

CHAPTER 5

ANALYZING STUDIES

This chapter describes how to use calIntegrator2 tools to analyze data in clinical or genomic studies that have been deployed in calIntegrator2.

Topics in this chapter include the following:

- [Data Analysis Overview](#) on this page
- [Creating Kaplan-Meier Plots](#) on page 54
- [Creating Gene Expression Plots](#) on page 60
- [Analyzing Data with GenePattern](#) on page 73

Data Analysis Overview

Once a study has been deployed, you can analyze the data using calIntegrator2 analysis tools.

You can verify that the study is in “Deployed” status by selecting the study name in the My Studies dropdown selector. After selecting the study name, click **Home** in the left sidebar of the calIntegrator2 Menu. A study summary should appear, including a status field. If the status is not deployed, or if the study summary does not appear, then the study is not deployed and available for analysis.

If the study is ready for analysis, you will see an **Analysis Tools** menu in the left sidebar with the following options:

- **K-M Plot:** This tool analyzes clinical data, generating a Kaplan-Meier (K-M) plot based on survival data sets. See [Creating Kaplan-Meier Plots](#) on page 54.
- **Gene Expression Plot:** This tool analyzes annotation, clinical or genomic data based on gene expression values. See [Creating Gene Expression Plots](#) on page 60.

- **GenePattern:** This feature provides an express link to GenePattern where you can perform analyses on selected calIntegrator2 studies, or it enables you to perform several GenePattern analyses on the grid. See [Analyzing Data with GenePattern](#) on page 73 .

After defining or running the analysis on selected data sets, analysis results display on the same page, allowing you to review the analysis method parameters you defined.

Creating Kaplan-Meier Plots

The Kaplan-Maier method analyzes comparative groups of patients or samples. In calIntegrator2, the K-M method compares survival statistics among comparative groups. You can configure the survival data in the application. For example, you might identify a group of patients with smoking history and compare survival rates with a group of non-smoking patients, or compare the survival data for two groups of patients with a specific disease type and based on Karnofsky scores . You could compare groups of patients with varying gene expression levels. You can also identify data sets using the query feature in the application, saving the queries, then configuring the K-M to compare groups identified by the queries.

The key is to first identify subsets of patients or samples that meet criteria you want to establish, thus filtering the data you want to compare. Next, generate a K-M plot based on their survival probability as a function of time. Survival differences are analyzed by the log-rank test.

Note: To perform a K-M plot analysis, survival data must have been identified for the study you want to analyze. For more information, see [Defining Survival Values](#) on page 23.

K-M Plot for Annotations

The groups identified for this K-M plot generation are based on clinical annotations.

1. Select the study whose data you want to analyze in the upper right portion of the calIntegrator2 page.
2. Under Analysis Tools on the left sidebar, select **K-M Plot**.
3. Select the **For Annotation** tab at the top of the page ([Figure 5.1](#)).

Kaplan-Meier Survival Plots (draft)

For Annotation For Gene Expression For Queries

Annotation Based Kaplan-Meier Survival Plots

	Annotation Type	Annotation	Values
1.) Patient Groups:	Select Annotation Type	Select Annotation	
Survival Value			
2.) Select Survival Measure:	Survival from enrollment		
Reset			

Figure 5.1 Fields for defining annotation data for a K-M plot

4. The groups to be compared in the K-M plot originate from one patient group. Varying data sets are based upon multiple values corresponding to the selected annotation. Define Patient Groups using these options:
 - **Annotation Type** – Select the annotation type that identifies the patient group. Selections are based on the data in the chosen study.
 - **Annotation** – Select an annotation. Fields are based on the annotation type you select. For example, if you choose **Subject**, then you could select **Gender** or **Radiation Type** or any field that would distinguish the patients into groups based upon their values.
 - **Values** – Using conventional selection techniques, select two or more values which will be the basis for the K-M plot. Permissible (available) values or “No Values” correspond to the selected annotation.
5. **Survival value** is the length of time the patient lived. For **Survival Value**, select the survival measure which is the unit of measurement for the survival value to be used for the plot.
6. Click the **Create Plot** button.

Note: The Create Plot button displays only after you have selected appropriate criteria.

calIntegrator2 generates the plot which then displays below the plot criteria ([Figure 5.2](#)).

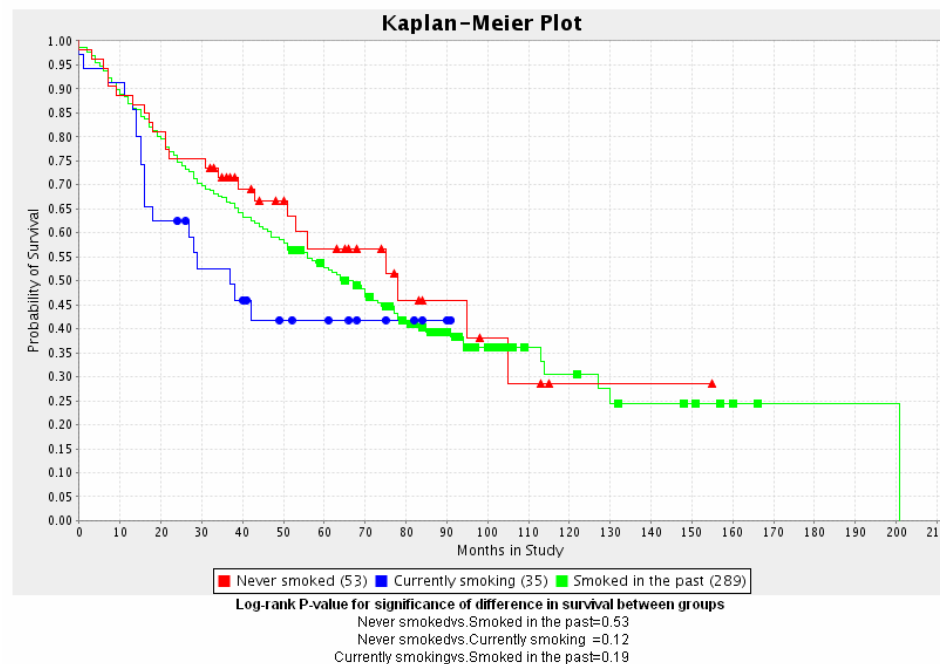


Figure 5.2 A K-M plot generated for groups based on clinical annotations

The number of subjects for each group appears embedded in the legend of the graph below the plot.

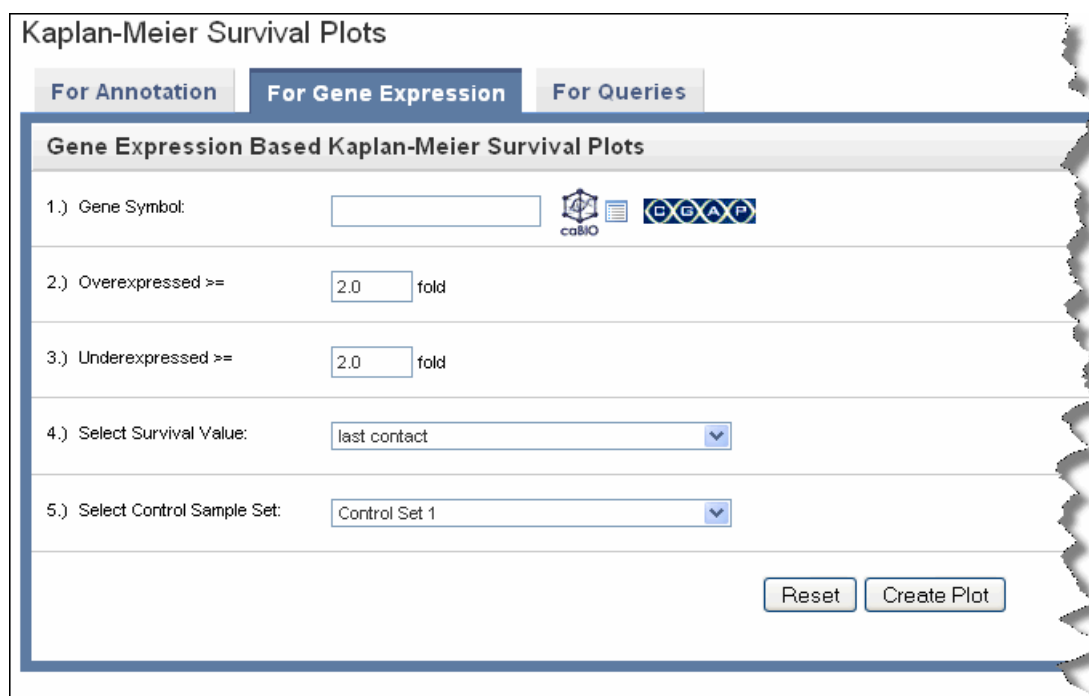
calIntegrator2 generates a P-value for the selected groups; it displays at the bottom of the page. A low P-value generally has more significance than a high P-value.

K-M Plot for Gene Expression

calIntegrator2 allows you to compare expression levels for one given gene at a time. The relative expression level is referred to as “fold change” and the numeric value for a given sample and reporter combination is the ratio of the expression value for that particular reporter for the given sample to a reference value calculated for that reporter across all control samples. The reference value is calculated by taking the mean of the \log_2 of the expression values for all control samples for the reporter in question. The \log_2 mean value (n) is then converted back to a comparable expression signal by returning 2 to the exponent n .

To create a K-M plot illustrating gene expression values, follow these steps:



1. Select the study whose data you want to analyze in the upper right portion of the calIntegrator page. You must select a study with gene expression data.
2. Under Analysis Tools on the left sidebar, select **K-M Plot**.
3. Select the **For Gene Expression** tab ([Figure 5.3](#)).



Kaplan-Meier Survival Plots

For Annotation For Gene Expression For Queries

Gene Expression Based Kaplan-Meier Survival Plots

1.) Gene Symbol:  

2.) Overexpressed >= fold

3.) Underexpressed >= fold


4.) Select Survival Value:

5.) Select Control Sample Set:

Figure 5.3 Fields for defining gene expression data for a K-M plot

4. For **Gene Symbol**, enter one or more gene symbols in the text box or click the icons to locate genes in the following databases. If you enter more than one gene in the text box, separate the entries by commas.

calIntegrator2 provides three methods whereby you can obtain gene names for calculating a KM plot:

- **caBio** – This link searches caBIO, then pulls identified genes into calIntegrator2 for analysis. Click the **caBIO** icon ().
 - a. Enter **Search Terms**.

- b. Select if you want to search in **Gene Keywords**, **Gene Symbols** or **Pathways** (from the drop-down list). Selecting **Gene Keywords** searches only the Full Name field in caBio. Selecting **Gene Symbols** searches only the Unigene and HUGO gene symbols in caBio. Selecting **Pathways** searches only the pathway names in caBio. Note that searching in Pathways is a two step process. First, the initial Pathway search produces search results which are pathways. Second, from the pathway search results screen, you must select pathways of interest, then click **Search Pathways for Genes** to obtain a list of genes related to the selected pathways.
- c. Select the **Any** or **All** choice to determine how your search terms will be matched. **Any** finds any match for any search term you entered. **All** finds only results that match all of the search terms.
- d. Choose the **Taxon** from the drop-down list and click **Search**. The search results display in the same dialog box (*Figure 5.4*).

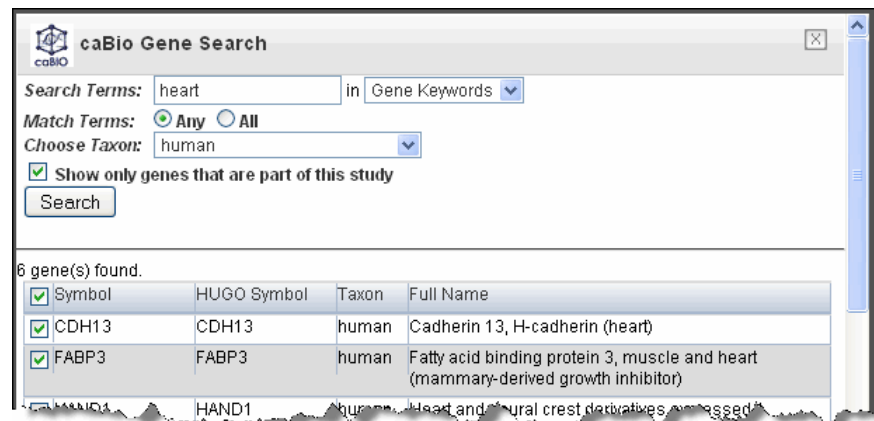



Figure 5.4 Example caBIO gene search criteria and results

- e. In the search results, use the check boxes to identify the genes whose symbols you want to use in the plot calculation.
- f. Click **Use Genes** at the bottom of the page. This pulls the checked genes into the For Gene Expression tab (*Figure 5.5*).



Figure 5.5 Genes pulled in from caBIO display on the selected tab

- **Genes List** – This link locates genes lists saved in caIntegrator2.
 - a. Click the Genes List icon () to open a small dialog that lists prior-saved gene lists in caIntegrator2.
 - b. In the drop-down menu, select a gene list. In the list that appears, use the check boxes to identify the genes whose symbols you want to use in the plot analysis.

- c. Click **Use Genes** at the bottom of the dialog. This pulls the checked genes into the For Gene Expression tab.
- **CGAP** – Use this directory to identify genes. Before clicking this link you must enter gene symbols in the text box. This link does not pull anything into calIntegrator2 but does provide information about the gene(s) whose names you entered.
5. **Over-expressed/Under-expressed** – Define the over- and under-expression criteria, expressed in terms of fold-change. Fold change is the ratio of the measured gene expression value for an experimental sample to the expression value for the control sample.
6. **Survival value** – The length of time the patient lived. For **Survival Value**, select the survival measure which is the unit of measurement for the survival value to be used for the plot.
7. **Control Sample Sets** – One or more are created by the study manager when a study is deployed. Select the **Control Sample Set** you would like to use to calculate fold-change.
8. Click the **Create Plot** button. calIntegrator2 generates the plot which then displays below the plot criteria (*Figure 5.6*).

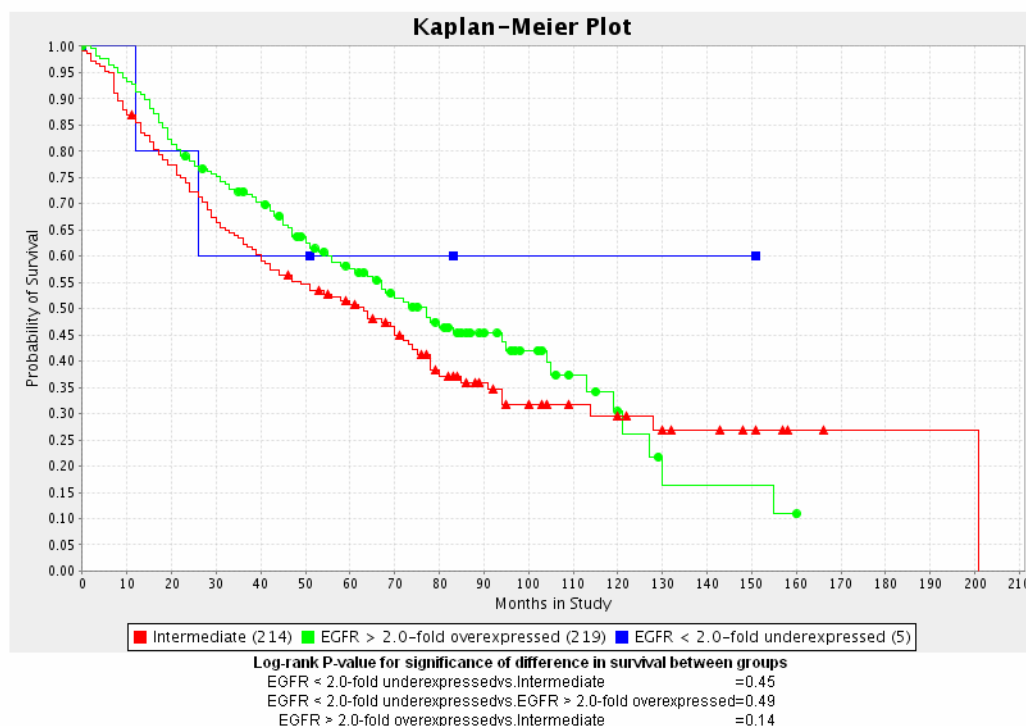


Figure 5.6 K-M plot generated from gene expression data.

The number of subjects for each group appears embedded in the legend of the graph below the plot. Note the appearance of an intermediate group (red entries), which is a group with gene expression values that are not up-regulated nor down-regulated.

In queries that include a fold change criterion and that are configured to return genomic data, the raw expression values are replaced with the calculated fold change value.

A P-value is also generated for the selected groups; it displays at the bottom of the page. A low P-value generally has more significance than a high P-value.

K-M Plot for Queries

You can identify data sets using the query feature in the application. You can manipulate the queries to find the groups you want to compare, save the queries, then configure the K-M to compare the query groups. This is one method of limiting the data considered in the K-M plot calculation.

1. Select the study whose data you want to analyze in the upper right portion of the caIntegrator page. You must select a study for which the queries you will identify for the K-M plot have been saved.
2. Under Analysis Tools on the left sidebar, select **K-M Plot**.
3. Select the **For Queries** tab ([Figure 5.7](#)).

Kaplan-Meier Survival Plots (draft)

For Annotation For Gene Expression **For Queries**

Query Based Kaplan-Meier Survival Plots

1.) Select Queries:

All Available Queries

gender female
equal to or > 60
equal to or > 70
never smoke
equal to or > 40
equal to or > 30

Add >

< Remove

Selected Queries

2.) ☐ Exclusive Subjects in Queries (Subjects in upper queries are removed from subsequent queries)

3.) ☐ Add additional group containing all other subjects not found in selected queries.

4.) Select Survival Value: Survival from enrollment

Reset Create Plot

Figure 5.7 Fields for defining K-M plot parameters based on saved queries in caIntegrator2

4. **Queries** – Select **Queries** whose data you want to analyze from the **All Available Queries** panel and move them to the **Selected Queries** panel using the **Add >>** button.

Note: Genomic queries do not appear in the lists; they cannot be selected for this type of K-M plot.

5. **Exclusive Subject in Queries** – Check the box if you want to exclude any subjects that appear in both (or all) queries selected for the plot, thus eliminating overlap.
6. **Add Additional Group...all other subjects** – Check the box to create an additional group of all other subjects that are not in selected query groups.

7. **Survival value** – The length of time the patient lived. Select the survival measure which is the unit of measurement for the survival value to be used for the plot.
8. Click the **Create Plot** button. calIntegrator2 generates the plot which then displays below the plot criteria ([Figure 5.8](#)).

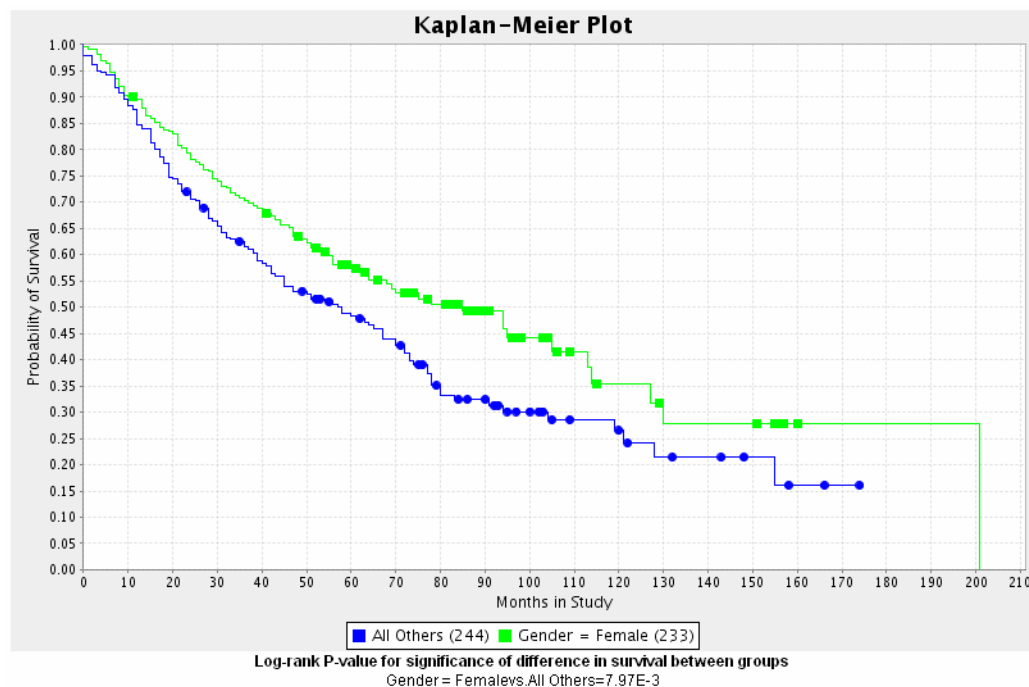


Figure 5.8 K-M Plot comparing statistics between subjects in two queries

The number of subjects for each group appears embedded in the legend of the graph below the plot.

A P-value is also generated for the selected groups; it displays at the bottom of the page. A low P-value generally has more significance than a high P-value.

Creating Gene Expression Plots

Gene expression plots compare signal values from reporters or genes. This statistical tool allows you to compare values for multiple genes at a time, but it does not require only two sets of data to be compared. It also allows you to compare expression levels for selected genes against expression levels for a set of control samples designated at the time of study definition.

calIntegrator2 provides three ways to generate meaningful gene expression plots, indicated by tabs on the page. The tabs are independent of each other and allow you to select the genes, reporters and sample groups to be analyzed on the plot.

- [Gene Expression Value Plot for Annotation](#) – You can locate genes in the caBio directories or calIntegrator 2 Gene Lists. You can learn more about the genes in

the CGAP directory. You can define criteria for the plot using clinical and image annotations.

- [Gene Expression Value Plot for Genomic Queries](#) – You can select data based on saved genomic queries.
- [Gene Expression Value Plot for Clinical Queries](#) – You can select data based on saved clinical queries. You can locate genes in the caBio directories or caIntegrator 2 Gene Lists.

See also [Understanding a Gene Expression Plot](#) on page 70.

Gene Expression Value Plot for Annotation

To generate a gene expression plot, follow these steps:

1. Select the study whose data you want to analyze in the upper right portion of the caIntegrator2 page. (You must select a study which has genomic data.)
2. Under Analysis Tools on the left sidebar, select **Gene Expression Plot**. This opens a page with three tabs
3. Select the **For Annotation** tab ([Figure 5.9](#)).


The screenshot shows the 'Gene Expression Value Plots' interface with the 'For Annotation' tab selected. The interface includes a title bar, three tabs ('For Annotation', 'For Genomic Queries', 'For Clinical Queries'), and a main content area titled 'Annotation Based Gene Expression Plots'. The content area contains five numbered steps for configuring the plot: 1. Gene Symbol(s) (comma separated list) with a text box containing 'DLEC1,PLUNC,KLF2,MYCL' and icons for caBio, CGAP, and another database; 2. Select Reporter Type with radio buttons for 'Reporter Id' (selected) and 'Gene'; 3. Sample Groups with a table for selecting annotation types and values; 4. Add additional group containing all other subjects not found in selected queries; 5. Add additional group containing all control samples for this study with a dropdown for 'Control Set 1'. A 'Reset' button is at the bottom right.

Annotation Type	Annotation	Values
Select Annotation Type	Select Annotation	

Figure 5.9 Gene expression value tab for configuring gene expression annotation value plot

4. **Gene Symbol** – Enter one or more gene symbols in the text box or click the icons to locate genes in the following databases. If you enter more than one gene in the text box, separate the entries by commas.

caIntegrator2 provides three methods whereby you can obtain gene names for calculating a gene expression plot:

- **caBio** – This link searches caBIO, then pulls identified genes into caIntegrator2 for analysis. Click the **caBIO** icon ().

- a. Enter **Search Terms**.

- b. Select if you want to search in **Gene Keywords**, **Gene Symbols** or **Pathways** (from the drop-down list). Selecting **Gene Keywords** searches only the Full Name field in caBio. Selecting **Gene Symbols** searches only the Unigene and HUGO gene symbols in caBio. Selecting **Pathways** searches only the pathway names in caBio. Note that searching in Pathways is a two step process. First, the initial Pathway search produces search results which are pathways. Second, from the pathway search results screen, you must select pathways of interest, then click **Search Pathways for Genes** to obtain a list of genes related to the selected pathways.
- c. Select the **Any** or **All** choice to determine how your search terms will be matched. **Any** finds any match for any search term you entered. **All** finds only results that match all of the search terms.
- d. Choose the **Taxon** from the drop-down list and click **Search**. The search results display in the same dialog box ([Figure 5.4](#)).

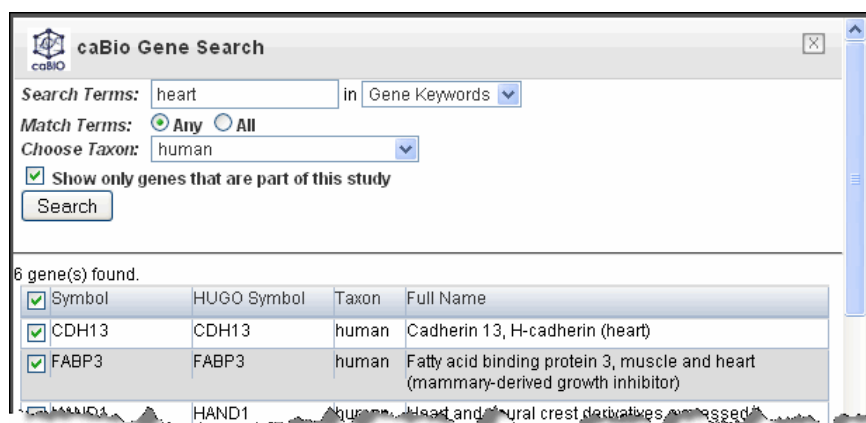



Figure 5.10 Example caBIO gene search results

- e. In the search results, use the check boxes to identify the genes whose symbols you want to use in the gene expression plot calculation.
- f. Click **Use Genes** at the bottom of the page. This pulls the checked genes into the For Annotation tab ([Figure 5.5](#)).



Figure 5.11 Genes pulled in from caBIO display on the tab

- **Genes List** – This link locates genes lists saved in caIntegrator2.
 - a. Click the Genes List icon () to open a small dialog that lists prior-saved gene lists in caIntegrator2.
 - b. In the drop-down menu, select a gene list. In the list that appears, use the check boxes to identify the genes whose symbols you want to use in the plot analysis.

- c. Click **Use Genes** at the bottom of the dialog. This pulls the checked genes into the For Annotation tab.
- **CGAP** – Use this directory to identify genes. Before clicking this link you must enter gene symbols in the text box. This link does not pull anything into calIntegrator2 but does provide information about the gene(s) whose names you entered.
5. **Reporter Type** – Select the radio button that describes the reporter type:
 - **Reporter ID** – Summarizes expression levels for all reporters you specify.
 - **Gene Name** – Summarizes expression levels at the gene level.
6. **Sample Groups** – Choose among the following options:
 - **Annotation Type** – Select the annotation type. Selections are based on the data in the chosen study
 - **Annotation** – Select an annotation. Fields are based on the annotation type you select. For example, if you choose Subject, then you could select Gender or Radiation Type or any field that would distinguish the patients into groups based upon study values.
 - **Values** – Using conventional selection techniques, select one or more values which will be the basis for the plot. Permissible (available) values or “No Values” correspond to the selected annotation.
 - **Add Additional Group...** – Define as follows:
 - **...all other subjects** – Check the box to create an additional group of all other subjects that are not in selected query groups.
 - **...control group** – Check the box to display an additional group of control samples for this study.
7. Click the **Create Plot** button. calIntegrator2 generates the plot which then displays below the plot criteria in bar graph format ([Figure 5.6](#)).

Legends below the plot indicate the plot input. By default, the plot shows the mean of the data. [Figure 5.12](#) displays a plot with gene expression median calculation summaries.

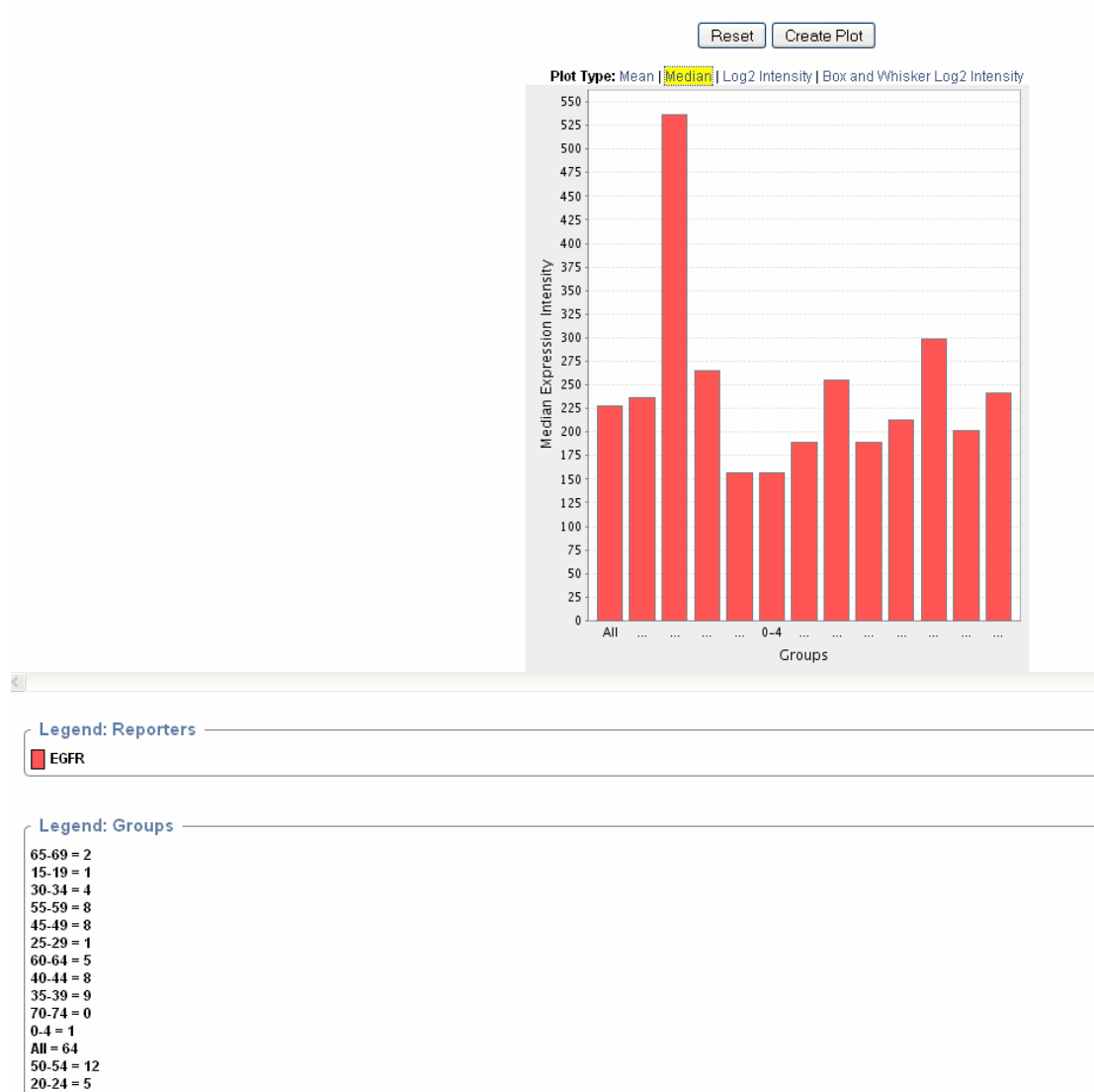


Figure 5.12 Gene expression plot based on selected annotations

- You can recalculate the data display by clicking the **Plot Type** above the graph. See *Understanding a Gene Expression Plot* on page 70.
- You can modify the plot parameters and click the **Reset** button to recalculate the plot.

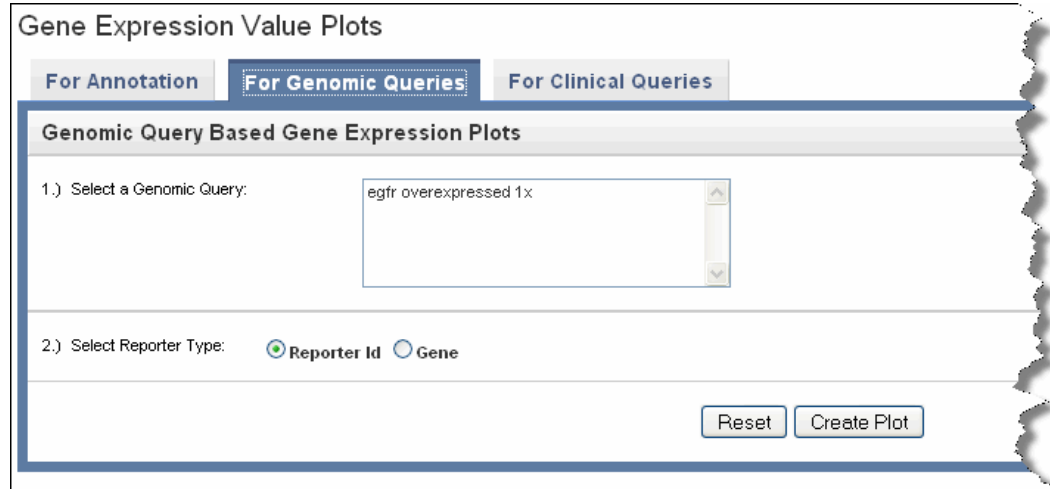
Gene Expression Value Plot for Genomic Queries

Data to be analyzed on this tab must have been saved as a genomic query. For more information, see [Saving a Query](#) on page 43.

To generate a gene expression plot using a genomic query, follow these steps:

1. Select the study whose data you want to analyze in the upper right portion of the calIntegrator page. (You must select a study which has genomic data.)

2. Under Analysis Tools on the left sidebar, select **Gene Expression Plot**.
3. Select the **For Genomic Queries** tab (*Figure 5.13*).



Gene Expression Value Plots

For Annotation For Genomic Queries For Clinical Queries

Genomic Query Based Gene Expression Plots

1.) Select a Genomic Query: egfr overexpressed 1x

2.) Select Reporter Type: ☒ Reporter Id ☐ Gene

Reset Create Plot

Figure 5.13 Gene expression value tab for configuring gene expression genomic queries plot

4. **Genomic Query** – Click on the genomic query upon which the plot is to be based.
5. **Reporter Type** – Select the radio button that describes the reporter type:
 - **Reporter ID** – Summarizes expression levels for all reporters you specify.
 - **Gene Name** – Summarizes expression levels at the gene level..

- Click the **Create Plot** button. calIntegrator2 generates the plot which then displays below the plot criteria. Legends below the plot indicate the plot input ([Figure 5.14](#)).



Figure 5.14 A gene expression plot (Mean) based on a genomic query.

- You can recalculate the data display by clicking the **Plot Type** above the graph. See [Understanding a Gene Expression Plot](#) on page 70.
- You can modify the plot parameters and click the **Reset** button to recalculate the plot.

Gene Expression Value Plot for Clinical Queries

Data to be analyzed on this tab must have been saved as a clinical query, but it must have genomic data identified in the query. For more information, see [Adding/Editing Genomic Data](#) on page 24. For the genomic data, you must identify genes whose expression values are used to calculate the plot.

To generate a gene expression plot using a clinical query, follow these steps:

- Select the study whose data you want to analyze in the upper right portion of the calIntegrator page. You must select a study saved as a clinical study, but which has genomic data.
- Under Analysis Tools on the left sidebar, select **Gene Expression Plot**.


3. Select the **For Clinical Queries** tab (Figure 5.15).

Gene Expression Value Plots

Figure 5.15 Gene expression value tab for configuring gene expression clinical queries plot

4. **Gene Symbol** – Enter one or more gene symbols in the text box or click the icons to locate genes in the following databases. If you enter more than one gene in the text box, separate the entries by commas.

caIntegrator2 provides three methods whereby you can obtain gene names for calculating a gene expression plot:

- **caBio** – This link searches caBIO, then pulls identified genes into caIntegrator2 for analysis. Click the caBIO icon ().

- a. Enter **Search Terms**.
- b. Select if you want to search in **Gene Keywords**, **Gene Symbols** or **Pathways** (from the drop-down list). Selecting **Gene Keywords** searches only the Full Name field in caBio. Selecting **Gene Symbols** searches only the Unigene and HUGO gene symbols in caBio. Selecting **Pathways** searches only the pathway names in caBio. Note that searching in Pathways is a two step process. First, the initial Pathway search produces search results which are pathways. Second, from the pathway search results screen, you must select pathways of interest, then click **Search Pathways for Genes** to obtain a list of genes related to the selected pathways.
- c. Select the **Any** or **All** choice to determine how your search terms will be matched. **Any** finds any match for any search term you entered. **All** finds only results that match all of the search terms.

- d. Choose the **Taxon** from the drop-down list and click **Search**. The search results display in the same dialog box ([Figure 5.4](#)).

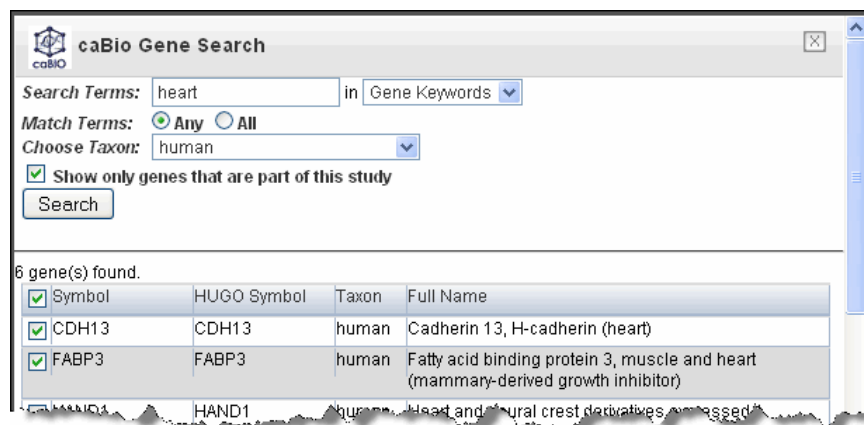



Figure 5.16 Example caBIO gene search results

- e. In the search results, use the check boxes to identify the genes whose symbols you want to use in the plot calculation.
- f. Click **Use Genes** at the bottom of the page. This pulls the checked genes into the tab ([Figure 5.5](#)).



Figure 5.17 Genes pulled in from caBIO display on the tab

- **Genes List** – This link locates genes lists saved in caIntegrator2.
 - a. Click the Genes List icon () to open a small dialog that lists prior-saved gene lists in caIntegrator2.
 - b. In the drop-down menu, select a gene list. In the list that appears, use the check boxes to identify the genes whose symbols you want to use in the gene expression analysis.
 - c. Click **Use Genes** at the bottom of the dialog. This pulls the checked genes into the tab.
 - **CGAP** – Use this directory to identify genes. Before clicking this link you must enter gene symbols in the text box. This link does not pull anything into caIntegrator2 but does provide information about the gene(s) whose names you entered.
5. For **Reporter Type**, select the radio button that describes the reporter type:
 - **Reporter ID** – Summarizes expression levels for all reporters you specify.
 - **Gene Name** – Summarizes expression levels at the gene level.
 6. For **Sample Groups**, choose among the following options:

- **Annotation Type** – Select the annotation type. Selections are based on the data in the chosen study
 - **Annotation** – Select an annotation. Fields are based on the annotation type you select. For example, if you choose Subject, then you could select Gender or Radiation Type or any field that would distinguish the patients into groups based upon study values.
 - **Values** – Using conventional selection techniques, select one or more values which will be the basis for the plot. Permissible (available) values or “No Values” correspond to the selected annotation.
 - For the **Add Additional Group...** options, define as follows:
 - **...all other subjects** – Check the box to create an additional group of all other subjects that are not in selected query groups.
 - **...control group** – Check the box to display an additional group of control samples for this study.
7. Click the **Create Plot** button. caIntegrator2 generates the plot which then displays below the plot criteria in bar graph format (*Figure 5.6*).

By default, caIntegrator2 displays the mean of the data below the plot criteria. Legends below the plot indicate the plot input.

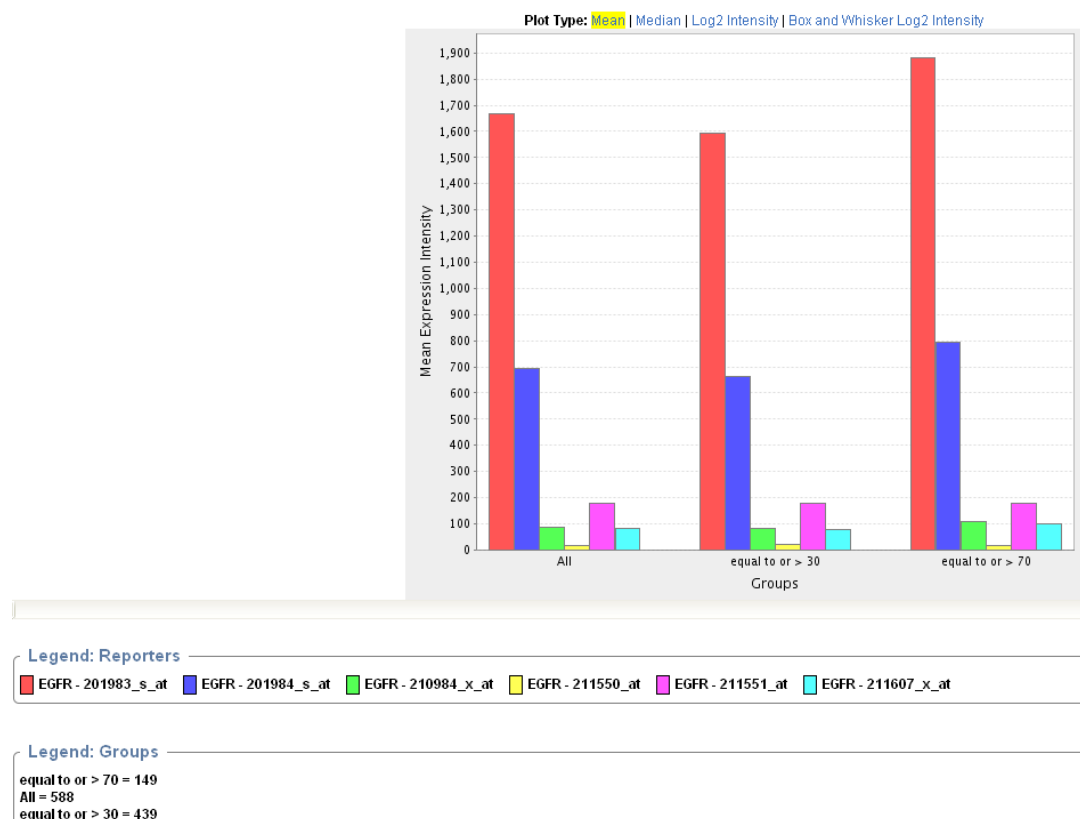


Figure 5.18 Gene expression plot based on clinical queries gene expression values

- You can recalculate the data display by clicking the **Plot Type** above the graph. See [Understanding a Gene Expression Plot](#) on page 70.
- You can modify the plot parameters and click the **Reset** button to recalculate the plot.

Understanding a Gene Expression Plot

Above the plot, you can select various plot types. When you do so, the plot is recalculated. Although all of the plots in this section appear similar, note the differences in calculation results and legends between the Y axis on each of the plots.

When you perform a Gene Expression simple search, by default the **Mean** Gene Expression Plot ([Figure 5.19](#)) appears.

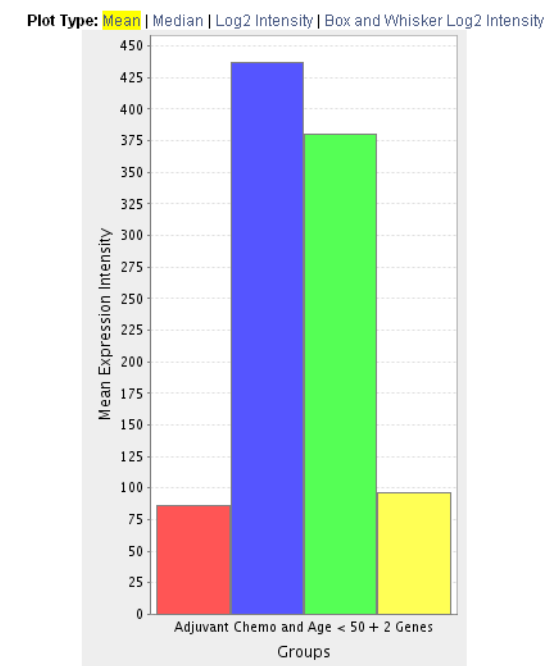


Figure 5.19 Gene expression plot calculating the mean

The **Mean** Gene Expression Plot ([Figure 5.19](#)) displays mean expression intensity (Geometric mean) versus Groups.

The **Median** Gene Expression Plot ([Figure 5.20](#)) displays the median expression intensity versus Groups..

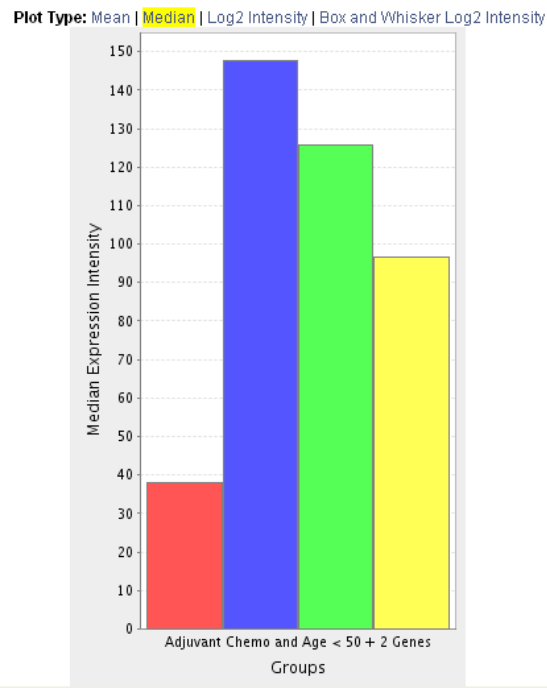


Figure 5.20 Gene expression plot calculating the median

The **Log2 Intensity** Gene Expression Plot ([Figure 5.21](#)) displays average expression intensities for the gene of interest based on Affymetrix GeneChip arrays (U133 Plus 2.0 arrays).

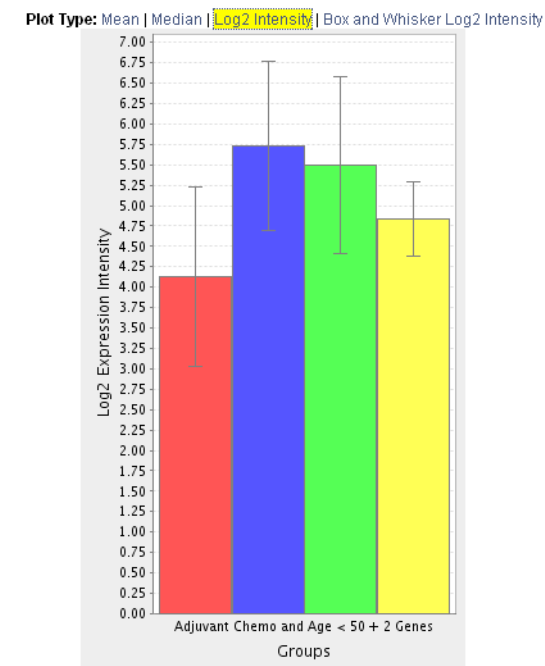


Figure 5.21 Gene expression plot displaying log2 intensity values

The box and whisker log2 expression intensity plot displays a box plot ([Figure 5.22](#), [Figure 5.23](#)). Example uses of box and whisker plots include the following:

- Indicate whether a distribution is skewed and whether there are potential unusual observations (outliers) in the data set.
- Perform a large number of observations.
- Compare two or more data sets.
- Compare distributions because the center, spread, and overall range are immediately apparent.

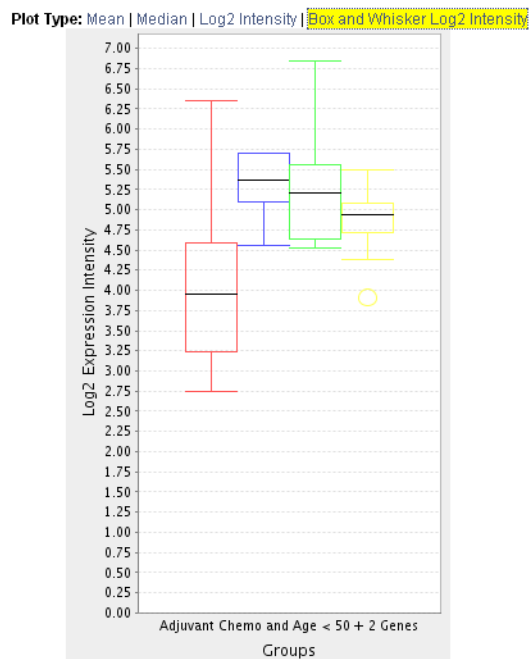


Figure 5.22 Box and whisker plot based on the same data set as represented in [Figure 5.19](#), [Figure 5.20](#), [Figure 5.21](#)

In descriptive statistics, a box plot or boxplot, also known as a box-and-whisker diagram or plot, is a convenient way of graphically depicting groups of numerical data through their five-number summaries (the smallest observation excluding outliers, lower quartile [Q1], median [Q2], upper quartile [Q3], and largest observation excluding outliers).

The box is defined by Q1 and Q3 with a line in the middle for Q2. The interquartile range, or IQR, is defined as Q3-Q1. The lines above and below the box, or 'whiskers', are at the largest and smallest non-outliers. Outliers are defined as values that are

more than $1.5 \times \text{IQR}$ greater than Q3 and less than $1.5 \times \text{IQR}$ than Q1. Outliers, if present, are shown as open circles (*Figure 5.23*).

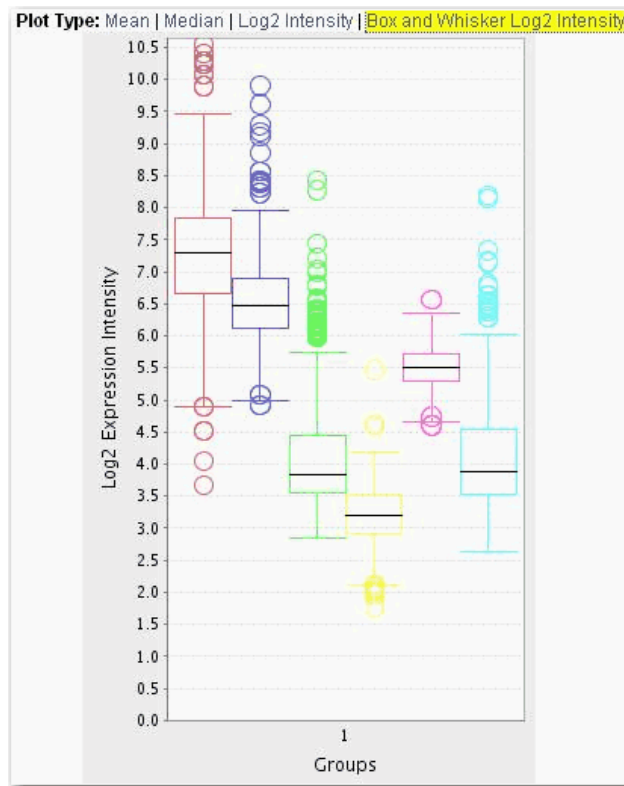


Figure 5.23 Box and whisker plot showing outliers

Boxplots can be useful to display differences between populations without making any assumptions of the underlying statistical distribution: they are non-parametric. The spacings between the different parts of the box help indicate the degree of dispersion (spread) and skewness in the data.

Analyzing Data with GenePattern

GenePattern is an application developed at the Broad Institute that enables researchers to access various methods to analyze genomic data. caIntegrator2 provides an express link to GenePattern where you can analyze data in any caIntegrator2 study.

Information is included in this section for connecting to GenePattern from caIntegrator2. Specifics for launching GenePattern tools from caIntegrator2 are included as well, but you may want to refer to additional GenePattern documentation, available at this website: http://www.broadinstitute.org/cancer/software/genepattern/tutorial/gp_concepts.html.

You have two options for using GenePattern from calIntegrator2:

- Option 1 – Use the web-interface of any available GenePattern instances.
 - a. To use the public instance from Broad, first register for an account at <http://genepattern.broad.mit.edu/gp/pages/login.jsf>
 - b. In calIntegrator2, enter the URL for connecting: <http://genepattern.broad.mit.edu/gp/services/>, then enter your userId and password.
- Option 2 – Use GenePattern on the grid.

The GenePattern feature in calIntegrator2 currently supports three analyses on the grid: Comparative Marker Selection (CMS), Principal Component Analysis (PCA) and GISTIC-supported analysis.

Tip: If you are using the web interface to access GenePattern (option #1 listed above), then you can run other GenePattern tools in addition to CMS, PCA and GISTIC.

1. Select the study whose data you want to analyze in the upper right portion of the calIntegrator2 page.
2. Click **GenePattern Analysis** in the left sidebar of calIntegrator2. This opens the GenePattern Analysis Status page (*Figure 5.24*).

GenePattern Analysis Status

Gene Pattern Modules		New Analysis Job		
Job Name	Job Type	Status	Creation Date	Status
JP - CMS - 2	Comparative Marker Selection	Completed - Download	2009/08/26 21:38:03	2009/08
CMS 1	Comparative Marker Selection	Completed - Download	2009/08/26 11:43:44	2009/08
PCA1	Principal Component Analysis	Completed - Download	2009/08/26 11:38:41	2009/08

Figure 5.24 GenePattern Analysis Status page

3. Select from the drop-down list the type of GenePattern analysis you want to run on the data.
 - **GenePattern Modules** – This option launches a session within GenePattern from which you can launch analyses. See [GenePattern Modules](#) on page 75.
 - **Comparative Marker Selection (Grid Service)**. This option enables you to run this GenePattern analysis on the grid. See [Comparative Marker Selection \(CMS\) Analysis](#) on page 76.
 - **Principal Component Analysis (Grid Service)**. This option enables you to run this GenePattern analysis on the grid. See [Principal Component Analysis \(PCA\)](#) on page 78.
 - **GISTIC (Grid Service)**. This option enables you to run this GenePattern analysis on the grid. See [GISTIC-Supported Analysis](#) on page 81.

- Click the **New Analysis Job** button to open a corresponding page where you can configure the analysis parameters.

GenePattern Modules

Note: To launch the analyses described in this section, you must have a registered GenePattern account. For more information, see <http://genepattern.broad.mit.edu/gp/pages/login.jsf>.

To configure the link for accessing GenePattern from caIntegrator2, open the appropriate page as described in *Analyzing Data with GenePattern* on page 73.

- Select the study whose data you want to analyze in the upper right portion of the caIntegrator2 page.
- Click **GenePattern Analysis** in the left sidebar of caIntegrator2. This opens the GenePattern Analysis Status page.
- Make sure **GenePattern Modules** is selected in the drop down list. Click **New Analysis Job**.
- In the GenePattern Analysis dialog box (*Figure 5.25*), specify connection information, , described *Table 5.1*.

GenePattern Analysis

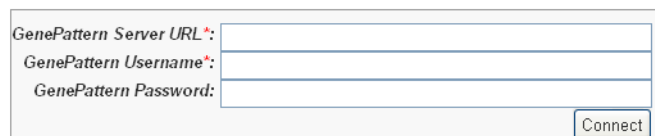


Figure 5.25 Dialog box for configuring the link to GenePattern

Fields	Description
Server URL	Enter any GenePattern publicly available URL, such as http://genepattern.broad.mit.edu/gp/services/Analysis .
GenePattern Username	Enter your GenePattern user name.
GenePattern Password	Enter your GenePattern password.

Table 5.1 Fields for selecting GenePattern configurations

If you choose to access GenePattern in this way, you can continue to use GenePattern tools from within that application. See GenePattern user documentation for more information.

Tip: If you run these analysis within GenePattern itself, you may be able to view results in the GenePattern visualization module. If you run them on the grid from caIntegrator2, your results will be available only in spreadsheet and XML format.

You can run GenePattern analyses for Comparative Marker Selection, Principal Component Analysis and GISTIC-based analysis on the grid if you choose.

Comparative Marker Selection (CMS) Analysis

The Comparative Marker Selection (CMS) module implements several methods to look for expression values that correlate with the differences between classes of samples. Given two classes of samples, CMS finds expression values that correlate with the difference between those two classes. If there are more than two classes, CMS can perform one-vs-all or all-pairs comparisons, depending on which option is chosen.

For more information, see the GenePattern website: http://www.broad.mit.edu/cgi-bin/cancer/software/genepattern/modules/gp_modules.cgi.

To perform a CMS analysis, follow these steps:

1. Select the study whose data you want to analyze in the upper right portion of the calIntegrator2 page. You must select a study saved as a clinical study, but which has genomic data.
2. Click **GenePattern Analysis** in the left sidebar of calIntegrator2. This opens the GenePattern Analysis Status page.
3. In the GenePattern Analysis Status page, select **Comparative Marker Selection (Grid Service)** from the drop down list and click **New Analysis Job**. This opens the Comparative Marker Selection Analysis page (*Figure 5.26*).

Comparative Marker Selection Analysis

The screenshot shows the 'Comparative Marker Selection Analysis' form. It contains the following fields and controls:

- Job Name:** A text input field.
- Preprocess Server:** A dropdown menu showing 'Default Broad service - http://node255.broad.mit.edu:6060/wsrf/services/cagrid/PreprocessDatasetMAGEService'.
- Comparative Server:** A dropdown menu showing 'Default Broad service - http://node255.broadinstitute.org:11010/wsrf/services/cagrid/ComparativeMarkerSelMAGESvc'.
- Clinical Queries:** A section with a text box containing instructions: 'Must select two clinical queries, which are used to group the samples into two separate classifications to run against ComparativeMarkerSelection. The queries selected here have been previously saved by the user. Selected queries will result in the processing of only those samples which are mapped to patients in the saved query result.' Below this is a list of 'All Available Queries' (empty) and a 'Selected Queries' list (empty), with 'Add >' and '< Remove' buttons between them.
- Filter flag:** A checkbox.
- Preprocessing Flag:** A dropdown menu showing 'no-disc-or-norm'.
- Min Change:** A text input field with value '3.0'.
- Min Delta:** A text input field with value '100.0'.
- Threshold:** A text input field with value '20.0'.
- Ceiling:** A text input field with value '2.1'.
- Max Sigma Binning:** A text input field with value '1'.
- Probability Threshold:** A text input field with value '1.0'.
- Num Exclude:** A text input field with value '0'.
- Log Base Two:** A checkbox.
- Number Of Columns Above Threshold:** A text input field with value '1'.
- Test Direction:** A dropdown menu showing 'two-sided'.
- Test Statistic:** A dropdown menu showing 'T-test'.
- Min Std:** A text input field with value '1.0'.
- Number Of Permutations:** A text input field with value '1000'.
- Complete:** A checkbox.
- Balanced:** A checkbox.
- Random Seed:** A text input field with value '779948241'.
- Smooth Pvalues:** A checkbox.
- Phenotype Test:** A dropdown menu showing 'one-versus-all'.
- Perform Analysis:** A button at the bottom right.

Figure 5.26 Comparative Marker Selection analysis parameters

4. Select or define CMS analysis parameters, described in [Table 5.2](#). An asterisk indicates required fields. The default settings are valid; they should provide valid results.

CMS Parameter	Description
Job Name*	Assign a unique name to the analysis you are configuring.
Preprocess Server*	A server which hosts the grid-enabled data GenePattern PreProcess Dataset module. Select one from the list and calIntegrator2 will use the selected server for this portion of the processing.
Comparative Server*	A server which hosts the grid-enabled data GenePattern Comparative Marker Selection module. Select one from the list and calIntegrator2 will use the selected server for this portion of the processing.
Clinical Queries*	All clinical queries with appropriate data for the analysis are listed. Select and move 2 or more queries from the All Available Queries panel to the Selected Queries panel. Note: If a query has a genomic component (e.g. gene criteria), it does not display in the queries field.
Filter Flag	Variation filter and thresholding flag
Preprocessing Flag*	Discretization and normalization flag
Min Change*	Minimum fold change for filter
Min Delta*	Minimum delta for filter
Threshold*	Value for threshold
Ceiling*	Value for ceiling
Max Sigma Binning*	Maximum sigma for binning
Probability Threshold*	Value for uniform probability threshold filter
Num Exclude*	Number of experiments to exclude (max & min) before applying variation filter
Log Base Two	Whether to take the log base two after thresholding
Number of Columns Above Threshold*	Remove row if n columns no \geq than the given threshold
Test Direction*	The test to perform (up-regulated for class0; up-regulated for class1, two sided). By default, Comparative Marker Selection performs the two-sided test.
Test Statistic*	Select the statistic to use.
Min Std*	The minimum standard deviation if test statistic includes the min std option. Used only if test statistic includes the min std option.

Table 5.2 Comparative Marker Selection analysis options

CMS Parameter	Description
Number of Permutations*	<p>The number of permutations to perform. (Use 0 to calculate asymptotic P-values.) The number of permutations you specify depends on the number of hypotheses being tested and the significance level that you want to achieve (3). The greater the number of permutations, the more accurate the P-value.</p> <p>Complete – Perform all possible permutations. By default, complete is set to No and Number of Permutations determines the number of permutations performed. If you have a small number of samples, you might want to perform all possible permutations.</p> <p>Balanced – Perform balanced permutations</p>
Random Seed*	The seed for the random number generator.
Smooth Pvalues	Whether to smooth P-values by using the Laplace's Rule of Succession. By default, Smooth Pvalues is set to Yes , which means P-values are always less than 1.0 and greater than 0.0.
Phenotype Test*	<p>Tests to perform when class membership has more than 2 classes: one versus-all, all pairs.</p> <p>Note: The P-values obtained from the one-versus-all comparison are not fully corrected for multiple hypothesis testing.</p>

Table 5.2 Comparative Marker Selection analysis options

- When you have completed the form, click **Perform Analysis**.

calIntegrator2 takes you to the JobStatus/Launch page where you will see the job and its status in the Status column of the list ([Figure 5.27](#)).

GenePattern Analysis Status (draft)

Gene Pattern Modules


Job Name	Job Type	Status	Creation Date	Status Update Date
Well-diff vs adjuvant chemo	Comparative Marker Selection	 Processing Locally	2009/08/14 11:48:35	2009/08/14 11:48:35
Filter out non-interesting genes	Gene Pattern	Completed - View 122444	2009/08/14 10:16:29	2009/08/14 10:19:47

Figure 5.27 The progress of a GenePattern analysis that has been launched displays in the status column of page

- When the job is complete, the system displays a completion date on the GenePattern Analysis status page. Click the **Download** link. This downloads zipped result files to your local work station. The number of files and their file type will vary according to the processing. The results format is compatible with GenePattern visualizers and can be uploaded within GenePattern.

Principal Component Analysis (PCA)

Principal Component Analysis is typically used to transform a collection of correlated variables into a smaller number of uncorrelated variables, or components. Those components are typically sorted so that the first one captures most of the underlying variability and each succeeding component captures as much of the remaining variability as possible.

You can configure GenePattern grid parameters for preprocessing the dataset in addition to PCA module parameters. For more information, see the GenePattern website: http://www.broad.mit.edu/cgi-bin/cancer/software/genepattern/modules/gp_modules.cgi.

To perform a PCA analysis, follow these steps:

1. Select the study whose data you want to analyze in the upper right portion of the calIntegrator page. You must select a study with gene expression data.
2. Click **GenePattern Analysis** in the left sidebar of calIntegrator2. This opens the GenePattern Analysis Status page.
3. In the GenePattern Analysis Status page, select **Principal Component Analysis (Grid Service)** from the drop down list and click **New Analysis Job**. This opens the Principal Component Analysis page (*Figure 5.28*).

Principal Component Analysis

(draft)

This form submits a job which analyzes samples using the GenePattern Principal Component Analysis module.

Job Name - Please enter a job name.
Principal Component Analysis Server - Select a PCA grid service from the dropdown.
Clinical Queries - Select saved Clinical queries to specify which samples will be processed.
Enable Preprocess Dataset - (Optional) Check this to display and configure preprocessing parameters.

* Job Name:

* Principal Component Analysis Server: Default Broad service - <http://node255.broad.mit.edu:8060/awstf/services/cagrid/PCA>

* Clinical Queries: Clinical Queries enable the user to specify which samples will be processed using PCA. The queries selected here have been previously saved by the user. Selected queries will result in the processing of only those samples which are mapped to subjects in the saved query result. If multiple queries are selected, all of the sample from each saved query are processed PLUS the results set will be classified according to those queries. (One class per selected query.)

All Available Queries

- gender female
- Never smoke

Selected Queries

Add >

< Remove

Enable Preprocess Dataset: ☐
(check to display preprocess parameters)

Perform Analysis

Figure 5.28 Principal Component Analysis parameters

4. Select or define PCA analysis parameters, described in *Table 5.3*. You must enter a job name and select a clinical query, but you can accept the other default settings..

PCA Parameters	Description
Job Name*	Assign a unique name to the analysis you are configuring.
Principal Component Analysis Server*	A server which hosts the grid-enabled data GenePattern Principal Component Analysis module. Select one from the list and calIntegrator2 will use the selected server for this portion of the processing.
Clinical Queries*	All clinical queries display in this list. Select one or more of these queries to define which samples are analyzed using PCA. If you select more than one query, then the union of the samples returned by the multiple queries is analyzed.

Table 5.3 PCA analysis options

PCA Parameters	Description
Cluster By*	Selecting rows looks for principal components across all expression values, and selecting columns looks for principal components across all samples.

Table 5.3 PCA analysis options

- If you want to preprocess the data set, click **Enable the Preprocess Dataset**. This opens an additional set of parameters ([Figure 5.29](#)), discussed in [Table 5.4](#). The preprocessing is executed prior to running the PCA.

Figure 5.29 Parameters for pre-processing parameters for PCA

PCA Preprocessing Parameters	Description
Preprocess Server*	A server which hosts the grid-enabled data GenePattern PreProcess Dataset module. Select one from the list and caIntegrator2 will use the selected server for this portion of the processing.
Filter Flag	Variation filter and thresholding flag
Preprocessing Flag	Discretization and normalization flag
Min Change	Minimum fold change for filter
Min Delta	Minimum delta for filter
Threshold	Value for threshold
Ceiling	Value for ceiling
Max Sigma Binning	Maximum sigma for binning
Probability Threshold	Value for uniform probability threshold filter

Table 5.4 Parameters for preprocessing data sets for PCA

PCA Preprocessing Parameters	Description
Num Exclude	Number of experiments to exclude (max & min) before applying variation filter
Log Base Two	Whether to take the log base two after thresholding
Number of Columns Above Threshold	Remove row if n columns no \geq than the given threshold

Table 5.4 Parameters for preprocessing data sets for PCA

- When you have completed the form, click **Perform Analysis**.
- When the job is complete, the system displays a completion date on the GenePattern Analysis status page. Click the **Download** link. This downloads zipped result files to your local work station. The number of files and their file type will vary according to the processing. The results format is compatible with GenePattern visualizers and can be uploaded within GenePattern.

GISTIC-Supported Analysis

The GISTIC Module is a GenePattern tool that identifies regions of the genome that are significantly amplified or deleted across a set of samples. For more information, see http://www.broad.mit.edu/cgi-bin/cancer/software/genepattern/modules/gp_modules.cgi.

To perform a GISTIC-supported analysis, follow these steps:

- Select the study whose data you want to analyze in the upper right portion of the calIntegrator2 page. You must select a study with copy number (either Affymetrix SNP or Agilent Copy Number) data.
- Click **GenePattern Analysis** in the left sidebar of calIntegrator2. This opens the GenePattern Analysis Status page.

3. In the GenePattern Analysis Status page, select **GISTIC (Grid Service)** from the drop down list and click **New Analysis Job**. This opens the GISTIC Analysis page (*Figure 5.30*).

GISTIC Analysis

This form submits a job which analyzes samples using the GenePattern GISTIC module.

Job Name - Please enter a job name.

GenePattern Server URL / GISTIC Server - Select whether to use the GISTIC web service or grid service and provide or select the service address. If the web service is selected, authentication information is also required.

Clinical Queries - (Optional) Select a saved Clinical query to specify which samples will be processed.

Exclude Sample Control Set - (Optional) Select a Control Sample Set to be excluded from the Clinical Query.

* Job Name:

* GenePattern Server URL:

* GenePattern Username:

GenePattern Password:

* GISTIC Server:

Use GenePattern GISTIC Web Service ☒ Use GISTIC Grid Service ☐

For the Clinical query parameter below, choose either "All Samples" or a clinical query. If "All Samples" is selected, then all samples will be used. If a clinical query is selected, only those samples which map to the subjects in the clinical query results will be used. The clinical queries in this list have been previously saved by the user. Control samples can be excluded from this processing by selecting a control set name in the Exclude Sample Control Set dropdown.

Clinical query:

* Exclude Sample Control Set:

* Amplifications Threshold:

* Deletions Threshold:

* Join Segment Size:

* QV Thresh:

* Remove X:

cnv File:

Figure 5.30 GISTIC analysis criteria

4. Select or define GISTIC analysis parameters, as described in *Table 5.2*. You must indicate a Job Name, but you can accept the other defaults settings, which are valid and should produce valid results.

GISTIC Parameters	Description
Job Name*	Assign a unique name to the analysis you are configuring.
GISTIC Server*	A server which hosts the grid-enabled data GISTIC-based analysis module. Select one from the list and calIntegrator2 will use the selected server for this portion of the processing.
Refgene File*	Enter or select the cytoband file to use in the analysis. Allowed values: {Human Hg18, Human Hg17, Human Hg16}. Default = Human Hg16.
Clinical Query	All clinical queries display in this list as well as an option to select all non-control samples. Select a clinical query if you wish to run GISTIC on a subset of the data and select all non-control samples if wish to include all samples.
Amplifications Threshold*	Threshold for copy number amplifications. Regions with a log2 ratio above this value are considered amplified. Default = 0.1.
Deletions Threshold*	Threshold for copy number deletions. Regions with a log2 ratio below the negative of this value are considered deletions. Default = 0.1.

Table 5.5 GISTIC analysis parameters

GISTIC Parameters	Description
Join Segment Size*	Smallest number of markers to allow in segments from the segmented data. Segments that contain fewer than this number of markers are joined to the neighboring segment that is closest in copy number. Default = 4.
QV Thresh[hold]*	Threshold for q-values. Regions with q-values below this number are considered significant. Default = 0.25.
Remove X*	Flag indicating whether to remove data from the X-chromosome before analysis. Allowed values = {1,0}. Default = 1(yes).
cnv File	<p>This selection is optional.</p> <p>Browse for the file. There are two options for the cnv file.</p> <p>Option #1 enables you to identify CNVs by marker name. Permissible file format is described as follows:</p> <p>A two column, tab-delimited file with an optional header row. The marker names given in this file must match the marker names given in the markers_file. The CNV identifiers are for user use and can be arbitrary. The column headers are:</p> <ol style="list-style-type: none"> 1. Marker Name 2. CNV Identifier <p>Option #2 enables you to identify CNVs by genomic location. Permissible file format is described as follows:</p> <p>A 6 column, tab-delimited file with an optional header row. The 'CNV Identifier', 'Narrow Region Start' and 'Narrow Region End' are for user use and can be arbitrary. The column headers are:</p> <ol style="list-style-type: none"> 1. CNV Identifier 2. Chromosome 3. Narrow Region Start 4. Narrow Region End 5. Wide Region Start 6. Wide Region End

Table 5.5 GISTIC analysis parameters

5. When you have completed the form, click **Perform Analysis**.
6. When the job is complete, the system displays a completion date on the GenePattern Analysis status page. Click the **Download** link. This downloads zipped result files to your local work station. The number of files and their file type will vary according to the processing. The results format is compatible with GenePattern visualizers and can be uploaded within GenePattern.

