

CHAPTER 5

ANALYZING STUDIES

This [chaptersection](#) describes how to use calIntegrator2 tools to analyze data in clinical or genomic studies that have been deployed in calIntegrator2.

Topics in this [chaptersection](#) include the following:

- [Data Analysis Overview on this page](#)
- [Creating Kaplan-Meier Plots](#) on page 50

Data Analysis Overview

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

Note: You can verify that the study is in “Deployed” status by clicking the Manage Studies link in the left sidebar of the calIntegrator2 home page.

- **K-M Plot:** This tool analyzes clinical data, generating a K-M plot based on survival data sets.
- **Gene Expression Plot:** This tool analyzes annotation, clinical or genomic data based on gene expression values.
- **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.

Query and KM plot functionality in calIntegrator2 allows you to compare expression levels for a given gene against expression levels for a set of control samples designated at study definition. 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. 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 .

Include GenePattern analyses overviews?

After defining or running the desired query to use to populate the data, 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 is used for analysis of comparative groups of patients or samples. In calIntegrator2, the KM 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 KM 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 KM plot based on their survival probability as a function of time. Survival differences are analyzed by the log-rank test.

Note: To perform a KM 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.

KM Plot for Annotations

In the Rembrandt doc, “KM Plot” always has a hyphen: “K-M Plot”. which is correct?

The groups identified for this KM plot generation are based on clinical annotations.

1. Select the study whose data you want to analyze in the upper right portion of the calIntegrator page.
2. Under Analysis Tools on the left sidebar, select **KM Plot**.

3. Select the **For Annotation** tab at the top of the page.

Kaplan-Meier Survival Plots

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 Start Date		

Reset

Figure 5.1 Fields for defining annotation data for a KM plot

4. Select fields as described in the following table.

	Description
Patient Groups	<p>The groups to be compared in the KM plot originate from one patient group. Varying data sets are based upon multiple values corresponding to the selected annotation.</p> <p>Annotation Type – Select the annotation type that identifies the patient group. Selections are based on the data in the chosen study.</p> <p>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.</p> <p>Values – Using conventional selection techniques, select two or more values which will be the basis for the KM plot. Permissible (available) values or “No Values” correspond to the selected annotation.</p>
Survival Value	<p>Survival value is 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 KM plot.</p>

Table 5.1 Fields for selecting KM annotations plot values

5. 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 KM plot criteria..

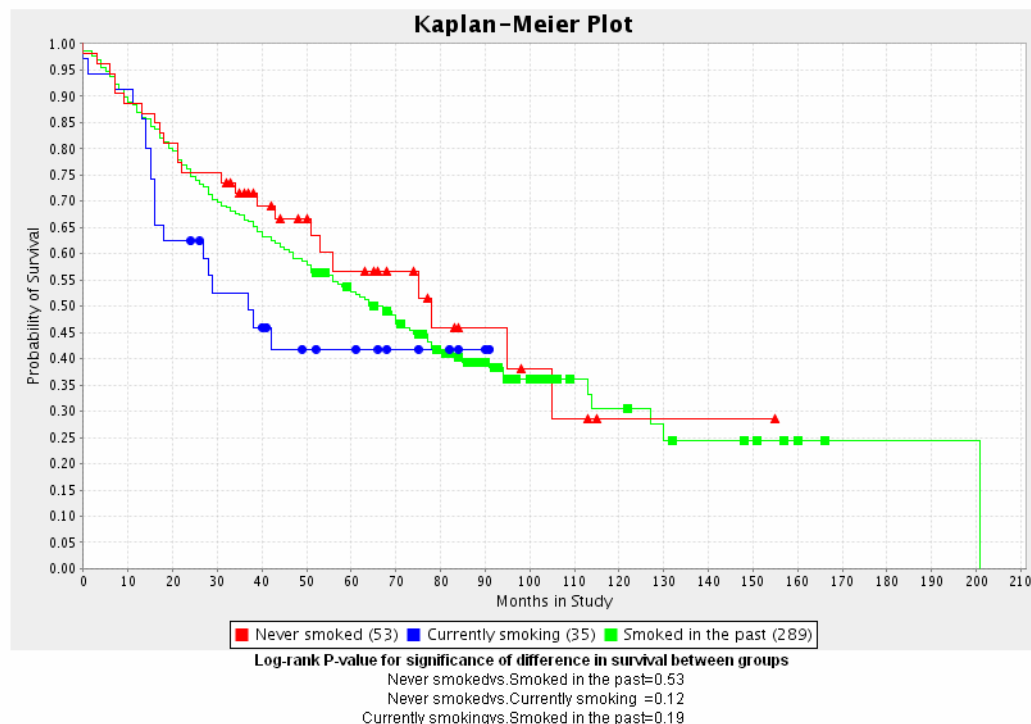


Figure 5.2 A KM plot generated for groups based on clinical annotations

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

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.

KM 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 KM 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 **KM Plot**.

3. Select the **For Gene Expression** tab.

Kaplan-Meier Survival Plots ?

For Annotation **For Gene Expression** For Queries

Gene Expression Based Kaplan-Meier Survival Plots

1.) Gene Symbol:

2.) Overexpressed \geq fold

3.) Underexpressed \geq fold

Survival Value

4.) Select Survival Measure:

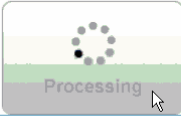
 Processing

Figure 5.3 Fields for defining gene expression data for a KM plot

4. Enter or select fields as described in the following table.

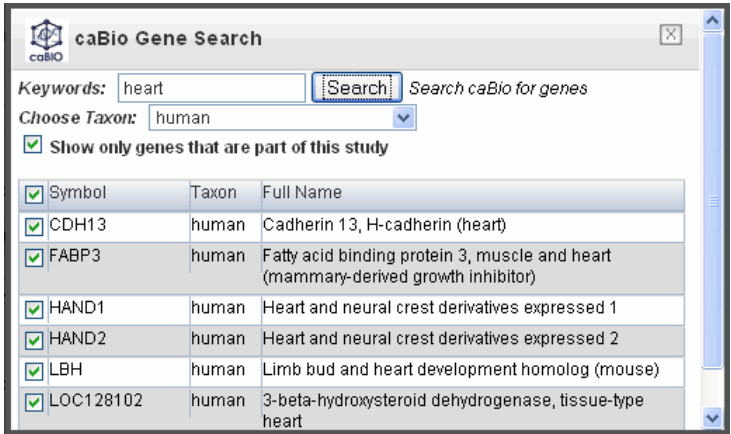


	Description
Gene Symbol	<p>In the text box, specify a single gene whose expression values can be used to split the subjects into three categories: high, low and intermediate expression or click the icons to locate genes in the following databases. If entering more than one gene in the text box, separate entries by commas.</p> <p>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).</p> <p>caBio – This link pulls identified genes into caIntegrator2 for analysis. Click the icon, enter Keyword(s) in the text box that opens and click Search.</p>  <p>Use the checkboxes to identify the genes whose symbols you want to use in the gene expression analysis.</p> <p>Click the Use Genes button at the bottom of the page. This pulls the checked genes into the Gene Expression ... tab.</p> <p>Annotation Based Gene Expression Plots</p> <p>1.) Gene Symbol(s) (comma separated list): CDH13,FABP3,HAND1,HA  </p>
Over-expressed/ Under-expressed	<p>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.</p>
Survival Value	<p>Survival value is 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 KM plot.</p>

Table 5.2 Fields for selecting KM gene expression plot values

5. Click the **Create Plot** button. caIntegrator2 generates the plot which then displays below the KM plot criteria.

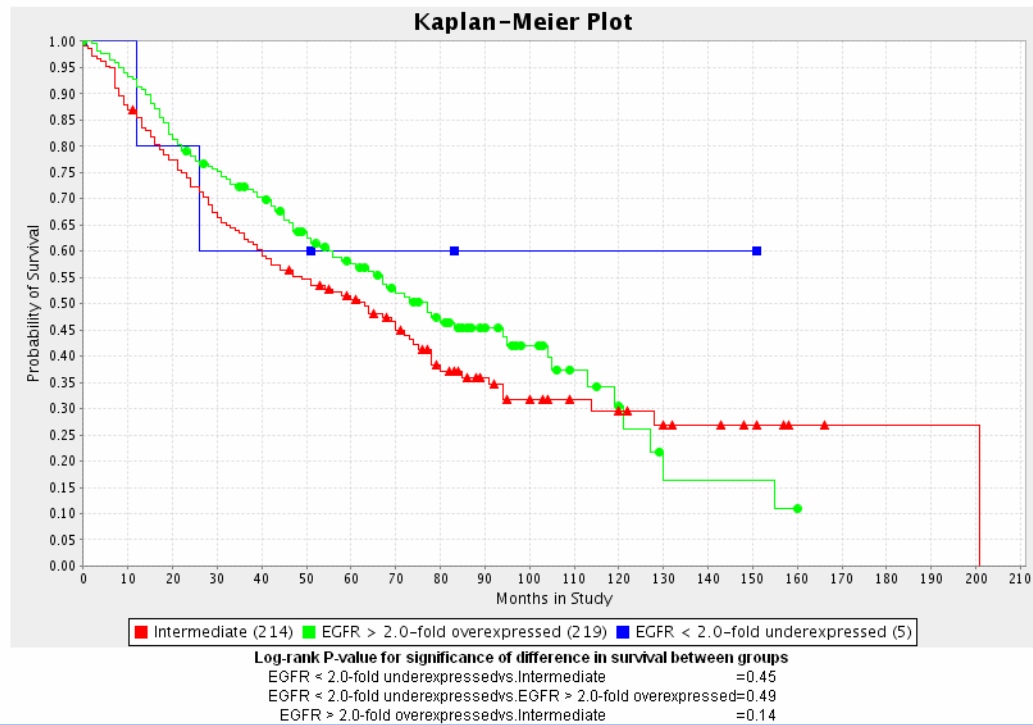


Figure 5.4 KM plot generated from gene expression data.

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, which is a group with gene expression values that are not up-regulated nor down-regulated. Don't see intermediate group in Fig. 5.4.

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.

KM 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 KM to compare the query groups. This is one method of limiting the data considered in the KM plot calculation.

1. Select the study whose data you want to analyze in the upper right portion of the calIntegrator page. You must select a study for which the queries you will identify for the KM plot have been saved.
2. Under Analysis Tools on the left sidebar, select **KM Plot**.

3. Select the **For Queries** tab.

Kaplan-Meier Survival Plots

For Annotation For Gene Expression **For Queries**

Query Based Kaplan-Meier Survival Plots

1.) Select Queries:

All Available Queries

- ADJV_CHEM_YES_FEMALES
- Gender = Male
- JP - female and BRCA1
- ADJV_MALES+NO_CHEMO
- TJ - ASP Patients
- ADJV_CHEMO_YES_MALES
- EGFR+CHEMO_YES
- Gender = Female

< Remove Add >

Query Groups

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 queries.

Survival Value

4.) Select Survival Measure: Survival From Start Date

Reset Create Plot

Figure 5.5 Fields for defining KM plot parameters based on saved queries in caIntegrator2

4. Enter or select fields as described in the following table.

	Description
Queries	Select the 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 KM plot.
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.
Add additional group...other subjects...	Check the box to create an additional group of subjects that are not in your other selected query groups.
Survival Value	Survival value is 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 KM plot.

Table 5.3 Fields for selecting KM plot values based upon caIntegrator2 queries

5. Click the **Create Plot** button. caIntegrator2 generates the plot which then displays below the KM plot criteria.

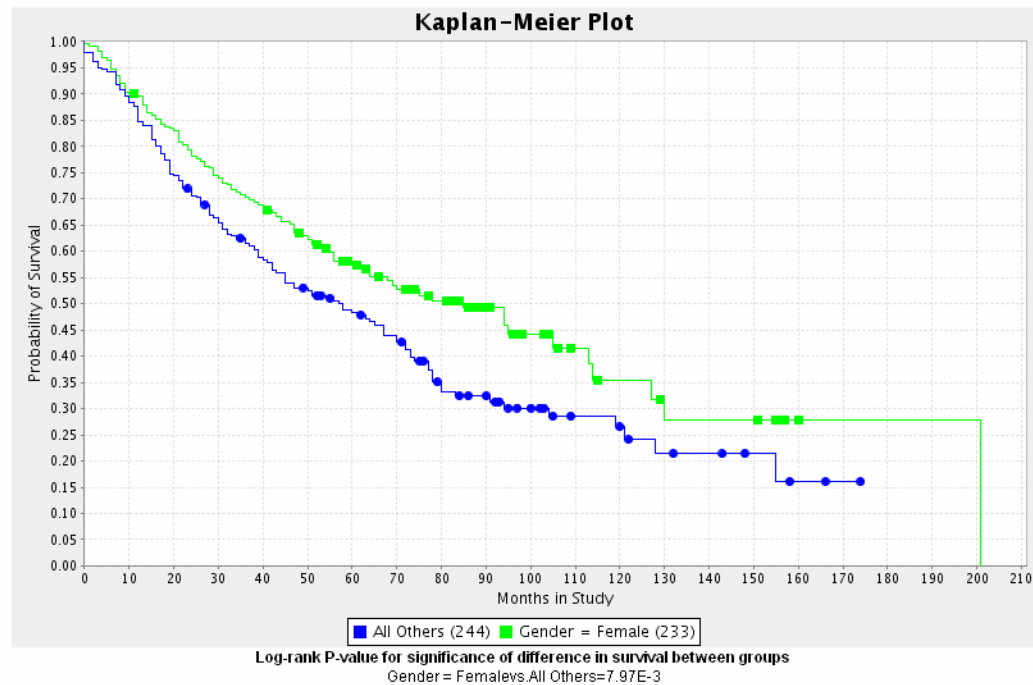


Figure 5.6

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, which you may not have selected. Intermediate group does not show in this example. Only with GE KM 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 CGAP and caBio directories and criteria can be defined 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.

See also *Understanding a Gene Expression Plot* on page 66.

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 calIntegrator2 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.

Gene Expression Value Plots

For Annotation For Genomic Queries For Clinical Queries

Annotation Based Gene Expression Plots

1.) Gene Symbol(s) (comma separated list): CEXAP CASIO

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

Annotation Type	Annotation	Values
Select Annotation Type	Select Annotation	<input type="text"/>

3.) Sample Groups:

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

5.) ☐ Add additional group containing all control samples for this study.

Reset

Figure 5.7 Gene expression value tab for configuring gene expression annotation value plot

4. Enter or select fields as described in [Table 5.4](#) the following table..

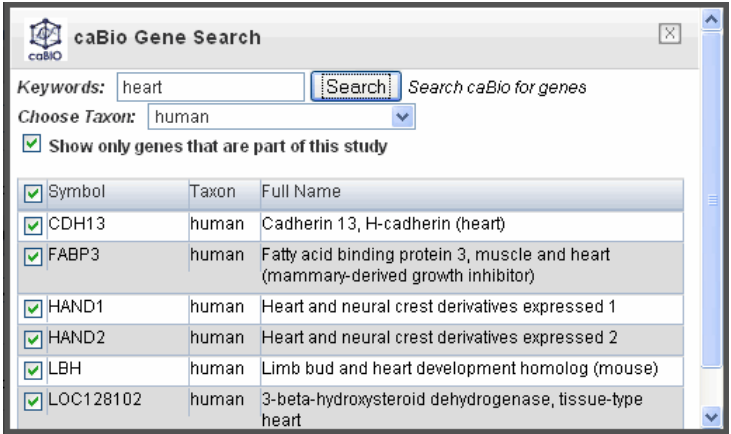


Field	Description
Gene Symbol	<p>Enter one or more gene symbols in the text box or click the icons to locate genes in the following databases. If entering more than one gene in the text box, separate entries by commas.</p> <p>CGAP – Use this directory to identify genes. Before clicking this link you must enter gene symbols in the text box. This link does not pull anything into caIntegrator2 but does provide information about the gene(s).</p> <p>caBio – This link pulls identified genes into caIntegrator2 for analysis. Click the icon, enter Keyword(s) in the text box that opens and click Search.</p>  <p>Use the checkboxes to identify the genes whose symbols you want to use in the gene expression analysis.</p> <p>Click the Use Genes button at the bottom of the page. This pulls the checked genes into the Gene Expression ... tab.</p> <p>Annotation Based Gene Expression Plots</p> <p>1.) Gene Symbol(s) (comma separated list): <input type="text" value="CDH13,FABP3,HAND1,HA"/>  </p>
Reporter Type	<p>Select the radio button that describes the reporter type:</p> <p>Reporter ID – Summarizes expression levels for all reporters you specify.</p> <p>Gene Name – Summarizes expression levels at the gene level.</p>

Table 5.4 Fields for selecting gene expression plot values based upon annotations

Field	Description
Sample Groups	<p>Annotation Type – Select the annotation type. Selections are based on the data in the chosen study</p> <p>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.</p> <p>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.</p>
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.
Add additional group...control group	Check the box to create an additional group of control samples for this study..

Table 5.4 Fields for selecting gene expression plot values based upon annotations

- Click the **Create Plot** button. calIntegrator2 generates the plot which then displays below the Gene Expression Plot criteria in bar graph format. Legends below the plot indicate the plot input.

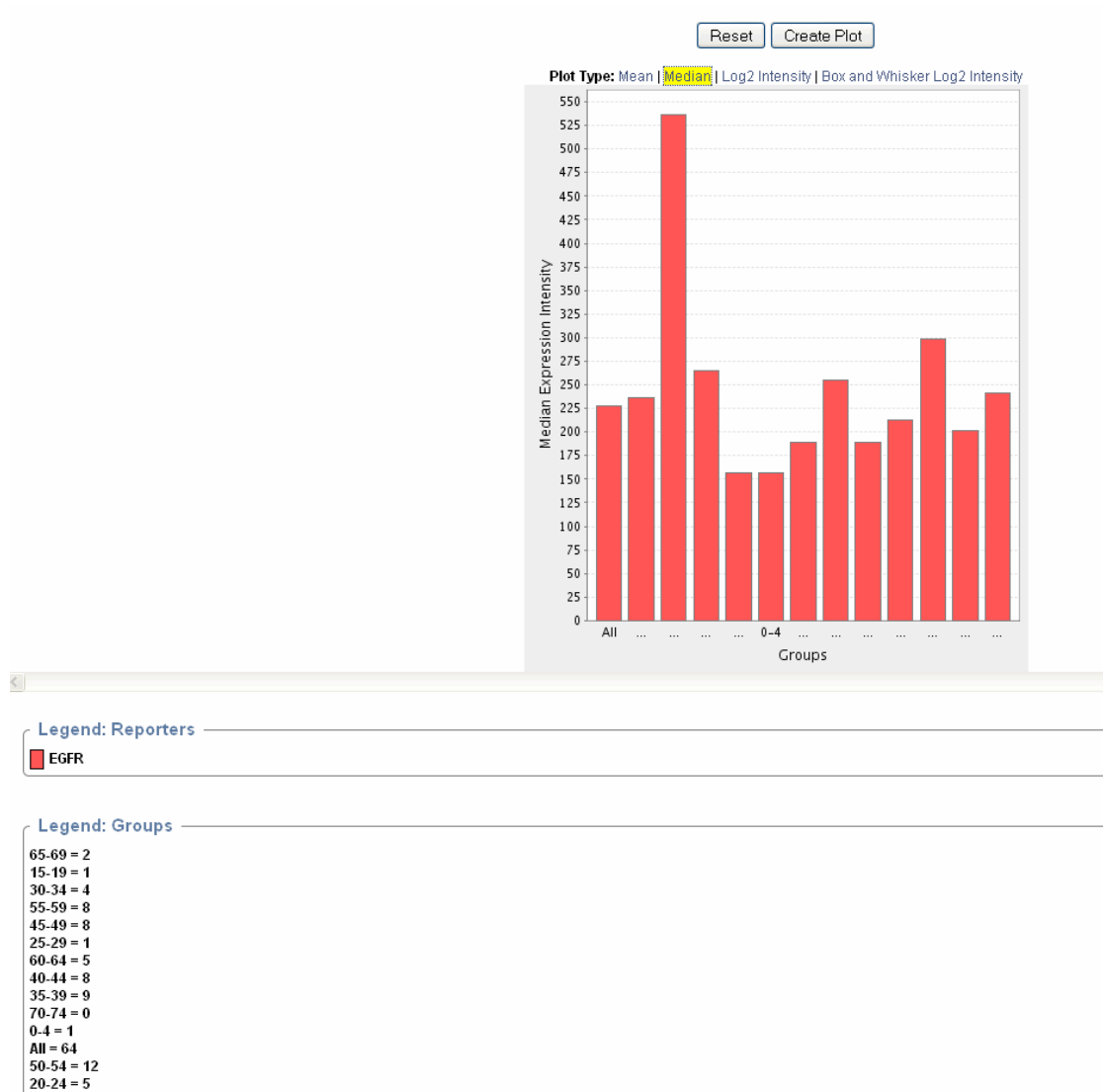


Figure 5.8 Gene expression plot based on selected annotations

- By default, calIntegrator2 displays a plot showing the mean of the data. You can recalculate the data display by clicking the **Plot Type** above the graph. See [Understanding a Gene Expression Plot](#) on page 66

[Figure 5.8](#) displays a plot with gene expression median calculation summaries. Legends below the plot indicate the plot input.

- 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 37.

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.

Gene Expression Value Plots

The screenshot shows the 'Gene Expression Value Plots' interface. At the top, there are three tabs: 'For Annotation', 'For Genomic Queries' (which is selected), and 'For Clinical Queries'. Below the tabs, the title 'Genomic Query Based Gene Expression Plots' is displayed, followed by a '(draft)' status icon. The main area contains two steps: '1.) Select a Genomic Query:' with a dropdown menu showing 'P - egfr down regulated 2', and '2.) Select Reporter Type:' with two radio buttons, 'Reporter Id' (selected) and 'Gene'. At the bottom right, there are 'Reset' and 'Create Plot' buttons.

Figure 5.9 Gene expression value tab for configuring gene expression genomic queries plot

4. Enter or select fields as described in the following table..

	Description
Genomic Query	Click on the genomic query upon which the plot is to be based.
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.

Table 5.5 Fields for selecting gene expression plot values based upon genomic queries

- Click the **Create Plot** button. caIntegrator2 generates the plot which then displays below the Gene Expression Plot criteria. Legends below the plot indicate the plot input.



Figure 5.10 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 66. Legends below the plot indicate the plot input.
- 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.

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 caIntegrator 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.

Clinical Query Based Gene Expression Plots (draft)

1.) Gene Symbol(s) (comma separated list):  

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

3.) Select Queries:

All Available Queries

Add >

< Remove

Selected Queries

v

^

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

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

6.) ☐ Add additional group containing all control samples for this study. control set 1

Reset
Create Plot

Figure 5.11 Gene expression value tab for configuring gene expression clinical queries plot

4. Enter or select fields as described in the following table..

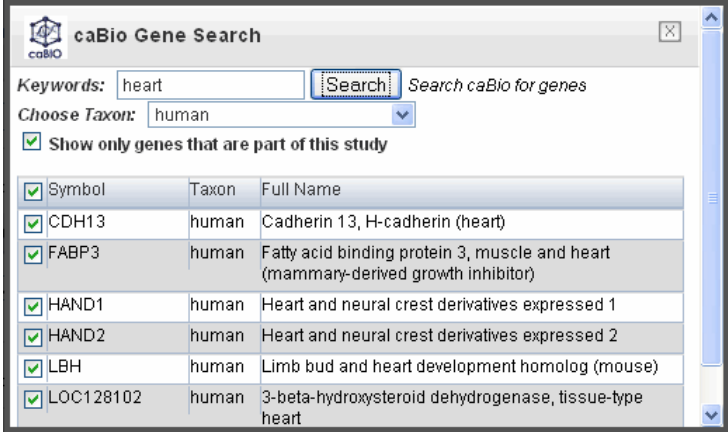


Field	Description
Gene Symbol	<p>Enter one or more gene symbols in the text box or click the icons to locate genes in the following databases. If entering more than one gene in the text box, separate entries by commas.</p> <p>CGAP – Use this directory to identify genes. Before clicking this link you must enter gene symbols in the text box. This link does not pull anything into caIntegrator2 but does provide information about the gene(s).</p> <p>caBio – This link pulls identified genes into caIntegrator2 for analysis. Click the icon, enter Keyword(s) in the text box that opens and click Search.</p>  <p>Use the checkboxes to identify the genes whose symbols you want to use in the gene expression analysis.</p> <p>Click the Use Genes button at the bottom of the page. This pulls the checked genes into the Gene Expression ... tab.</p> <p>Annotation Based Gene Expression Plots</p> <p>1.) Gene Symbol(s) (comma separated list): <input type="text" value="CDH13,FABP3,HAND1,HA"/>  </p>
Reporter Type	<p>Select the radio button that describes the reporter type:</p> <p>Reporter ID – Summarizes expression levels for all reporters you specify.</p> <p>Gene Name – Summarizes expression levels at the gene level.</p>

Table 5.6 Fields for selecting gene expression plot values based upon clinical queries

Field	Description
Sample Groups	<p>Annotation Type – Select the annotation type. Selections are based on the data in the chosen study</p> <p>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 its values.</p> <p>Values – Using conventional selection techniques, select two or more values which will be the basis for the KM plot. Permissible (available) values or “No Values” correspond to the selected annotation.</p>
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.
Add additional group...control group	Check the box to create an additional group of control samples for this study..

Table 5.6 Fields for selecting gene expression plot values based upon clinical queries

- Click the **Create Plot** button. By default, calIntegrator2 generates the plot which displays the mean of the data below the Gene Expression Plot criteria. Legends below the plot indicate the plot input.
- [screen shot](#)
- You can recalculate the data display by clicking the **Plot Type** above the graph. See [Understanding a Gene Expression Plot](#) on page 66.
- You can modify the plot parameters and click the **Reset** button to recalculate the plot.

Understanding a Gene Expression Plot

I adapted these descriptions from Rembrandt User Guide. Please modify as appropriate. Rembrandt plots are more complex, justifying this section.

When you perform a Gene Expression simple search, by default the **Mean** Gene Expression Plot ([Figure 2.2](#)) appears.

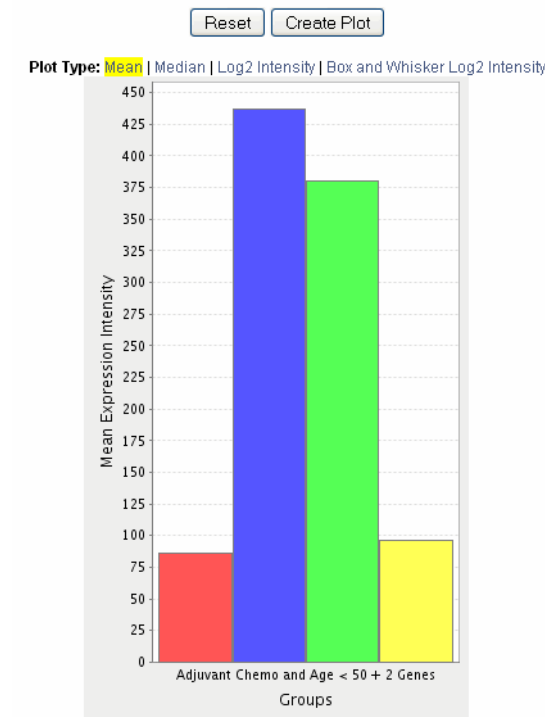


Figure 5.12 Gene expression plot calculating the mean

The **Mean** Gene Expression Plot ([Figure 5.12](#)) displays mean expression intensity (Geometric mean) versus Groups.

Above the plot, you can select other plot types. When you do so, the plot is recalculated.

The **Median** Gene Expression Plot ([Figure 5.13](#)) displays the median expression intensity versus Groups..

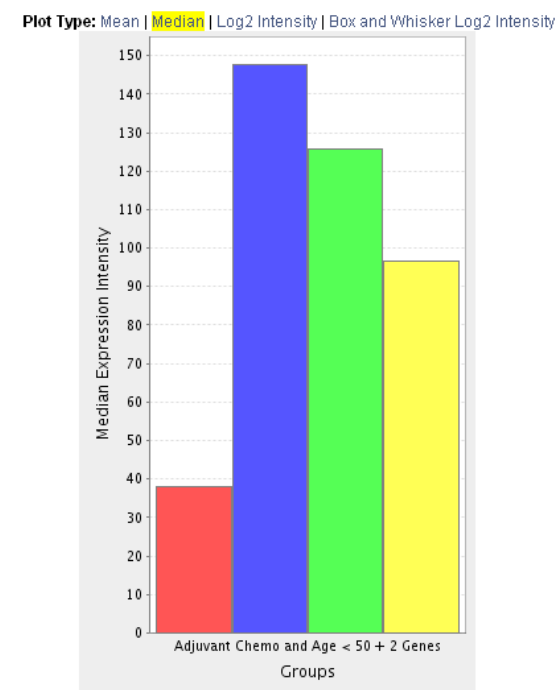


Figure 5.13 Gene expression plot calculating the median

The **Log2 Intensity** Gene Expression Plot ([Figure 2.5](#)) displays average expression intensities for the gene of interest based on Affymetrix GeneChip arrays (U133 Plus 2.0 arrays).

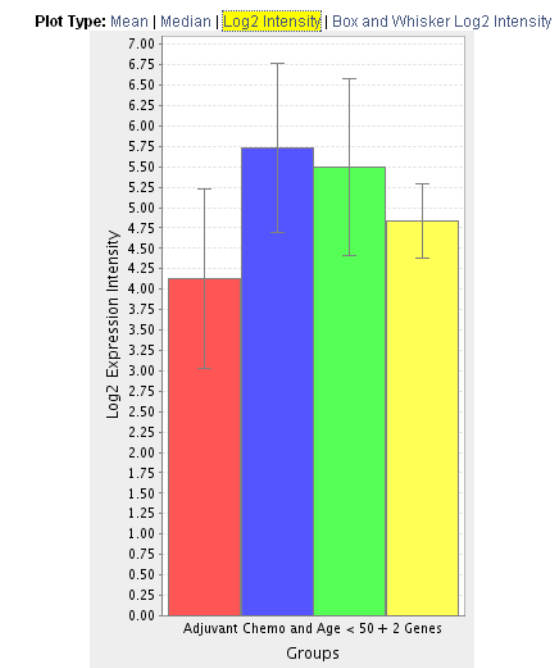
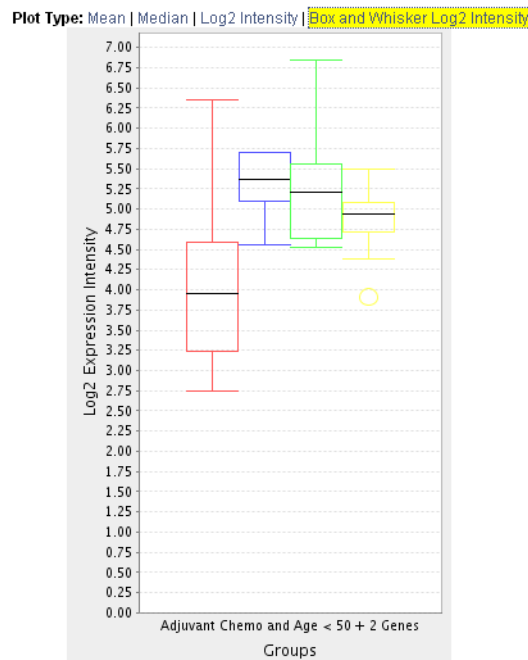


Figure 5.14 Gene expression plot displaying log2 intensity values

The box and whisker log2 expression intensity plot displays a box plot without all the individual data points. Although the one I sent you, JP, showed a “coin plot” as well. What determines whether that displays? 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 dataset.
- Perform a large number of observations.
- Compare two or more datasets.
- Compare distributions because the centre, spread, and overall range are immediately apparent.



Use anything in these 2 paragraphs?: 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 (sample minimum), lower quartile (Q1), median (Q2), upper quartile (Q3), and largest observation (sample maximum)). A boxplot may also indicate which observations, if any, might be considered 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, and identify outliers.

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 calIntegrator2. Specifics for launching GenePattern tools from calIntegrator2 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.
 - c. **For Jill:** Development version of GenePattern: <http://cainegrator2-dev.nci.nih.gov/gp/services/> - no password (for your testing use)
- 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.

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.



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](#)
 - **Comparative Marker Selection (Grid Service)**. This option enables you to run this GenePattern analysis on the grid. See [Comparative Marker Selection \(CMS\) Analysis](#).
 - **Principal Component Analysis (Grid Service)**. This option enables you to run this GenePattern analysis on the grid. See [Principal Component Analysis \(PCA\)](#).

- **GISTIC (Grid Service)**. This option enables you to run this GenePattern analysis on the grid. See [GISTIC-Supported Analysis](#).
- 4. 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](#).

1. In the GenePattern Analysis dialog box, specify connection information.

GenePattern Analysis

GenePattern Server URL*:	<input type="text"/>
GenePattern Username*:	<input type="text"/>
GenePattern Password:	<input type="password"/>
<input type="button" value="Connect"/>	

Figure 5.15 Dialog box for configuring the link to GenePattern

	Description
Server URL	Enter any GenePattern publicly available URL., such as <i>genepattern.broad.mit.edu</i> .
GenePattern Username	Enter your GenePattern user name.
GenePattern Password	Enter your GenePattern password.

Table 5.7 Fields for selecting GenePattern configurations

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

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

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.

Comparative Marker Selection (CMS) Analysis

The Comparative Marker Selection (CMS) module includes several approaches to determine the features that most closely correlate with a class template. I find this sentence really hard to decipher. It really doesn't explain in understandable terms what this analysis accomplishes. Help? It also determines the significance of that correlation.

If the input class template has more than two classes, than a one-versus-all comparison is performed for each class. Note that the p-values obtained from the one-versus-all comparison are not fully corrected for multiple hypothesis testing.

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 calIntegrator page. You must select a study saved as a clinical study, but which has genomic data.
2. 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.

Comparative Marker Selection Analysis

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

- Job Name*:** Text input field with 'jhh1' entered.
- Preprocess Server*:** Dropdown menu showing 'Default Broad service - http://node255.broad.mit.edu:6060/wsrf/services/cagrid/PreprocessDatasetMAGEService'.
- Comparative Server*:** Dropdown menu showing 'Default Broad service - http://node255.broad.mit.edu:6060/wsrf/services/cagrid/ComparativeMarkerSelMAGESvc'.
- Clinical Queries:** Two list boxes. 'All Available Queries' contains 'age 55-59'. 'Selected Queries' is empty. Between them are 'Add >' and '< Remove' buttons.
- Filter flag:** Check box, currently unchecked.
- Preprocessing Flag*:** Dropdown menu showing 'no-disc-or-norm'.
- Min Change*:** Text input field with '3.0'.
- Min Delta*:** Text input field with '100.0'.
- Threshold*:** Text input field with '20.0'.
- Ceiling*:** Text input field with '2.1'.
- Max Sigma Binning*:** Text input field with '1'.
- Probability Threshold*:** Text input field with '1.0'.
- Num Exclude*:** Text input field with '0'.
- Log Base Two:** Check box, currently unchecked.
- Number Of Columns Above Threshold*:** Text input field with '1'.
- Test Direction*:** Dropdown menu showing 'two-sided'.
- Test Statistic*:** Dropdown menu showing 'T-test'.
- Min Std*:** Text input field with '1.0'.
- Number Of Permutations*:** Text input field with '1000'.
- Complete:** Check box, currently unchecked.
- Balanced:** Check box, currently unchecked.
- Random Seed*:** Text input field with '779948241'.
- Smooth Pvalues:** Check box, currently unchecked.
- Phenotype Test*:** Dropdown menu showing 'one-versus-all'.
- Perform Analysis:** Button at the bottom right.

3. Select or define CMS analysis parameters, described in [Table 5.8](#). An asterisk indicates required fields. [Suggest leave default settings?](#)

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.

Table 5.8 Comparative Marker Selection analysis options

CMS Parameter	Description
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
Probablility 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 >= 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*	The statistic to use expand--see reference
Min Std*	The minimum standard deviation if test statistic includes the min std option. Used only if test statistic includes the min std option. If o (sigma symbol?) is less than min std, o is set to min std.
Number of Permutations*	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. 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. Balanced – Perform balanced permutations
Random Seed*	The seed for the random number generator.

Table 5.8 Comparative Marker Selection analysis options

CMS Parameter	Description
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*	Tests to perform when class membership has more than 2 classes: one versus-all, all pairs. Note: The P-values obtained from the one-versus-all comparison are not fully corrected for multiple hypothesis testing.

Table 5.8 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.

GenePattern Analysis Status (draft)

Gene Pattern Modules New Analysis Job

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 5.16 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 the zipped (always zipped? always 3 files? always in same format?) result files to the destination you select. One of the files may be an XML file and another can be opened in spreadsheet format.

Principal Component Analysis (PCA)

GenePatterns PCA documentation is useless. Need an introductory paragraph. USBAT optionally configure genepattern grid PreprocessDataset parameters in addition to PCA module parameters. All PreprocessDataset parameters are available on screen, but will be collapsed by default and will be disabled by default. When selected, PreprocessDataset will be executed prior to PCA. (Similar behavior to grid based comparative marker selection. Also, for the pca service settings, the user will see Rows selected by default.

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:

- 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.
- In the GenePattern Analysis Status page, select **Principle Component Analysis (Grid Service)** from the drop down list and click **New Analysis Job**. This opens the Principle Component Analysis page.

Principal Component Analysis

Job Name*:

Principal Component Analysis Server*: Default Broad service - <http://node255.broad.mit.edu:6060/wsrf/services/cagrid/PCA> ▼

Clinical Queries*:

All Available Queries

- AIIF
- My Query2
- AIIM

< Remove

Add >

Selected Queries

Cluster By*: columns ▼

Perform Analysis

3. Select or define PCA analysis parameters, described in [Table 5.9](#). Suggest leave default settings?.

	<i>Description</i>
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 with appropriate data <u>what is appropriate data??</u> for the analysis are listed. Select and move them (<u>1? or 1 or more??</u>) from the All Available Queries panel to the Selected panel.
Cluster By*	<u>description??</u>

Table 5.9 PCA analysis options

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

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:

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.

- 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.

GISTIC Analysis

Job Name*:

GISTIC Server*: Default GISTIC service - http://node255.broadinstitute.org:10010/wsrf/services/cagrid/Gistic

Clinical query: All non-control Samples

Amplifications Threshold*:

Deletions Threshold*:

Join Segment Size*:

QV Thresh*:

Remove X*: Yes

cnv File:

- Select or define GISTIC analysis parameters, as described in [Table 5.8](#).
[Suggest leave default settings?](#)

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 with appropriate data <u>what is appropriate data??</u> for the analysis are listed. Select and move them (<u>1? or 1 or more?</u>) from the All Available Queries panel to the Selected panel.
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.
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).

Table 5.10 GISTIC analysis options

GISTIC Parameters	Description
cnv File	<p>This selection is optional. <u>Did I hear this correctly?</u></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 5.10 GISTIC analysis options

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

