6 - Analyzing Studies

This chapter describes how to use caIntegrator tools to analyze data in subject annotation or genomic studies that have been deployed in caIntegrator.

Topics in this chapter include the following:

Data Analysis Overview

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

You can verify that the study has "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 caIntegrator 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 nor 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 subject annotation data, generating a Kaplan-Meier (K-M) plot based on survival data sets. See Creating Kaplan-Meier Plots.
- **Gene Expression Plot**: This tool analyzes annotation, subject annotation or genomic data based on gene expression values. See Creating Gene Expression Plots.
- **GenePattern**: This feature provides an express link to GenePattern where you can perform analyses on selected caIntegrator studies, or it enables you to perform several GenePattern analyses on the grid. See Analyzing Data with GenePattern.

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

This help topic opens from any of the three K-M plot tabs. For specific details about working with these tabs, see the following topics:

- K-M Plot for Annotations
- K-M Plot for Gene Expression
- K-M Plot for Queries and Saved Lists

The Kaplan-Meier method analyzes comparative groups of subjects or samples. In caIntegrator, the K-M method can compare 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, based on Karnofsky scores. You could compare groups of subjects 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 subjects 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.

caIntegrator calculates the log-rank p-value for the data, indicating the significance of the difference in survival between any two groups of samples. The log rank p-value is calculated using the Mantel-Haenszel method. The p-values are recalculated every time a new plot is generated.

ⓘ

K-M Plot for Annotations

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

1. Select the study whose data you want to analyze in the upper right portion of the caIntegrator 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, shown in the following figure.

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 the options described in the following table:

| Field	Description
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. ✓ **Tip** Only annotations that are defined with permissible values display in the drop-down list.
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.
Survival Value	Survival value is the length of time the patient lived. caIntegrator displays valid survival values entered for this study. Select the survival measure which is the unit of measurement for the survival value to be used for the plot.

5. Click the **Create Plot** button.

caIntegrator generates the plot, as shown in the following figure. The plot displays below the plot criteria.
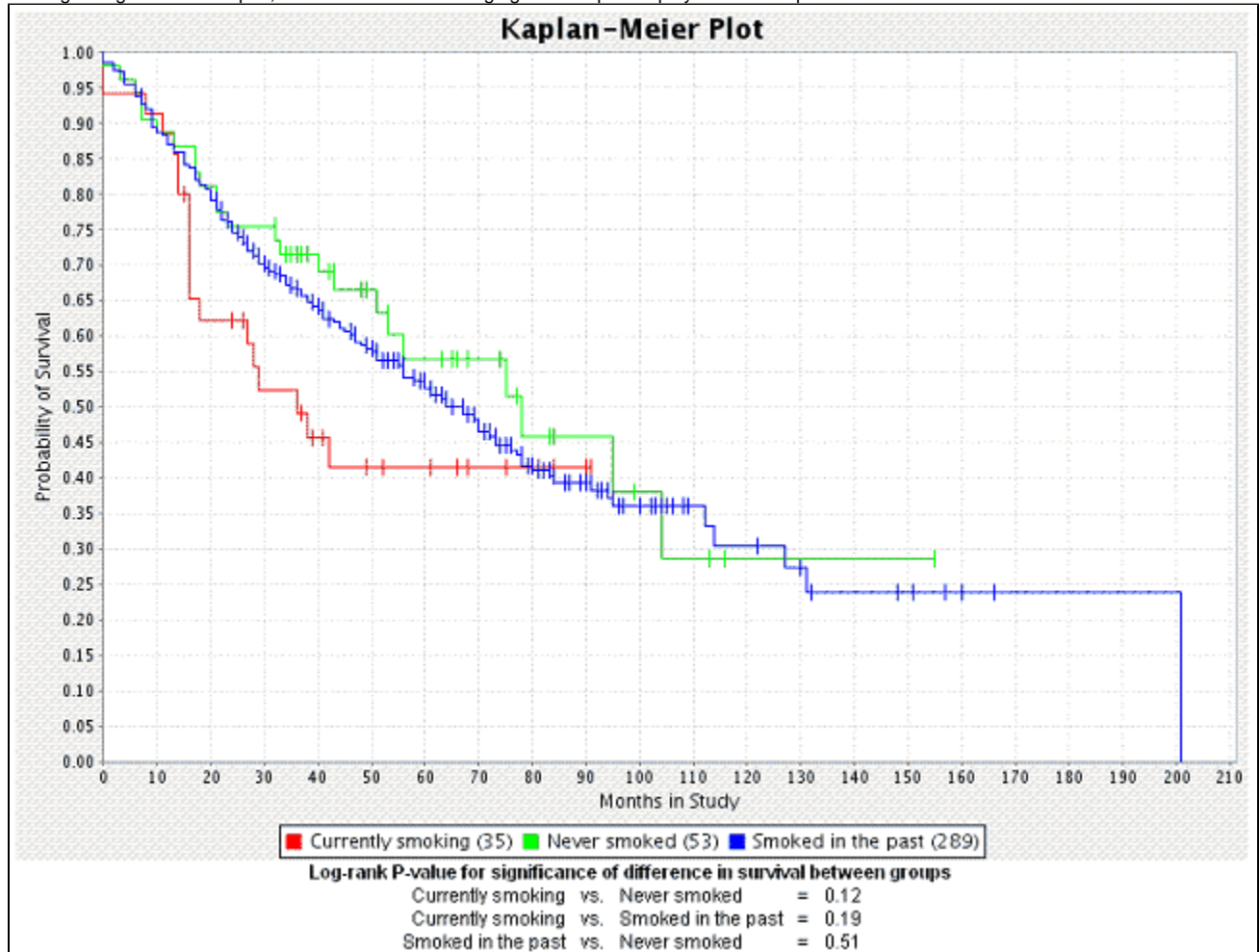
- The number of subjects for each group is embedded in the legend below the plot.
- caIntegrator generates a P-value for the selected groups; it displays at the bottom of the page. A low P-value generally has more significance than a high P-value.
- For information regarding the P-value calculation, see Creating Kaplan-Meier Plots.

K-M Plot for Gene Expression

caIntegrator allows you to compare expression levels for one given gene in different representative groups. The relative expression level is referred to as "fold change". Fold change is the ratio of the measured gene expression value in an experimental sample as determined by a reporter to a reference value calculated for that reporter against 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 caIntegrator 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, shown in the following figure. Plot criteria are described in the table following the figure.

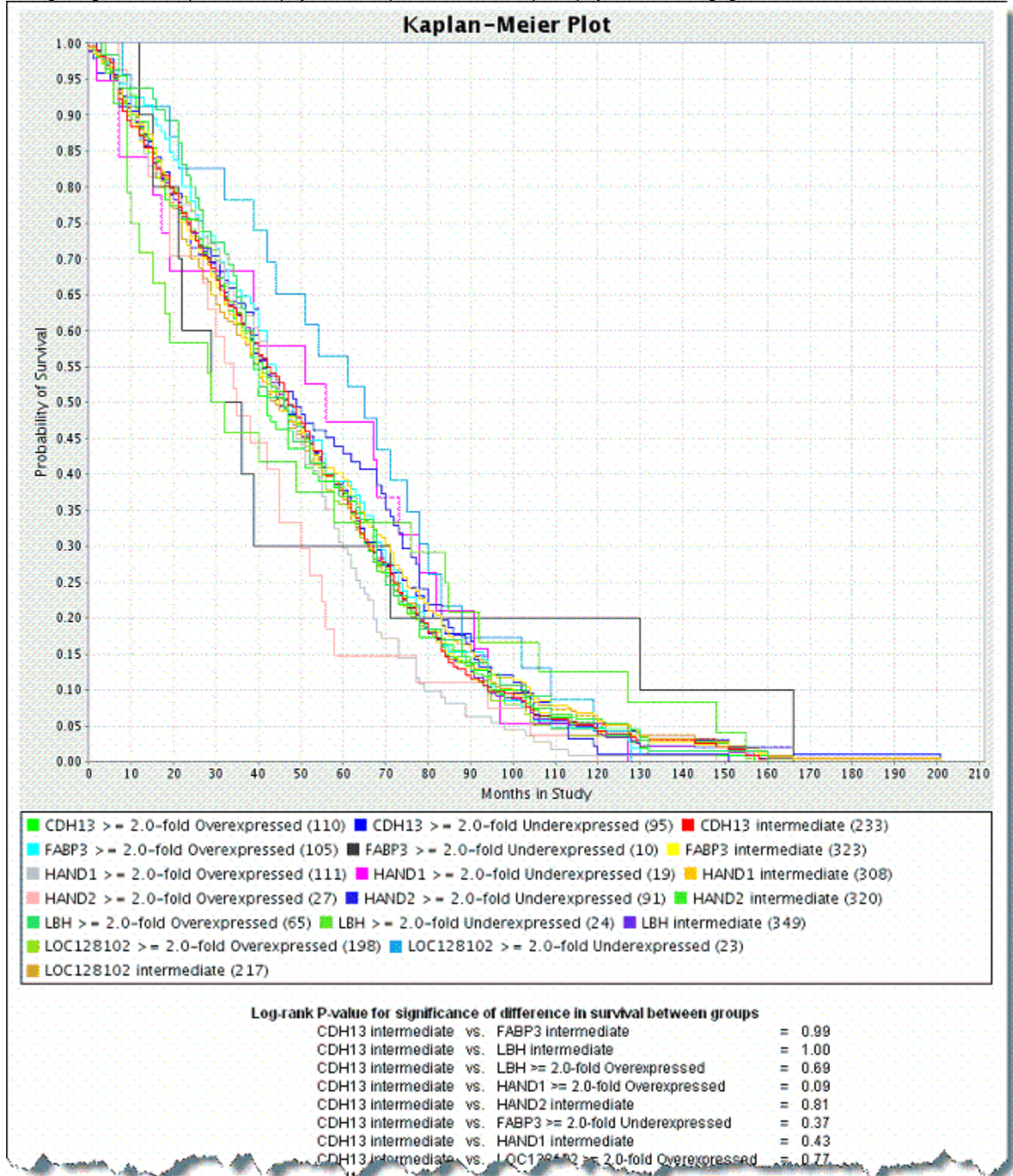| Field	Description
Gene Symbol	Enter one or more gene symbols in the text box or click the icons to locate genes in the following databases. If you enter more than one gene in the text box, separate the entries by commas.
	caIntegrator provides three methods whereby you can obtain gene symbols for calculating a KM or gene expression plot. For more information, see Choosing Genes.
	.
Expression Type	**By Fold Change:** This option allows you to define in the next two fields 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. **By Expression Level:** This option allows you to run a KM gene expression plot when there is no control group nor reference data set. In the next two fields, enter values **Overexpressed** (above expression level) or **Underexpressed** (below expression level).
Survival Value	Survival value, the length of time the patient lived, is required for both expression types. Select the survival measure which is the unit of measurement for the survival value to be used for the plot.
Control Sample Sets	This field is required only for fold change data. One or more control sets 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.
	ⓘ **Platform for control set**
	If the study has more than one platform associated with it, the platform is inherently selected when you select the control set. Control sets are comprised of samples from only one platform.

4. Click the **Create Plot** button.

caIntegrator generates the plot which displays below the plot criteria. An example displays in the following figure.

Kaplan–Meier Plot

Legend:
- ■ CDH13 >= 2.0-fold Overexpressed (110) ■ CDH13 >= 2.0-fold Underexpressed (95) ■ CDH13 intermediate (233)
- ■ FABP3 >= 2.0-fold Overexpressed (105) ■ FABP3 >= 2.0-fold Underexpressed (10) ■ FABP3 intermediate (323)
- ■ HAND1 >= 2.0-fold Overexpressed (111) ■ HAND1 >= 2.0-fold Underexpressed (19) ■ HAND1 intermediate (308)
- ■ HAND2 >= 2.0-fold Overexpressed (27) ■ HAND2 >= 2.0-fold Underexpressed (91) ■ HAND2 intermediate (320)
- ■ LBH >= 2.0-fold Overexpressed (65) ■ LBH >= 2.0-fold Underexpressed (24) ■ LBH intermediate (349)
- ■ LOC128102 >= 2.0-fold Overexpressed (198) ■ LOC128102 >= 2.0-fold Underexpressed (23)
- ■ LOC128102 intermediate (217)

Log-rank P-value for significance of difference in survival between groups

CDH13 intermediate vs. FABP3 intermediate	=	0.99
CDH13 intermediate vs. LBH intermediate	=	1.00
CDH13 intermediate vs. LBH >= 2.0-fold Overexpressed	=	0.69
CDH13 intermediate vs. HAND1 >= 2.0-fold Overexpressed	=	0.09
CDH13 intermediate vs. HAND2 intermediate	=	0.81
CDH13 intermediate vs. FABP3 >= 2.0-fold Underexpressed	=	0.37
CDH13 intermediate vs. HAND1 intermediate	=	0.43
CDH13 intermediate vs. LOC128102 >= 2.0-fold Overexpressed	=	0.77

- The gene symbol for each group represented in the data appears with its color correlation to the plot embedded in the legend below the plot. Three lines on this plot represent each gene symbol entered for the plot. Each line of the three represents a subgroup of people carrying the gene--one line for overexpressed values, one line for under expressed values and one line for intermediate values which represents gene 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, raw expression values are replaced with calculated fold change values.
- 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.
- For information regarding the P-value calculation, see Creating Kaplan-Meier Plots.

K-M Plot for Queries and Saved Lists

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. The queries you identify for the K-M plot must have been saved previously in caIntegrator.
2. Under Analysis Tools on the left sidebar, select **K-M Plot**.
3. Select the **For Queries and Saved Lists** tab, shown in the following figure. The criteria for the plot are described in the table below the figure.

Field	Description
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.
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...all other subjects	Check the box to create an additional group of all other subjects that are not in 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 plot.

4. Click the **Create Plot** button.

caIntegrator generates the plot which then displays below the plot criteria. An example displays in the following figure.

Kaplan-Meier Plot

Log-rank P-value for significance of difference in survival between groups
Never Smoked vs. Currently Smoking = 0.12
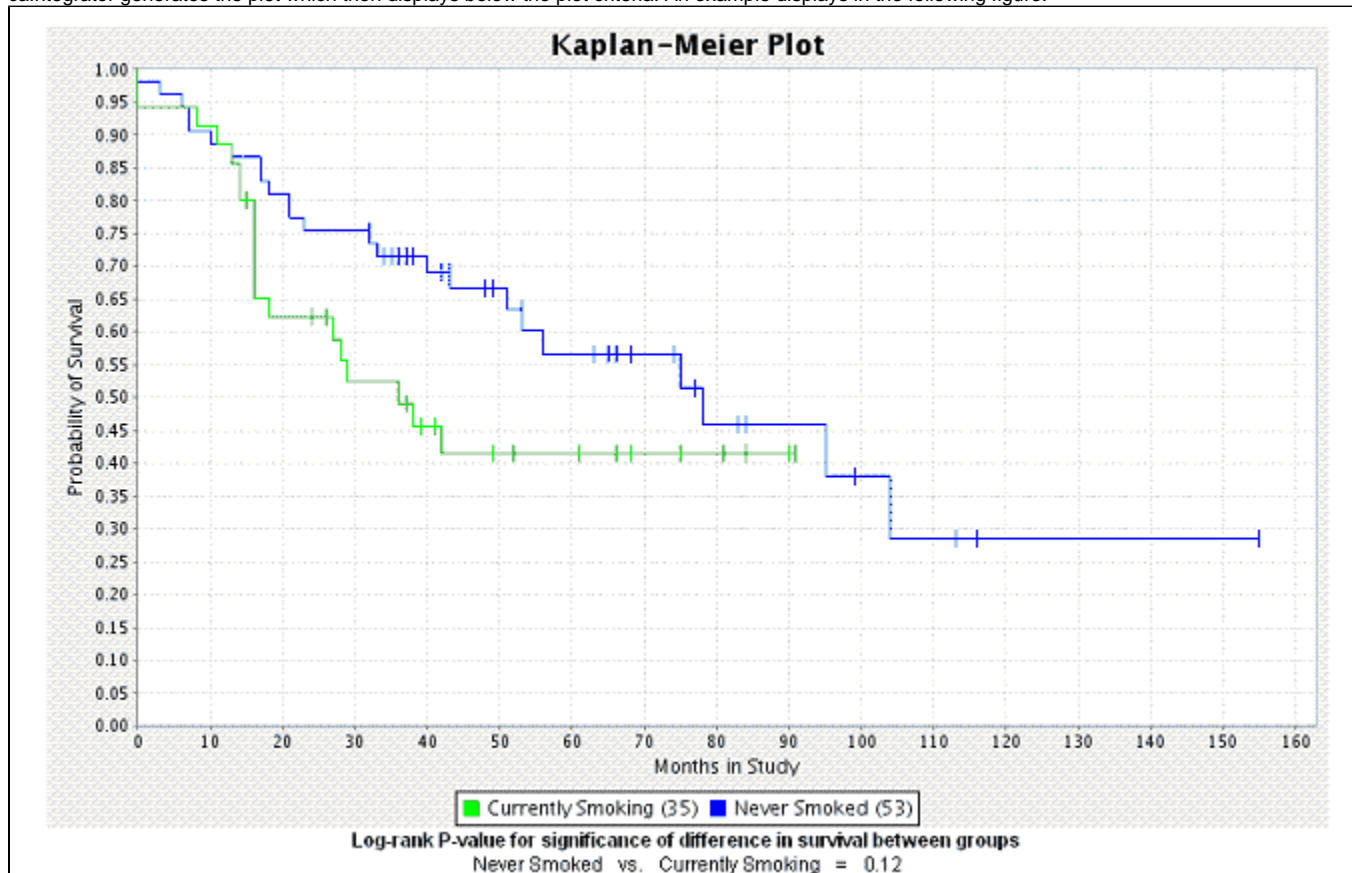
- The number of subjects for each group is embedded in the legend below the plot.
- A P-value is also generated for the selected groups; it displays at the bottom of the page. A low P-value generally has more significance than a high P-value.
- For information regarding the P-value calculation, see Creating Kaplan-Meier Plots.

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; it does not limit your comparison to only two sets of data. 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.

caIntegrator 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 caIntegrator Gene Lists. You can learn more about the genes in the CGAP directory. You can define criteria for the plot using subject annotation and image annotations.
- Gene Expression Value Plot for Genomic Queries – You can select data based on saved genomic queries.
- Gene Expression Value Plot for Annotation and Saved List Queries – You can select data based on saved subject annotation queries. You can locate genes in the caBIO directories or caIntegrator Gene Lists.

See also Understanding a Gene Expression Plot.

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 caIntegrator 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, shown in the following figure. Criteria for the plot are described in the table following the figure.

Field	Description
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. caIntegrator provides three methods whereby you can obtain gene symbols for calculating a KM or gene expression plot. For more information, see Choosing Genes.
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. **Platform**--This field displays only if the study has multiple platforms. Select the appropriate platform for the plot. The platform you select determines the genes used for the plot.
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 organize the subjects 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. The control set should be composed of only samples which are mapped to subjects. See Uploading Control Samples.

4. Click the **Create Plot** button.

caIntegrator generates the plot which displays below the plot criteria in bar graph format. Legends below the plot indicate the plot input. By default, the plot shows the median calculation summaries of gene expression. An example displays in the following figure.

- You can recalculate the data display by changing the **Plot Type** above the graph.
- You can modify the plot parameters and click the **Reset** button to recalculate the plot.

See also See Understanding a Gene Expression 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.

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 caIntegrator 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, shown in the following figure. Criteria for the plot are described in the table following the figure.

Field	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.

4. Click **Create Plot**. caIntegrator generates the plot, as shown in the following figure. The plot displays below the plot criteria. Legends below the plot indicate the plot input.
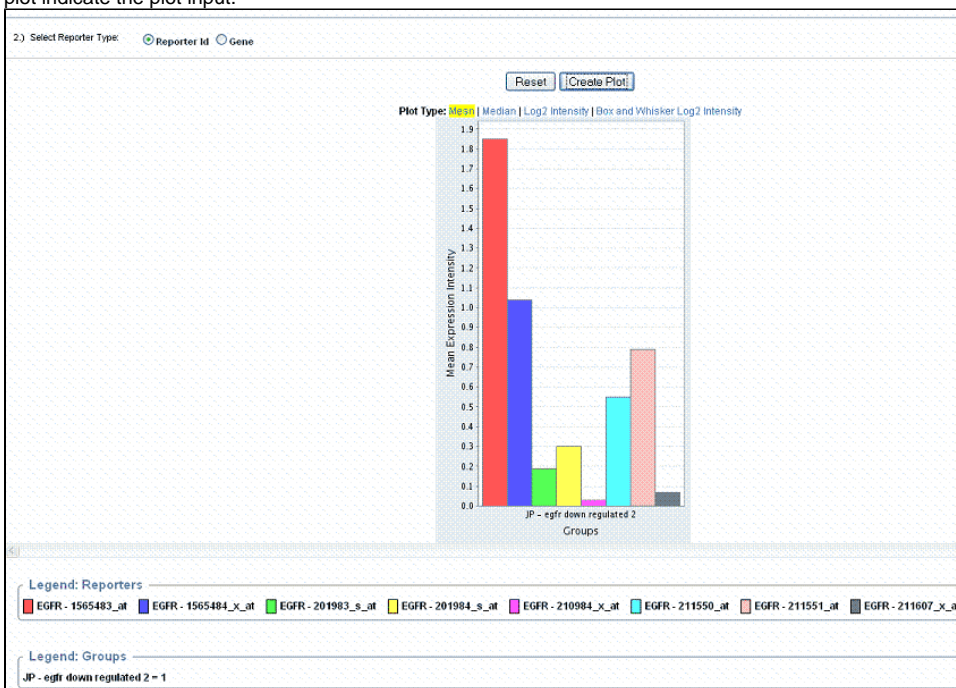
- You can recalculate the data display by changing the **Plot Type** above the graph.
- You can modify the plot parameters and click the **Reset** button to recalculate the plot.

See also Understanding a Gene Expression Plot.

Gene Expression Value Plot for Annotation and Saved List Queries

Data to be analyzed on this tab must have been saved as a subject annotation query, but it must have genomic data identified in the query. For more information, see Adding or Editing Genomic Data. For the genomic data, you must identify genes whose expression values are used to calculate the plot.

To generate the plot, follow these steps:

1. 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 subject annotation study, but which has genomic data.
2. Under Analysis Tools on the left sidebar, select **Gene Expression Plot**.

3. Select the **For Annotation Queries and Saved Lists** tab, shown in the following figure. Plot criteria are described in the table following the figure.

Field	Description
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. caIntegrator provides three methods whereby you can obtain gene symbols for calculating a KM or gene expression plot. For more information, see Choosing Genes.
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.
Platform	This field displays only if the study has multiple platforms. Select the appropriate platform for the plot. The platform you select determines the genes used for the plot.
Saved Queries	Choose among the available saved queries and lists. Build your selections in the right panel by using the **Add >** and **Remove <** buttons. ⓘ **Prefixes for lists** The [SL] and [Q] prefixes to list names indicate "Subject Lists" or "Saved Queries". A "G" in the prefix indicates the list is Global. For more information, see Creating a Gene or Subject List.
Exclusive Subjects...	To remove subjects in your queries and lists selection from queries or lists you use subsequently for analysis, check the button. This allows you to use them exclusively for the current analysis.
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. The control set should be composed of only samples which are mapped to subjects. See Uploading Control Samples.

4. Click the **Create Plot** button.

caIntegrator generates the plot in bar graph format which displays below the plot criteria. By default, legends below the plot indicate the plot input. An example displays in the following figure.
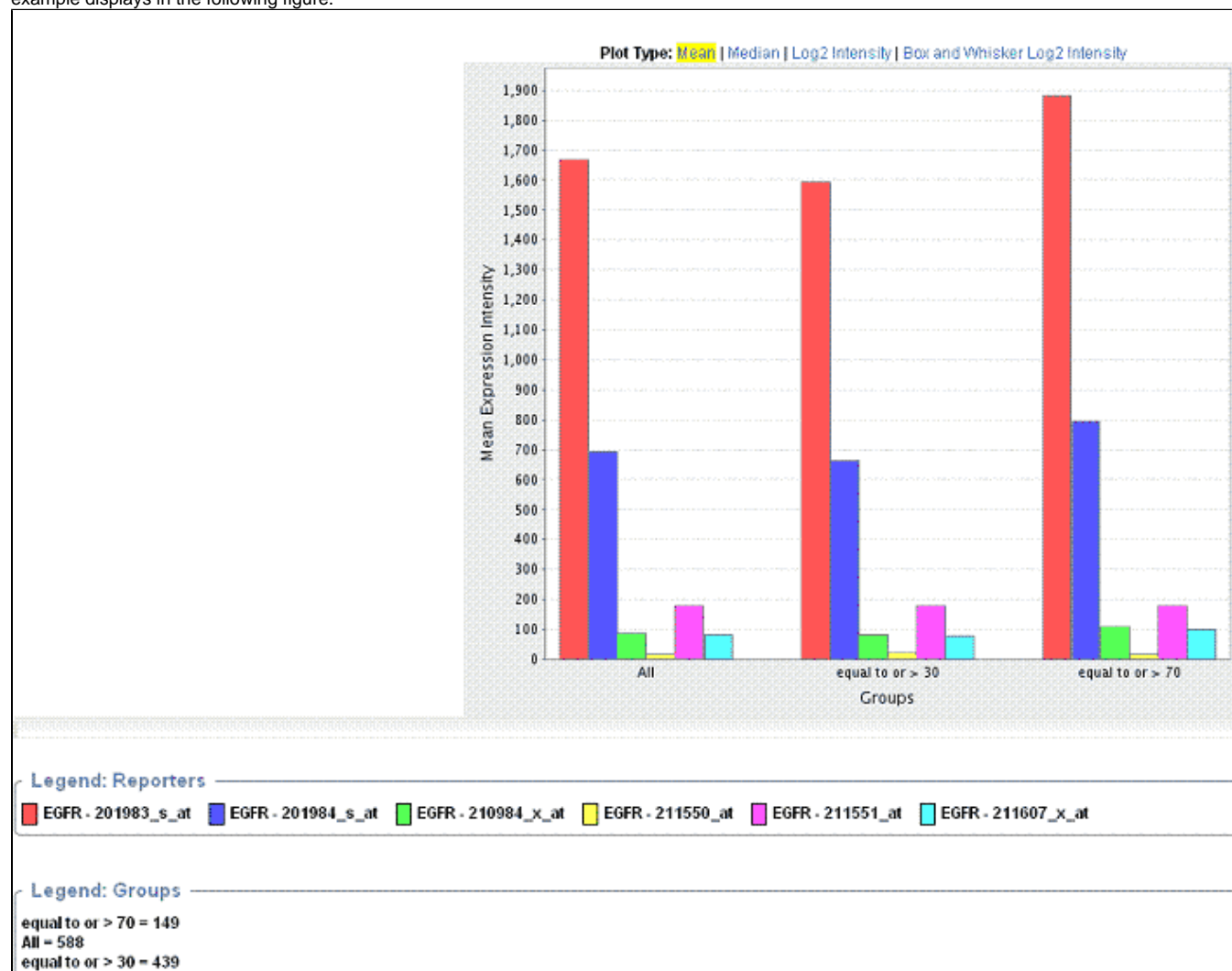

- You can recalculate the data display by changing the **Plot Type** above the graph.
- You can modify the plot parameters and click the **Reset** button to recalculate the plot.

See also See Understanding a Gene Expression 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 Gene Expression Plot appears, as shown in the following figure.
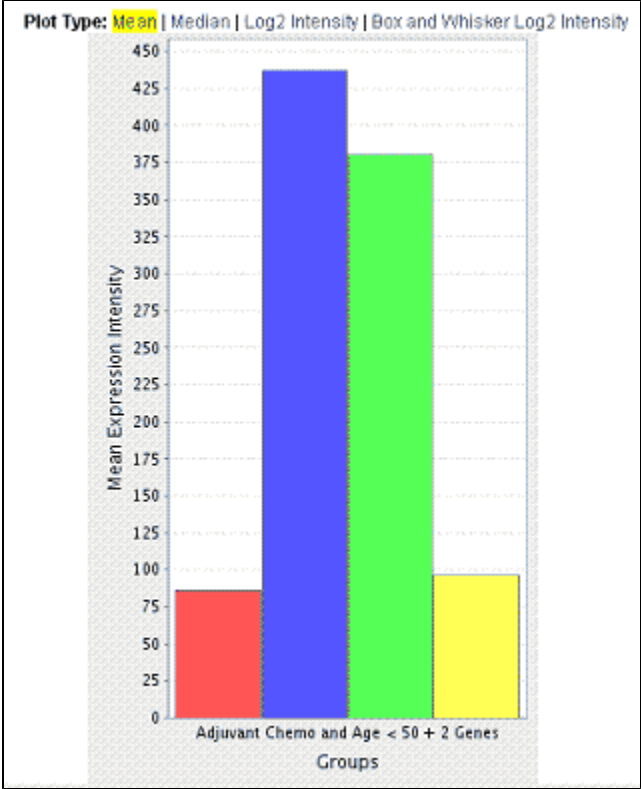
The **Mean** Gene Expression Plot shown in the following figure displays mean expression intensity (Geometric mean) versus Groups.
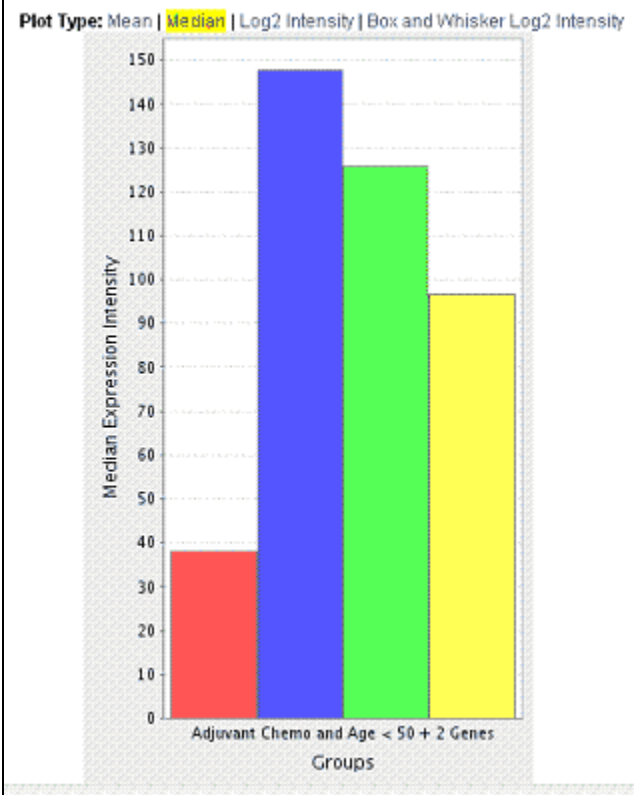
The **log2 intensity** Gene Expression Plot, shown in the following figure, displays average expression intensities for the gene of interest based on Affymetrix GeneChip arrays (U133 Plus 2.0 arrays).
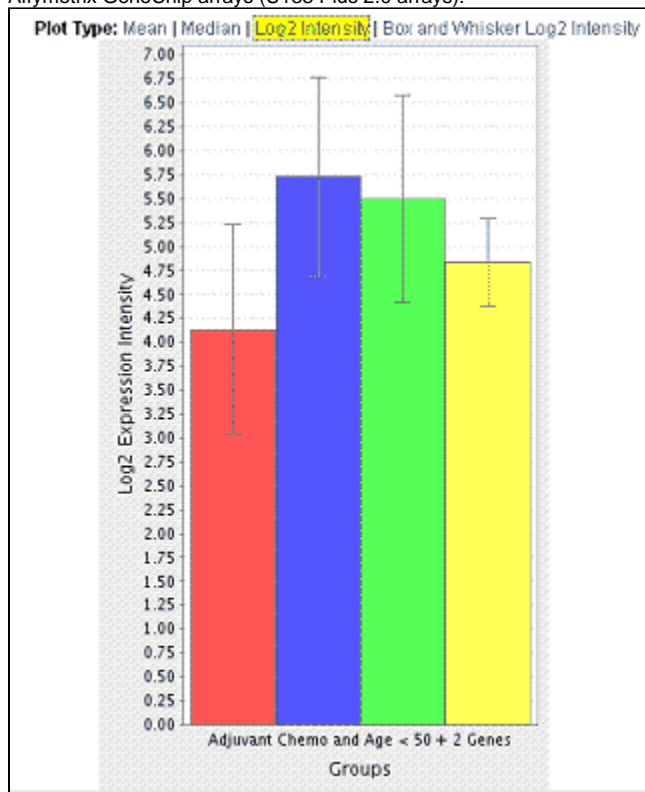
The box and whisker log2 expression intensity plot displays a box plot. A box and whisker plot can display the following data:

- 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

The box and whisker plot in the following figure is based on the same data set represented in the three previous figures.
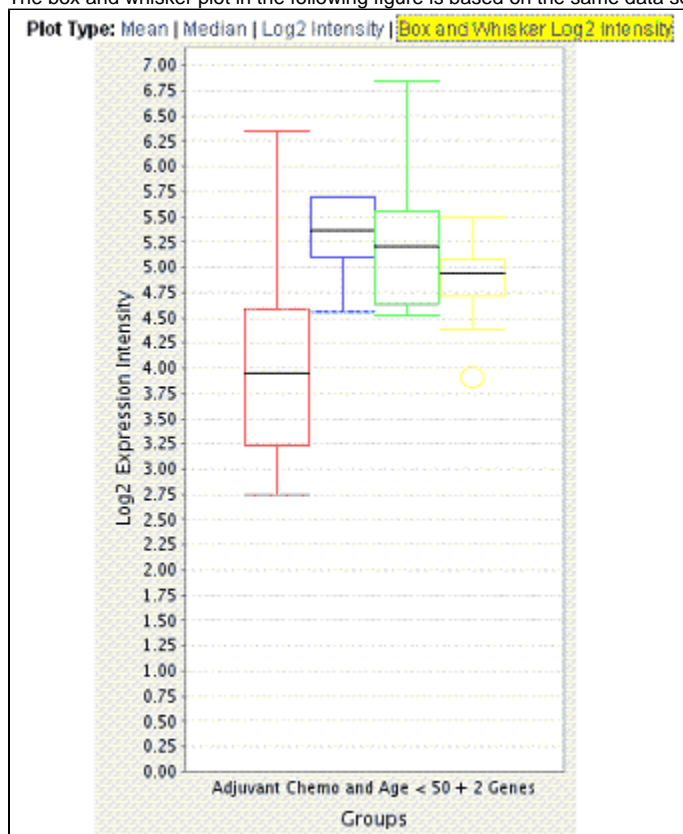
In descriptive statistics, a box plot or boxplot, also known as a box-and-whisker diagram or plot, is a convenient way of graphically depicting groups of numerical data through their five-number summaries (the smallest observation excluding outliers, lower quartile [Q1], median [Q2], upper quartile [Q3], 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 * IQR greater than Q3 and less than 1.5 * IQR than Q1. Outliers, if present, are shown as open circles, shown in the following figure.
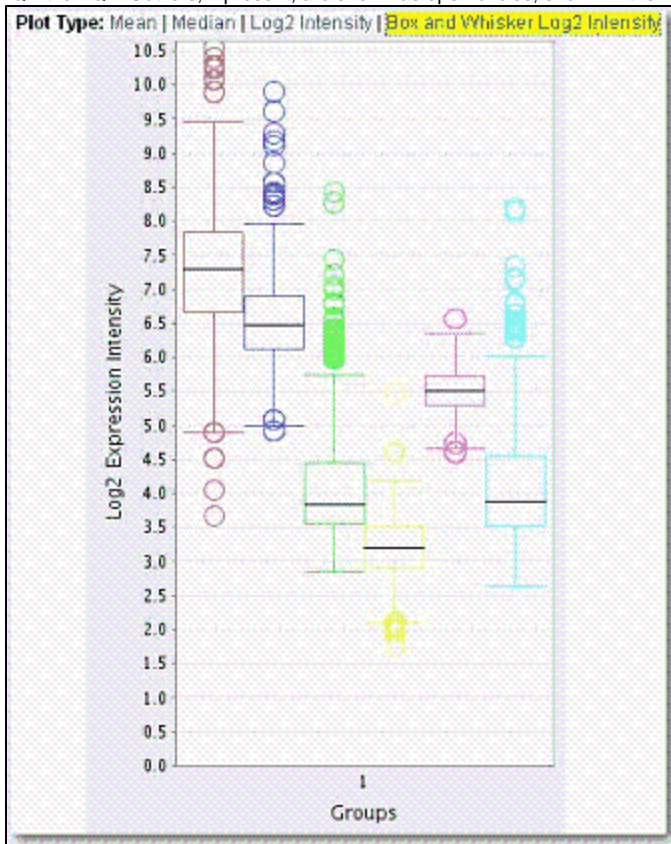
Boxplots can be useful to display differences between populations without making any assumptions of the underlying statistical distribution: they are non-parametric. The spacings between the different parts of the box help indicate the degree of dispersion (spread) and skewness in the data.

Choosing Genes

To obtain gene names for a gene expression search or analysis, use one of the following three methods described in this section: bioDBnet, Gene List or CGAP.

- **bioDBnet**– This link searches bioDBnet for gene IDs, symbols or genes within pathways. Then caIntegrator pulls identified genes into the application for analysis.
    1. Click **bioDBnet**.
    2. Enter **Search Terms**. Note that caIntegrator can perform a search on a partial HUGO symbol. For example, as search using *ACH** would find matches with 'achalasia' and 'arachidonate'.
    3. Select if you want to search in **Gene IDs**, **Gene Symbols, Gene Aliases, Pathways** (from the drop-down list), or **Search Pathways for Genes**.
        - **Gene IDs** searches the exact gene ID(s) you enter.
        - **Gene Symbols** searches only the Unigene and HUGO gene symbols in bioDBnet.
        - **Gene Aliases** searches for one or more gene symbols which are synonymous for the current gene symbol.
        - **Pathways** searches only the pathway names in bioDBnet.
        - **Search Pathways for Genes** searches for pathways containing gene(s) you specify for the search.
    4. Select **Show only genes that are part of this (caIntegrator) study** or **Case Sensitive Search** if either of these criteria are to be applied to the search. (By default, the search is case insensitive.)

5. Choose the **Taxon** from the drop-down list and click **Search.** (The Taxon criterion defaults to Human.) The search results display on the same page below the search criteria. The following figure shows search criteria and a few of the listed search results.

6. In the search results, use the check boxes to identify the genes whose symbols you want to use in the gene expression analysis.
7. Click **Use Genes** at the bottom of the page. This pulls the checked genes into the Gene Symbol text box on the Criteria tab. The following figure reveals some of the genes pulled into the Gene Symbol text box.

- **Gene List**– This link locates gene lists saved in caIntegrator.

  1. Click the Genes List icon ( ) to open a Gene List Picker dialog. For more information, see Creating a Gene or Subject List. If a GISTIC analysis has been run, you may see the following options:
     - GISTIC Amplified genes is a list of gene symbols in which the corresponding regions of the genome are significantly amplified.
     - GISTIC Deleted genes is a list of gene symbols in which the corresponding regions of the genome are significantly deleted.
  2. In the drop-down menu that lists previously saved gene lists, 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.
  3. Click **Use Genes** at the bottom of the dialog. This pulls the checked genes into the Search Criteria tab.

- **CGAP** – Use this directory to identify genes. Before clicking the CGAP icon ( ) you must enter gene symbols in the text box. This link does not pull anything into caIntegrator but does provide information about the gene(s) whose names you entered.

Analyzing Data with GenePattern

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

Information is included in this section for connecting to GenePattern from caIntegrator. Directions for launching specific GenePattern tools from caIntegrator are included as well, but you may want to refer to additional GenePattern documentation .

You have two options for using GenePattern from caIntegrator:

- Option 1 – Use the web-interface of any available GenePattern instances. To use the public instance from Broad, first register for an account. In caIntegrator, enter the URL for connecting: http://genepattern.broadinstitute.org/gp/services/Analysis, then enter your user ID and password.
- Option 2 – Use GenePattern on the grid.

The GenePattern feature in caIntegrator currently supports three analyses on the grid: Comparative Marker Selection (CMS) Analysis, Principal Component Analysis (PCA) and GISTIC-Supported Analysis.
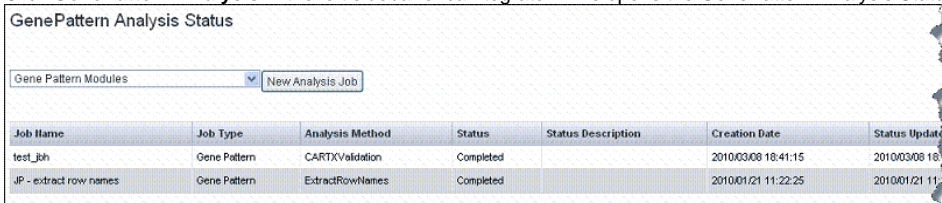
> ⊘ **Tip**
>
> If you are using the web interface to access GenePattern (option #1 listed above), then you can run other GenePattern tools in addition to CMS, PCA and GISTIC.

1. Select the study whose data you want to analyze in the upper right portion of the caIntegrator page.

2. Click **GenePattern Analysis** in the left sidebar of caIntegrator. This opens the GenePattern Analysis Status page, shown in the following figure.

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 Gene Pattern 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|.
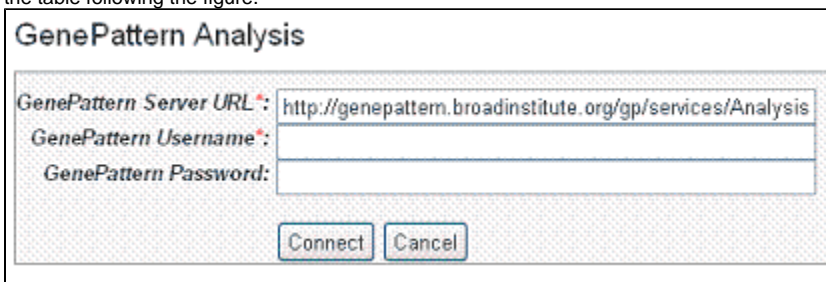4. Click the **New Analysis Job** button to open a corresponding page where you can configure the analysis parameters.

GenePattern Modules
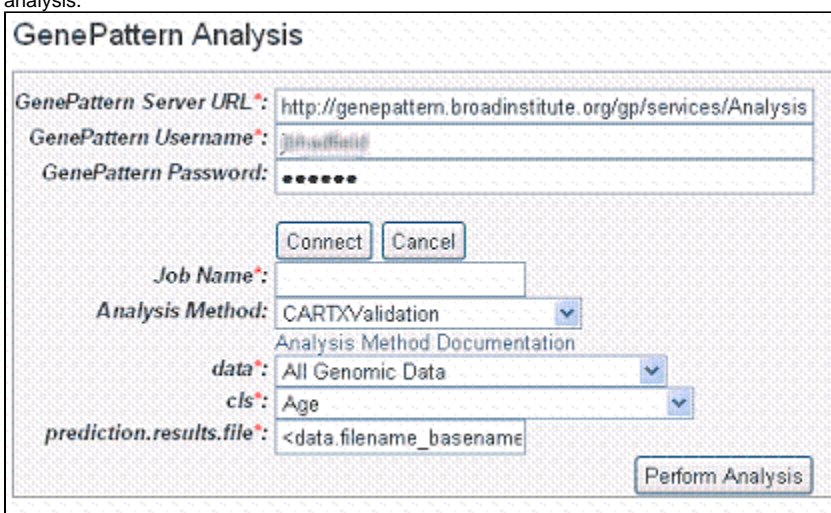
> ⓘ **Note**
>
> To launch the analyses described in this section, you must have a registered GenePattern account. For more information, see http://genepattern. broadinstitute.org/gp/pages/login.jsf 🖾 .

1. To configure the link for accessing GenePattern from caIntegrator, open the appropriate page as described in Analyzing Data with GenePattern.
2. Select the study whose data you want to analyze in the upper right portion of the caIntegrator page.
3. Click **GenePattern Analysis** in the left sidebar of caIntegrator. This opens the GenePattern Analysis Status page.
4. Make sure **GenePattern Modules** is selected in the drop down list. Click **New Analysis** Job.
5. In the GenePattern Analysis dialog box, shown in the following figure, specify connection information and click **Connect**. Fields are described in the table following the figure.

Field	Description
Server URL	Enter any GenePattern publicly available URL, such as http://genepattern.broadinstitute.org/gp/pages/login.jsf 🖾 .
GenePattern Username	Enter your GenePattern user name.
GenePattern Password	Enter your GenePattern password.

6. After logging in with the GenePattern profile, the dialog box, shown in the following figure, expands to include fields for defining the GenePattern analysis.

GenePattern Analysis
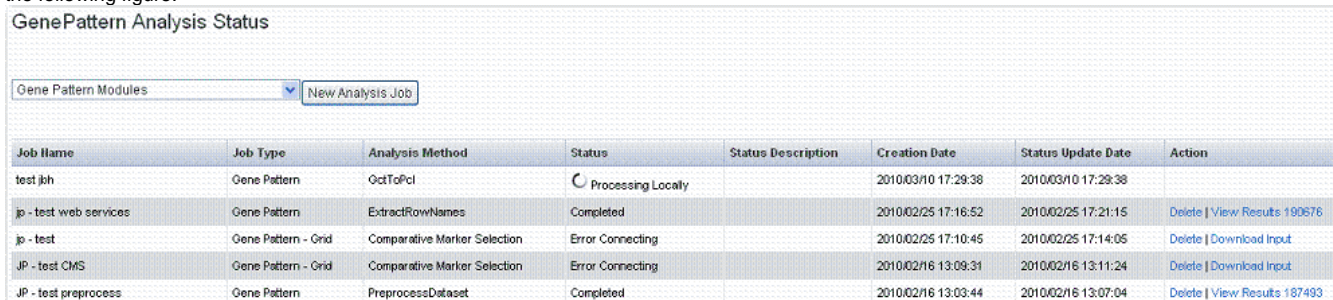
Field	Value
GenePattern Server URL*:	http://genepattern.broadinstitute.org/gp/services/Analysis
GenePattern Username*:	[obscured]
GenePattern Password:	••••••

[Connect] [Cancel]

Job Name*:	
Analysis Method:	CARTXValidation
	Analysis Method Documentation
data*:	All Genomic Data
cls*:	Age
prediction.results.file*:	<data.filename_basename

[Perform Analysis]

7. Enter information for the following fields. Fields with a red asterisk are required.

Field	Description
Job Name*	Enter a unique name for the analysis
Analysis Method	Select any method from the drop down list. Click Analysis Method Documentation for descriptions of the different analysis methods.
Data	All genomic data is selected by default. Select from the list any list that has been created for this study.
cls	Select any annotation field. The CLS file format defines phenotype (class or template) labels and associates each sample in the expression data with a label. It uses spaces or tabs to separate the fields. The CLS file format differs somewhat depending on whether you are defining categorical or continuous phenotypes: *Categorical labels* define discrete phenotypes; for example, normal vs tumor). *Continuous phenotypes* are used for time series experiments or to define the profile of a gene of interest (gene neighbors). ⓘ **Note** Most GenePattern modules are intended for use with categorical phenotypes. Therefore, unless the module documentation explicitly states otherwise, a CLS file should define categorical labels.
prediction. results. file	Enter the name of this file which is part of the output from a GenePattern module.

8. Click **Perform Analysis**. Based on the analysis method you select, you may be asked to add more information for the analysis. For more information, refer to the GenePattern Help site: http://www.broadinstitute.org/cancer/software/genepattern/tutorial/gp_concepts 🖾 .

Once the analysis is launched, caIntegrator returns to the GenePattern Analysis Status page where you can monitor the status of your current study, which is listed in the Analysis Method column. You can also view information about other GP analyses that have been run on this study. An example displays in the following figure.

GenePattern Analysis Status

Gene Pattern Modules [▼] [New Analysis Job]

Job Name	Job Type	Analysis Method	Status	Status Description	Creation Date	Status Update Date	Action
test jbh	Gene Pattern	GctToPcl	◯ Processing Locally		2010/03/10 17:29:38	2010/03/10 17:29:38	
jp - test web services	Gene Pattern	ExtractRowNames	Completed		2010/02/25 17:16:52	2010/02/25 17:21:15	Delete	View Results 190676
jp - test	Gene Pattern - Grid	Comparative Marker Selection	Error Connecting		2010/02/25 17:10:45	2010/02/25 17:14:05	Delete	Download Input
JP - test CMS	Gene Pattern - Grid	Comparative Marker Selection	Error Connecting		2010/02/16 13:09:31	2010/02/16 13:11:24	Delete	Download Input
JP - test preprocess	Gene Pattern	PreprocessDataset	Completed		2010/02/16 13:03:44	2010/02/16 13:07:04	Delete	View Results 187493

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 analyses within GenePattern itself, you may be able to view results in the GenePattern visualization module. Click **View Results** on the row where the results are listed. If you run them on the grid from caIntegrator, 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 GenePattern 🗗 .

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 caIntegrator page. You must select a study saved as a subject annotation study, but which has genomic data.
2. Click **GenePattern Analysis** in the left sidebar of caIntegrator. 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, shown in the following figure.

4. Select or define CMS analysis parameters, described in the following table. 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 caIntegrator 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 caIntegrator will use the selected server for this portion of the processing.
Annotation Queries and Lists*	All subject annotation queries and gene lists with appropriate data for the analysis are listed. Select and move two or more queries from the **All Available Queries** panel to the **Selected Queries** panel using the **Add >** and **Remove <** buttons. **Note:** The [SL] and [Q] prefixes to list names indicate "Subject Lists" or "Saved Queries". A "G" in the prefix indicates the list is Global. For more information, see Creating a Gene or Subject List.

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; default setting is "Yes".
Number of Columns Above Threshold*	Remove row if n columns are not >= than the given threshold In other words, the module can remove rows in which the given number of columns does not contain a value greater or equal to a user defined 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.
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.
Smooth P-values	Whether to smooth P-values by using the Laplace's Rule of Succession. By default, Smooth P-values 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.

5. When you have completed the form, click **Perform Analysis**.
caIntegrator takes you to the JobStatus/Launch page where you will see the job and its status in the Status column of the list, shown in the following figure.

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.

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 .

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 caIntegrator page. You must select a study with gene expression data.
2. Click **GenePattern Analysis** in the left sidebar of caIntegrator. This opens the GenePattern Analysis Status page.
3. Select **Principal Component Analysis (Grid Service)** from the drop down list and click **New Analysis Job**. This opens the Principal Component Analysis page, shown in the following figure.

4. Select or define PCA analysis parameters, described in the following table. An asterisk indicates required fields. You must enter a job name and select an annotation query, but you can accept the default settings for other options.

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 caIntegrator will use the selected server for this portion of the processing.
Annotation Queries*	All annotation 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.
Cluster By*	Selecting rows looks for principal components across all expression values, and selecting columns looks for principal components across all samples.

5. If you want to preprocess the data set, click **Enable the Preprocess Dataset**. This opens an additional set of parameters, shown in the following figure and described in the following table. The preprocessing is executed prior to running the 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 caIntegrator will use the selected server for this portion of the processing.
Filter Flag	Variation filter and thresholding flag
Preprocessing Flag	Discretization and normalization flag
	Minimum fold change for filter

Min Change	
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 >= than the given threshold

6. When you have completed the form, click **Perform Analysis**.
7. 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

> ⓘ **Note**
>
> The GISTIC test option displays only if the study contains copy number or SNP data. For more information, see Configuring Copy Number Data.

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 GenePattern Module documentation ⧉ .

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 caIntegrator page. You must select a study with copy number (either Affymetrix SNP or Agilent Copy Number) data.
2. Click **GenePattern Analysis** in the left sidebar of caIntegrator. 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, shown in the following figure.

4. Select or define GISTIC analysis parameters, described in the following table. You must indicate a Job Name, but you can accept the other default settings, which are valid and should produce valid results. Asterisks identify required fields.

GISTIC Parameters	Description
Job Name*	Assign a unique name to the analysis you are configuring.
GISTIC Service Type*	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
GenePattern User Name /Password	Include these to log into GenePattern for the analysis.
Annotation Queries and Lists	All annotation queries display in this list as well as an option to select all non-control samples. Select an annotation 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.
Select Platform	This option appears only if more than one copy number platform exists in the study. Select the appropriate platform from the drop-down list.
Exclude Sample Control Set*	From the drop-down list, select the name of the control set you want to exclude from the analysis. Click **None** if that is applicable.
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).
https://wiki.nci.nih.gov/pages /editpage.action? pageId=48726029# **cnv File**	This selection is optional. Browse for the file. There are two options for the CNV file. **Option #1** enables you to identify CNVs by marker name. Permissible file format is described as follows: 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: *Marker Name* *CNV Identifier* **Option #2** enables you to identify CNVs by genomic location. Permissible file format is described as follows: 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: *CNV Identifier* *Chromosome* *Narrow Region Start* *Narrow Region End* *Wide Region Start* *Wide Region End*

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.

Additionally, upon completion of a successful GISTIC anaylsis, caIntegrator automatically displays the two gene lists that it generates in the Gene List Picker so that you can use them in a caIntegrator query or plot calculation. The lists are visible only to your userID. For more information, see Choosing Genes. The genes will also display in **Saved Copy Number Analyses** in the left sidebar. See Editing a GISTIC Analysis.

> ⊘ **Caution**
>
> If samples from a copy number source are deleted, the GISTIC job in which they are appear is also deleted.

Editing a GISTIC Analysis

To view a GISTIC analysis page in caIntegrator where you can review or edit analysis parameters and results, under **Study Data** in the left sidebar, click **S aved Copy Number Analysis**. Select the analysis you want to open. The system displays analysis parameters and gene lists that that were retrieved from the analysis, as shown in the following figure.

Edit GISTIC Analysis

Click **Rename or Delete** to perform editing on this GISTIC Analysis.

Gistic Analysis Name:	jbh-text GISTIC
GISTIC URL used:	http://genepattern.broadinstitute.org/gp/services/Analysis
Amplifications Threshold:	0.1
Deletions Threshold:	0.1
Genome Build:	hg18_with_miR_20080407.mat
Join Segment Size:	4
Query or List Name:	jbh CN chr7
QV Threshold:	0.25
Samples used in analysis calculation:	E09471 E09948 E09896 E09945 E09800
Amplified Genes:	7A5 A2BP1 A2BP1 A2M AATK
Deleted Genes:	ZRANB3 ZSCAN29 ZSWIM2 ZZEF1 ZZZ3

Rename   Delete   Cancel

⊘ **Tip**

In the context of copy number data, 'Amplified genes', shown in this figure, refers to a list of gene symbols in which the corresponding regions of the genome are significantly amplified. 'Deleted genes', also shown in this figure, is a list of gene symbols in which the corresponding regions of the genome are significantly deleted.

From this page you can rename or delete the analysis.

- To rename the analysis, click the **Rename** button.
- To delete the analysis, click the **Delete** button.

As long as you leave this analysis in the study, caIntegrator lists the genes retrieved from the analysis in the Gene Picker dialog box when you open it.

See also Creating a Gene or Subject List and Editing a Gene or Subject List.

Viewing Data with the Integrative Genomics Viewer

Once you have run a query for gene expression, or have run analyses for copy number, or analyses for genomic data, you can view results in the Integrative Genomics Viewer (IGV).

The IGV is a high-performance visualization tool for interactive exploration of large, integrated datasets. It supports a wide variety of data types including sequence alignments, microarrays, and genomic annotations.

ⓘ **IGV information**

For more information about the Integrative Genomics Viewer or to connect independently to the IGV home page, see Integrative Genomics Viewer ⮹ .

You may also want to refer to the IGV User Guide ⮹

There are two ways to integrate caIntegrator with the IGV. To configure the connection to IGV, follow one of these methods.

Method 1 IGV

1. With the appropriate study open, at the bottom of the Query Results page, click the **View in Integrative Genomics Viewer** button.
2. If you click the button at the bottom of the page with any of the query results line items selected, caIntegrator creates IGV files, with a monitor informing you of this. After the files are created, click the **Launch Integrative Viewer** hypertext link.
3. Follow the instructions through the intermediate dialog boxes. After clicking **Open** with the Java program listed, the IGV.jnlp opens, displaying the dataset in the computer screen. An example displays in the following figure.
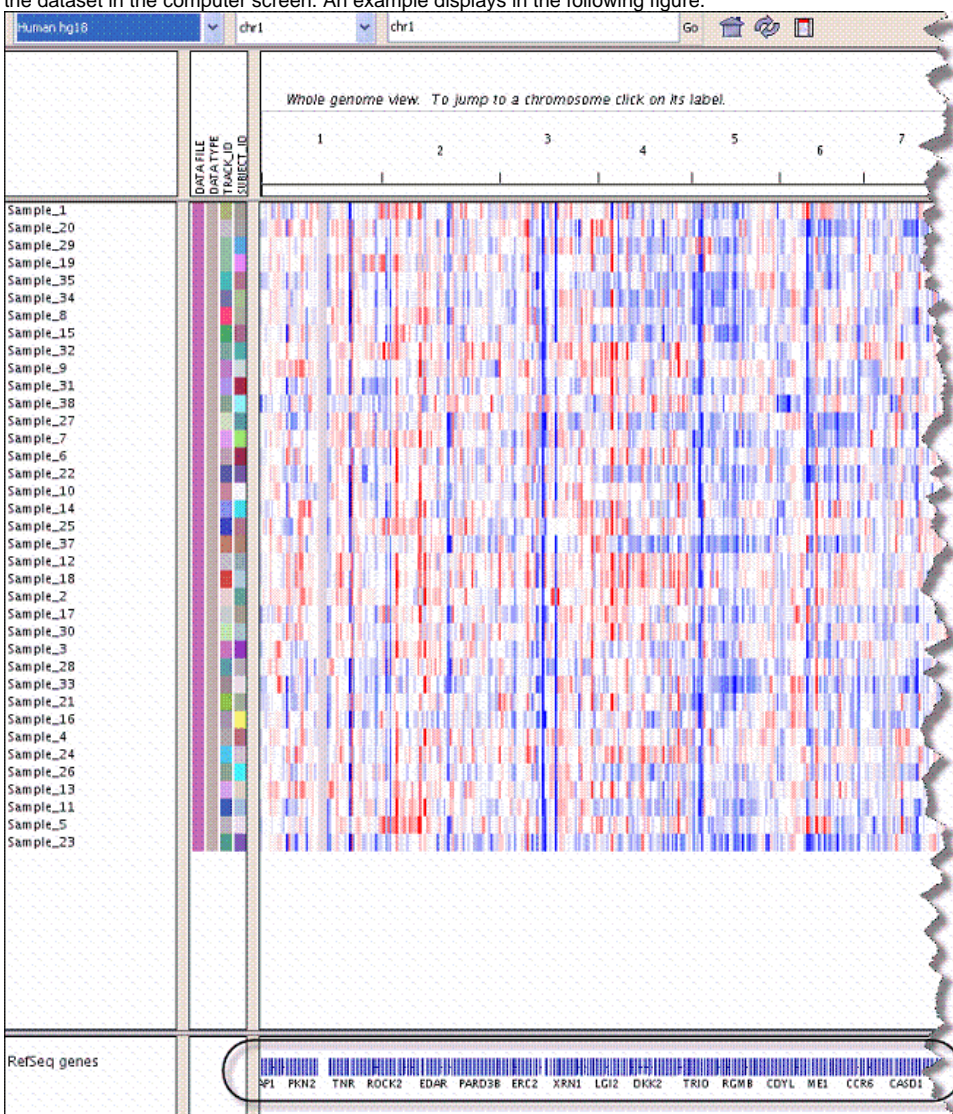
4. Move your mouse to hover over the genes graphic at the bottom of the page, indicated in the figure.

5. Click the mouse when you've identified a gene of interest.
   This opens the genome site at UCSC, where you can learn more about the gene. The following figures exhibits the kind of metadata you can expect from the UCSC genome site.

| Home | Genomes | Genome Browser | Blat | Tables | Gene Sorter | PCR | Session | FAQ | Help |

Human Gene GABRG1 (uc003gxb.1) Description and Page Index

Description: gamma-aminobutyric acid A receptor, gamma 1
RefSeq Summary (NM_173536): The protein encoded by this gene belongs to the ligand-gated ionic channel family. I
and opening an integral chloride channel. This gene is clustered with three other family members on chromosome 4. [provid
Strand: - Genomic Size: 88296 Exon Count: 9 Coding Exon Count: 9

Page Index	Sequence and Links	UniProtKB Comments	Genetic Associations	CTD		Microarray
RNA Structure	Protein Structure	Other Species	GO Annotations	mRNA Descriptions	Pathways	
Other Names	Model Information	Methods				

− Sequence and Links to Tools and Databases

Genomic Sequence (chr4:45,732,544-45,820,839)		mRNA (may differ from genome)	Protein (465 aa)	
Gene Sorter	Genome Browser	Protein FASTA	Proteome Browser	VisiGene	Table Schema
CGAP	Ensembl	Entrez Gene	ExonPrimer	GeneCards	GeneNetwork
Gepis Tissue	H-INV	HGNC	HPRD	HuGE	Jackson Lab
OMIM	PubMed	Stanford SOURCE	Treefam	UniProtKB	User annotations

− Comments and Description Text from UniProtKB

ID: GBRG1_HUMAN
DESCRIPTION: RecName: Full=Gamma-aminobutyric acid receptor subunit gamma-1; AltName: Full=GABA(A) rece
FUNCTION: GABA, the major inhibitory neurotransmitter in the vertebrate brain, mediates neuronal inhibition by binding
SUBUNIT: Generally pentameric. There are five types of GABA(A) receptor chains: alpha, beta, gamma, delta, and rho
SUBCELLULAR LOCATION: Cell junction, synapse, postsynaptic cell membrane; Multi-pass membrane protein. Cell i
PTM: May be palmitoylated (By similarity).
MISCELLANEOUS: This subunit carries the benzodiazepine binding site.
SIMILARITY: Belongs to the ligand-gated ionic channel (TC 1.A.9) family.
WEB RESOURCE: Name=Protein Spotlight; Note=Forbidden fruit - Issue 56 of March 2005; URL="http://www.ex

− Genetic Association Studies of Complex Diseases and Disorders

Genetic Association Database: GABRG1
CDC HuGE Published Literature: GABRG1

− Comparative Toxicogenomics Database (CTD)

The following chemicals interact with this gene

- D004026

Method 2 IGV

1. With the appropriate study open, click **Integrative Genomics Viewer** on the left sidebar. This opens the View IGV Selector page, shown in the following figure.

2. In the drop-down list, select the **Gene Expression Platform** for the data you want to view.
3. Select the **Copy Number Platform** ID.
4. The Annotations - Default panel displays existing annotation fields for the gene expression data in the open study. Select those fields you want to view when you open the IGV. Use the buttons for convenience if you want to **Select All** or **Unselect All**, when all are checked.
5. Click **View** to see the data in the Integrative Genomic Viewer. caIntegrator creates IGV files of the data.
6. After the files are created, click the **Launch Integrative Viewer** hypertext link that appears.
7. Continue with Step 3 in Method 1 IGV.

Viewing Data with Heat Map Viewer

Once you have run a query for gene expression, or for copy number, or have run analyses on genomic data, you can view results in the Heat Map Viewer (HMV).

> ### ⓘ HMV information
>
> For more information about the Heat Map Viewer or to connect independently to the HMV home page, see Heat Map Viewer documentation or HMV documentation. The IGV viewer and the NCI Heat Map viewer both require you to install a version of Java containing Java Web Start. For more information, see Java for IGV and Heat Map Viewer..

There are two ways to integrate caIntegrator with the Heat Map Viewer. To configure the connection, follow one of these methods.

Method 1 HMV

1. With the appropriate study open, at the bottom of the Query Results page, click the **View in Heat Map Viewer** button.
2. If you click the button at the bottom of the page with any of the query results line items selected, caIntegrator creates HMV files, with a monitor informing you of this. After the files are created, click the **Launch Heat Map Viewer** hypertext link.

3. Follow the instructions through the intermediate dialog boxes. After clicking **Open** with the Java program listed, the runs, displaying the dataset in the computer screen. An example displays in the following figure.
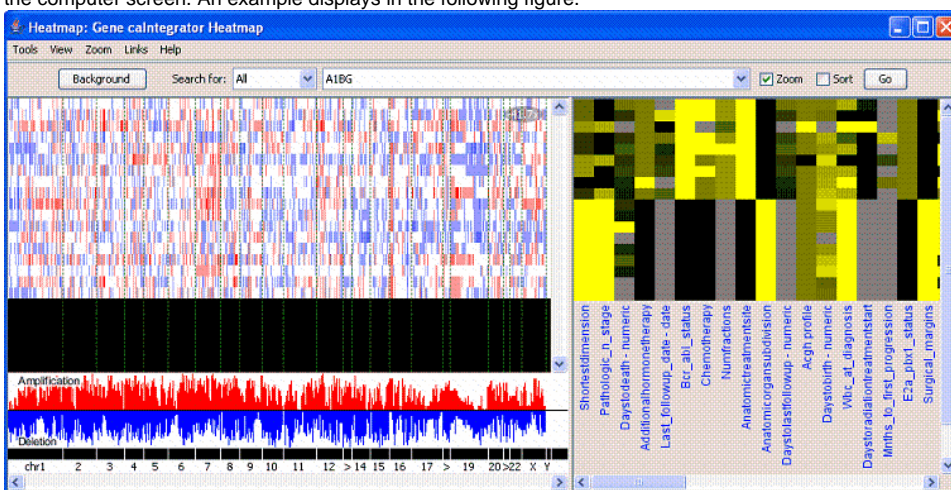
Method 2 HMV

1. With the appropriate study open, click **Heat Map Viewer** on the left sidebar. This opens the View Heat Map Viewer Selector page, shown in the following figure.

2. Select the appropriate **Copy Number Platform** in the drop down list.
3. The Annotations - Default panel displays existing annotation fields for the gene expression data in the open study. Select one or more annotations in the annotation list. For convenience, you can use the **Select All** or **Unselect All buttons**.
4. Click **View** to view the data you select in Heat Map Viewer. caIntegrator creates Heat Map Viewer files of the data.
5. After the files are created, click the **Launch Heat Map Viewer** hypertext link that appears.
6. Continue with Step 3 in Method 1 HMV.

> ⓘ **HMV help files**
>
> For interpretation of the results and using HMV features, see the help files opened from HMV.

Java for IGV and Heat Map Viewer

To use the IGV and the NCI Heat Map viewer, described in Viewing Data with the Integrative Genomics Viewer and Viewing Data with Heat Map Viewer, you must install a version of Java containing Java Web Start. You must install recent versions of the Java Development Kit (JDK 1.5.0 aka JDK 5.0 or newer) or Java Runtime Environment (JRE 1.5.0 aka JRE 5.0 or newer). The easiest option is to install JRE 5.0 ⎘

Without Java Web Start, when you click **Launch Integrative Genomics Viewer** or **Launch Heat Map Viewer**, a dialog box displays in your browser giving you the option to save or open with `igv.jnlp` (IGV) or `retrieveFile.jnlp` (HMV). Clicking the **Open** option starts the Java Web Start Launcher (default), installing the Java app so that you can view the files.

> ⓘ **Upon first launch**

The first time you launch the IGV or HMV with Java properly installed, regardless of browser type, a warning may appear: the "the digital signature cannot be verified". Click **Run** to proceed with opening the viewer.