page appears (refer to *Figure 3.4* on page 22), select amplification and deletion criteria, choose a reporter, and click **Redraw Graph.**
5. If a message displays, stating that one or more genes or their aliases have been found, select the appropriate option from the drop-down list and click **Go.** To end the query, click **Cancel** (*Figure 3.2*).

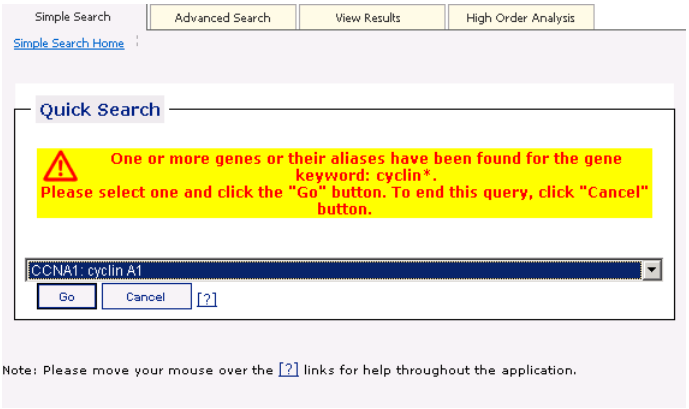


Figure 3.2 Alias message

Creating Gene Expression Plots

To create gene expression plots, follow these steps:

1. From the **Simple Search** tab, select **Gene Expression plot.**
2. Enter a Gene Keyword in the text box.

   Enter a HUGO gene symbol such as **EGFR** or **WT1** to plot a gene expression profile based on the expression of your gene of interest.
3. Click **Go.** The Gene Expression plot appears (*Figure 3.3*).

   The tumor group average expression value is plotted for each probe set. Multiple probe sets designed to measure the expression of the gene of interest are shown in different colors on the bar graph.

   Point to each bar on the graph to see *p*-values that indicate the confidence levels for obtaining the differences in expression values between tumor and non-tumor samples.

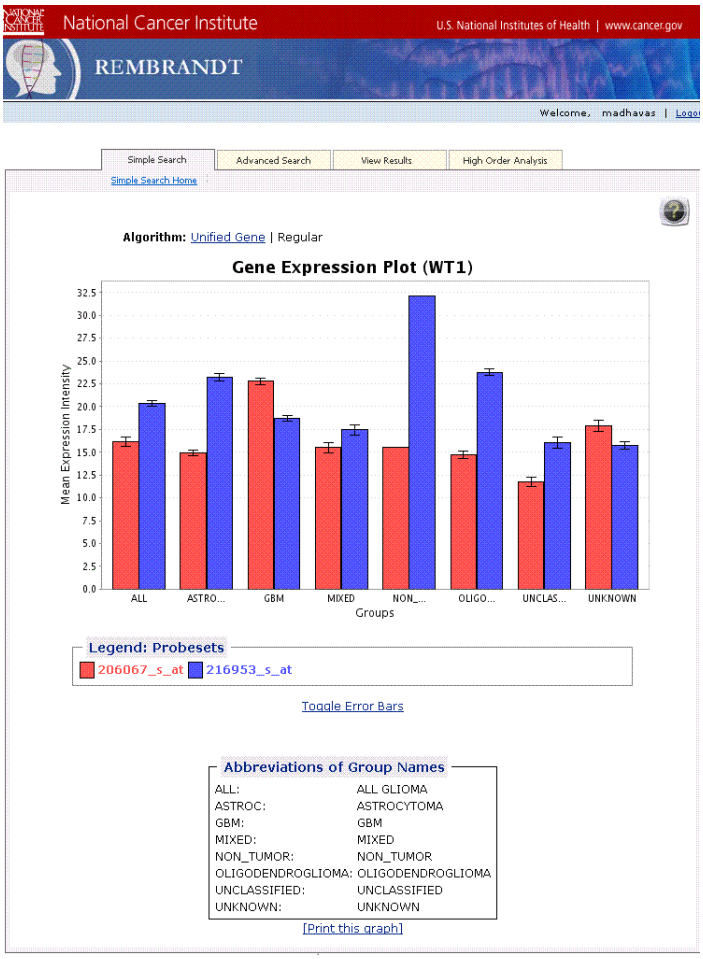


Figure 3.3 Gene Expression plot

- The Gene expression plot displays average expression intensities for the gene of interest based on Affymetrix GeneChip arrays (U133 Plus 2.0 arrays).

  Note: Click the **Toggle Error Bars** link to turn on/off the display of error bars.

- Multiple probe sets (for some genes) are designed to measure the expression of the gene of interest.

  Note: Refer to http://www.affymetrix.com for more information on probe set design strategy for human genes.

- Group average (samples were averaged based on tumor subtypes into six categories, Glioblastoma Multiforme, Oligodendroglioma, Astrocytoma, Mixed tumors, Unclassified and Unknown) was calculated for each probe set and is plotted on the *Y*-axis for each tumor type. As an indication of probabilities of obtaining the differences in expression values between tumor (or a subtype of tumor) and non-tumor samples, a *p*-value is displayed when pointing to a bar on the graph.

- You can toggle to a unified gene expression view with lesser reporters to get a gene-based view of the expression data. To obtain the unified gene expression values, the probe-level data is processed with custom CDF (Chip Definition Files) that rearranges Affymetrix probes into gene-based probe sets. Probes

mapped to alternatively spliced exons are grouped into a distinct probesets. The most 3` probes are selected for processing. Non-specific probes are masked before processing. Selecting the “unified algorithm” repaints the graph with lesser probe sets calculated using the custom CDF as described above.

Note: See the REMBRANDT information portal at http://rembrandt.nci.nih.gov for detailed information on data processing.

Note: To print the graph, click the **Print this graph** link.

Creating Kaplan-Meier Survival Plots for Gene Expression Data

To create Kaplan-Meier survival plots for gene expression data, follow these steps:

1. From the **Simple Search** tab, select **Create Kaplan-Meier survival plot for Gene Expression Data.**
2. Enter a Gene Keyword.

   Enter a HUGO gene symbol (such as **EGFR** or **WT1**) to plot a Kaplan-Meier survival plot based on the expression of your gene of interest.

   Note: If a message displays, stating that one or more genes or their aliases have been found, select the appropriate option from the drop-down list and click **Go.** To end the query, click **Cancel** (refer to *Figure 3.2* on page 19).
3. Click **Go.**

   The Kaplan-Meier survival plot appears *(Figure 3.4)*.

   You can toggle to a unified gene expression view with lesser reporters to get a gene-based view of the expression data. To obtain the unified gene expression values, the probe-level data is processed with custom CDF (Chip Definition Files) that rearranges Affymetrix probes into gene-based probe sets. Probes mapped to alternatively spliced exons are grouped into a distinct probesets. The most 3` probes are selected for processing. Non-specific probes are masked before processing. Select **unified** from the algorithm drop-down and select a unified probeset from the **Reporters** drop-down list to visualize the K-M plot for the unified probeset.

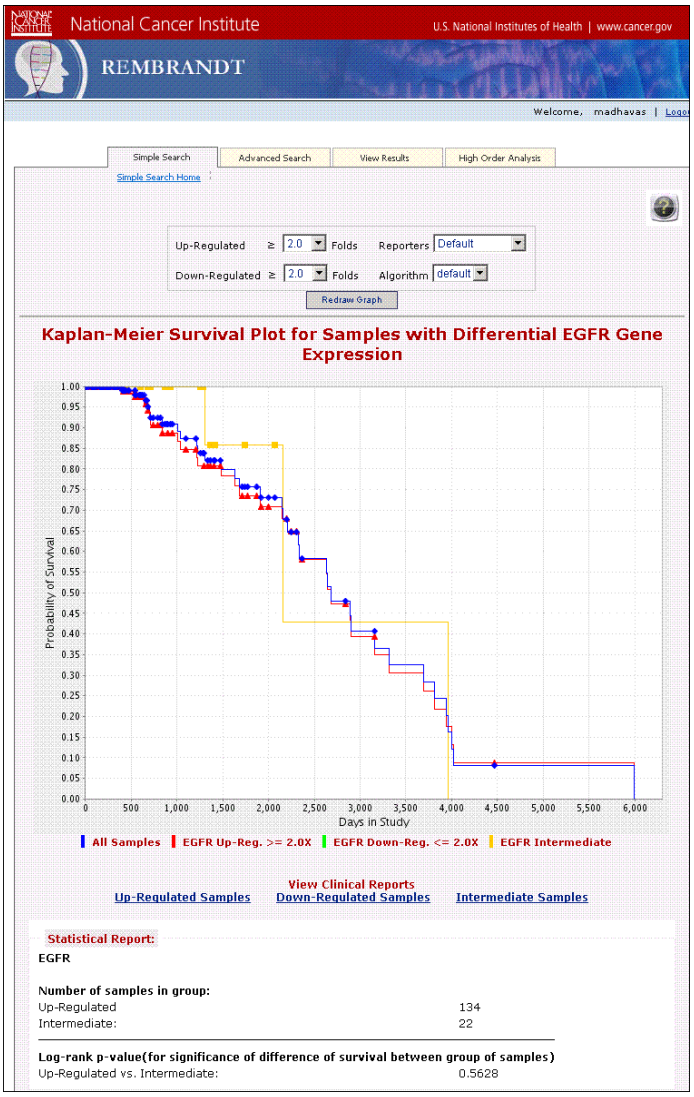


Figure 3.4 Kaplan-Meier survival plot for gene expression samples

Creating Copy Number-based Graphs

There are two types of copy number-based graphs available. A Kaplan-Meier survival plot can be created that is based on the gene copy number or the SNP reporter.

Creating Kaplan-Meier Survival Plots for a Gene Keyword

To create a Kaplan-Meier survival plot for Copy Number Data with a Gene Keyword, follow these steps:

1. From the **Simple Search** tab, select **Kaplan-Meier survival plot for Copy Number Data.**
2. Choose **Gene Keyword** from the drop-down list.
3. Enter a HUGO gene symbol (such as **EGFR** or **WT1**).

Note: If a message displays, stating that one or more genes or their aliases have been found, select the appropriate option from the drop-down list and click **Go.** To end the query, click **Cancel** (refer to *Figure 3.2* on page 19).

4. Click **Go.**
5. Select amplification and deletion criteria.
6. Choose a reporter *(Figure 3.5).*

Figure 3.5 Selecting a Reporter

7. Click **Redraw Graph.**

Note: If necessary, you can alter the amplification and/or deletion values and/or change the Reporter, then regenerate the graph.

The Kaplan-Meier survival plot for a sample with copy number analysis for the selected criteria appears *(Figure 3.6).*

There is a legend below the graph that describes the plot in detail.

A statistical report is available below the legend. This report provides additional information about the plot.

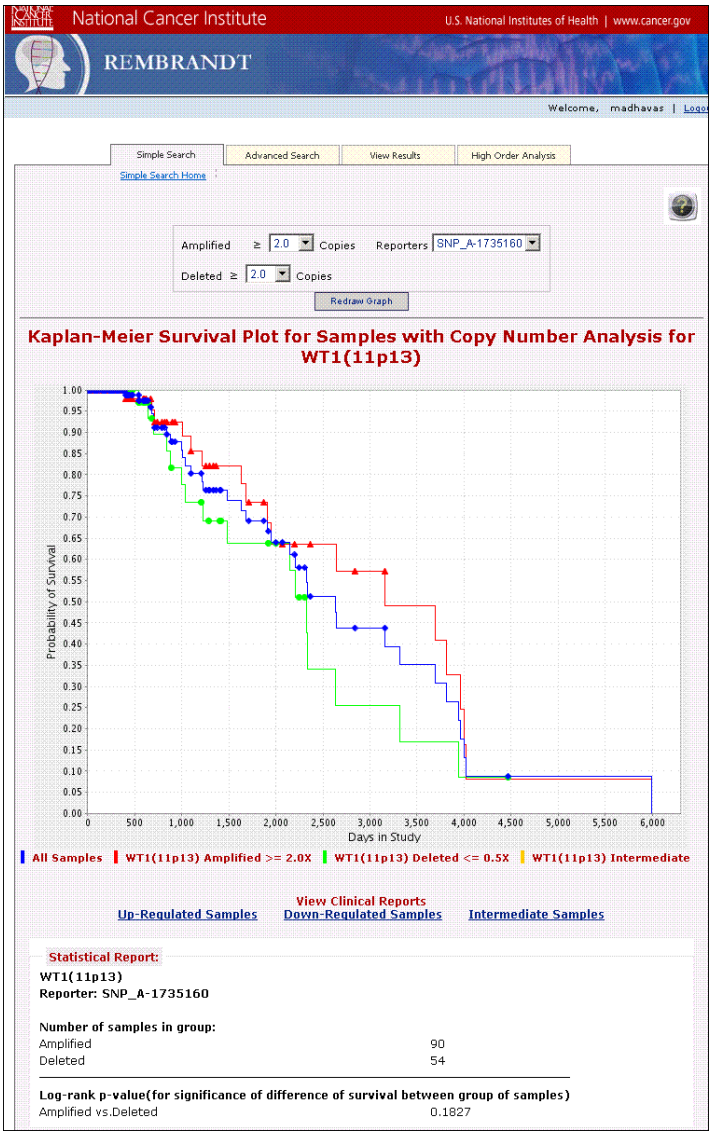


Figure 3.6 Kaplan-Meier survival plot for samples with copy number analysis

- For samples with certain amplification/deletion characteristics (such as amplification of the cytoband that EGFR maps to, 7p11.2), a Kaplan-Meier plot is provided for each SNP probe set associated with the gene of interest to show the survival rate at each time point.
- 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.
- You can dynamically modify the fold change (up and down regulation) thresholds and redraw the plot.
- A log-rank *p*-value is provided as an indication of significance of the difference in survival between any two groups of samples segregated based on amplification/deletion of the gene or SNP of interest. The log rank *p*-value is calculated

using Mantel-Haenszel procedure. The *p*-values are recalculated every time a new threshold is selected.

Creating Kaplan-Meier Survival Plots for a SNP Probe Set ID

To create a Kaplan-Meier survival plot for a SNP Probe Set ID, follow these steps:

1. From the **Simple Search** tab, select **Kaplan-Meier survival plot for Copy Number Data.**
2. Choose **SNP Probe set ID** from the drop-down list.
3. Enter a SNP Probe set ID.
4. Click **Go.**

   The Kaplan-Meier survival plot for samples with copy number analysis for the specified SNP probe set ID appear. A graph similar to *Figure 3.6* (on page 24) will appear.

- For samples with certain amplification/deletion characteristics (such as amplification of the cytoband that EGFR maps to, 7p11.2), a Kaplan-Meier plot is provided for each SNP probe set associated with the gene of interest to show the survival rate at each time point.
- 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.
- You can dynamically modify the fold change (up and down regulation) thresholds and redraw the plot.
- A log-rank *p*-value is provided as an indication of significance of the difference in survival between any two groups of samples segregated based on amplification/deletion of the gene or SNP of interest. The log rank *p*-value is calculated using Mantel-Haenszel procedure. The *p*-values are recalculated every time a new threshold is selected.

CHAPTER

4

CONDUCTING ADVANCED SEARCHES

This chapter describes how to use REMBRANDT to conduct advanced searches of the REMBRANDT repository and to create graphs from the results obtained.

Topics in this chapter include:

Advanced Searches Overview

Advanced searches can be conducted by adding gene expression analyses, copy number data analyses and/or clinical study analyses to queries.

Advanced searches are initiated on the **Advanced Search** tab *(Figure 4.1).* The **Advanced Search** tab provides three search categories:

- **Gene Expression Analysis**
- **Copy Number Data Analysis**
- **Clinical Study Analysis**

When you are finished searching in these categories, clicking the **Finalize Query** button located on the bottom of the tab brings up a page with options that enable you to validate your queries, select a report view and run the report.

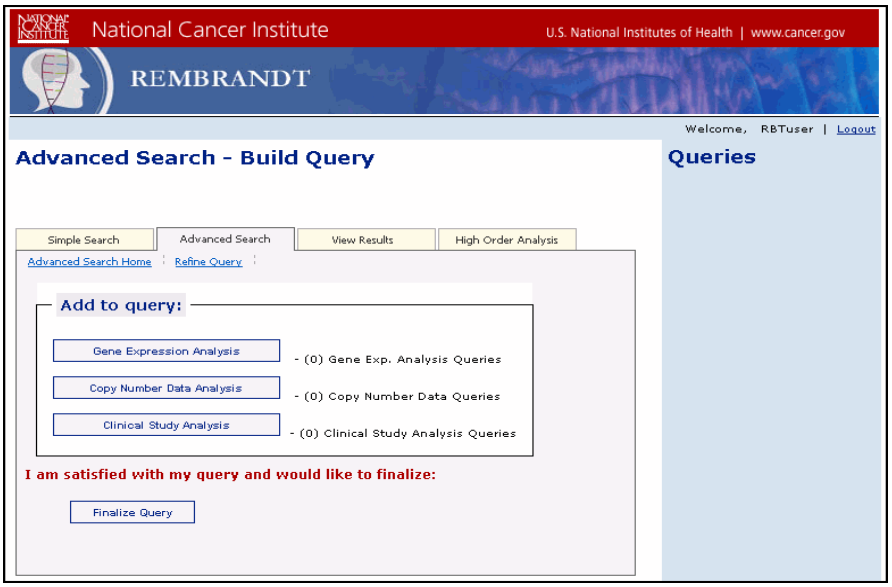


Figure 4.1 Advanced Search tab

Advanced Searches Workflow

When conducting an advanced search, complete the following steps in this suggested order:

1. Click the **Advanced Search** tab.
2. Choose the query you wish to add.
3. Enter search criteria on the next page that opens.

   Note: You can either enter search criteria or just preview.
4. Submit the query.
5. Refine your query on the Refine Query page.
6. Run the report.
7. View the report, filtering, selecting samples, and so on, as needed.
8. Save the report.

Adding Gene Expression Analyses to Queries

To add a gene expression analysis to a query, follow these steps:

1. From the Advanced Search - Build Query page, click **Gene Expression Analysis** *(Figure 4.2)*.

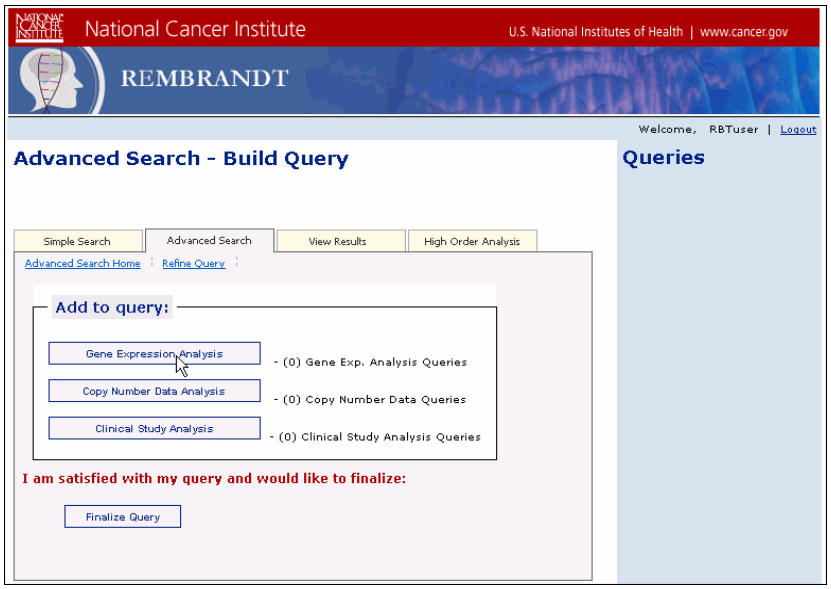


Figure 4.2 Selecting Gene Expression Analysis

The Gene Expression page appears *(Figure 4.3)*.

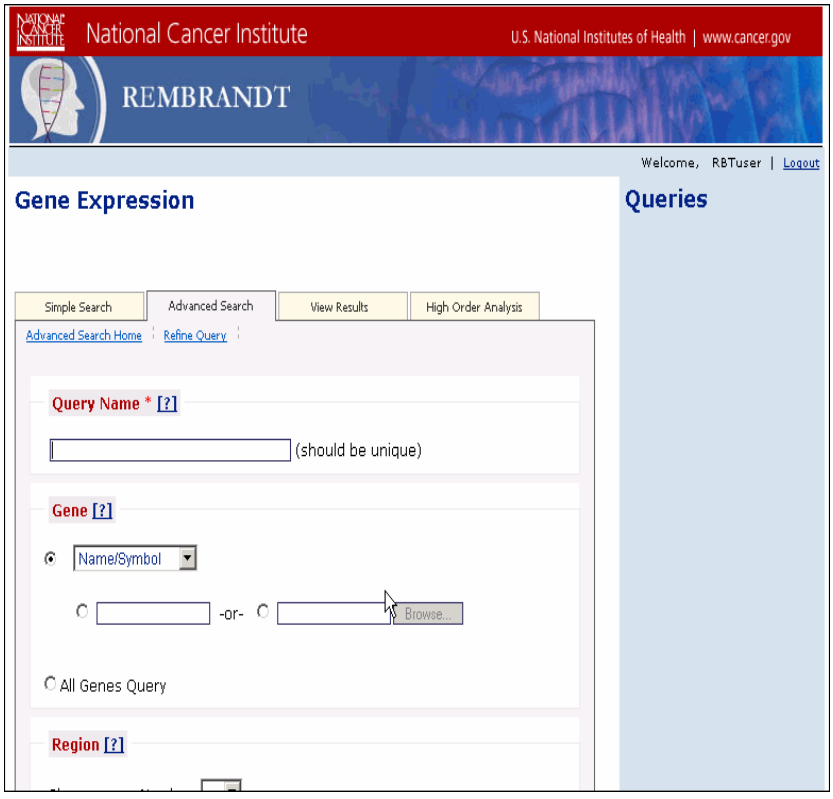


Figure 4.3 Gene Expression page (top portion)

Enter the desired search criteria, beginning with step 2.

Note: Only the **Query Name** field (described in step 2) is required, the other fields are optional; however, ensure that at least one of the remaining fields (as described below in steps 3-12) is filled out.

2. **Query Name (required):** Enter a title/name for this query into the **Query Name** text box.

   This name must be unique among all your queries in this session.

3. **Gene (optional):** Choose either:

   a. A gene identifier option (**Name/Symbol, Locus Link ID** or **GenBank AccNo.**) from the drop-down list, then enter the corresponding comma delimited value or IDs for the genes to be searched in the text box. Optionally, you can upload a text file containing Gene identifiers by clicking **Browse.**

   Note: There must only be one entry per line and a return must be added at the end of the file.

   b. The **All Genes Query** if you do not wish to specify a list of genes, but would like to see the data for all the genes analyzed.

   Note: You must apply the **All Genes Query** to a pre-existing result set.

4. **Region (optional):** Specify the chromosomal region of interest to search by choosing a chromosome of interest (**1-22, X** or **Y**) from the **Chromosome Number** drop-down list.

   Then either:

   - Choose a **Cytoband** range. The **Cytoband** drop-down list is context sensitive and lists only the relevant cytobands for a particular chromosome. Click **MAP Browser** to conduct a search.

   - OR -

   - Enter the start and end **Base Pair Positions.**

5. **Clone Id/Probe Set Id (optional):** Choose either **IMAGE Id** or **Probe Set Id** from the drop-down list.

   Then paste a comma delimited IMAGE Clone Id/Affymetrix probe set Id list, or upload a text file containing the IDs using the **Browse** button.

   Note: There must only be one entry per line and a return must be added at the end of the file. Image IDs must start with "IMAGE:".

6. **Gene Ontology (GO) Classifications (optional):** Enter a Gene Ontology (GO) ID in the text box to search for one or more branches of the GO hierarchy, such as ***GO:0005006.***

   To search for a classification, click **GO Browser.** The AmiGO browser opens *(Figure 4.4).*

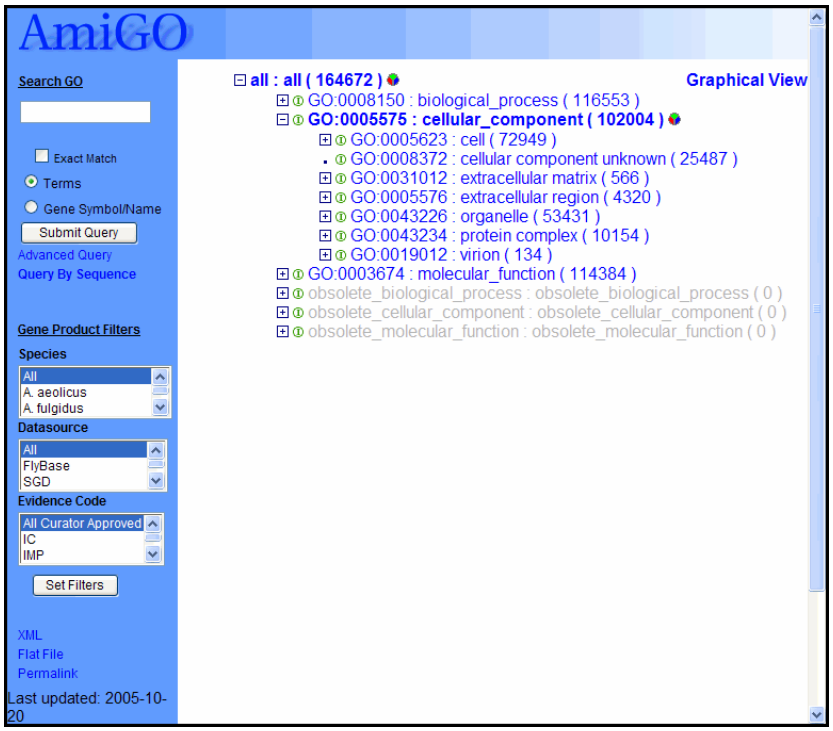


Figure 4.4 AmiGO browser

Enter your GO number in the search box, select your search criteria, and then click **Submit Query**. Results from the search will display *(Figure 4.5).*

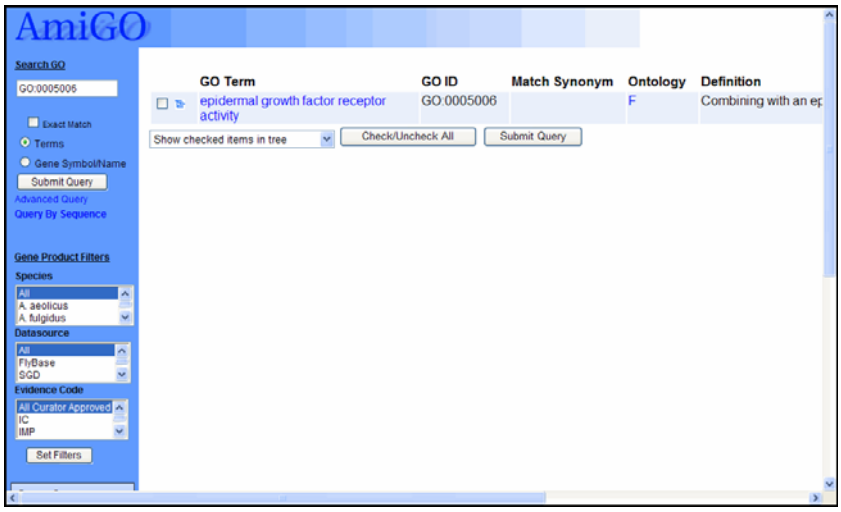


Figure 4.5 AmiGO search results

7. **Pathways (optional):** Enter the pathways name in the text box by clicking **browse caBIO** and selecting the pathway of interest from the pop-up window *(Figure 4.6).* Biocarda pathways are displayed in the pop-up window.

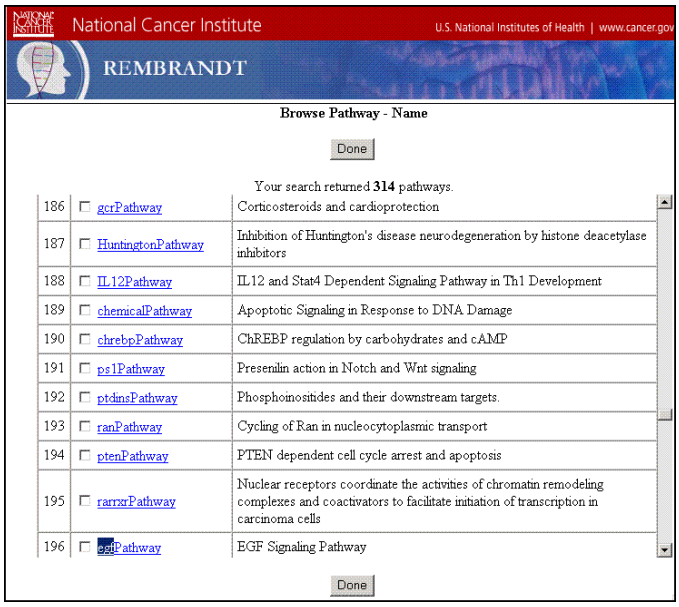


Figure 4.6 caBIO pop-up window

Note: While browsing pathways, to quickly find the pathway of interest type CTR-F to open a search dialog box *(Figure 4.7)*. Enter the pathway into the search text box and click **Find Next.**

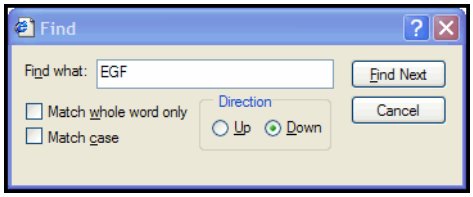


Figure 4.7 Search dialog box

From the caBIO pop-up window, select the check box next to the pathway of interest and click **Done**. The selected pathway will display on the Gene Expression page in the Pathways text box *(Figure 4.8)*.

Figure 4.8 Completed Pathways text box

Note: To remove the name of the pathway and re-enter information, click **clear text area.**

Combine a query and disease type

To combine a query with a disease type, select criteria from the **AND** group box by continuing with these steps:

8. **Disease Type (optional):** Select the type of disease from the drop-down list. You can see the context-sensitive Grade drop-down list for the disease type selected.

   Click the **sub-types** link to view tumor sub-types *(Table 4.1)* for the selected disease type. The data base currently houses data from the brain tumor types shown in Table *4.1*.

Disease Type	*Tumor Sub-Types*
Astrocytic	Astrocytoma (WHO grade II) Variants: protoplasmic, gemistocytic, fibrillary, mixed
	Anaplastic (malignant) astrocytoma (WHO grade III)
	Glioblastoma multiforme (WHO grade IV)
	Pilocytic astrocytoma (non-invasive, WHO grade I)
	Subependymal giant cell astrocytoma (non-invasive, WHO grade I)
	Pleomorphic xanthoastrocytoma (non-invasive, WHO grade I)
Oligodendroglial	Oligodendroglioma (WHO grade II)
	Anaplastic (malignant) oligodendroglioma (WHO grade III)
Mixed gliomas	Ependymoma (WHO grade II)
	Anaplastic ependymoma (WHO grade III)
	Myxopapillary ependymoma
	Subependymoma (WHO grade I)
Glioblastoma	Mixed oligoastrocytoma (WHO grade II)
	Anaplastic (malignant) oligoastrocytoma (WHO grade III)
	Others (e.g. ependymo-astrocytomas)
Unclassified	
Unknown	

Table 4.1 Tumor sub-types

9. **Sample Identifier (optional):** Manually enter Sample Id(s) or select Sample Id(s) from a file using the **Browse** button.

   Note: There must only be one entry per line and a return must be added at the end of the file.

10. **Fold change (optional):** Specify the threshold for the differential regulation by indicating “up," "down" or "unchanged" criteria. This will return differential

expression ratios between tumor and non-tumor samples for a particular reporter. If you are creating an "All Genes" query, you must select a fold change threshold of 4 or above.

11. **Array Platform (optional):** Select the array platform (**All, Oligo (Affymetrix)** or **cDNA**) from the drop-down list.
12. When you have finished selecting your criteria, choose one of the following buttons located at the bottom of the page.
    a. **Clear:** Restores the report to its original state and clears any highlighting.
    b. **Cancel:** Eliminates all data currently entered in the form; a query will not be added
    c. Preview: Displays a preview of your report.
    d. Submit: Returns you to the Advanced Search-Build Query page where you can either finalize your query or add another query.
13. Continue with the next procedure in *Finalizing Queries* on page 34.

Finalizing Queries

To finalize a query, follow these steps:

1. After clicking **submit** in the last step of the procedure to add a query (see *Adding Gene Expression Analyses to Queries* on page 28), you will have been returned to the Advanced Search - Build Query page.
   - To add another query, click one of the buttons in the **Add to query** group box and repeat the procedure to add a query.
   - If you are ready to finalize your query, continue with Step 2.
2. Click **Finalize Query.**
   The Refine Query page appears. This page helps you group your queries and select the report type in a systematic way using five steps.
3. Continue with the next procedure in the following section, Working with the Refine Query Page.

Working with the Refine Query Page

The Refine Query page *(Figure 4.9)* allows you to select and/or combine the queries that you have saved in the current session. You can also select the report you would like to view for the queries you have selected. You can combine complex ad hoc queries to create specific reports.

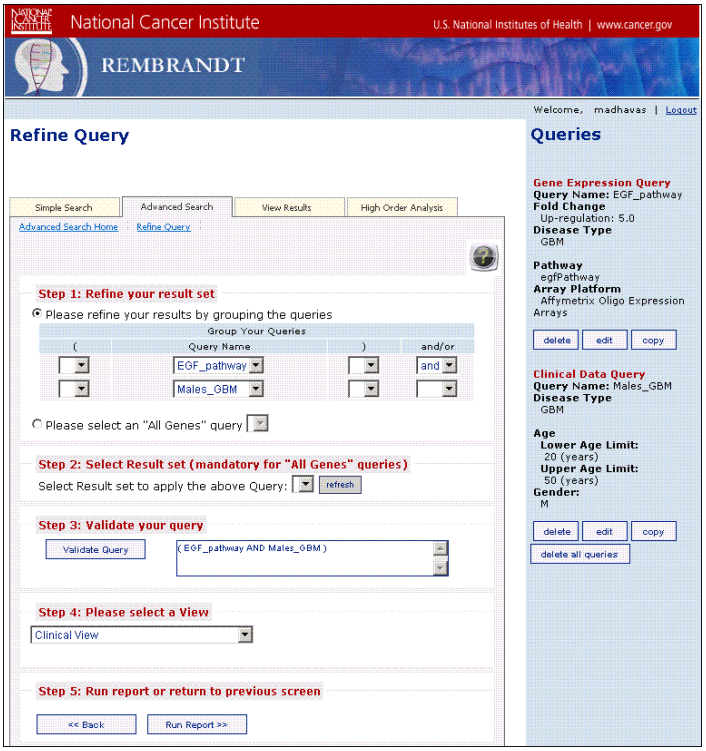


Figure 4.9 Refine Query page

Steps on the Refine Query page are filled out as described below:

1. **Step 1: Refine your result set.** Step 1 helps you select the queries that you have built from the drop-down list. Selecting an **and/or** operator at the end of a query row enables the next query row where you can select another query of interest.

   Choose one of the following options:

   - **Please refine your results by grouping the queries**

     Select an option from the Query Name drop-down, select a closing parentheses, from the third drop-down and then choose “and” or “or” from the last drop-down on that row.

     Repeat for each query name to be grouped.

   - **Please select an “All Genes” query.**

     Selecting this option activates the drop-down list from which you can choose an All Genes query.

   Note: Choose the **All Genes** query if you do not wish to specify a list of genes, but would like to see the data for all the genes analyzed. You must apply the **All Genes** query to a pre-existing result set.

2. **Step 2: Select Result set (mandatory for “All Genes” queries).** Step 2 lets you select a previously saved result to which to apply these queries. You will not see any result sets if you have not saved them from a previously viewed report.

Clicking on the **refresh** button brings up previously saved result sets in the drop-down list.

3. **Step 3: Validate your query.** Step 3 helps validate your query collection. When you click **Validate Query,** the number of parentheses you may have added to the query grouping will be validated and the name of your query will display in the text box.
4. **Step 4: Please select a View.** Step 4 allows you to select a report from the drop-down list. The available reports vary, depending upon the queries that you selected in **Step 1. Refine your result set.**
5. **Step 5: Run report or return to previous screen.** Step 5 allows you to run the report; for example, a Gene Expression Sample report *(Figure 4.10).*
   - To filter, select samples from and/or annotate your report, continue with the procedures in one or more of the following sections: *Filtering Reports* on page 36, *Selecting Samples* on page 37 or *Differentiating Data* on page 39.
   - To run your report, continue with the procedures in *Saving Samples and Displaying Results* on page 42.

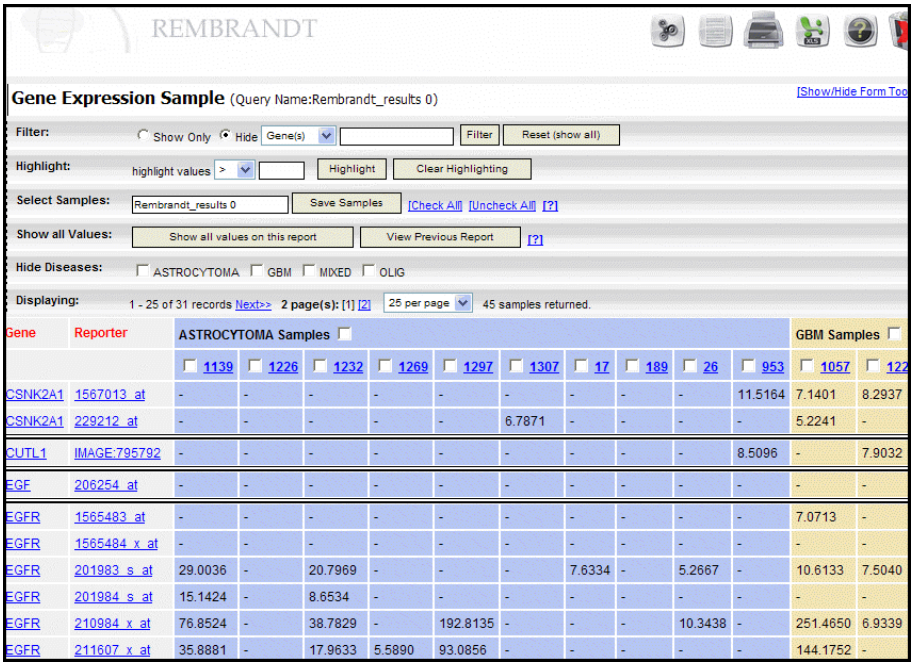


Figure 4.10 Gene Expression Sample

Filtering Reports

To filter a report, follow these steps:

Filter Toolbar

1. From the **Filter** toolbar *(Figure 4.11),* select the filter mode (**show only** or **hide**).
2. Select the element(s) that are to be shown or hidden and click **Filter.**

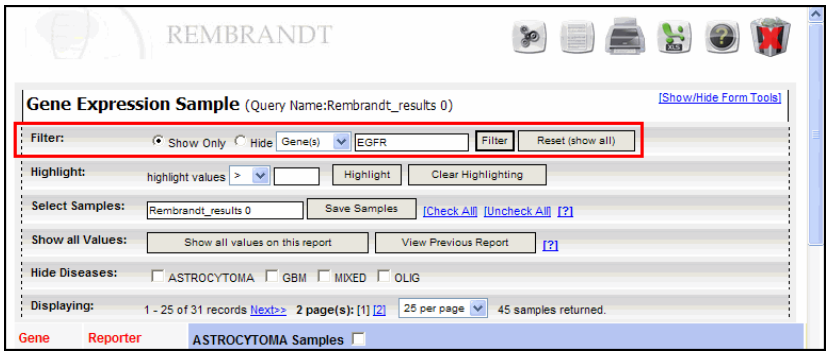


Figure 4.11 Filter toolbar

In this example, only the EGFR gene is selected to display on this report. All other genes *(Figure 4.12)* have been removed from the report and only EGFR is displayed. This allows you to efficiently target specific elements of interest.

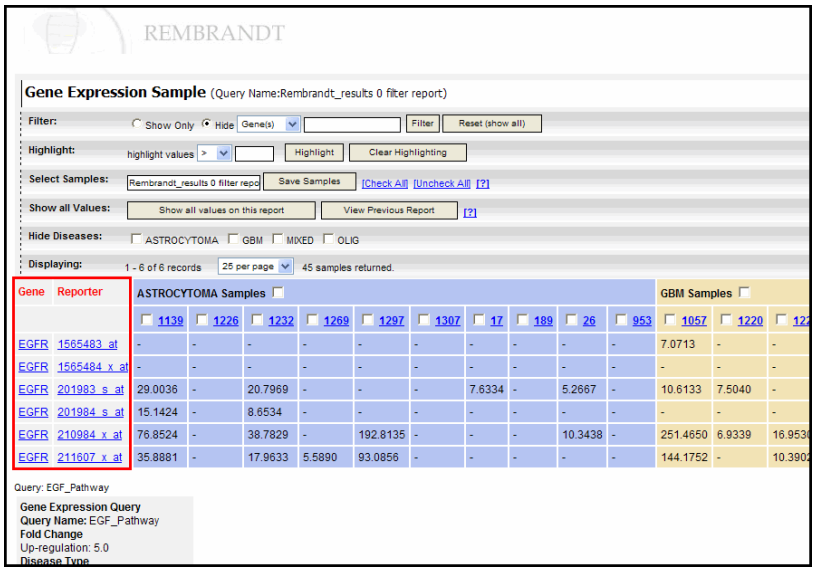


Figure 4.12 Only selected gene displays

3. To reset your filter, click **Reset (show all),** which will restore your report to its original state.

Selecting Samples

To highlight certain data, follow these steps:

Highlights toolbar

1. Choose an operator and a threshold value from the **Highlights** toolbar *(Figure 4.13)*.

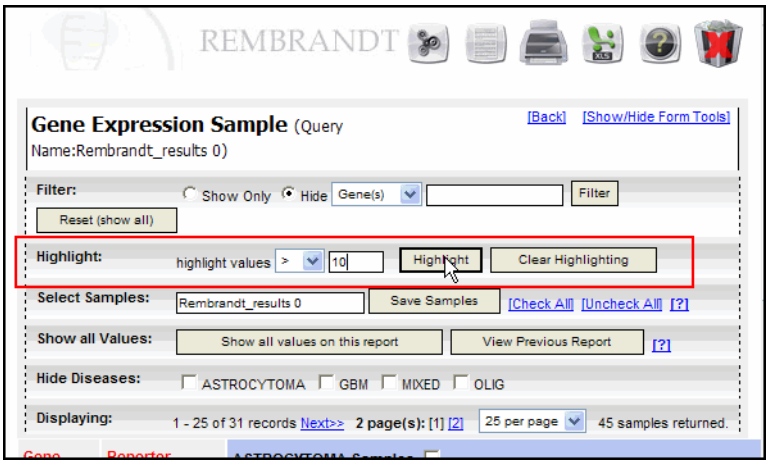
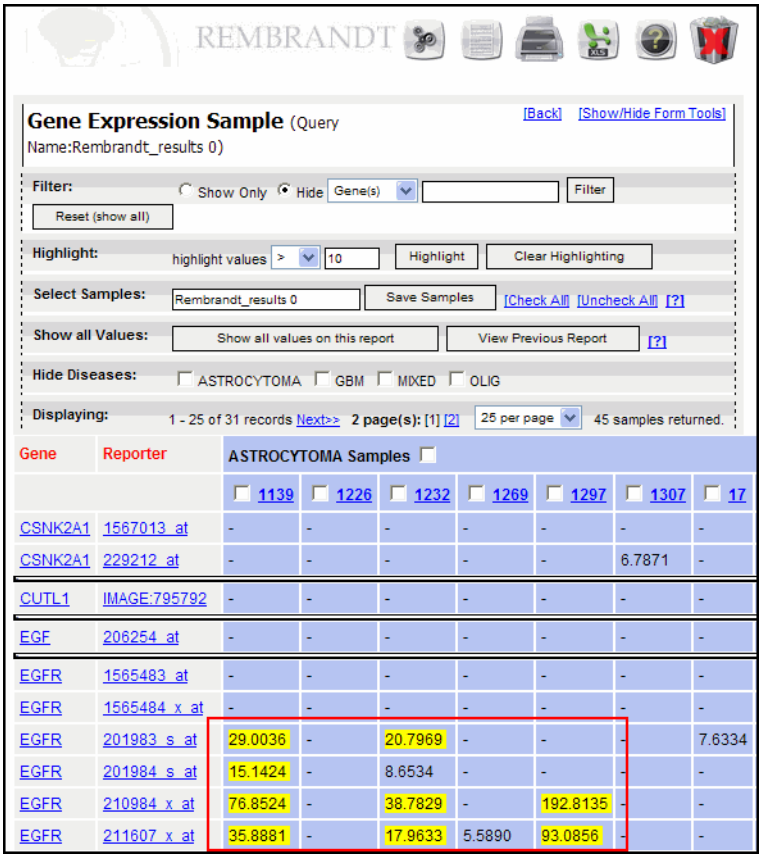


Figure 4.13 Highlights toolbar

2. Click **Highlight.**
   The values that meet this criteria will be highlighted in yellow *(Figure 4.14).*



Gene	Reporter	ASTROCYTOMA Samples						
		1139	1226	1232	1269	1297	1307	17
CSNK2A1	1567013_at	-	-	-	-	-	-	-
CSNK2A1	229212_at	-	-	-	-	-	6.7871	-
CUTL1	IMAGE:795792	-	-	-	-	-	-	-
EGF	206254_at	-	-	-	-	-	-	-
EGFR	1565483_at	-	-	-	-	-	-	-
EGFR	1565484_x_at	-	-	-	-	-	-	-
EGFR	201983_s_at	29.0036	-	20.7969	-	-	-	7.6334
EGFR	201984_s_at	15.1424	-	8.6534	-	-	-	-
EGFR	210984_x_at	76.8524	-	38.7829	-	192.8135	-	-
EGFR	211607_x_at	35.8881	-	17.9633	5.5890	93.0856	-	-

Figure 4.14 Selected highlights display

3. To restore the report to its original state and clear the highlighting, click **Clear Highlighting.**

Differentiating Data

Data can be differentiated in several ways:

Show All Values toolbar

- To select samples, refer to *Saving Samples and Displaying Results* on page 42.
- To differentiate between missing values in the array and data that did not meet your search criteria, click **Show All Values on this Report** on the **Show all Values** toolbar *(Figure 4.15)*.

  This displays in gray all available data that did not meet your criteria. A value of **Null** indicates a missing value for that reporter.

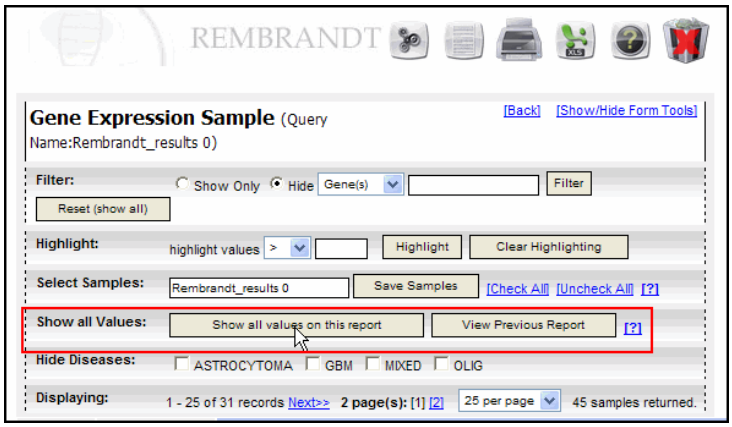


Figure 4.15 Show All Values toolbar

- To pop up a new page with additional information about a gene or reporter, click on the **Gene** or **Reporter** column.
  - The **Gene** column is shown in *Figure 4.16*.

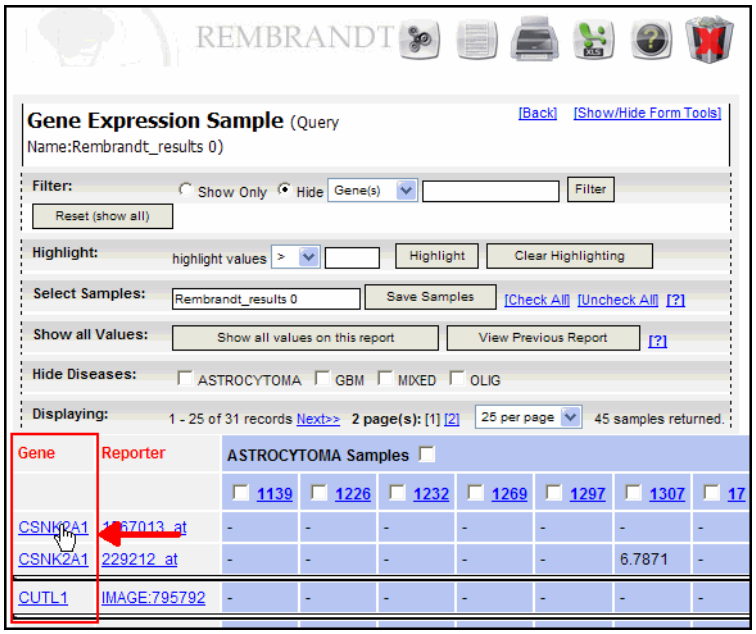


Figure 4.16 The Gene column

To display more information about a gene, click on a selected gene in the **Gene** column, which opens the Cancer Genome Anatomy Project (CGAP) browser *(Figure 4.17).*

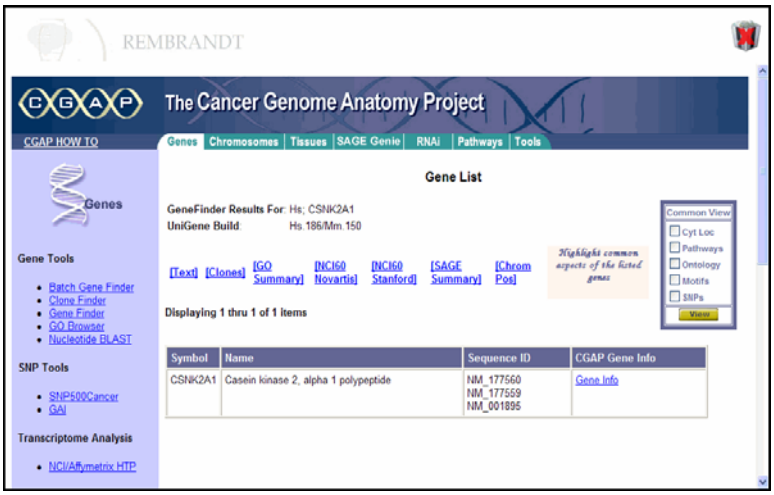


Figure 4.17 Cancer Genome Anatomy Project browser

- The **Reporter** column is shown in *Figure 4.18*.

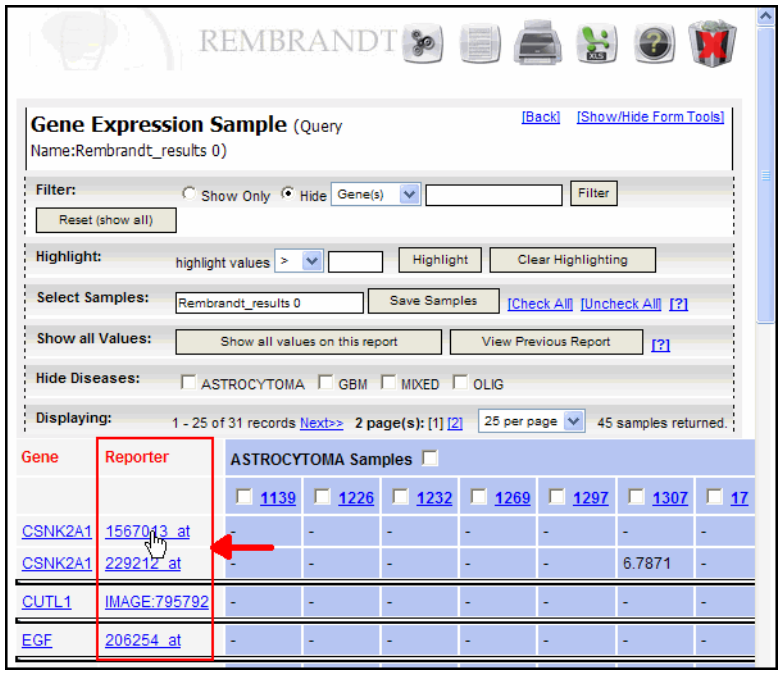


Figure 4.18 The Reporter column

To display more information about a reporter, click the reporter in the **Reporter** column, which opens the Affymetrix Probe Viewer from NCI's Laboratory of Population Genetics *(Figure 4.19),* which tells you where the probes are lined up.

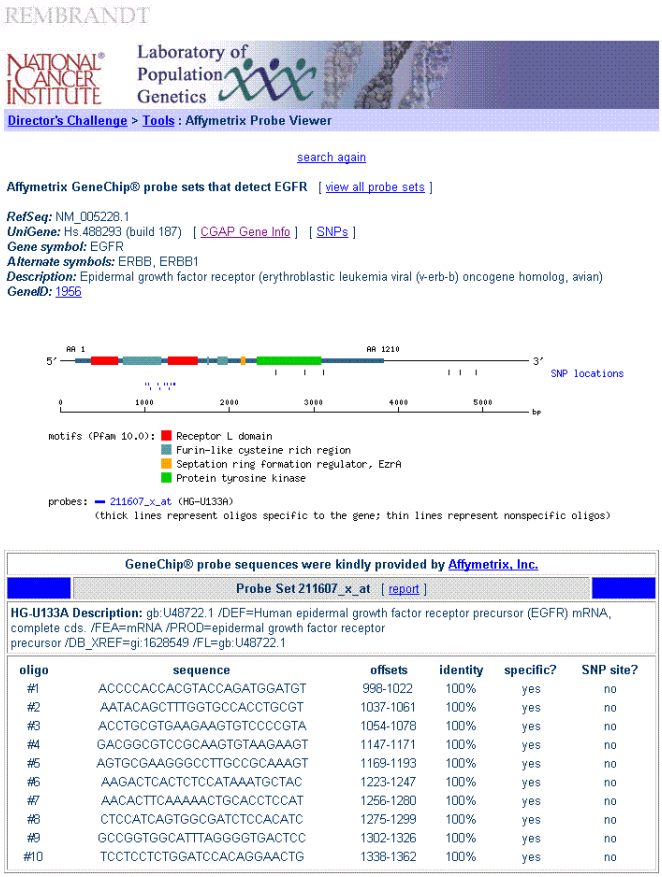


Figure 4.19 Affymetrix Probe Viewer from NCI's Laboratory of Population Genetics

Hide Diseases toolbar

- To remove an entire group of columns by removing a specific disease from the report, select the check box for the disease in the **Hide Diseases** toolbar *(Figure 4.20)*.

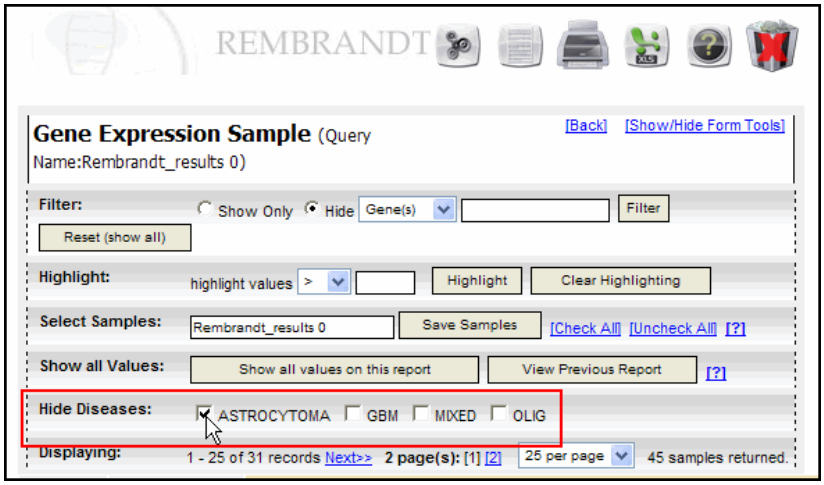


Figure 4.20 Hide Diseases toolbar

The selected disease samples will be removed from your report.

Saving Samples and Displaying Results

To save samples for use with an additional query, follow these steps:

1. Select the samples in which you are interested *(Figure 4.21)*.

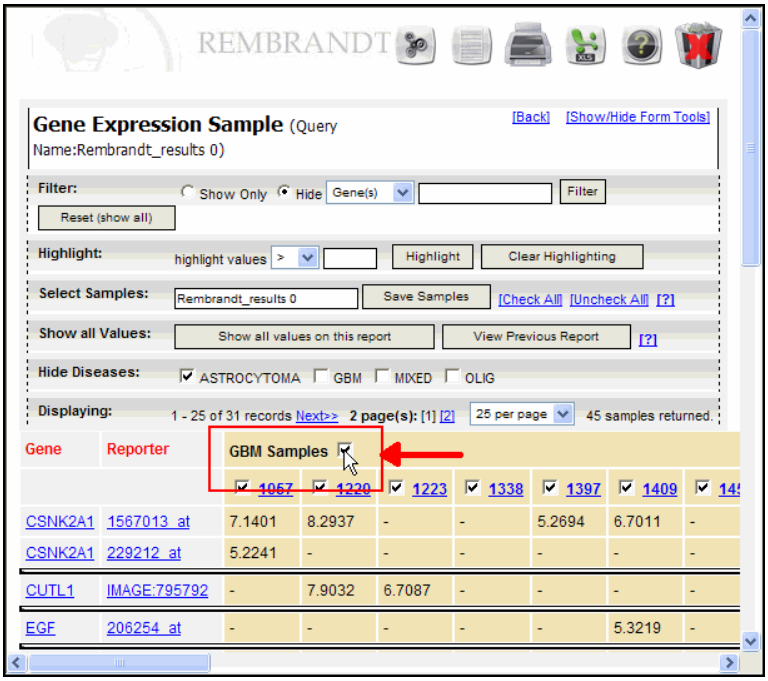


Figure 4.21 Selecting a sample to save

Select Samples toolbar

2. Give the sample set a unique name by entering the name into the **Select Samples** text box, located on the **Select Samples** toolbar *(Figure 4.22)*.
3. Click **Save Samples.**

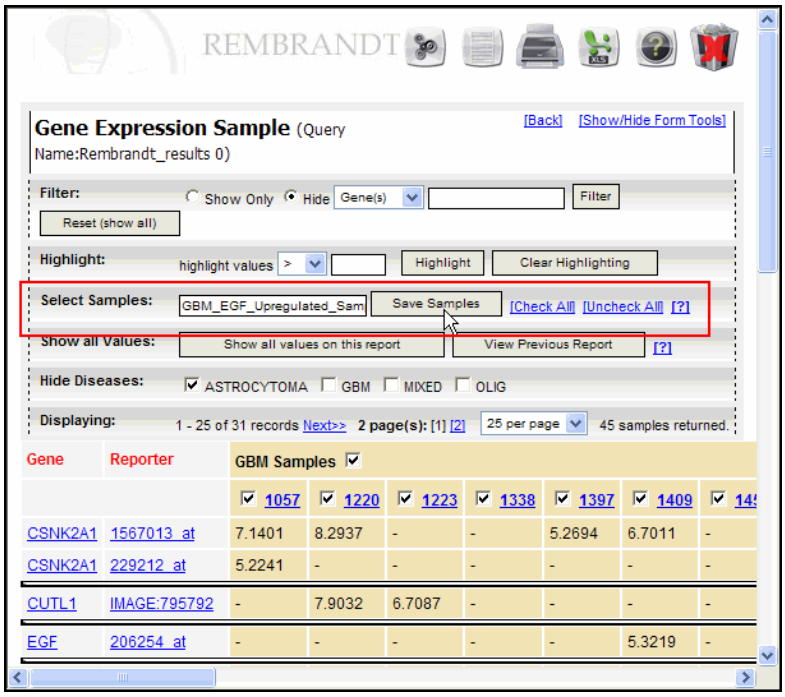
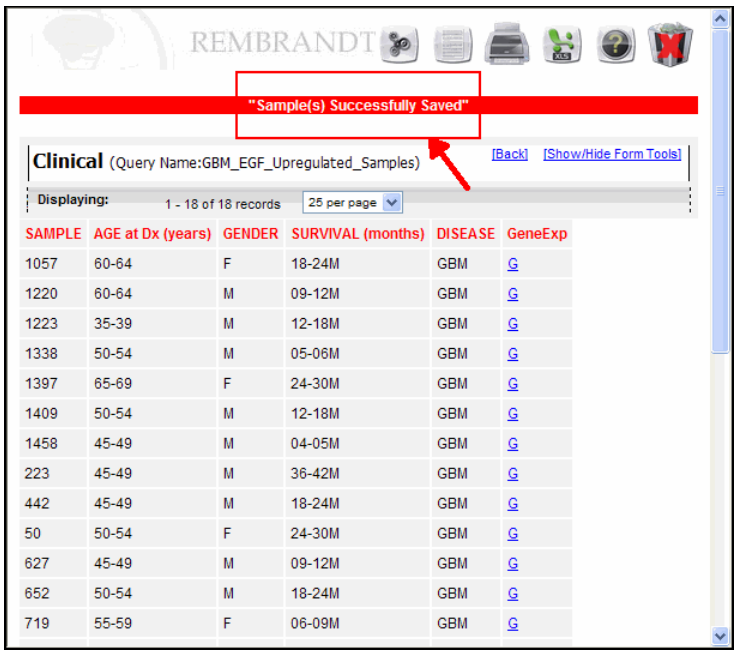


Figure 4.22 Select Samples toolbar

4. A new page appears, showing the clinical data for the samples that were chosen to be saved. A confirmation message, “Sample(s) Successfully Saved” displays at the top of the page *(Figure 4.23)*.



SAMPLE	AGE at Dx (years)	GENDER	SURVIVAL (months)	DISEASE	GeneExp
1057	60-64	F	18-24M	GBM	G
1220	60-64	M	09-12M	GBM	G
1223	35-39	M	12-18M	GBM	G
1338	50-54	M	05-06M	GBM	G
1397	65-69	F	24-30M	GBM	G
1409	50-54	M	12-18M	GBM	G
1458	45-49	M	04-05M	GBM	G
223	45-49	M	36-42M	GBM	G
442	45-49	M	18-24M	GBM	G
50	50-54	F	24-30M	GBM	G
627	45-49	M	09-12M	GBM	G
652	50-54	M	18-24M	GBM	G
719	55-59	F	06-09M	GBM	G

Figure 4.23 A successfully saved sample

Making a result set available to another query

5. To refresh your list of available Result sets so that your newly saved sample set will be available for use with your additional query, click the "X" button in the top right-hand corner to close the page. The Refine Query page will reappear.
6. From the Refine Query page, click the **Refresh** button *(Figure 4.24)* under "Step 2" to refresh your list of available Results sets so that your newly saved sample set will be available for use.

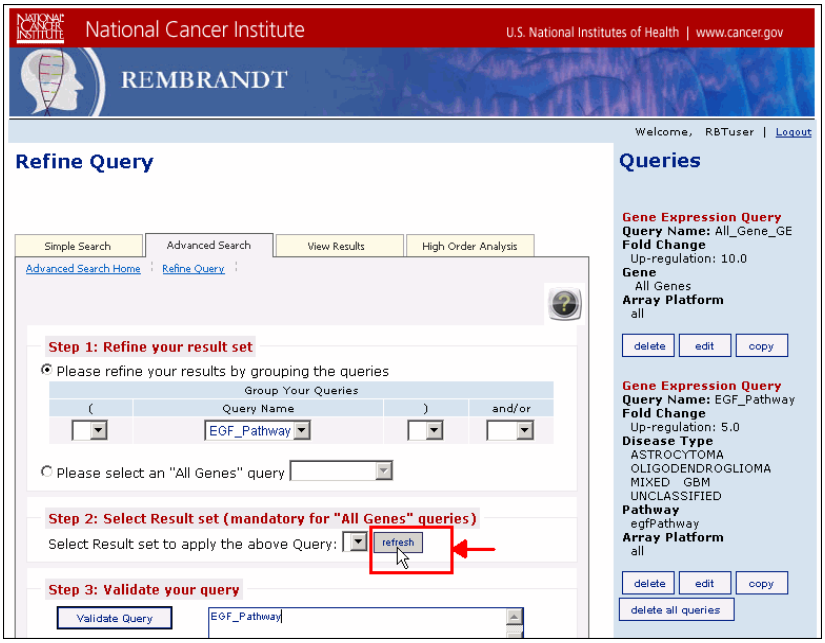


Figure 4.24 Clicking the Refresh button

7. To run the additional query, select the radio button adjacent to the query and select the query from the drop-down list *(Figure 4.25)*.

Figure 4.25 Selecting the additional query

8. To apply your saved result set for use with your additional query, click **refresh** and then select the saved result set from the drop-down list.

9. Under “Step 3” click **Validate Query.**
10. Under “Step 4” select a view.
11. Under “Step 5” click **Run Report** *(Figure 4.26).*

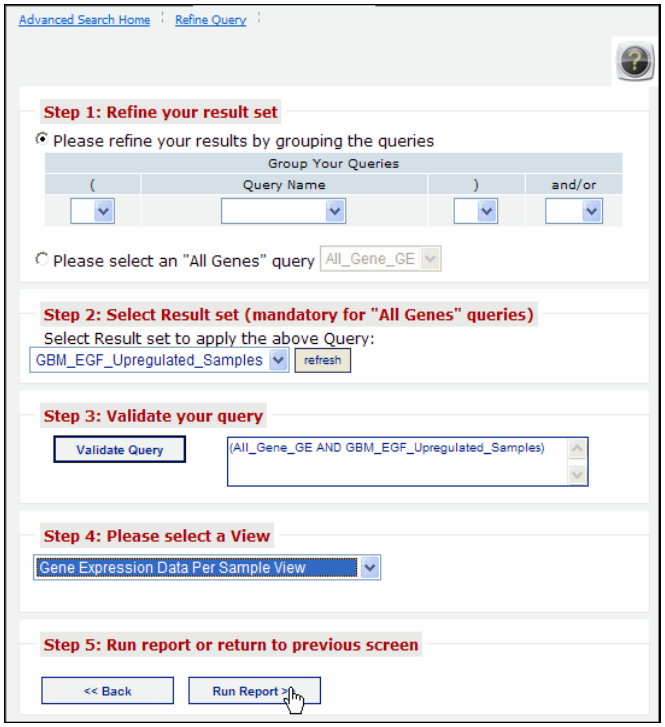


Figure 4.26 Running a report

12. The report now displays the results for all genes applied to the saved sample set *(Figure 4.27).* From this report, you can print, download for Excel or view query details by clicking the navigation icons in the top right-hand corner.

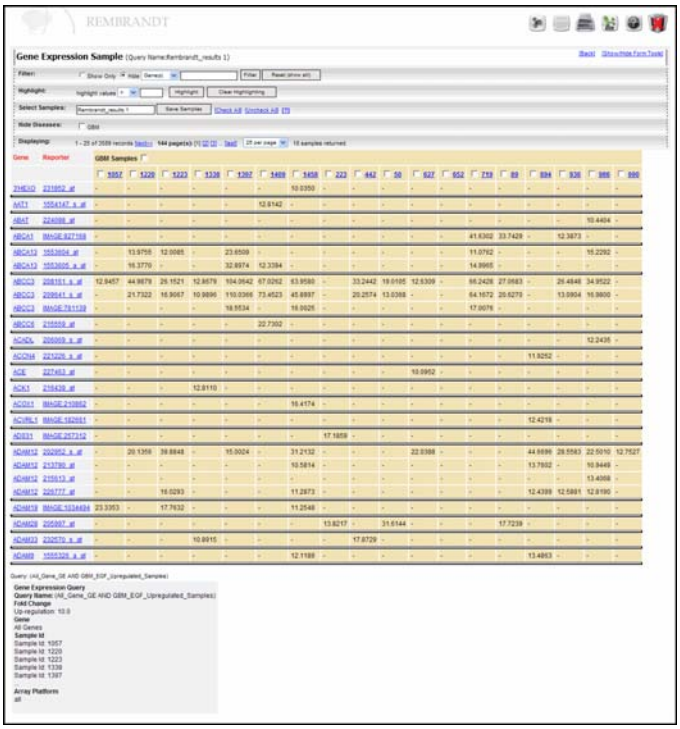


Figure 4.27 Report for two samples

Adding Copy Number Data Analyses to Queries

To add a copy number data analysis to a query, follow these steps:

1. From the Advanced Search - Build Query page, click **Copy Number Data Analysis** *(Figure 4.28)*.

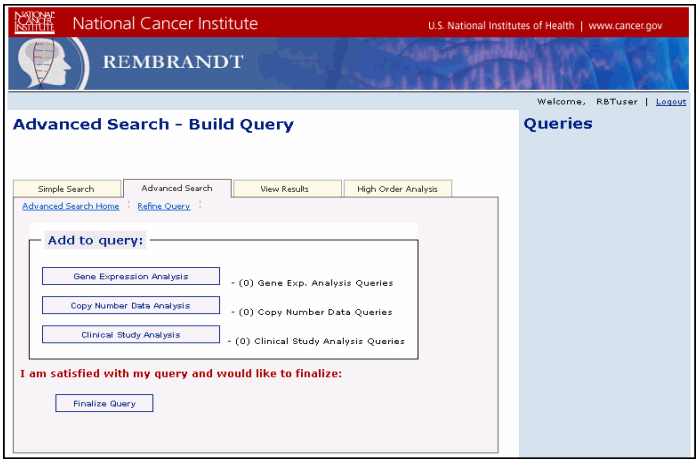


Figure 4.28 Selecting Copy Number Data Analysis option

The Copy Number Data page *(Figure 4.29)* appears.

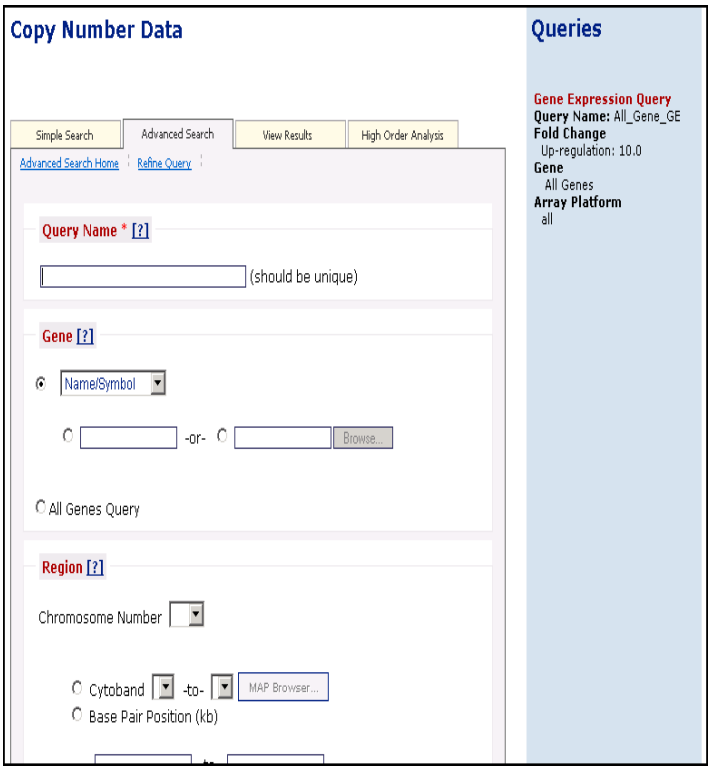


Figure 4.29 Copy Number Data page (top portion)

Enter the desired search criteria as described in the following steps.

Note: Only the **Query Name** field and the **Disease Type** field are required, the rest are optional.

2. **Query Name (required):** Enter a title/name for this query into the **Query Name** text box.

   This name must be unique among all your queries in this session.

3. **Gene (optional):** Choose one of the following Gene identifier options from the drop-down list.

   a. **Genbank ID**

   b. **LocusLink ID**

   c. **Gene symbol**

   Then enter the corresponding comma delimited value or IDs for the genes to be searched in the text box. Optionally, you can upload a text file containing Gene identifiers by clicking **Browse.**

   Note: There must only be one entry per line and a return must be added at the end of the file.

   Select **All Genes query** if you do not wish to specify a list of genes, but would like to see the data for all the genes analyzed. You must apply the **All Genes query** to a pre-existing result set.

Note: You will not see any result sets if you have not saved them from a previously viewed report. Clicking on the **refresh** button brings up previously saved result sets in the drop-down list.

4. **Region (optional):** Specify the chromosomal region of interest to search by choosing a chromosome of interest and either choosing a cytoband range or entering the start and end base pair positions. The **Cytoband** drop-down list is context sensitive and lists only the relevant cytobands for a particular chromosome.
    a. **Chromosome Number:** Select an option from the drop-down list (1-22, X, Y).
    b. Choose either the **Cytoband** range or indicate the start and end **Base Pair Positions.**
5. **SNP Id (optional):**
    a. Choose one type of SNP identifiers (**dbSNP ID** or **SNP Probe set ID**) from the drop-down list.
    b. Then enter the corresponding comma delimited IDs for SNPs to be searched in the text box. Optionally, you can upload a text file containing SNP ID list by clicking the browse button.

    Note: There must only be one entry per line and a return must be added at the end of the file.

    c. Choose one type of Validated SNPs: **All, Excluded, Included,** or **Only.**
6. **Disease Type (required):** Select the type of disease from the drop-down list. You can see the context-sensitive Grades menu for the disease type selected. Filtering the data based on grades will be available in the next release.

    Click the **sub-types** link to view tumor sub-types for the selected disease type. (Refer to *Tumor Sub-Types* on page 33.)
7. **Sample Identifier (optional):** Manually enter Sample Id(s) or select Sample Id(s) from a file using **Browse** button.

    Note: There must only be one entry per line and a return must be added at the end of the file.
8. **Copy Number (optional):** Specify the threshold for the copy number by indicating the **Amplified, Deleted** or **Unchanged** criteria. If you are creating an "All Genes" query, you must select an amplification threshold > 10 or a deletion threshold < 1.
9. **Assay Platform (optional):** Indicate the platform that was used for the comparative genomic study.
10. When you have finished selecting your criteria, choose one of the following buttons located at the bottom of the page.
    a. **Clear:** Restores the report to its original state and clears any highlighting.
    b. **Cancel:** Eliminates all data currently entered in the form; a query will not be added
    c. **Preview:** Displays a preview of your report.

11. **Submit:** Returns you to the Advanced Search-Build Query page where you can either finalize your query or add another query.

    Note: See *Finalizing Queries* on page 34 for more information.

Copy Number Report

This report displays the copy number data from Affymetrix 100K SNP arrays

The CHP files from the Affymetrix Gene Chip Operating System were processed using the GDAS (GeneChip® DNA Analysis Software) and copy number data was collected for each mapping SNP reporter on the Chip, for all the tumor samples using the Affymetrix CNAT (Copy Number Analysis Tool).

Each column represents a sample and the samples are grouped based on the tumor sub-type.

By selecting copy number in the **filter** drop-down list, you can filter the report for samples that have a user-defined number of consecutive SNPs or % SNPs that match the criteria.

You can visualize the copy number data in a graphing application called webGenome by clicking on the **View plots** link in the copy number report.

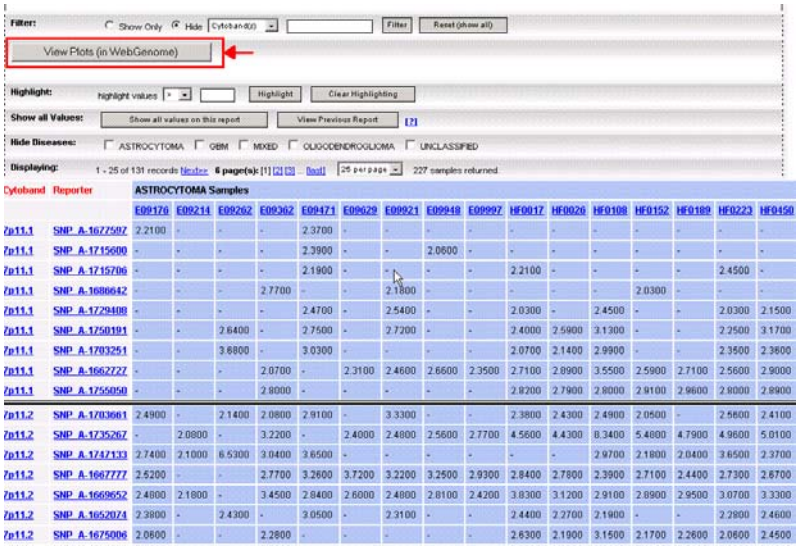


Figure 4.30 Copy Number Report

Scatter Plot

Scatter plots display measured copy number against physical genome location. You can view data at arbitrary resolutions from the entire genome on down. When moving the mouse over specific probes the system provides mouse-over probe names. Clicking on the name of an experiment or bioassay in the plot legend will highlight the corresponding data.

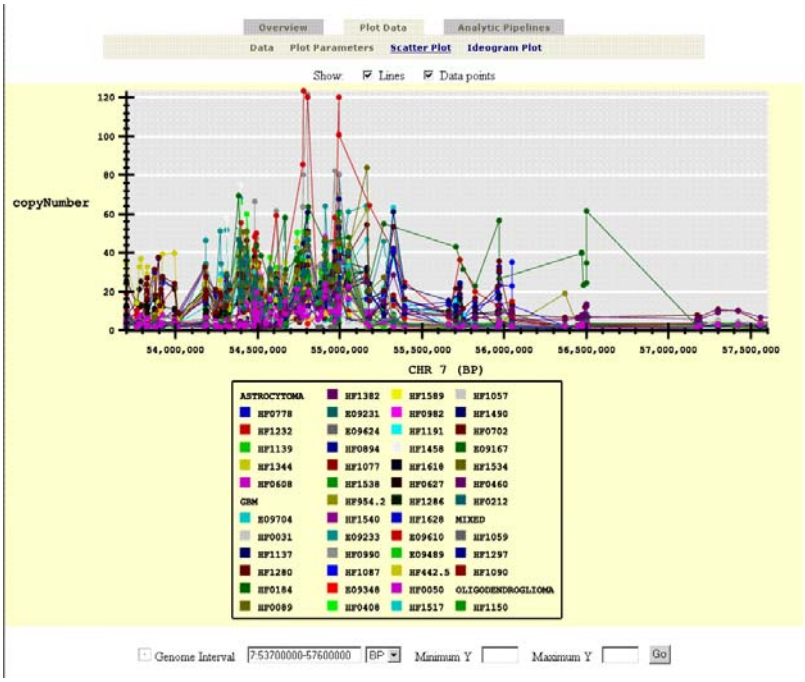


Figure 4.31 WebGenome scatter plot

Ideogram Plot

This plot shows color-coded data values in relation to chromosome ideograms.

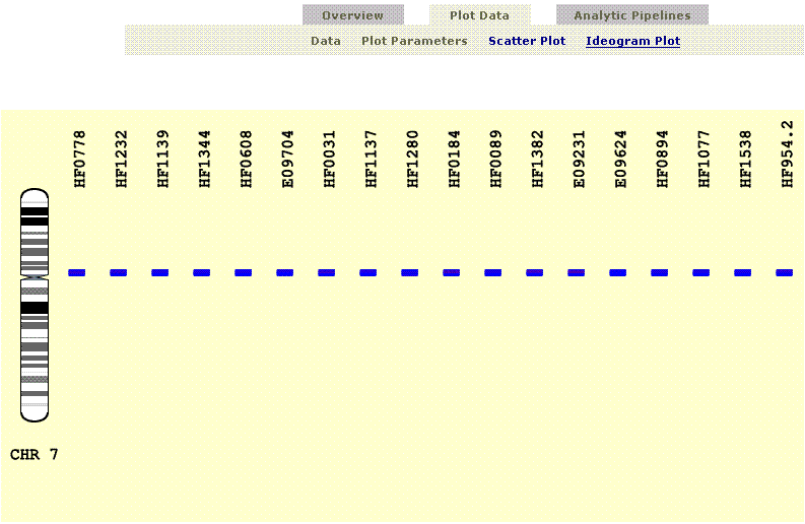


Figure 4.32 WebGenome ideogram plot

Adding Clinical Study Analyses to Queries

To add a clinical study analysis to a query, follow these steps:

1. From the Advanced Search - Build Query page *(Figure 4.33)*, click **Clinical Study Analysis.**

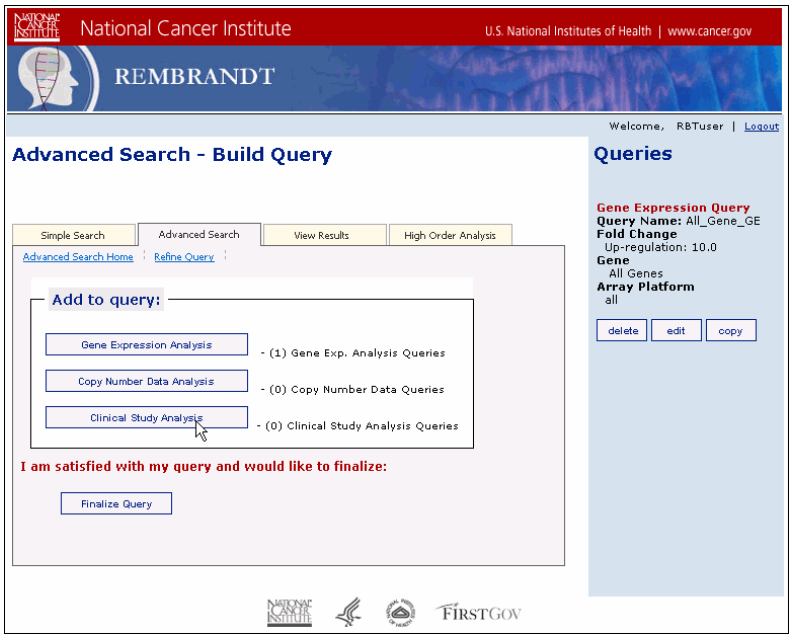


Figure 4.33 Selecting the Clinical Study Analysis option

The Clinical Data page *(Figure 4.34)* appears.

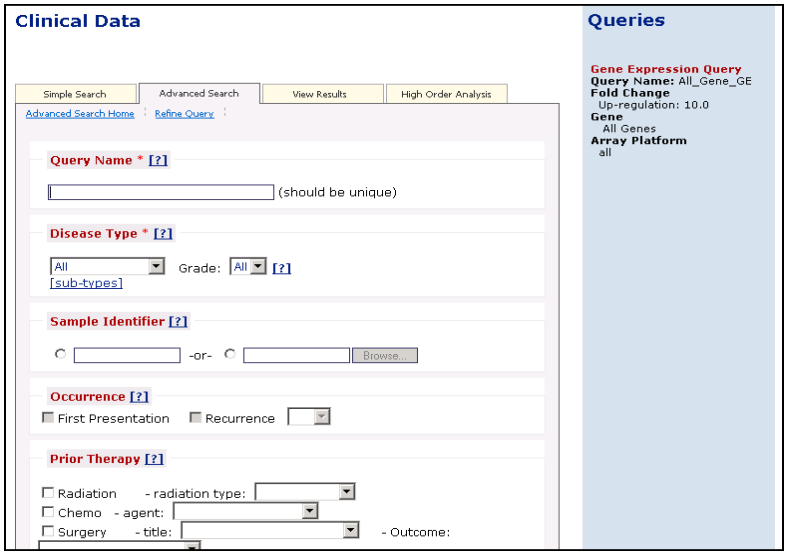


Figure 4.34 Clinical Data page (top portion)

Enter the desired search criteria as described in the following steps.

2. **Query Name (required):** Enter a title/name for this query into the **Query Name** text box.

   This name must be unique among all your queries in this session.

3. **Disease Type (required):** Select the type of disease from the drop-down list options. You can see the context-sensitive Grades menu for the disease type selected.

   Click the **sub-types** link to view tumor sub-types for the selected disease type. Refer to *Tumor Sub-Types* on page 33.

4. **Sample Identifier (optional):** Manually enter Sample Id(s) or select Sample Id(s) from a file using **browse** button.

   There must only be one entry per line and a return must be added at the end of the file.

5. **Survival Range (optional):** Indicate the range of the patient survival length (in months) after the first diagnosis by selecting an option from the **lower** and **upper** drop-down lists.

6. **Age at Dx (optional):** Specify the upper and lower limits for filtering the clinical data based on the age (in years) at which a patient was diagnosed with the disease by selecting an option from the **lower** and **upper** drop-down lists.

7. **Gender (optional):** Select either **Male, Female** or **Other** from the drop-down list.

8. **Race (optional):** Choose either **Caucasian, African American, Latino, Asian American** or **Native American.**

9. When you have finished selecting your criteria, choose one of the following buttons located at the bottom of the page.

   a. **Clear:** Restores the report to its original state and clears any highlighting.

   b. **Cancel:** Eliminates all data currently entered in the form; a query will not be added

   c. **Preview:** Displays a preview of your report.

10. **Submit:** Returns you to the Advanced Search-Build Query page where you can either finalize your query or add another query.

    Note: See *Finalizing Queries* on page 34 for more information.

Clinical Report and Graph

You can plot the clinical data from a report on two kinds of graphs:

1. **Survival (months) vs Age at diagnosis (years):** The data points are colored by disease type. You can select the samples of interest by clicking on the graph and drawing a rectangle around the samples that you would like to save for future use.

2. **Karnowsky score (Neurological assessment) vs Age at diagnosis (years):** The data points are colored by disease type. You can select the samples of interest by clicking on the graph and drawing a rectangle around the samples that you would like to save for future use.

CHAPTER

5

HIGH ORDER ANALYSIS

This chapter describes how to use REMBRANDT to run higher order analyses, including class comparisons, hierarchical clustering and principal component analyses.

Topics in this chapter include:

High Order Analysis Overview

After data preprocessing (filtering and normalization), further statistical analysis ("High Order Analysis") of gene expression data are performed, including class comparison and class discovery.

These analyses are initiated on the **Higher Order Analysis** tab *(Figure 5.1).*

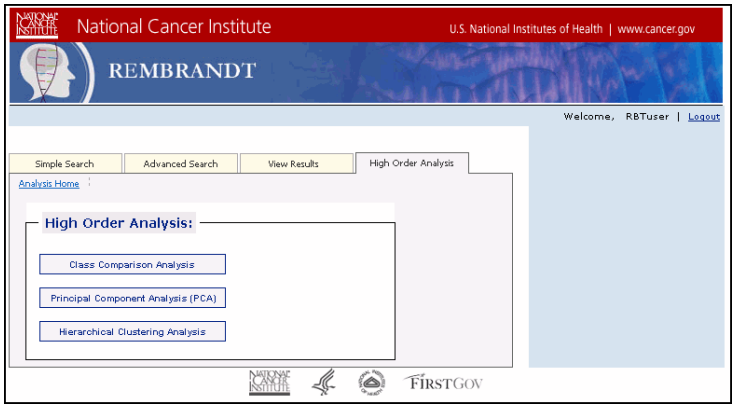


Figure 5.1 High Order Analysis tab

High Order Analysis Workflow

To create a High Order Analysis, complete the following steps in this suggested order:

1. Click the **High Order Analysis** tab.
2. Choose one of the following options from the **Analysis Tools** menu:
   a. ***Class Comparison*** on page 54
   b. ***Principal Component Analysis (PCA)*** on page 57
   c. ***Hierarchical Clustering*** on page 59
3. Enter criteria on the next page that opens.
4. Click **submit** to run your report.
5. Filter your report, if desired.

Using Analysis Tools

To create a High Order Analysis, choose one of the analysis tools from the **High Order Analysis** tab and follow the steps on the next page that opens to complete the analysis, as detailed in the following sections:

- *Class Comparison* on page 54
- *Principal Component Analysis (PCA)* on page 57
- *Hierarchical Clustering* on page 59

Class Comparison

To create a High Order Analysis with Class Comparisons, follow these steps:

1. Click the **High Order Analysis** tab.
2. Choose **Class Comparison** from the **Analysis Tools** menu.
   Select criteria from the Class Comparison Analysis Form *(Figure 5.2)* by continuing with the next step.

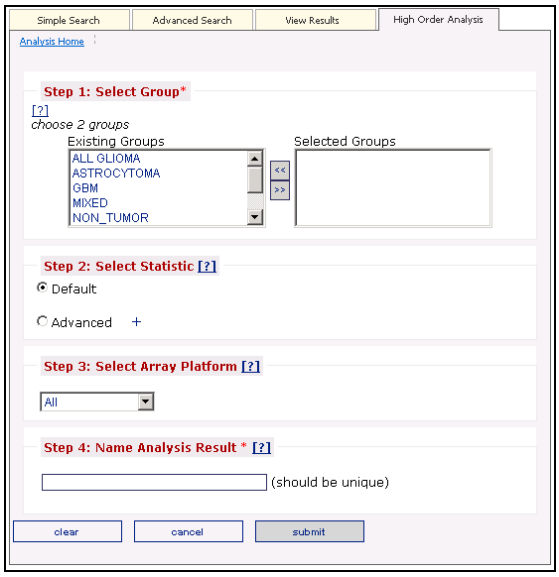


Figure 5.2 Selecting Class Comparison criteria

3. **Step 1: Select Group (required):** Choose two groups from the **Existing Groups** box and click the right-arrow button to move your selection(s) to the **Selected Groups** box *(Figure 5.3).*

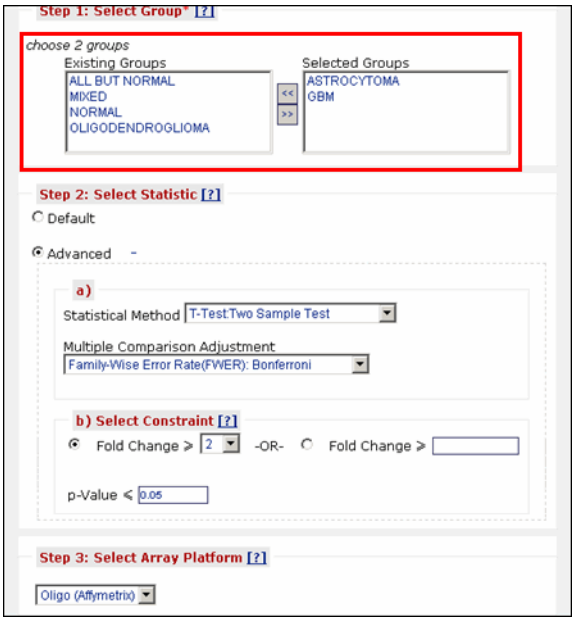


Figure 5.3 Selecting a group

4. **Step 2: Select Statistic (optional):** Leave **Default** as is or select **Advanced** by clicking the pertinent radio button. If you choose **Advanced,** click the "**+**" to access the following options (see *Figure 5.4*):
    a. **Statistical Method:** Choose a test (**t-test:two sample test** or **Wilcoxin Test: Mann-Whitney Test**) from the drop-down list.

b. **Multiple Comparison Adjustment:** Choose an adjustment (**Family-wise Error Rate (FWER): Bonferroni** or **False Discovery Rate (FDR): Benjamini-Hochberg**) from the drop-down list.

c. **Select Constraint:** Select the **Fold Change** (**2, 3, 4, 5, 6, 7, 8, 9** or **10**) from the drop-down list or enter another fold change into the adjacent text box. Enter the ***p*-value** into the text box.

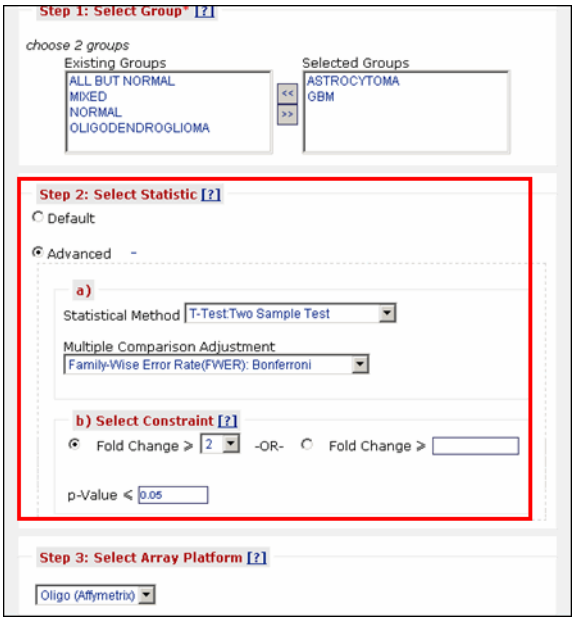


Figure 5.4 Selecting statistics

5. **Step 3: Select Array Platform (optional):** Select a platform (**All, Oligo (Affymetrix)** or **cDNA**) from the drop-down list.
6. **Step 4: Name Analysis Result (required):** Enter a title/name for this analysis into the text box. This name must be unique among all your queries in this session.
7. When you have finished selecting your criteria, click **submit** to submit your criteria and open the report page *(Figure 5.5).*

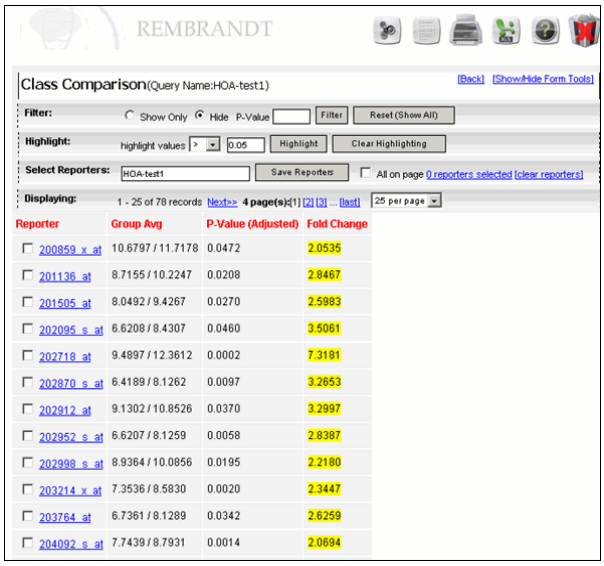


Figure 5.5 Class Comparison report

Figure 5.5 shows a Class Comparison report. This report shows the results from the class comparison performed using the parameters (statistical methods and constraints) set by the user. The report displays group average, where the numerator is the mean of log(base 2) expression signals (geometric mean) from the samples in the first group and the denominator is the mean of log(base 2) expression signals (geometric mean) from the samples in the second group.

Absolute fold change for the reporter between the selected groups is also displayed along with *p*-value. If multiple-comparison adjustment is chosen, then adjusted *p*-value is displayed. Gene symbol annotations are displayed for each reporter. Extensive annotations can be obtained by clicking the **excel download** button on the upper right-hand corner of the report.

Principal Component Analysis (PCA)

To create a High Order Analysis with Principal Component Analysis, follow these steps:

1. Click the **High Order Analysis** tab.
2. Choose **Principal Component Analysis** from the **Analysis Tools** menu.

   Select criteria from the Principal Component Analysis (PCA) Form *(Figure 5.6)* by continuing with the next step.

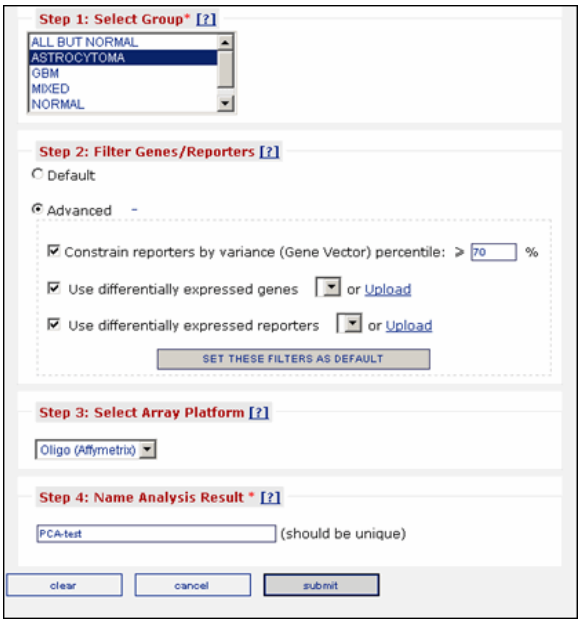


Figure 5.6 Selecting Principal Component Analysis criteria

3. **Step 1: Select Group (required):** Select either **Show all samples** or **Select samples** by clicking the pertinent radio button. If you chose **Select samples,** choose one or more groups from the **Existing Groups** box and click the right-arrow button to move your selection(s) to the **Selected Groups** box.
4. **Step 2: Filter Genes/Reporters (optional):** Select either **Default** or **Advanced** by clicking the pertinent radio button. If you choose **Advanced,** click the "**+**" to access the following options:
    a. **Constrain reporters by variance (Gene Vector) percentile:** The reporters whose variances of the log ratio (or Log2 signals) across all experiments were among the top percentile of variance of all reporters were identified. Select the check box and enter the percentage into the text box. If you enter 70% into the text box, it means that you choose reporters with top 30 (100 - 70) percentile of variance of all reporters.
    b. **Use differentially expressed genes:** Choose saved differentially expressed genes identified by class comparison. Choose an option (**gene list1** or **gene list2**) from the drop-down list or click the **Upload** link to upload a file.
    c. **Use differentially expressed reporters:** Choose saved differentially expressed reporters identified by class comparison. Choose an option (**reporter list1** or **reporter list2**) from the drop-down list or click the **Upload** link to upload a file.
    d. If desired, click the **Set These Filters As Default** button.
5. **Step 3: Select Platform: (optional):** Select either **All, Oligo (Affymetrix)** or **cDNA** from the drop-down list.
6. **Step 4: Name Analysis Result (required):** Enter a title/name for this analysis into the text box. This name must be unique among all your queries in this session.

7. When you have finished selecting your criteria, click **submit** to submit your criteria and open the report page *(Figure 5.7)*.

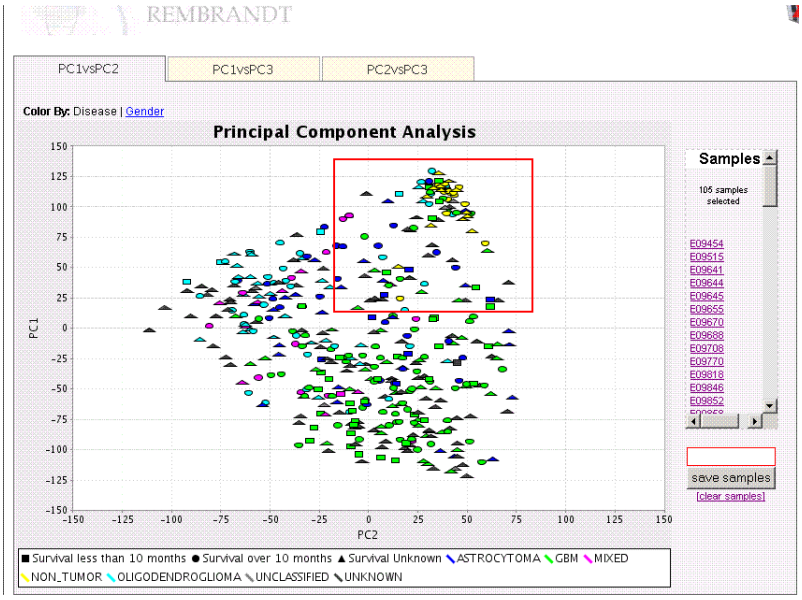


Figure 5.7 Principal Component Analysis report

Figure 5.7 shows the Principal Component Analysis report. This two-dimensional graph plots the various principal components from the analyses. You can click on the three tabs at the top of the graph to display either PC1 vs PC2, or PC1 vs PC3, or PC2 vs PC3. Each point on the graph represents a sample. The samples are colored by disease type. You can color by gender by clicking on the link on the upper left-hand corner of the graph. Patients with different survival ranges are indicated by different shapes on the graph. You can select the samples of interest by clicking on the graph and drawing a rectangle around the samples that you would like to save for future use.

Hierarchical Clustering

To create a High Order Analysis with Hierarchical Clustering, follow these steps:

1. Click the **High Order Analysis** tab.
2. Choose **Hierarchical Clustering** from the **Analysis Tools** menu.

   Select criteria from the Hierarchical Clustering Analysis Form *(Figure 5.8)* by continuing with the next step.

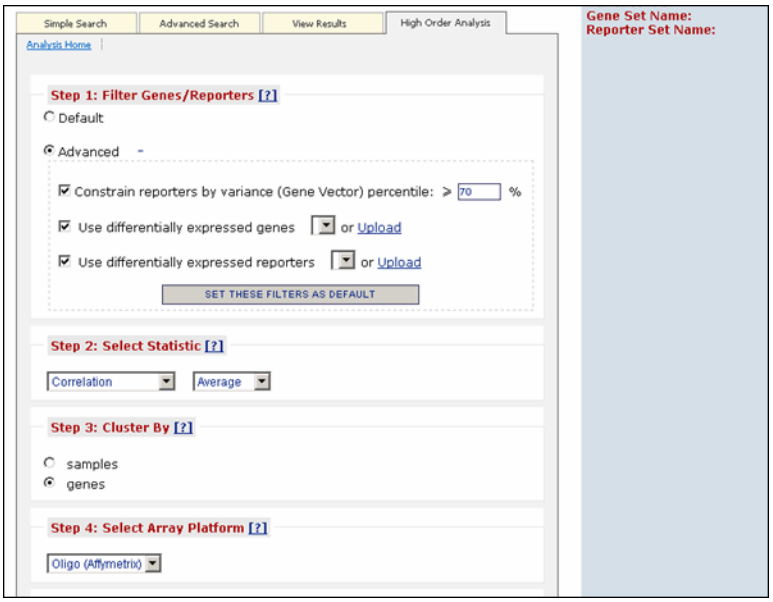
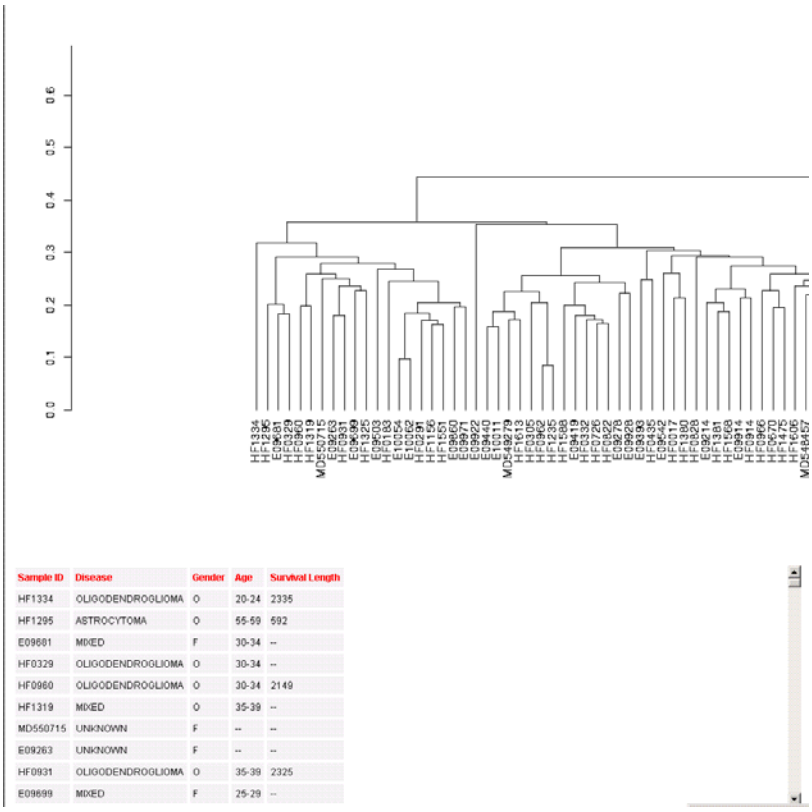


Figure 5.8 Selecting Hierarchical Clustering criteria

3. **Step 1: Filter Genes/Reporters (required):** Select either **Default** or **Advanced** by clicking the pertinent radio button. If you choose **Advanced,** click the "**+**" to access the following options:
    a. **Constrain reporters by variance (Gene Vector) percentile:** The reporters whose variances of the log ratio (or Log2 signals) across all experiments were among the top percentile of variance of all reporters were identified. Select the check box and enter the percentage into the text box. If you enter 70% into the text box, it means that you choose reporters with top 30 (100 - 70) percentile of variance of all reporters.
    b. **Use differentially expressed genes:** Choose saved differentially expressed genes identified by class comparison. Choose an option (**gene list1** or **gene list2**) from the drop-down list or click the **Upload** link to upload a file.
    c. Use differentially expressed reporters: Choose saved differentially expressed reporters identified by class comparison. Choose an option (**reporter list1** or **reporter list2**) from the drop-down list or click the **Upload** link to upload a file.
    d. If desired, click the **Set These Filters As Default** button.
4. **Step 2: Select Statistic (optional):** Select the following options.
    a. **Distance Matrix:** Select an option (**Correlation** or **Euclidean distance**) from the drop-down list. Pearson correlation measures the relative shape of the gene regulations rather than the absolute levels. This is a natural choice because it is widely used to measure gene correlations. Euclidean distance is the most common distance measure. It measures the absolute level of gene regulation.
    b. **Linkage Method:** Select an option (**Average, Single** or **Complete**) from the drop-down list. Different linkage methods affect the shape of the resulting clusters. **Average linkage:** The linking distance is the aver-

age of all pair-wise distances between members of the two clusters.Selecting Hierarchical Clustering criteria. **Single linkage:** The linking distance is the minimum distance between two clusters. **Complete linkage:** The linking distance is the maximum distance between two clusters.

5. **Step 3: Cluster by:** Leave the default to cluster on samples or change the radio button to cluster by genes.
6. **Step 4: Select Platform (optional):** Select either **All, Oligo (Affymetrix)** or **cDNA** from the drop-down list.
7. **Step 5: Name Analysis Result (required):** Enter a title/name for this analysis into the text box. This name must be unique among all your queries in this session.
8. When you have finished selecting your criteria, click **submit** to submit your criteria and open the report page *(Figure 5.9).*



Sample ID	Disease	Gender	Age	Survival Length
HF1334	OLIGODENDROGLIOMA	O	20-24	2335
HF1295	ASTROCYTOMA	O	55-59	592
E09681	MIXED	F	30-34	--
HF0329	OLIGODENDROGLIOMA	O	30-34	--
HF0960	OLIGODENDROGLIOMA	O	30-34	2149
HF1319	MIXED	O	35-39	--
MD550715	UNKNOWN	F	--	--
E09263	UNKNOWN	F	--	--
HF0931	OLIGODENDROGLIOMA	O	35-39	2325
E09699	MIXED	F	25-29	--

Figure 5.9 Hierarchical Clustering report

Figure 5.9 displays the dendrogram from hierarchical clustering analysis. Clicking on **full size** at the top left-hand corner of the graph displays the image at full resolution. Based on the cluster parameter selected by the user, either sample or reporter annotations are displayed beneath the dendrogram.

GLOSSARY

Acronyms and other terms referred to in the chapters of this User's Guide are described in this glossary.

Term	*Definition*
allele	Mutually exclusive alternative forms of the same gene occupying the same locus on homologous chromosomes, differing in DNA sequence and governing the same biochemical and developmental process.
anaplastic	Cancer cells that divide rapidly and have little or no resemblance to normal cells.
Astrocytic tumors	Neoplasms of the brain and spinal cord derived from glial cells.
Benjamini-Hochberg Multiple Testing Correction	The concept of False Discovery Rate (FDR) was introduced in multiple testing by Benjamini and Hochberg (1995).
CCR	Center of Cancer Research
CCR-NOB	CCR Neuro-Oncology Branch
CGAP	Cancer Genome Anatomy Project
Class Comparison	Differential gene expression across the tumor types will be evaluated by calculating the typical *t*-statistic for each reporter. Both parametric and non-parametric *p*-value will be computed.
ependymoma	A type of brain tumor that may arise in the ventricles of the brain or in the spinal cord. Also called an ependymal tumor.
False Discovery Rate (FDR)	The expected proportion of Type I errors among rejected hypotheses in simultaneous testing of multiple null hypotheses.
Family-wise Error Rate (FWER)	Denotes the probability of having at least one false significant test result within the set of tested hypotheses.

Term	Definition
fibrillary astrocytoma	Most frequent histological variant of Diffuse Astrocytoma; predominantly composed of fibrillary neoplastic astrocytes.
gemistocytic astrocytoma	Rare variant of Diffuse Astrocytoma. It is characterized by the presence of a conspicuous, though variable, fraction of gemistocytic neoplastic astrocytes.
Gene Ontology (GO) Classification	The Gene Ontology (GO) project is a collaborative effort to address the need for consistent descriptions of gene products in different databases. The goal of the Gene Ontology project is to produce a controlled vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing. GO provides three structured networks of defined terms, molecular function, biological process, and cellular component, to describe gene product attributes. (from http://www.geneontology.org)
Glioblastoma	Malignant form of astrocytoma histologically characterized by pleomorphism of cells, nuclear atypia, microhemorrhage, and necrosis.
Gliomas	Any of the largest group of primary tumors of the brain, composed of malignant glial cells. Kinds of gliomas are astrocytoma, ependymoma, glioblastoma multiforme, medulloblastoma, and oligodendroglioma.
Hierarchical Clustering	Hierarchical cluster analysis is a statistical method for finding relatively homogeneous clusters of cases based on measured characteristics. It starts with each case in a separate cluster and then combines the clusters sequentially, reducing the number of clusters at each step until only one cluster is left.
High Order Analysis	After data preprocessing (filtering and normalization), further statistical analysis of gene expression data are performed, including class comparison, class discovery and class prediction.
HUGO [gene symbol] [Human Genome Organisation]	HUGO is an international organization of scientists involved in human genetics. Established in 1989 by a collection of the world's leading human geneticists, the primary ethos of the Human Genome Organisation is to promote and sustain international collaboration in the field of human genetics.

Term	*Definition*
Kaplan-Maier	The Kaplan Maier method is used for survival analysis. Kaplan-Maier curves are used to estimate survival probability as a function of time, and survival differences are analyzed by the log-rank test.
Karnofsky Performance Status	A standard way of measuring the ability of cancer patients to perform ordinary tasks. The scores range from 0 to 100, with a higher score indicating a better ability to carry out daily activities. KPS may be used to determine a patient's prognosis, to measure changes in functioning, or to decide if a patient could be included in a clinical trial.
Lansky Play-Performance Status	The play-performance scale for children is a parent-rated instrument which records usual play activity as the index of performance. It is similar to the Karnofsky Performance Scale for adults.
Mann-Whitney Test	A nonparametric test (distribution-free) used to compare two independent groups of sampled data. Unlike the parametric *t*-test, this non-parametric makes no assumptions about the distribution of the data (e.g., normality).
Multiple Comparison Adjustment	Since tens of thousands of genes are compared, many genes can be false positives. However, genes are not all independent and genes in the same pathway could have similar *t*-statistics or *p*-values. Multiple-comparison adjusted *p*-values have been proposed to handle the multiple comparison issues in the context of microarray data.
myxopapillary ependymoma	Slow growing gliomas which generally occur in young adults and arise almost exclusively in the conus-cauda-filum terminale region. It generally has a favorable prognosis and is characterized histologically by tumor cells arranged in a papillary manner around vascularized mucoid stromal cores.
NCI	National Cancer Institute
NCICB	National Cancer Institute Center for Bioinformatics
NINDS	National Institute of Neurological Disorders and Stroke
Oligodendroglial Tumor; Oligodendroglioma	Rare, slow growing tumor that begins in the oligodendrocytes (brain cells that provide support and nourishment for nerve cells). Also called an oligodendroglioma.

Term	*Definition*
pleomorphic xanthoastrocytoma	Astrocytic tumor with a relatively favorable prognosis (WHO grade II) and is typically encountered in children and young adults. It has a superficial location in the cerebral hemispheres and involvement of the meninges.
Principal Component Analysis	PCA is commonly used in microarray research as a cluster analysis tool. It is designed to capture the variance in a dataset in terms of principle components. In effect, one is trying to reduce the dimensionality of the data to summarize the most important (i.e. defining) parts while simultaneously filtering out noise.
protoplasmic [astrocytoma]	Rare variant of Diffuse Astrocytoma. It is predominantly composed of neoplastic astrocytes showing a small cell body with few, flaccid processes with a low content of glial filaments and scant GFAP expression.
SNP	Single nucleotide polymorphisms or SNPs (pronounced "snips") are DNA sequence variations that occur when a single nucleotide (A,T,C or G) in the genome sequence is altered.
subependymal giant cell astrocytoma;	Benign, slowly growing tumor (WHO grade I) typically arising in the wall of the lateral ventricles and composed of large ganglioid astrocytes. It is the most common CNS neoplasm in patients with Tuberous Sclerosis Complex and typically occurs during the first two decades of life.
Wilcoxin Test	Nonparametric statistics for testing hypotheses about whether two samples differ.

INDEX