

CAINTEGRATOR2 v.1.0

User's Guide



Center for Biomedical Informatics
and Information Technology

CREDITS AND RESOURCES

<i>Development</i>	<i>Quality Assurance</i>	<i>Documentation</i>	<i>Project and Product Management</i>
JP Marple ²	Tom Boal ⁴	JP Marple ²	Shine Jacob ⁶
Will Fitzhugh ²	Quy Phung ⁴	Jill Hadfield ¹	Anand Basu ¹
Eric Tavela ²			Juli Klemm ¹
TJ Andrews ⁵			
Ngoc Nguyen ⁴			
Matt Reh fuss ²			
Huaitian Liu ⁷			
Yuri Kotliarov ¹			
		<i>Training</i>	
Cuong Nguyen ²			
Deanna Siemaszko ³			
¹ NCI Center for Biomedical Informatics and Information Technology (CBIIT)		² 5AM Solutions	³ Terrapin Systems
⁴ NARtech	⁵ ScenPro	⁶ Enterprise Solutions And Consulting (ESAC)	⁷ Science Application International Corporation (SAIC)

<i>Contacts and Support</i>	
NCICB Application Support	http://ncicbsupport.nci.nih.gov/sw/ Telephone: 301-451-4384 Toll free: 888-478-4423

TABLE OF CONTENTS

Credits and Resources	i
Using the caIntegrator2 v.1.0 User's Guide	1
Introduction to the caIntegrator2 User's Guide	1
Organization of this Guide	1
User's Guide Text Conventions	2
Chapter 1	
Getting Started with caIntegrator2	5
About caIntegrator2	5
Registering as a New caIntegrator2 User	6
Logging In	8
Using the caIntegrator2 Workspace	8
caIntegrator2 Functions	9
Using Online Help	10
Logging Out	10
Application Support	11
Chapter 2	
Creating a New Study	13
Creating a Study – Overview	13
Configuring and Deploying Study	14
Creating/Editing a Study	15
Adding Clinical Data	16
Adding/Editing Genomic Data	24
Adding Imaging Data	29
Deploying the Study	31
Managing a Study	31
Managing Platforms	32

Chapter 3

Searching a caIntegrator2 Study35

Search Overview	35
Searching a Study	36
Results Type and Sorting Tabs	40
Managing Queries	42
Saving a Query	42
Editing a Query	43
Exporting Query Results	43

Chapter 4

Viewing Query Results45

Query Results Overview	45
Browsing Query Results	46
Clinical and Imaging Data	46
Genomic Data	46
Expanding Imaging Data Results	48
Relationship of Patient to Study to Series to Images	51
Exporting Data	52

Chapter 5

Analyzing Studies53

Data Analysis Overview	53
Creating Kaplan-Meier Plots	54
K-M Plot for Annotations	54
K-M Plot for Gene Expression	56
K-M Plot for Queries	58
Creating Gene Expression Plots	60
Gene Expression Value Plot for Annotation	61
Gene Expression Value Plot for Genomic Queries	64
Gene Expression Value Plot for Clinical Queries	66
Understanding a Gene Expression Plot	70
Analyzing Data with GenePattern	73
GenePattern Modules	75
Comparative Marker Selection (CMS) Analysis	76
Principal Component Analysis (PCA)	78
GISTIC-Supported Analysis	81

Appendix A

Data Import Configurations85

Subject Clinical Data Configuration	85
---	----

Delimited-Text Annotation Import	85
Annotation Field Configuration	86
Sample Data Configuration	86
Genomic Data Configuration	87
Imaging Data Configuration	87
Index	89

USING THE caINTEGRATOR2 v.1.0 USER'S GUIDE

This chapter introduces you to the *calIntegrator2 v.1.0 User's Guide* and suggests ways you can maximize its use.

Topics in this chapter include:

- [Introduction to the calIntegrator2 User's Guide](#) on this page
- [Organization of this Guide](#) on this page
- [User's Guide Text Conventions](#) on page 2

Introduction to the caIntegrator2 User's Guide

The *calIntegrator2 v.1.0 User's Guide* is the companion documentation to the calIntegrator2 software application. The user's guide includes information and instructions for the end user about using calIntegrator2.

Organization of this Guide

The *calIntegrator2 v.1.0 User's Guide* contains the following chapters and appendices:

Using the calIntegrator2 User's Guide — This chapter introduces you to the *calIntegrator2 v.1.0 User's Guide* and suggests ways you can maximize its use.

Chapter 1 Getting Started in calIntegrator2 — This chapter describes the processes for creating and managing studies in calIntegrator2.

Chapter 2 Creating a Study — This chapter describes the processes for creating and managing studies in calIntegrator2.

Chapter 3 Searching a calIntegrator2 Study — This chapter describes the processes for searching studies within calIntegrator2 using the search and browse tools.

Chapter 4 Viewing Search Results — This chapter describes search results that calIntegrator2 returns after queries.

Chapter 5 Analyzing Studies — This chapter describes how to use calIntegrator2 tools to analyze data in clinical or genomic studies that have been deployed in calIntegrator2.

Appendix A Exporting Data — This appendix describes how MAGE-TAB documents are parsed, validated and imported into caIntegrator2. It also provides examples of the types of MAGE-TAB documents that are expected by caIntegrator2.

Index—This section of the guide provides a complete index.

User's Guide Text Conventions

Table 2.1 illustrates how text conventions are represented in this guide. The various typefaces differentiate between regular text and menu commands, keyboard keys, toolbar buttons, dialog box options and text that you type.


Convention	Description	Example
Bold & Capitalized Command Capitalized command > Capitalized command	Indicates a Menu command Indicates Sequential Menu commands	Admin > Refresh
TEXT IN SMALL CAPS	Keyboard key that you press	Press ENTER
TEXT IN SMALL CAPS + TEXT IN SMALL CAPS	Keyboard keys that you press simultaneously	Press SHIFT + CTRL and then release both.
Monospace type	Used for filenames, directory names, commands, file listings, and anything that would appear in a Java program, such as methods, variables, and classes.	URL_definition ::= url_string
Icon	A toolbar button that you click	Click the Paste button () to paste the copied text.
Boldface type	Options that you select in dialog boxes or drop-down menus. Buttons or icons that you click.	In the Open dialog box, select the file and click the Open button.
<i>Italics</i>	Used to reference other documents, sections, figures, and tables.	<i>caCORE Software Development Kit 1.0 Programmer's Guide</i>
<i>Italic boldface monospace type</i>	Text that you type	In the New Subset text box, enter <i>Proprietary Proteins.</i>
Note:	Highlights a concept of particular interest	Note: This concept is used throughout the installation manual.
Warning!	Highlights information of which you should be particularly aware.	Warning! Deleting an object will permanently delete it from the database.

Table 2.1 caIntegrator2 User's Guide Text Conventions

Convention	Description	Example
{ }	Curly brackets are used for replaceable items.	Replace {root directory} with its proper value, such as c:\cabio

Table 2.1 caIntegrator2 User's Guide Text Conventions (Continued)

CHAPTER

1

GETTING STARTED WITH CAINTEGRATOR2

This chapter introduces general calIntegrator2 procedures and how to obtain help to use calIntegrator2.

Topics in this chapter include:

- [*About calIntegrator2*](#) on this page
- [*Registering as a New calIntegrator2 User*](#) on page 6
- [*Logging In*](#) on page 8
- [*Using the calIntegrator2 Workspace*](#) on page 8
- [*Using Online Help*](#) on page 10
- [*Logging Out*](#) on page 10
- [*Application Support*](#) on page 11

About calIntegrator2

NCI, Center for Biomedical informatics and Information Technology (CBIIT) is developing a novel translational informatics platform called calIntegrator that allows researchers and bioinformaticians to access and analyze clinical and experimental data across multiple clinical trials and studies. The calIntegrator framework provides a mechanism for integrating and aggregating biomedical research data and provides access to a variety of data types (e.g. Immunohistochemistry (IHC), microarray-based gene expression, SNPs, clinical trials data, etc.) in a cohesive fashion.

calIntegrator2 is a web based or locally installed tool that allows integration of clinical data with genomic and/or imaging data. The calIntegrator2 user can import data of various types in a predefined flat format, and create new studies with multiple study

data. The user can update an existing study to add new attributes or to add/modify data. The user can also perform analyses on study data.

Registering as a New caIntegrator2 User

To request a caIntegrator2 user account, you must register as a new user, completing the following steps:

1. Go to the CBIIT caIntegrator2 login page <http://caintegrator2.nci.nih.gov> or use the URL provided by your System Administrator for the caIntegrator2 instance at your institution.
2. Click the **Register Now** hypertext link, under the caIntegrator2 login section in the upper left of the page. This opens the account registration form (*Figure 1.1*).

Register

Figure 1.1 New user account registration form

3. In the Register form, enter the appropriate information¹.

- **Security Information**

- **Do you have an LDAP account** [a user profile with your institution] at [NCICB or your institution]?

If **Yes**, enter your username and case-sensitive password for the purposes of verifying that it is correct. After you submit your request, you can continue to use caIntegrator2 without an account to browse and search available experiments and download data while your account is verified and activated.

—**Username***

—**LDAP Password***

1. Items with an asterisk or highlight are required.

–Requested role(s)* – Select one or more of the roles. Roles are described in [Table 1.1](#).

If your LDAP profile is not validated, calIntegrator2 indicates that the LDAP credentials do not check out. You are asked to reenter them, but you can choose to answer no, and the System Administrator will manually ensure you don't get a duplicate LDAP account during provisioning. You can **Cancel** or talk with your System Administrator about the problem.

If you select **No** [you do not have an LDAP account], the text boxes for entering the LDAP account information disappear. You must indicate the role you would like to be assigned in calIntegrator2, and continue entering the appropriate information in the **Account Details** section.

	<i>Description</i>	<i>Permissible 1.0 Actions</i>
Study Manager	Creates, owns and manages studies	Create studies Assign annotations studies Edit studies Search studies Perform analyses on study data
Study Investigator	Investigates and queries the study data	Query study data Save queries Analyze using K-M Plot Analyze using Gene Expression Plots Analyze using GenePattern
Platform Manager	Uploads and maintains the array platform annotation	Uploads array platform annotation files Ensures that if study requires a certain platform, then it is available within the installation

Table 1.1 calIntegrator2 role descriptions

◦ **Account Details**

- **First Name***
- **Last Name***
- **Email [address]***
- **Organization***
- **Address [Lines 1* and 2]**
- **City***
- **State***
- **Country***

- **Postal [or Zip] Code***
- **Phone***
- **Fax**

4. Click **Submit Registration Request** to execute the request, or click **Cancel** to abort the registration.

After registration is sent, the screen displays a confirmation message.

At this point, an email containing all of the information you specified in the new user request form is sent to the caIntegrator2 system administrator and an account request confirmation email is also sent to you, the prospective user, at your specified email address. In response, the caIntegrator2 system administrator uses UPT to create your user account and assign the requested roles (in predefined groups like Study Investigator). When your account is created, the system administrator sends you an email to alert you, after which you can login to caIntegrator2.

When your account is registered, the UserID and password you are assigned determines your access rights for the software.

Logging In

To log into caIntegrator2, follow these steps:

1. On the login page, enter your **username** and **password**.
2. Click the **Login** button. If your login is successful, the Welcome to Browse/Study page appears (*Figure 1.2*).

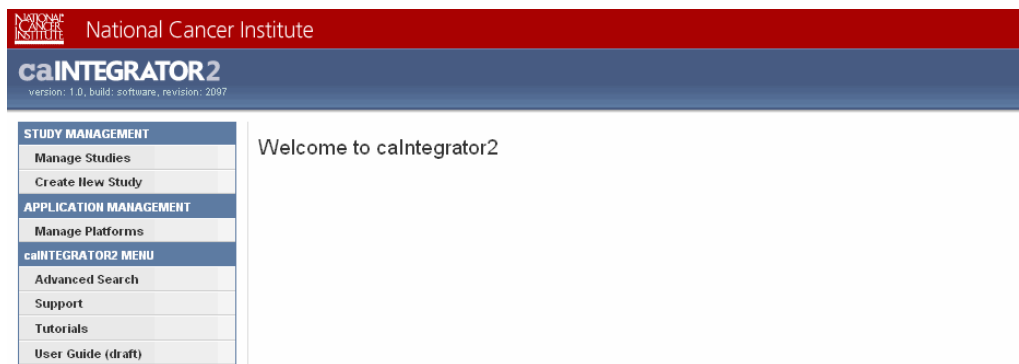


Figure 1.2 The caIntegrator2 workspace before any studies have been deployed

See also *Using the caIntegrator2 Workspace* on page 8.

Using the caIntegrator2 Workspace

The caIntegrator2 workspace enables quick access to all caIntegrator2 functions and information. To access caIntegrator2 functions, use the options listed on the left sidebar of the workspace.

caIntegrator2 Functions

When you log into caIntegrator2, before any studies have been created the workspace opens with a Welcome page, as shown in ([Figure 1.2](#)). Once a study is created, its name is listed at the top of the left sidebar.

[Table 1.2](#) describes each caIntegrator2 option in the workspace ([Figure 1.2](#)).

	Function
[Study Name]	When you log in, one study displays in the left sidebar by default. Any study that you select in the My Studies drop-down list in the upper right of the page replaces this default selection.
Home	Click this to return to the home page for the selected study.
Search [study name]	Click this option to open the Search [study name] page from which you can launch queries into your selected study. For more information, see Searching a caIntegrator2 Study .
Study Data	Click Queries > My Queries to open the list of previous queries you saved. Click any item in the list to open the saved query, which displays on the Criteria, Columns and Sorting tabs. From those tabs, you can modify criteria and/or launch the query again. For more information, see Saving a Query on page 42.
Analysis Tools	Click either of the listed options listed to open a page where you can launch an analysis of the data in the selected study. <ul style="list-style-type: none"> • Generate a K-M Plot. See Creating Kaplan-Meier Plots on page 54. • Generate a Gene Expression Plots. See Creating Gene Expression Plots on page 60. • Launch GenePattern Analysis. Analyzing Data with GenePattern on page 73.
Study Management	Click either of the listed options to manage the selected study through editing or deleting it or by creating a new study. <ul style="list-style-type: none"> • Click Manage Studies. See Managing a Study on page 31. • Click Create a New Study. See Configuring and Deploying Study on page 14.
Application Management	Click Manage Platforms to identify, add or remove platforms that caIntegrator2 supports . For more information, see Managing Platforms on page 32.
caIntegrator2 Menu	<ul style="list-style-type: none"> • Click Support to view contact information for Application Support. • Click Tutorials to view a tutorial to help you get started using caIntegrator2. • Click User Guide to open the caIntegrator2 v.1.0 User's Guide in PDF format.

Table 1.2 caIntegrator2 tabs

In the **My Studies** drop-down list in the upper right of the page, select the study you want to use for your current session. (The list includes all studies to which you are subscribed.) As you do so, the following left sidebar contents change to reflect options relevant to your study selection:

- the logo for the selected study (if it exists)

- the name for the selected study
- the list of saved queries for that study

Using Online Help

The online help explains how to use all of the features.

To access online help, click the help icon at the top of each page to open a context-sensitive topic. Context-sensitive help displays information that corresponds to the page from which help was opened.

Help opens displaying the table of contents in the left panel.

Once you are in online help, several buttons and/or options help you locate topics of interest.


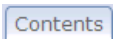
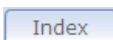
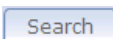
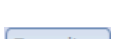



	<i>Description</i>
	Locates and highlights your current topic in the table of contents.
	Select a topic from the complete online help table of contents.
	Select a topic from the online help index.
	Perform word searches of Help by entering query text in the search text box.
	Create a list of your frequently-accessed topics.
 Related Topics 	Opens other closely related topics.
	Prints the current topic.
Topic Name > Topic Name	The breadcrumb trail shows the relative location of the current help topic relative to neighboring topics. Click a breadcrumb link to display that help topic.
Back Forward	Navigates through previously viewed topics.

Table 1.3 Online help tips

Logging Out

To log out of calIntegrator2, click the **logout** link in the upper right-hand corner of the page.

Application Support

For any general information about the application, application support or to report a bug, contact NCICB Application Support.

Email: ncicb@pop.nci.nih.gov	When submitting support requests via email, please include: <ul style="list-style-type: none">• Your contact information, including your telephone number.• The name of the application/tool you are using• The URL if it is a Web-based application• A description of the problem and steps to recreate it.• The text of any error messages you have received
Application Support URL	http://ncicb.nci.nih.gov/NCICB/support
Telephone: 301-451-4384 Toll free: 888-478-4423	Telephone support is available: Monday to Friday, 8 am – 8 pm Eastern Time, excluding government holidays.

CHAPTER 2

CREATING A NEW STUDY

This chapter describes the processes for creating and managing studies in caIntegrator2.

Topics in this chapter include:

- *Creating a Study – Overview* on this page
- *Configuring and Deploying Study* on page 14
- *Managing a Study* on page 31

Creating a Study – Overview

You can create a caIntegrator2 study by importing clinical study data, genomics data and imaging data, using a combination of spreadsheet/files and existing caGrid applications as source data. Each instance of caIntegrator2 can support multiple studies. As the manager creating a study, it is important that you understand the study well and that the data you wish to aggregate has been submitted to the applications whose data can be integrated in caIntegrator2.

- **Clinical** – The clinical data should be available in CSV files, with a unique patient identifier in one column, one patient per row. Other relevant data can be supplied in other columns to be identified as annotations in the file from within caIntegrator2. You, as the study creator, must have access to the clinical data file, as the file does not come from a caBIG[®] repository.
- **Genomic** – To use caIntegrator2 to integrate array data, the data should be imported into caArray, either locally or the CBIIT installation, using that system's data file import functionality. You must also have a mapping file in CSV format. This file indicates correlations between array files and the clinical subjects in the clinical data files. A mapping file consists of two columns: one with the patient ID, and one with the sample ID.

- **Imaging** – Imaging data should have been submitted to the NBIA grid node as public data, either locally or as part of the CBIIT installation. Image annotations, which includes information about images provided by radiologists or other researchers can include such information as tumor size, tumor location, etc. It must be in CSV format, with unique image series IDs in one column and annotation IDs in the second column. You must also have an image mapping file in CSV format. This file indicates correlations between clinical subjects or images in NBIA and clinical subjects in the clinical data files. A mapping file consists of two columns: one with the patient ID, and one with the NBIA image series ID in the other column.

As you create the study, you define its structure in the process, identifying the data sources and mapping the data between different source data. After the study has been created and deployed, the study can then be used to perform analyses.

Configuring and Deploying Study

Note: Only a user with a Study Manager role can create a study.

When you create a study, you must specify different data-types (clinical, array, image, etc), data sources (caGrid applications – caArray and NBIA) and map the data, (patient to sample, image series, etc.).

To create a new study, follow these steps:

1. In the Study Management section of the left sidebar, click **Create New Study**.
2. In the Create New Study dialog box that opens, provide a name and description for the study you are creating ([Figure 2.1](#)).

Create New Study

Figure 2.1 Create Study page

3. Click **Save**.

This opens an Edit Study page where you can add identify data files for your study.

Creating/Editing a Study

The Edit Study page displays the Name and Description that you entered for a new study, or for an existing study that you are editing ([Figure 2.2](#)).

Figure 2.2 Edit Study page

To continue creating a study or to modify a study, on the Edit Study page complete these steps:

1. Change the name and or description, if you so choose. Click **Save**.
Note: You can save the study at any point in the process of creating it. You can resume the definition and deployment process later.
2. If you choose to add a logo for the study, click the **Browse** button corresponding to **Logo File**. Navigate for the file, then click **Upload Now**. Once you save the study (or its edit), the logo displays in the center of the page ([Figure 2.3](#)). On the home page for the study, the logo displays in the upper left, above the sidebar.



Figure 2.3 Example of a logo added to the caIntegrator2 browser on the Edit Study page

To continue, you can add clinical data sources, genomic data or imaging data.

Adding Clinical Data

The Edit Study page opens after you save a new study or click to edit an existing study.

Note: To edit information for an existing study, follow the same basic directions in this section. Instead of entering new information, you will modify existing information.

To add or edit clinical metadata in this page, follow these steps:

1. On the Edit Study page, click the **Browse** button in the Clinical Data Sources section. Navigate to locate the file. Files must be in CSV file format.

In the Clinical Data Sources section, if a file has already been selected, its information displays in the varying fields.

2. Click **Add Clinical Data Source**. This opens the Define Fields for Clinical Data page ([Figure 2.4](#)).

Define Fields for Clinical Data

Study Name: jbh test		
Field Definition	Field Header from File	Data from File
Change Assignment	PateintID	X56123
Change Assignment	Age	45
GENDER Change Assignment	Gender	Male
Change Assignment	Diagnosis	Crohn's Disease
Change Assignment	Diagnosis Date	1/23/2007
Change Assignment	Karnofsky Score	80

Figure 2.4 Define Fields for Clinical Data page

The Field Header from File column on the Define Fields... page displays column headers taken from the source *CSV file. The page also displays data values in the file you have designated. You must map each column name to an existing

column name in the caIntegrator2 database or in caDSR. If it doesn't yet exist, you can create a custom column name ([Figure 2.5](#)).

	A	B	C	D	E	F	G	H
1	Pa	Age	Gender	Survival	Disease	Grade	Race	
2	ASP221	50-54	M		ASTROCYTOMA		WHITE	
3	ASP308	50-54	M		GBM		WHITE	
4	FPH113	20-24	M		UNKNOWN		WHITE	
5	FPH114	40-44	M		UNKNOWN		WHITE	
6	FPH118	55-59	M		GBM		WHITE	
7	FPH309	50-54	M		GBM		WHITE	
8	E09238	45-49	M	18-24M	GBM		WHITE	
9	E09239	25-29	M		UNKNOWN		WHITE	
10	E09262	35-39	M		ASTROCYTOMA		WHITE	
11	E09278	30-34	M		UNKNOWN		WHITE	
12	E09331	35-39	M		UNKNOWN		ASIAN NOS	
13	E09332	55-59	M		GBM		WHITE	
14	E09336	30-34	M		GBM		WHITE	
15	E09348	60-64	M		GBM		WHITE	
16	E09378	45-49	M		UNKNOWN		WHITE	
17	E09449	50-54	M		UNKNOWN		OTHER	
18	E09454	0-4	M		UNKNOWN		WHITE	
19	E09489	55-59	M		GBM		WHITE	
20	E09515	35-39	M		UNKNOWN		WHITE	
21	E09569	45-49	M		UNKNOWN		WHITE	
22	E09587	35-39	M		UNKNOWN		OTHER	
23	E09601	40-44	M		GBM		WHITE	
24	E09610	55-59	M		GBM		WHITE	
25	E09611	60-64	M		UNKNOWN		ASIAN NOS	
26	E09615	45-49	M		UNKNOWN		WHITE	
27	E09624	35-39	M		GBM		WHITE	
28	E09645	45-49	M		UNKNOWN		WHITE	
29	E09657	50-54	M		UNKNOWN		WHITE	
30	E09730	40-44	M		UNKNOWN		WHITE	

Figure 2.5 Example of a source CSV file whose data you are mapping in caIntegrator2

The MOST important steps in creating a new study in caIntegrator2:

- You MUST designate one column in the file as a unique “identifier” column type.
- You MUST review and define column annotation definitions for each column header in the file.

If caIntegrator “recognizes” the same column header in other files already in the system, a term, for example “age” or “survival”, which is the current definition appears in the **Field Definition** column above the blue **Change Assignment** link. If the area above the blue **Change Assignment** link is blank, no correlating term exists in the database; you must specify the field type, and then the term will populate the space.

3. To indicate the unique identifier of choice, on the row showing the column header (PatientID in the figure, but other examples are subject identifier, sample identifier, etc), click **Change Assignment** in the **Field Definition** column.

Assigning An Identifier or Annotation

When you click **Change Assignment** on the Define Fields... page, the Assign Annotation Definition for Column dialog box opens ([Figure 2.6](#)). On this page you can change the column type and the field definition for the specific data field you selected.

Note: When you change an assignment, you must make sure the data types match--numeric, etc.

Assign Annotation Definition for Column: PatientID

Column Type: Annotation

New

Save

Search For an Annotation Definition:

Search Search existing studies and caDSR for definitions.

Matching Annotation Definitions			
Name	CDE Public ID	Data Type	Definition
No matches found for your Search.			

Matches from caDSR					
Name	Actions	CDE Public ID	Context	Status	Definition
No matches found in caDSR for your Search.					

Figure 2.6 The Assign Annotation Definition dialog box

1. For the column (PatientID) that you choose to be the one and only Identifier column, in the **Column Type** drop-down list, select **Identifier**.

When the definition becomes an identifier, the rest of the page disappears, showing only the Identifier definition ([Figure 2.7](#)).

Assign Annotation Definition for Column: PATIENT_ID

Column Type: Identifier

Save

Figure 2.7 Identifier definition

Note: If you select **Annotation** after you have already selected **Identifier**, the rest of the page reappears.

2. Click **Save** to save the identifier. This returns you to the Define Fields for Clinical Data page where the Identifier is noted in the Field Definition column.
3. After you have defined which field is the Identifier, you must ensure that ALL other fields also have a field definition assignment. For those fields without a Field Definition assignment or for those whose Field Definition you want to review, click **Change Assignment**.
4. In the Assign Annotation Definition for Column: [column header] dialog box, select **Annotation** in the drop-down list.

As you select the column type, you can work with column headers in one of four ways in this dialog box.

- You can accept existing default definitions (those that are inherent in the data file you selected). See [Step 5](#).
 - You can create your own definitions manually. See [Step 6](#).
 - You can search for and use definitions in other calIntegrator2 studies. See [Searching for Annotation Definitions](#) on page 20.
 - You can search for and use definitions found in caDSR. See [Searching for Annotation Definitions](#) on page 20.
5. If there is anything you want to change about an existing annotation definition of the field such as its name, or if you want to view or edit its definition, click the **Change Assignment** link on the Define Fields... page. The Assign Definition

page opens, expanded now to include a Current Annotation Definition section above a section where you can still initiate a search for an annotation definition (Figure 2.8).

Note: If the column header you are working with already has a designated Field Definition, the Current Annotation Definition section of the Assign Annotation Definition for Column page is already visible when you open this dialog box.

Assign Annotation Definition for Column: SITE

Column Type: Annotation

Current Annotation Definition:

Name: SITE

Definition: Created via selenium for DC Lung Full on 07/15/09 13:54:57.

Keywords: SITE

Data Type: string

Permissible Values:

Non-Permissible	Permissible
	MSKCC
	MI
	HLM
	DFCI

Add >

< Remove

New

Save

Search For an Annotation Definition:

Search existing studies and caDSR for definitions.

Figure 2.8 Current Annotation Definition

- To enter a new name annotation, or any other information about the annotation definition, click the **New** button and enter the information described in Table 2.1

Annotation Field	Field Description
Name	Enter the name for the annotation.
Definition	Enter the term(s) that define the annotation.
Keywords	Insert keyword(s) that can be used to find the annotation in a search, separated by commas.
Data Type	Enter a string (default), numeric, or date

Table 2.1 Annotation fields for new definitions

Annotation Field	Field Description
Permissible/Non-permissible Values	<p>Note: The first time you load a file, before you assign annotation definitions (step 3 on page 17), these panels may be blank. If the column header for the data is already “recognizable” by calIntegrator2, the system makes a “guess” about the data type and assigns the values to the data type in the newly uploaded file. They will display in the Non-permissible values sections initially. Use the Add and Remove buttons to move the values shown from one list to the other, as appropriate.</p> <p>When you select or change annotation definitions by selecting matching definitions (described in Searching for Annotation Definitions on page 20), this may add (or change) the list of non-permissible values in this section.</p> <p>If you leave all values for a field in the Non-permissible panel, then when you do a study search, you can enter free text in the query criteria for this field.</p> <p>If there are items in the Permissible values list, then the values for this annotation are restricted to only those values. When you perform a study search, you will select from a list of these values when querying this field. If there are no items in the permissible values list then the field is considered free to contain any value.</p> <p>To edit a field's permissible values, you must change the annotation definition. You can do this even after a study has been deployed.</p>

Table 2.1 Annotation fields for new definitions

Searching for Annotation Definitions

An alternative to creating a new definition is to search for annotation definitions already present in calIntegrator2 studies or in caDSR.

1. Enter search keyword(s) in the **Search** text box on the Assign Annotation Definition page. Click **Search**. After a few moments, the search results display on the page (*Figure 2.9*).

Search For an Annotation Definition:

age Search existing studies and caDSR for definitions.

Matching Annotation Definitions from caIntegrator2

Name	CDE Public ID	Data Type	Definition
Age		string	Created via selenium test for Rembrandt / VASARI.
Age		string	

Matching Annotation Definitions from caDSR

Name	Actions	CDE Public ID	Context	Status	Definition
Weight Person Age 20 Number	Select View	2443544	PS&CC	RELEASED	The numeric value to represent a person's weight in pounds at age 20 years.
Pregnancy Age Birth First Child Category	Select View	2442895	PS&CC	RELEASED	The age of the participant at the birth of her first child.
Weight Person Age 50 Number	Select View	2443527	PS&CC	RELEASED	The numeric value to represent a person's weight in pounds at age 50 years.
Uterus Removed Age Category	Select View	2442944	PS&CC	RELEASED	The age the participant had her uterus or womb removed.
Demographics Patient Diagnosis Age Value	Select View	2660065	PS&CC	RELEASED	The number that represents the age of the patient at diagnosis in complete years.
End-over-ride A Site	Select View	0697798	PS&CC	RAFT NEW	Some computer edits identify errors. Others indicate possible errors that remain.

Figure 2.9 Results for annotation definition search

2. To view the definitions corresponding to any of the “Matching Annotation Definitions”, which are those currently found in other caIntegrator2 studies, click the [term], such as “age”, hypertext link. The definition then appears in the Current Annotation Definition segment of the page just above.

In summary, when you click the link, that assigns the definition to the Define Fields for Clinical Data page, and it also closes the Annotation Definition page.

You can modify any portion of the definition, as described in [step 6](#) on page 19.

3. The matches from caDSR display some of the details of the search results. To view more details of a match, such as permissible values, click **View**, which opens caDSR to the term. If you click **Select**, the caDSR definition automatically replaces the annotation definition for this field with which you are working.

Caution: Take care before you add a caDSR definition that it says exactly what you want. caDSR definitions can have minor nuances that require specific and limited applications of their use.

4. Once you have settled on an appropriate field definition for the annotation, click **Save**. This returns you to the Define Field for Clinical Data page.

Note: If you have not clicked **Select** for alternate definitions in this dialog box, then click **Save** to return to the Define Field...dialog box without making any definition changes.

5. From the Define Fields for Clinical Data page, be sure and designate the annotations for each field in the file. Click **Save** on each page to save your entries or click **New** to clear the fields and start again. You will not be able to proceed until every Field Definition entry on the Fields for Clinical Data screen has a unique entry, one as an Identifier and the remainder as annotations.

The Data From File columns on the page display the column header *values* of the first three rows you designated as “annotations”.

6. Click **Done**. This saves the study by name and description, but does not deploy the study. See [Deploying the Study](#) on page 31.

Saving the study returns you to the Edit Study page where a “Not Loaded” status now appears for the file whose annotations (column headers) you have defined ([Figure 2.10](#)).

Type	Description	Status	Action
DELIMITED_TEXT	dc_lung_clinical_data.csv	Not Loaded	

Figure 2.10 Example file whose annotations have been defined

7. Click the **Load Clinical** link in the Action section to load the data file you configured. At this point, the Status changes to “Loaded”.

Note: You can add as many files as are necessary for a study. Patients 1-20 in first file, 21-40 in second file, or many patients in first file and annotations in second file, etc. As long as IDs are defined correctly, it works.

8. Once you have assigned data types to every column header in the data file and have loaded the clinical data, click **Save and Deploy**. At that point, calIntegrator2 loads data from the file to the calIntegrator2 database.

Note: You can change assignments even after the study is deployed, using the Edit feature. For more information, see [Creating/Editing a Study](#) on page 15.

The Manage Studies page opens when the study is deployed. The **Deployed** status is indicated on the Manage Studies page as well as the Edit Study page. For more information, see [Managing a Study](#) on page 31.

You can continue to perform other tasks in calIntegrator2 while deployment is in process.

See also [Deploying the Study](#) on page 31.

Note: You can repeatedly upload additional or updated subject annotations, samples, image data, array data to the study at later intervals. These later imports do not remove any existing data; they instead insert any new subjects or update annotations for existing subjects.

Defining Survival Values

Survival value is the length of time a patient lived. If you plan to analyze your data in calIntegrator2 to create a Kaplan-Meier (K-M) Plot, then during the Annotation Definition process described above, you must make sure that you have defined at least three fields set to the “date” Data Type. These will be matched to the following three properties during Survival Value definition.

- **Survival Start Date**
- **Death Date**
- **Last Followup Date**

Note: Setting survival values is optional if you do not plan to use the K-M plot analysis feature or if you do not have this kind of data (survival values) in the file.

For some applications, such as REMBRANDT and I-SPY, survival values are pre-defined in the databases when you load the data. In calIntegrator2, however, you can review and define survival value ranges in a data set you are uploading to a study. To be able to do so, you need to understand the kind of data that can comprise the survival values.

To set up survival values, follow these steps:

1. On the Edit Study page, click **Edit Survival Values**. This opens the Survival Value Definitions dialog box ([Figure 2.11](#)).

Survival Value Definitions for 'test jbh'

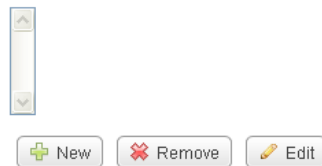


Figure 2.11 Survival Value Definition dialog box

2. Click **New** to enter new survival value definitions.
- OR -
- Click **Edit** to edit existing survival value definitions.

- The dialog box extends, now displaying three drop-down lists that show column headers for date metadata in the spreadsheet you have uploaded. [Figure 2.12](#) displays survival value ranges that have already been added to a study.

Survival Value Definitions for 'Demo Study based on [T11 Long-Full data](#)'

Survival from enrollment

New

Remove

Edit

Survival Value Definition Properties for 'Survival from enrollment'

Name:	Survival from enrollment
Survival Start Date:	ENROLLMENT_DATE
Death Date:	DEATH_DATE
Last Followup Date:	LAST_CONTACT_DATE

Save

Figure 2.12 Survival Definitions example

In the drop-down lists, select the appropriate survival value definitions for each field listed. You might want to refer to the column headers in the data file itself. Dates covered by the definitions are already in the data set. You cannot enter specific dates.

- **Name** – Enter a unique name that adequately describes the survival values you are defining here. *Example:* Survival from Enrollment Date or Survival from Treatment Start. The name you enter displays later when you are selecting survivals to create the K-M plot.
- **Survival Start Date** – Select the column header for this data.
- **Death Date** – Select the column header for this data.
- **Last Followup Date** – Select the column header for this data.

See also [Creating Kaplan-Meier Plots](#) on page 54.

Adding/Editing Genomic Data

Note: Genomic data must be parsed and stored in caArray to be able to use it in caIntegrator2.

Once you have loaded clinical data and identified patient IDs, you can add either array genomic sample data from caArray, which caIntegrator2 maps by sample IDs to the patient IDs in the clinical data, covered in this section, or you can load imaging files from NBIA, also mapped by IDs to the patient data, covered in [Adding Imaging Data](#) on page 29. You can also edit genomic data information that you have already added to the study. Genomic sample data and imaging data are independent of each other, so neither is required before loading the other.

It is essential that you are well acquainted with the data you are working with--the clinical data, and the corresponding array data in caArray.

caIntegrator2 supports a limited number of array platforms. For more information, see [Managing Platforms](#) on page 32.

To add genomic data to your caIntegrator2 study, follow these steps:

1. On the Edit Study page where you have selected and added the clinical data, click the **Add** button under Genomic Data Sources. You can upload genomic data only from caArray.

This opens the Edit Genomic Data Source dialog box (Enter the appropriate information in the fields (*Figure 2.13*).

Edit Genomic Data Source

caArray Server Hostname:	array.nci.nih.gov
caArray server JNDI Port:	0
caArrayUsername:	
caArrayPassword:	
caArray Experiment Id:	
Vendor:	Affymetrix
Data Type:	Expression
Platform (only needed for Agilent):	

Cancel Save

Figure 2.13 Edit Genomic Source dialog box

- **caArray Host Name** – Enter the hostname for your local installation or for the CBIIT installation of caArray, array.nci.nih.gov. If you misspell it, you will receive an error message.
- **caArray JNDI Port** – Enter the appropriate server port. See your administrator for more information. *Example:* For the CBIIT installation of caArray, enter **8080**.
- **caArray Username** and **caArray Password** – If the data is private, you must enter your caArray account user name and password; you must have been given permissions in caArray for the experiment. If the data is public, you can leave these fields blank.
- **caArray Experiment ID** – Enter the caArray Experiment ID which you know corresponds with the clinical data you uploaded. *Example:* Public experiment “beer-00196” on the CBIIT installation of caArray (array.nci.nih.gov). If you misspell your entry, you will receive an error message.
- **Vendor** – Select either Agilent or Affymetrix
- **Data Type** – Select Expression or Copy Number.

Note: If you select Expression, you can map samples to the data. If you select Copy Number, you cannot map samples.

- **Platform (needed only for Agilent)** – Select the Agilent platform.
2. Click **Save**.

caIntegrator2 goes to caArray, validates the information you have entered here, finds the experiment and retrieves all the sample IDs in the experiment. Once

this finishes, the experiment information displays on the Edit Study page under the Genomic Data Sources section ([Figure 2.14](#)).

Host Name	Experiment Identifier	File Description	Data Type	Status	Action
nci-0227-v.nci.nih.gov	admin-00001	Mapping File(s): nci_sample_mapping.csv Control Sample Mapping File(s): page_0034_control_samples.csv	Expression	Loaded	Edit Map Samples Delete

Figure 2.14 Genomic Data Sources section of the Edit Study page

3. If you want to redefine the caArray experiment information, you can edit it. Click the **Edit** link corresponding to the Experiment ID. The Edit Genomic Data Source dialog box reopens, allowing you to edit the information.

Mapping Genomic Data to Clinical Data

Because the goal of caIntegrator2 is to integrate data from clinical, genomic and imaging data sources, data from uploaded source files must be mapped to each other.

To map the samples from the caArray experiment to the patients in the clinical data you uploaded, follow these steps:

1. On the Edit Study page, click the **Map Samples** link. This opens the Edit Sample Mappings page ([Figure 2.15](#)).

caArray Server Hostname:

caArray server JNDI Port:

caArrayUsername:

caArrayPassword:

caArray Experiment Id:

Subject to Sample Mapping File: [Browse...](#) [Upload Mapping File](#)

Control Sample Set Name:

Control Samples File: [Browse...](#) [Upload Control Samples File](#)

Sample Name
GeneratedSample.UNKNOWN_DISEASE_L_E10216_U133P2
GeneratedSample.GBM_L_20070226_15-14-09-358_HF0936_U133P2
GeneratedSample.OLIGO_L_20070227_12-21-11-104_HF1380_U133P2
GeneratedSample.GBM_L_NOB1228_S_p3_3
GeneratedSample.GBM_L_20070226_14-05-29-569_HF1262_U133P2
GeneratedSample.UNKNOWN_DISEASE_L_E10076_U133P2
GeneratedSample.OLIGO_L_20070227_11-49-51-876_HF0899_U133P2
GeneratedSample.Ast_L_20070226_11-54-59-645_HF0026_U133P2
GeneratedSample.UNKNOWN_DISEASE_L_0308NT133_p7_1_U133P2
GeneratedSample.E10662B
GeneratedSample.OLIGO_L_20070227_11-49-51-876_HF1136_U133P2

Figure 2.15 Edit Sample Mappings page

When you first open this page, all of the samples in the caArray experiment you selected are listed as unmapped, because caIntegrator2 does not know how these sample names correlate to the patient data in the clinical file until you upload the mapping file.

2. At the top of the page, click **Browse** to navigate for the CSV file that identifies the mapping information. Click the **Upload Mapping File** button.

The mapping file has only two columns (typically without headers)—one that shows the subject ID (designated in caIntegrator2 as the “Identifier”) and one that has “Sample name” field from the linked caArray experiment, with one subject per row (*Figure 2.16*). This provides caIntegrator2 with the information for mapping patients to caArray samples.

	A	B	C	D	E	F	G	H
1	E10216	GeneratedSample.UNKNOW	DISEASE_L_E10216_U133P2					
2	E10144	GeneratedSample.UNKNOW	DISEASE_L_E10144_U133P2					
3	E09212	GeneratedSample.UNKNOW	L_20070227_16-22-37-238_E09212_U133P2					
4	E09369	GeneratedSample.UNKNOW	L_20070227_16-22-37-238_E09369_U133P2					
5	E10162	GeneratedSample.UNKNOW	DISEASE_L_E10162_U133P2					
6	E10318	GeneratedSample.UNKNOW	DISEASE_L_E10318_U133P2					
7	E09264	GeneratedSample.OLIGO_L_20070227_11-27-27-881_E09264_U133P2						
8	E10252	GeneratedSample.UNKNOW	DISEASE_L_E10252B_U133P2					
9	E09074	GeneratedSample.OLIGO_L_20070227_11-27-27-881_E09074_U133P2						

Figure 2.16 Example sample mapping file, in CSV format

Note: When you open the mapping file, make sure that the patient ID is used for mapping.

Unmapped samples continue to show at the top of the caIntegrator2 page. They were loaded from caArray, but they are not in the mapping file. These are not used for integration.

3. Scroll down the page to see samples that are mapped to the patients in the clinical data (*Figure 2.17*).

1336	GeneratedSample.OLIGO_L_20070227_11-49-51-876_HF0599_U133P2																														
1338	GeneratedSample.UNKNOW_DISEASE_L_E10029_U133P2																														
1339	GeneratedSample.GBM_L_20070226_14-05-29-569_HF1356_U133P2																														
1340	GeneratedSample.OLIGODENDROGLIOMA_L_HF0599_U133P2																														
1342	GeneratedSample.GBM_L_20070226_13-30-40-39_HF0142_U133P2																														
1345	GeneratedSample.GBM_L_20070226_14-31-29-427_HF1469_U133P2																														
<table border="1"> <thead> <tr> <th colspan="2">Samples Mapped to Subjects</th></tr> <tr> <th>Sample ID</th><th>Sample Name</th></tr> </thead> <tbody> <tr> <td>901</td><td>GeneratedSample.UNKNOW_DISEASE_L_E10216_U133P2</td></tr> <tr> <td>911</td><td>GeneratedSample.UNKNOW_DISEASE_L_E10144_U133P2</td></tr> <tr> <td>914</td><td>GeneratedSample.UNKNOW_L_20070227_16-22-37-238_E09212_U133P2</td></tr> <tr> <td>918</td><td>GeneratedSample.UNKNOW_L_20070227_16-22-37-238_E09369_U133P2</td></tr> <tr> <td>922</td><td>GeneratedSample.UNKNOW_DISEASE_L_E10162_U133P2</td></tr> <tr> <td>925</td><td>GeneratedSample.UNKNOW_DISEASE_L_E10318_U133P2</td></tr> <tr> <td>930</td><td>GeneratedSample.OLIGO_L_20070227_11-27-27-881_E09264_U133P2</td></tr> <tr> <td>940</td><td>GeneratedSample.UNKNOW_DISEASE_L_E10252B_U133P2</td></tr> <tr> <td>954</td><td>GeneratedSample.GBM_L_20070226_13-14-06-67_E09624_U133P2</td></tr> <tr> <td>957</td><td>GeneratedSample.ASTROCYTOMA_L_E09137_U133P2</td></tr> <tr> <td>958</td><td>GeneratedSample.UNKNOW_DISEASE_L_E09890_U133P2</td></tr> <tr> <td>958</td><td>GeneratedSample.UNKNOW_L_20070227_16-57-07-283_E09515_U133P2</td></tr> <tr> <td>1004</td><td>GeneratedSample.UNKNOW_L_20070227_17-26-09-910_E09722_U133P2</td></tr> </tbody> </table>		Samples Mapped to Subjects		Sample ID	Sample Name	901	GeneratedSample.UNKNOW_DISEASE_L_E10216_U133P2	911	GeneratedSample.UNKNOW_DISEASE_L_E10144_U133P2	914	GeneratedSample.UNKNOW_L_20070227_16-22-37-238_E09212_U133P2	918	GeneratedSample.UNKNOW_L_20070227_16-22-37-238_E09369_U133P2	922	GeneratedSample.UNKNOW_DISEASE_L_E10162_U133P2	925	GeneratedSample.UNKNOW_DISEASE_L_E10318_U133P2	930	GeneratedSample.OLIGO_L_20070227_11-27-27-881_E09264_U133P2	940	GeneratedSample.UNKNOW_DISEASE_L_E10252B_U133P2	954	GeneratedSample.GBM_L_20070226_13-14-06-67_E09624_U133P2	957	GeneratedSample.ASTROCYTOMA_L_E09137_U133P2	958	GeneratedSample.UNKNOW_DISEASE_L_E09890_U133P2	958	GeneratedSample.UNKNOW_L_20070227_16-57-07-283_E09515_U133P2	1004	GeneratedSample.UNKNOW_L_20070227_17-26-09-910_E09722_U133P2
Samples Mapped to Subjects																															
Sample ID	Sample Name																														
901	GeneratedSample.UNKNOW_DISEASE_L_E10216_U133P2																														
911	GeneratedSample.UNKNOW_DISEASE_L_E10144_U133P2																														
914	GeneratedSample.UNKNOW_L_20070227_16-22-37-238_E09212_U133P2																														
918	GeneratedSample.UNKNOW_L_20070227_16-22-37-238_E09369_U133P2																														
922	GeneratedSample.UNKNOW_DISEASE_L_E10162_U133P2																														
925	GeneratedSample.UNKNOW_DISEASE_L_E10318_U133P2																														
930	GeneratedSample.OLIGO_L_20070227_11-27-27-881_E09264_U133P2																														
940	GeneratedSample.UNKNOW_DISEASE_L_E10252B_U133P2																														
954	GeneratedSample.GBM_L_20070226_13-14-06-67_E09624_U133P2																														
957	GeneratedSample.ASTROCYTOMA_L_E09137_U133P2																														
958	GeneratedSample.UNKNOW_DISEASE_L_E09890_U133P2																														
958	GeneratedSample.UNKNOW_L_20070227_16-57-07-283_E09515_U133P2																														
1004	GeneratedSample.UNKNOW_L_20070227_17-26-09-910_E09722_U133P2																														
	Subject Identifier																														
	E10216																														
	E10144																														
	E09212																														
	E09369																														
	E10162																														
	E10318																														
	E09264																														
	E10252																														
	E09624																														
	E09137																														
	E09890																														
	E09515																														
	E09722																														

Figure 2.17 Example of samples mapped to patients' data

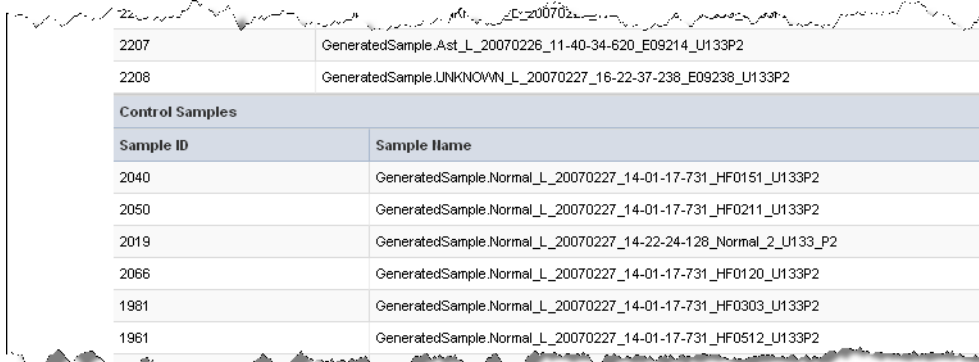
Uploading Control Samples

A Control Samples file is used to calculate fold change data, which compares “tumor” sample gene expression in the caArray experiment to the control samples to identify those that exhibit up or down gene regulation. Control samples can be the “normal” samples, but that is not necessarily the case.

To upload the control samples, follow these steps:

1. On the Edit Sample Mappings page, click the **Map Samples** link.

2. Click **Browse** to navigate for the control samples file, and click the **Upload Control Samples** File button. Scroll down the page to view the list of control samples that have been added (*Figure 2.18*).



Control Samples	
Sample ID	Sample Name
2207	GeneratedSample.Ast_L_20070226_11-40-34-620_E09214_U133P2
2208	GeneratedSample.UNKNOWN_L_20070227_16-22-37-238_E09238_U133P2
2040	GeneratedSample.Normal_L_20070227_14-01-17-731_HF0151_U133P2
2050	GeneratedSample.Normal_L_20070227_14-01-17-731_HF0211_U133P2
2019	GeneratedSample.Normal_L_20070227_14-22-24-128_Normal_2_U133_P2
2066	GeneratedSample.Normal_L_20070227_14-01-17-731_HF0120_U133P2
1981	GeneratedSample.Normal_L_20070227_14-01-17-731_HF0303_U133P2
1961	GeneratedSample.Normal_L_20070227_14-01-17-731_HF0512_U133P2

Figure 2.18 Example list of control samples

The control samples now display toward the bottom of the page.

3. This information will be used when performing other tasks in caIntegrator2, to be described in other sections.

Configuring Copy Number Data

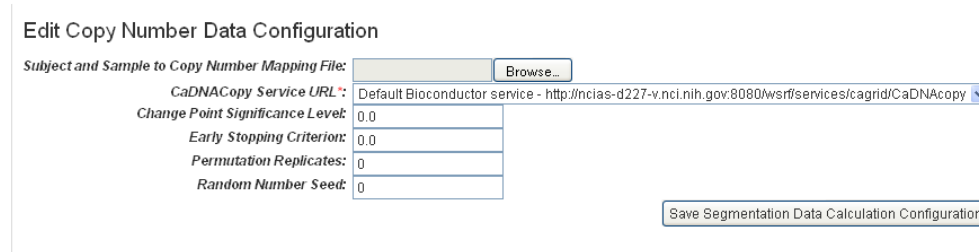
You can add copy number data for a genomic data source by uploading the mapping file. This allows you to configure parameters to be used when segmentation data is being configured.

To add copy number data relating to the genomic data you are adding, follow these steps:

1. In the Genomic Data Sources section, for the data you have already added, click **Configure Copy Number Data** hypertext link.

Note: This link is available only if you have uploaded copy number data and you are configuring a Copy Number data type (as indicated by the Data Type column on the Edit Study page).

The Edit Copy Number page opens (*Figure 2.19*).



Edit Copy Number Data Configuration

Subject and Sample to Copy Number Mapping File:

CaDNACopy Service URL: Default Bioconductor service - http://ncias-d227-v.nci.nih.gov:8080/wsr/services/cagrid/CaDNACopy

Change Point Significance Level:

Early Stopping Criterion:

Permutation Replicates:

Random Number Seed:

Figure 2.19 Edit Copy Number page

2. Browse for and enter appropriate information to identify the copy number mapping file. The fields are described in [Table 2.2](#). An asterisk* indicates a required field..

	<i>Description</i>
Subject and Sample to Copy Number Mapping File	Browse for the appropriate CN mapping file
caDNACopy Service URL*	Control for selecting the URL which hosts the caDNACopy grid service
Change Point Significance Level	Significance levels for the test to accept change-points
Early Stopping Criteria	The sequential boundary used to stop and declare a change
Permutation Replicates	The number of permutations used for p-value computation
Random Number Seed	The segmentation procedure uses a permutation reference distribution. This should be used if you plan to reproduce the results.

Table 2.2 Fields for retrieving a copy number mapping file.

3. Click **Configure copy number data** for a genomic data source. On the screen upload a copy number mapping file (format: subject id, sample id, file name) and configure the parameters to be sent when computing segmentation data.

Adding Imaging Data

Once you have loaded clinical data and identified patient IDs, you can add either array genomic sample data from caArray which caIntegrator2 maps by sample IDs to the patient IDs in the clinical data, or you can load imaging files from NBIA, also mapped by IDs to the patient data, covered in this section. Genomic sample data and imaging data are independent of each other, so neither is required before loading the other.

It is essential that you are well acquainted with the data you are working with--the clinical data, and the corresponding imaging data in NBIA.

Any data in NBIA can be uploaded to caIntegrator2. Imaging data consist of images and or annotations for images.

To add imaging data to the study you are creating or are editing, follow these steps:

1. On the Edit Study page, click the **Add** button under Imaging Data Sources section. Imaging data can be NBIA images or image annotations, which are uploaded in spreadsheet format.

This opens the Edit Imaging Data Source dialog box. Enter the appropriate information in the fields (*Figure 2.20*). Asterisks indicate required fields..

Edit Imaging Data Source

Figure 2.20 Edit Image Data Source dialog box

- **NBIA Server Grid URL*** – Enter the URL for the grid connection to NBIA
- **NBIA Username and NBIA Password.** This information is not required, as currently all data in the NBIA grid is Public data.
- **Collection Name*** – Enter the name/source for the collection.
- **Current Mapping** – If a mapping file has already been uploaded to the study to map imaging data, the file name displays here.
- **Select Mapping File Type*** – Click to select the file type:
 - **Auto** – No file required. Selecting this takes all clinical subject IDs and attempts to map them to the corresponding ID in the collection in NBIA. If the ID does not exist in NBIA, then no mapping is made for that ID.
 - **By Subject** – Requires a file to be uploaded. The “clinical to imaging mapping file” must be a two column mapping (CSV) from the caintegrator2 clinical subject ID to the NBIA subject ID.
 - **By Image Series** – Requires a file to be uploaded. The clinical to imaging mapping file needs to be a two column mapping (CSV) from the caintegrator2 clinical subject ID to the NBIA study instance UID.
- **Clinical to Imaging Mapping File** – Click **Browse** to navigate to the appropriate clinical to imaging mapping file. See **Select Mapping File Type*** field description.

2. Click **Add** to upload the data to calIntegrator.

The imaging data information displays on the Edit Study page under the Imaging Data Sources section (*Figure 2.21*).

Imaging Data Sources				
Host Name	Collection Name	File Description	Status	Action
imaging.nci.nih.gov	NCRI	Annotation File: ncr_image_annotations.csv Mapping File: ncr_image_mapping.csv	Loaded	Edit Edit Annotations Delete

Figure 2.21 Imaging Data Sources section of the Edit Study page

3. Once the data is uploaded, you must assign identifiers and annotations to the data in the same way you did with the clinical data. For more information, see [Assigning An Identifier or Annotation](#) on page 17 and [Searching for Annotation Definitions](#) on page 20.
4. To deploy the study, see [Deploying the Study](#).

Deploying the Study

When you are ready to deploy the study, click the **Save and Deploy** button on the Edit Study page. caIntegrator2 retrieves the selected data from the data service(s) you defined and makes the study available to a study manager or to anyone else who may want to analyze the study's data. Using the Manage Studies feature, you can then configure and share data queries and data lists with all investigators who access the study.

Note that you can continue to work in caIntegrator2 while study is being deployed.

Managing a Study

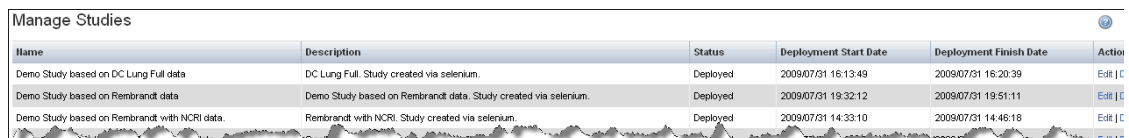
Note: A user without management privileges has no access to this section of caIntegrator2.

Once you have started to create a study or have deployed it, you can update an existing study in the following ways:

- Add new attributes (annotations) and upload relevant data to an existing study.
- Delete a study
- Modify existing annotation definitions
- Reload subset of study data and re-deploy the study and perform new analyses
- Re-deploy the entire study with new set of data and mappings.

To update, edit or delete a study, follow these steps:

1. On the left sidebar, click **Manage Studies**. The Manage Studies page appears ([Figure 2.22](#)).



Name	Description	Status	Deployment Start Date	Deployment Finish Date	Action
Demo Study based on DC Lung Full data	DC Lung Full. Study created via selenium.	Deployed	2009/07/31 16:13:49	2009/07/31 16:20:39	Edit C
Demo Study based on Rembrandt data	Demo Study based on Rembrandt data. Study created via selenium.	Deployed	2009/07/31 19:32:12	2009/07/31 19:51:11	Edit C
Demo Study based on Rembrandt with NCR data	Rembrandt with NCR. Study created via selenium.	Deployed	2009/07/31 14:33:10	2009/07/31 14:46:18	Edit C

Figure 2.22 Manage Studies page

All of the “in process” or “completed” studies display on this page.

2. Click the **Edit** link corresponding to your study of choice to open the Edit Studies page.

On this page you can edit any details such as adding or deleting files, survival values, and so forth. For information about working in the Edit Study, see [Creating/Editing a Study](#) on page 15.

3. Click the **Delete** link to delete the corresponding study.

Managing Platforms

calIntegrator2 supports a limited number of array platforms, all of which originate from Agilent or Affymetrix. While they do not represent all of the platforms supported by caArray, calIntegrator2 must have array definitions loaded for the platforms it supports, and be able to properly load the data from caArray and parse it.

You can create a study without genomic data, but you cannot add genomic data to a calIntegrator2 study without a corresponding supported array platform.

On the Manage Platforms page, you can identify, add or remove supported platforms.

To manage platforms in calIntegrator2, follow these steps:

1. Click **Manage Platforms** on the left sidebar.

The Manage Platforms page that opens lists the platforms calIntegrator2 currently supports, those that the system can pull from caArray ([Figure 2.23](#)). You can also add a new platform by entering information in the fields at the top of the page.

Name	Vendor	Reporter List	Action
AgilentG4502A_07_01	AGILENT	AgilentG4502A_07_01, AgilentG4502A_07_01	Delete
GeneChip Human Mapping 100K Set	AFFYMETRIX	Mapping50K_Hind240, Mapping50K_Xba240	None
HG-U133A	AFFYMETRIX	HG-U133A, HG-U133A	Delete
HG-U133A	AFFYMETRIX	HG-U133A, HG-U133A	None

Figure 2.23 Manage Platforms page

2. To add a platform, in the Platform Type field, select the appropriate platform type from the drop down list. Click **Browse** to navigate for the Affymetrix or Agilent file you want to add.
3. Enter a **Platform Name** if the file is a NON-GEML.xml file.
Depending on what Platform Type is selected, there may be other parameters to provide here as well. Once all parameters have been provided, click **Create Platform**.
4. Click the **Browse** button to browse for the appropriate annotation file. When you have located it, click **Add Annotation**. The system displays annotation files you select in the Associated File(s) Selected box.
5. Click the **Add** button.

CHAPTER 3

SEARCHING A CAINTEGRATOR2 STUDY

This chapter describes the processes for searching studies within calIntegrator2.

Topics in this chapter include:

- [Search Overview](#) on this page
- [Searching a Study](#) on page 36
- [Managing Queries](#) on page 42

Search Overview

The search and browse functions in calIntegrator2 allow you to search for clinical data, genomic or imaging data that were uploaded into the application as part of a study. When gene expression and imaging data are uploaded into a calIntegrator2 study, mapping files that correlate the data in those files to patient IDs in the clinical data file must also be uploaded. When you launch a search, calIntegrator2 finds and integrates the clinical, genomic and imaging data based on the mapping files and the criteria that you define in the search query.

In a search query, you can specify criteria for just one of the data types, or configure complex search criteria that join two or three data types. The available criteria for the query were defined when the study was deployed.

The basic workflow for a study search follows these steps:

1. Select the study to be searched.
2. Select one data type:
 - **Clinical** – searches one or more uploaded CSV files for data identifiers or annotations (column headers) specified when the study was created
 - **Genomic** – Searches caArray experiments samples uploaded in the study for gene expression data by gene name or reporter ID.

- **Image Series** – Searches NBIA files uploaded in the study for image annotations or links to images, identified by subject identifiers or image series IDs.
- 3. Define criteria for the search in the selected data type and run the search.
- 4. For a more complex search, select multiple criteria from more than one data type.
- 5. Specify whether you want clinical/imaging annotations to display or genomic data to display.
- 6. Review search results.
- 7. Configure results column and sorting display settings. You can do this before or after you run a search. If you choose to do it after, you must re-run the search.
- 8. Download annotation search results as a CSV file. The CSV file contains only the data you specified in the annotation and display configurations.
- 9. Follows links to NBIA in the search results to view or download images located in the search.

Searching a Study

To initiate a search of all annotations and/or other data in a study, follow these steps:

1. In calIntegrator2, in the upper right hand corner, select the study you want to browse or perform a simple search.
2. On the left sidebar, under the first section that displays the study name, click **Search [Study Name]**. This opens a simple search query page with five tabs (*Figure 3.1*).

Search Demo Study based on Rembrandt with NCRI data.

The screenshot shows the 'Criteria' tab of the search interface. It features a dropdown menu with 'Clinical' selected and an 'Add' button. A message below the dropdown states 'No criteria added. Please select criteria from the pulldown box.' At the bottom, there are radio buttons for 'or' and 'and', and a 'Run Query' button.

Figure 3.1 Search page

3. On the Criteria tab, in the drop-down list, select the type of data you want to search. The listed options reflect the type of data that have been uploaded to the study.

Note: You can perform a search using one or more criteria you set in one of the data types, or you can define criteria in more than one data type per query, creating a more complex search.

- **Clinical**
- **Gene Expression**
- **Image Series**

Note: NBIA submissions are organized in the following hierarchy, which illustrates the relationship of an image series to its parent study and patient, as well as to the images in the series.:

Clinical trial > Patient (Subject) > Study > Series > Images

4. Click the **Add** button to define annotation elements for the search.

Continue with:

Clinical and Image Series on page 37

Gene Expression on page 38

5. To add additional criteria for the search, repeat steps 3 and 4, as appropriate. You can set more than one data type or more than one criterion for a data type. The criteria become cumulative, thus refining the search.
 6. Once you have configured the query criteria, select the Boolean **Or** or **And** search operator at the bottom of the page.
 - **Or** finds a data subset with at least one of the search criteria
 - **And** finds a data subset with both/or all search criteria.
 7. Click the **Remove** button to clear any data elements you have defined.
 8. You can launch the search from this tab. Click the **Run Search** button. For information about the search results, see *Chapter 4 Viewing Query Results*. You may want to run the search first to see what kind of results you get before you configure the data display, described in step 9.
- or –**
9. On the Results Type tab, you can specify the columns you want to display in the search results data. On the Sorting tab, you can specify how the data is to be sorted. For more information, see *Results Type and Sorting Tabs* on page 40.

Note: As long as you are still in the current query session, you can return to the Criteria, Columns and Sorting tabs to add, modify or remove data and display criteria and re-run the search. If you configure another query without saving the first, the first query will be lost. If you save the query, your current search criteria are saved.

Clinical and Image Series

- If you select Clinical or Image Series data types, an additional drop-down list displays data elements that are annotation definitions specified when

the data was uploaded into the study ([Figure 3.2](#)). Select a search criterion from among the options. You can make only one selection at a time.

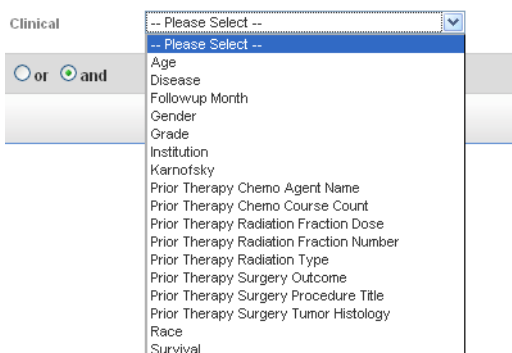


Figure 3.2 Additional clinical search criteria

- ° Each choice opens other fields relevant to the selection where you can further define your search query.
 - If permissible values were added when the annotation was defined, you must select among the values in a drop-list that displays on the right side of the page.
 - If no permissible values were defined as part of the annotation, you have the option to enter descriptive text in a text box on the right side of the page ([Figure 3.3](#)).

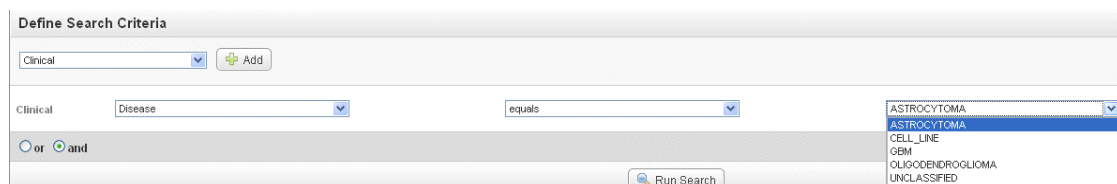



Figure 3.3 You may be able to further define search criteria when you select a specific clinical or imaging annotation element

Note: When working with image data, if only an Imaging Mapping file was uploaded when the study was created and not an Image Series Annotation file, you cannot enter image search criteria. The search results will, however, display a link that allows you to view the associated images in NBIA.

Continue with step 5 in [Searching a Study](#) on page 36.

Gene Expression

1. For the Gene Expression selection, select **Gene Name** or **Fold Change**.
2. For Gene Name or Fold Change, enter one or more gene symbols (up to 100 characters, separated by commas) 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.

- **CGAP** – Use this directory to identify genes. Before clicking this link you must enter gene symbols in the text box. This link does not pull anything into calIntegrator2 but does provide information about the gene(s) whose names you entered.
- **caBio** – This link searches caBIO, then pulls identified genes into calIntegrator2 for analysis. Click the caBIO icon (), enter **Keyword(s)** in the text box that opens and click **Search**. The search results display (Figure 3.4).

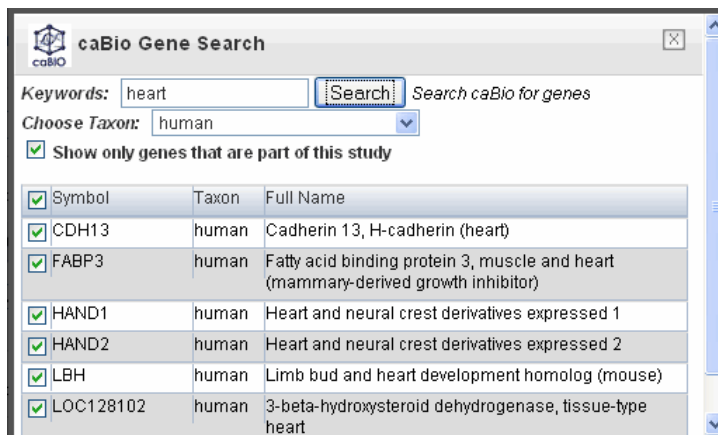


Figure 3.4 Example caBIO gene search results

3. Use the check boxes to identify the genes whose symbols you want to use in the gene expression analysis.
4. Click the **Use Genes** button at the bottom of the page. This pulls the checked genes into the Criteria tab (Figure 3.5).



Figure 3.5 Genes pulled in from caBIO display on the Criteria tab

Additional fields display for the Fold Change selection.

The fold change option appears only if genomic control samples have been uploaded to the study. Fold change identifies genes with expression differences compared to control samples, as defined when the study was deployed in calIntegrator2. You can enter query values in greater/lesser-than-or-equal-to arguments.

5. Select or enter data for the Fold change fields shown in [Figure 3.6](#) :

Figure 3.6 Fields for identifying fold change search criteria

- **Control Sample Set** – Select from the drop down list the name of the uploaded control sample set to serve as the fold change reference.
- **Regulation Type** – Select the term that describes the gene expression in comparison with the control samples: **Up** is increased expression; **Down** is decreased expression; **Up or Down** is increased or decreased; **Unchanged** means no change in expression.
- **Up-Regulation Folds** – The options here are dependent upon the Regulation Type you selected.
 - **Up** = Up Regulation Folds – Samples with a fold change greater than this value, when compared to the control samples, will be returned.
 - **Down** = Down Regulation Folds – Samples with a fold change less than this value, when compared to the control samples, will be returned.
 - **Up or Down** = Down Regulations Folds, Up Regulation Folds – Samples with a fold change either up or down, when compared to the control samples, will be returned.
 - **Unchanged** = Samples with a fold change between the two specified values will be returned.

Continue with step 5 in [Searching a Study](#) on page 36.

Results Type and Sorting Tabs

You can specify columns and sorting options for the way you want the search results to display either before or after you run the search. If you run the search directly from the Criteria tab before setting the results type/sorting features, by default only the Subject Identifiers display on the Search Results tab. You can then come back to the Results Type and Sorting tabs to expand the display options and re-run the search, having set the display parameters.

The selection you make on the Results Type tab determines whether caIntegrator2 displays search results for clinical or genomic data. It filters the search based on the criteria you set on the Criteria tab, whether it is clinical, gene expression or image series data type(s). In other words, if you select clinical criteria on the Criteria tab, but select Genomic on the Results Type tab, the data subset that displays on the Search Results tab is genomic data that is filtered by the clinical criteria you defined on the Criteria tab.

1. On the Results Type tab, select the **Clinical** or **Genomic** radio button to search clinical data (*Figure 3.7*).

Search Demo Study based on Rembrandt with NCRI data.

Criteria Results Type Sorting Query Results Save as...

Select Results Type: ☐ Genomic ☒ Clinical

Genomic result type - will display a gene expression data matrix.
Clinical result type - will display tabular data, including column selection.

Select Columns for Results

Subject Annotations	Image Annotations
<input checked="" type="checkbox"/> Age	<input type="checkbox"/> Comments
<input checked="" type="checkbox"/> Death Date	<input type="checkbox"/> Eloquent Cortex
<input type="checkbox"/> Disease	<input type="checkbox"/> Exam Number
<input type="checkbox"/> Followup Month	<input type="checkbox"/> Radiologist
<input checked="" type="checkbox"/> Gender	<input type="checkbox"/> Side of Tumor Epicenter
<input type="checkbox"/> Tumor Location	<input checked="" type="checkbox"/> Tumor Location

Select All Unselect All Select All Unselect All

Run Query

Figure 3.7 Results Type tab

Clinical – Select the annotation elements that you want to display in the search results. All elements listed are column headers in the data uploaded to the study. You can make multiple selections on this list.

Note: For Clinical Annotations, the Patient or Subject Identifier display by default in the search results.

Results display as tabular data.

Genomic – Select the Reporter Type:

- **Gene Name** – Finds and summarizes at the gene level all reporters that match criteria for the gene you defined on the Criteria tab.
- **Reporter ID** – Finds all reporters that map to the gene(s) you identified on the Criteria tab

Results display in a gene expression data matrix.

Imaging – If imaging annotations have been added to the study, annotation elements also display on the lower right section of this page when you select **Clinical**. All elements listed are column headers in the image annotation data uploaded to the study. You can make multiple selections on this list.

Note: If you select even one Image Annotation on the Results Type tab, the Image Series IDs display by default in the search results. If you select no Image Annotations on the Results Type tab, however, even if you have selected image series criteria on the Criteria tab, no image series IDs display in the search results. The fact that images can be located, however, in NBIA is indicated by two image-related buttons at the bottom of the Query Results page. You can open the images in NBIA, but

they will be at StudyInstance UID level. See [Relationship of Patient to Study to Series to Images](#) on page 51.

Results display as tabular data.

2. Use the **Select All** or **Unselect All** buttons to aid you in making your selections.

The column selection is saved as part of the query if you save it. See [Saving a Query](#) on page 42.

3. Select the Sorting tab and indicate the column order of the Search Results ([Figure 3.8](#)).

Search Demo Study based on Rembrandt with NCRI data.

Criteria Results Type **Sorting** Query Results Save as...

Set Sort Order for Selected Columns

Column	Order (L-R)
Tumor Location	1
Age	2
Gender	3
Death Date	4

Run Query

Figure 3.8 Sorting tab

Sorting parameters are saved as part of the query if you choose to save it using the Save Query feature. On the Search Results page, you can also sort the results by clicking on a column name.

4. Click **Run Search**. Search results display on the Search Results page. For information about the search results, see [Chapter 4 Viewing Query Results](#).

Managing Queries

When you create a search query in caIntegrator2, you can save the query for later use or edit it.

Exporting Query Results

Saving a Query


To save a query, follow these steps:

1. Click the **Save As** tab and enter a **Search Name** and **Search Description**, unique to the search. *Example: Batch ID 6 and female*
2. Click **Save**.

Once the query is saved, it is listed by its name under the **Study Data > Queries > My Queries** in the left sidebar, whenever the study to which the query applies is selected. Click on the saved query in this list to either edit or re-run the query. Click on the query name to retrieve query results. If you hover over the Name text for the query, a popup displays the query description.

Editing a Query

To edit a query, follow these steps:

1. To edit a query, select it in the left sidebar under the **Study Data > Queries > My Queries**.
2. Click the **Edit** icon () corresponding to the study.
3. Change the query and display criteria on the Criteria, Columns and Sorting tabs.
4. On the Save As tab, check the appropriate options and click **Save As**. You can use the same name as the original query or modify the name as needed.

Exporting Query Results

After running a search, you can export the result set or a subset as a tab-delimited text file. For more information, see [Exporting Data](#) on page 52.

CHAPTER 4

VIEWING QUERY RESULTS

This chapter describes search results that calIntegrator2 returns after queries.

Topics in this chapter include the following:

- [Query Results Overview](#) on this page
- [Browsing Query Results](#) on page 46

Query Results Overview

After you launch a search of a calIntegrator2 study, the system automatically opens the Query Results tab showing the results of your search.

If you have not configured column and sort display parameters before launching the search, by default the tab shows only the subject identifiers and a column that allows you to select each row of the data subset.

To display and/or sort additional data, you must return to the Columns and/or Sorting tabs to set display parameters, then re-run the search. The new search results will display the additional information, with the columns and data sorted as you specified. See [Results Type and Sorting Tabs](#) on page 40.

calIntegrator2 paginates search results into pages of configurable size (default 20) with standard paginated navigation controls. To sort columns by ascending or descending parameters for on any displayed field, click on the underlined column header.

You can download search results as a CSV file. The file contains the annotations, columns and data sort configurations you specified in the search query. See [Exporting Query Results](#) on page 43.

Browsing Query Results

The query results that can display depend upon the criteria you established for the search. Follow the links below for more information about the category of data you searched.

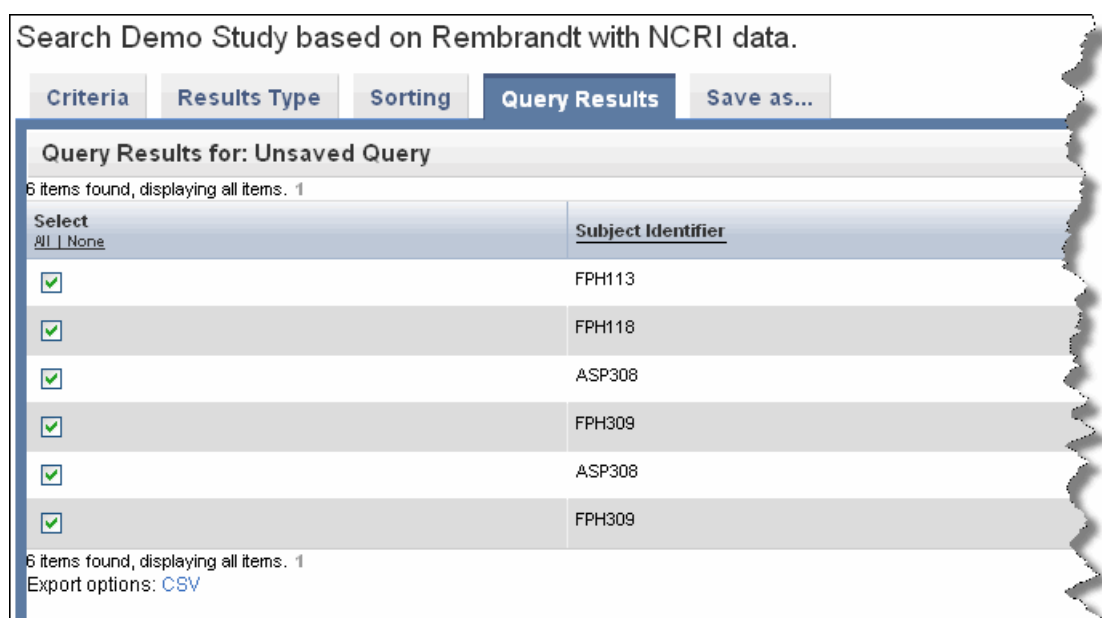
[Clinical and Imaging Data](#) on page 46

[Genomic Data](#) on page 46

[Expanding Imaging Data Results](#) on page 48

Clinical and Imaging Data

If you run the search before configuring column and sort display parameters, only the [subject] ID that meet the criteria and a column allowing you to select each row appear on the table ([Figure 4.1](#)). .



Select	Subject Identifier
<input checked="" type="checkbox"/>	FPH113
<input checked="" type="checkbox"/>	FPH118
<input checked="" type="checkbox"/>	ASP308
<input checked="" type="checkbox"/>	FPH309
<input checked="" type="checkbox"/>	ASP308
<input checked="" type="checkbox"/>	FPH309

Figure 4.1 Query Results page

You can add details for a one or more subjects by configuring them on the Results Type tab. Annotations listed there are the column headers in the CSV file(s) that were uploaded to the study. For information about using the Results Type tab, see [Results Type and Sorting Tabs](#) on page 40.

Genomic Data

If you select the Genomic result type on the Results Type tab, genomic data search results display in a gene expression data matrix. Because the data was downloaded from caArray, the data permissions granted there still apply. In other words, if you have been given access to the data in caArray, you can see it in caIntegrator2.

In the matrix, samples in the experiment form the column headings. If the rows display Gene Name, the cells display the median gene expression value for each gene. If the

rows display Probe Set, the cells display the normalized signal-based value for a given reporter for a given sample (*Figure 4.2, Figure 4.3*).

Search Demo Study based on DC Lung Full data

Criteria

Results Type

Sorting

Query Results

Save as...

Query Results for: Unsaved Query

Subject ID	321	326	158	53	180	350	545	54	142	199	539	89	193	107	305	244	74	361
Sample ID	321	326	158	53	180	350	545	54	142	199	539	89	193	107	305	244	74	361
Gene Name																		
EGFR	151.48	395.43	179.63	70.39	91.81	175.26	97.61	80.96	111.84	115.25	309.05	81.36	136.61	84.75	145.52	208.45	66.68	14

Export options: CSV

Figure 4.2 Gene Name search EGFR, Reporter Type: Gene

Search Demo Study based on DC Lung Full data

Criteria

Results Type

Sorting

Query Results

Save as...

Query Results for: Unsaved Query

	Subject ID	21	362	369	332	480	135	93	345	255	486	333	115
	Sample ID	21	362	369	332	480	135	93	345	255	486	333	115
Gene Name	Reporter ID												
EGFR	201983_s_at	7071.93	957.15	1747.74	2614.96	875.01	646.42	503.99	4350.84	1297.73	2425.24	1275.78	365.28
EGFR	201984_s_at	2625.52	402.21	1225.3	1629.62	569.71	477.81	311.24	673.6	611.8	637.53	347.0	376.42
EGFR	210984_x_at	420.18	44.28	56.72	214.84	133.44	24.43	20.37	36.96	40.0	31.66	45.54	40.23
EGFR	211550_at	9.43	18.68	33.64	11.73	23.74	15.93	12.4	31.44	29.33	20.11	33.91	14.49
EGFR	211551_at	180.43	157.4	349.1	208.0	161.24	173.67	59.78	213.68	319.14	189.92	250.89	131.74
EGFR	211607_x_at	378.93	96.02	48.57	108.57	209.22	13.32	29.6	61.36	62.75	52.88	87.53	47.32

Export options: CSV

Figure 4.3 Gene Name search EGFR, Reporter Type: Reporter ID

- Genomic data does not display in tandem with clinical and imaging data; it only displays when you select the Genomic result type on the Results Type tab. Genomic data is however, filtered by clinical and imaging query criteria configured on the Criteria tab.
- Click the Export Options CSV link to download the CSV file whose data displays on the Search Results tab. When you do so, the CSV file opens automatically in MS Excel or similar applications for working with spreadsheets, showing the columns and sorting as you defined them in calIntegrator2 on the appropriate tabs.

Expanding Imaging Data Results

In reviewing imaging search results, it is important to understand the hierarchy of submissions in NBIA. For more information, see [Relationship of Patient to Study to Series to Images](#) on page 51.

If you run the search before configuring column and sort display parameters, only the Subject Identifiers for the patients/images that meet the criteria and a column containing one check box per row display by default ([Figure 4.4](#)).

Search Demo Study based on Rembrandt with NCRI data.

Criteria Results Type Sorting **Query Results** Save as...

Query Results for: Unsaved Query

6 items found, displaying all items. 1

Select All None	Subject Identifier
<input checked="" type="checkbox"/>	FPH113
<input checked="" type="checkbox"/>	FPH118
<input checked="" type="checkbox"/>	ASP308
<input checked="" type="checkbox"/>	FPH309
<input checked="" type="checkbox"/>	ASP308
<input checked="" type="checkbox"/>	FPH309

6 items found, displaying all items. 1

Export options: [CSV](#)

Figure 4.4 With imaging criteria only and no column definition, only Subject IDs display

If your annotation choice on the Columns page identifies annotations such as tumor size or tumor location, the search results display image series subsets that have those annotations. The check boxes work in conjunction with buttons at the bottom of the results page ([Figure 4.5](#)).

Search Demo Study based on Rembrandt with NCRI data.

Criteria Results Type Sorting **Query Results** Save as...

Query Results for: Unsaved Query Results per Page: 20 Apply

6 items found, displaying all items. 1

Select All None	Subject Identifier	Image Series Identifier	Tumor Location
<input checked="" type="checkbox"/>	ASP308	2.16.124.113543.6003.121591217.13842.19801.1684612788 View in NBIA	Frontal
<input checked="" type="checkbox"/>	ASP308	2.16.124.113543.6003.2317586685.40219.20287.3012655789 View in NBIA	Frontal
<input checked="" type="checkbox"/>	FPH309	2.16.124.113543.6003.549598632.64081.17387.2785982861 View in NBIA	Frontal
<input checked="" type="checkbox"/>	FPH309	2.16.124.113543.6003.2205896078.6864.16978.1740991361 View in NBIA	Frontal
<input checked="" type="checkbox"/>	FPH113	2.16.124.113543.6003.2255697655.34510.18599.2966603150 View in NBIA	Frontal
<input checked="" type="checkbox"/>	FPH118	2.16.124.113543.6003.2241039616.45708.20383.2016653450 View in NBIA	Frontal

6 items found, displaying all items. 1

Export options: [CSV](#)

Figure 4.5 By expanding display parameters, you can view complete details for image search results

You can add more details for images by configuring image annotations on the Results Type tab. Annotations listed there are the column headers in the image series CSV file(s) that were uploaded to the study. Examples of image details include the following:

- All image details (name, size, etc.)
- The series that the image belongs to
- Image feature attributes
- The subject ID. Click the subject ID under Clinical Annotations on the Results Type tab to display this.

You can set display parameters for the results on the Columns and Sorting tabs. For more information, see [Results Type and Sorting Tabs](#) on page 40.

See also [caIntegrator2 and NBIA](#) and

caIntegrator2 and NBIA

Images can be accessed in NBIA if you see buttons on the Search Results page. See the Imaging Note in [Results Type and Sorting Tabs](#) on page 40. You can click links on the Search Results tab to view or download image data.

- **View in NBIA** – This link corresponds to each Image Series listed in the results table. If you click the link, NBIA opens to the login page. After you log in, NBIA brings up the first image in the corresponding image series ([Figure 4.6](#)). You must log into NBIA to see the data. On the NBIA page that opens, you can opt to view the entire series containing this image, or you can display the image as a large JPEG-formatted image. You can also add the image to the NBIA basket. For more information, see the NBIA online help or user's guide accessible from NBIA.



Figure 4.6 An example of displaying the first image in image series

- **Forward to NBIA** – This button is linked to results you have selected by row. Click the button to open NBIA, where the image series you select are loaded in the NBIA image basket. In the event that the caIntegrator2 study was NOT

configured with image annotation for an image series, calIntegrator2 sends NBIA a list of Study Instance UIDs, for which NBIA will add all corresponding image series to the basket. In the event that the calIntegrator2 study was configured with annotations for an image series, the system sends NBIA a list of Image Series IDs, for which NBIA adds all corresponding image series to the basket.

Retrieving Dicom Images

On the Imaging data Search Results page, you can click the **Retrieve DICOM Images** button which is linked to results you have selected by row. calIntegrator2 retrieves the corresponding image(s) from NBIA through the grid. NBIA organizes the download file by patientID, StudyInstanceUID, and ImageSeriesUID, and compresses it into a zip file. When calIntegrator2 notifies you that the file is retrieved, the DICOM Retrieval page indicates whether the retrieved files are Study Instance UIDs or Image Series UIDs (*Figure 4.7*). For more information, see the note below.



Figure 4.7 DICOM Retrieval result

Click the **Download DICOM** link to download and save the file. calIntegrator2 unzips the file and displays the list of images in the file. To open the DICOM images, you must have a DICOM image viewer application installed on your computer. For more information, see <http://dicom.online.fr/fr/download.htm>.

In the search results, not all of the patients in the data subset may be mapped to image series IDs. If you select a mixture of patients that have image annotations as indicated by an image series ID and patients that do not have image annotations (no image series ID), when you click the **Retrieve DICOM Images** button, NBIA retrieves the images for the entire *NBIA study instance UID* that includes the image seriesIDs you checked.

If on the Search Results tab you select only patients that have image annotations as indicated by an image series ID, when you click the **Retrieve DICOM Images** button, NBIA retrieves images for the *NBIA image series* that were matched in the search. If the results are a mixture, but you select one specific row with a valid image annotation, calIntegrator2 aggregates to the *images series*. If results are a mixture and you select multiple rows, calIntegrator2 aggregates to the NBIA study in which multiple image series you have selected in the search results are found.

If your query does not have image annotations and all check boxes are selected, results will go up to image series UID and gives all image series in it. Search results

may ultimately depend on how the study was created. For example, if no image series display in query results, it means they were not mapped in the study. In that case, the results “move” up to Study Instance UIDs.

To best understand this, it is important to review the hierarchy of submissions in NBIA. For more information, see [Relationship of Patient to Study to Series to Images](#) on page 51.

Example of Retrieving Images:

You are searching a study that has image data and image annotation(s) for at least one image series.

1. Open a study that has imaging data associated with it that points to the production NBIA server.
2. Make a query that will have image series or patients who are associated to Image Studies and select a few of those patients in the check box.
3. Click the **Retrieve Dicom Images** button.

Note that it aggregates to the image study.

4. Now go back to Results Type tab, select all image annotations and run the query again.
5. Select an image series type column and click the **Retrieve Dicom Images** button.

calIntegrator2 now aggregates to the Image Series that were selected and not the Image Study.

6. Select a row that doesn't have image series data, and a row that does, and push the button.

This should aggregate to the study for the rows selected.

7. Click **Forward to NBIA**. You should see the same types of aggregation for these tests.

When the image Study is in the checked boxes (regardless of image series being there or not), the system aggregates up to the Image Study level.

Relationship of Patient to Study to Series to Images

This flowchart illustrates the relationship of patient to study to series and lastly to images.

Clinical trial > Patient (Subject) > Study > Series > Images

For example, the Study Instance UID is the set of images resulting from one patient office visit. When you upload a spreadsheet of an image series, the hierarchy of images in an image series might look like this:

Study Instance UID (one office visit):

Brain (image series)

- Brain image 1
- Brain image 2
- Brain image 3

Leg (image series)

- Leg image 1
- Leg image 2
- Leg image 3

You can add details for images by configuring image annotations on the Results Type tab. Annotations listed there are the column headers in the image series CSV file(s) that were uploaded to the study. Examples of image details include the following:

- All image details (name, size, etc.)
- The series that the image belongs to
- Image feature attributes
- The subject ID. Click the subject ID under Clinical Annotations on the Results Type tab to display this.

Exporting Data

You can choose to download tabular search results as a CSV file. Click the **Export .csv** link at the bottom of the page. You may need to scroll the page to see it. The file contains the annotations, columns and data sort configurations you specified in the search query.

Note: You will not see the Export option when genomic data displays as query results.

CHAPTER 5

ANALYZING STUDIES

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

Topics in this chapter include the following:

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

Data Analysis Overview

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

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

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

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

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

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

Creating Kaplan-Meier Plots

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

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

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

K-M Plot for Annotations

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

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

Kaplan-Meier Survival Plots (draft)

For Annotation For Gene Expression For Queries

Annotation Based Kaplan-Meier Survival Plots

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

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

4. Select fields as described in [Table 5.1](#).

	Description
Patient Groups	<p>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.</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 K-M 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 K-M plot.</p>

Table 5.1 Fields for selecting K-M 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 K-M plot criteria ([Figure 5.2](#)).

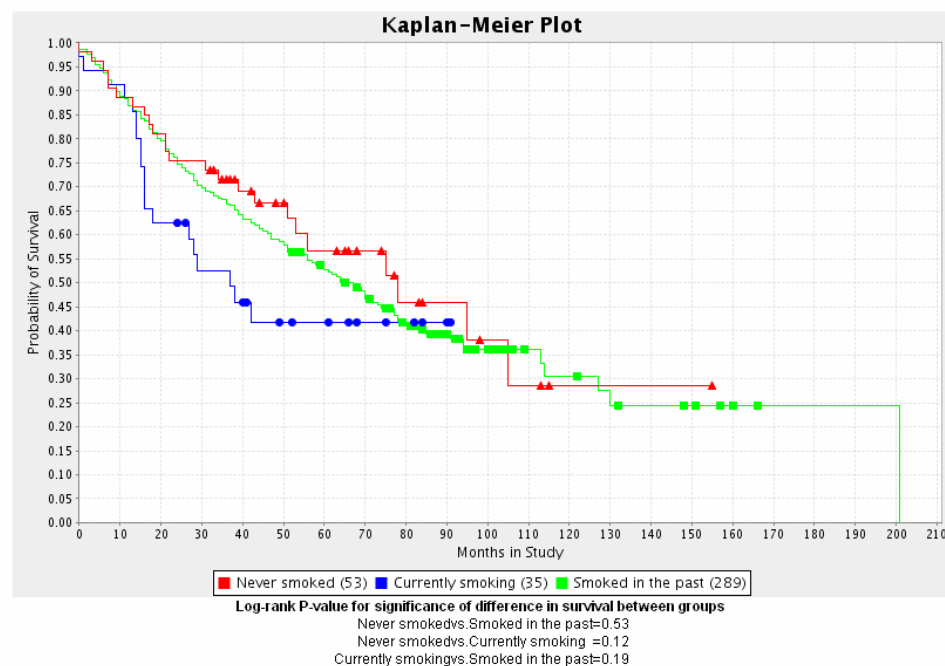


Figure 5.2 A K-M plot generated for groups based on clinical annotations

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

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

K-M Plot for Gene Expression

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

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

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

The screenshot shows the 'Kaplan-Meier Survival Plots' interface with three tabs: 'For Annotation', 'For Gene Expression' (selected), and 'For Queries'. Below the tabs is a section titled 'Gene Expression Based Kaplan-Meier Survival Plots'. This section contains five numbered input fields:

- 1.) Gene Symbol: A text input field.
- 2.) Overexpressed >=: A numeric input field with '2.0' and the unit 'fold'.
- 3.) Underexpressed >=: A numeric input field with '2.0' and the unit 'fold'.
- 4.) Select Survival Value: A dropdown menu with 'Survival from enrollment' selected.
- 5.) Select Control Sample Set: A dropdown menu with 'Control Set 1' selected.

At the bottom right of the form are two buttons: 'Reset' and 'Create Plot'. Logos for 'CGAP' and 'caBio' are visible in the top right corner of the form area.

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

4. Enter or select fields as described in [Table 5.2](#).

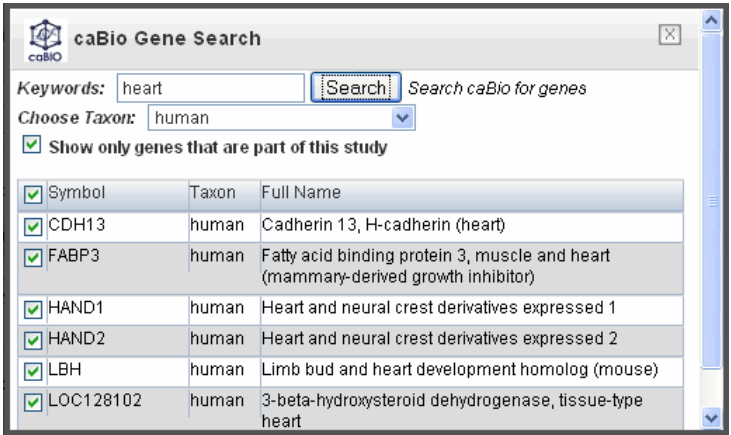


	Description
Gene Symbol	<p>In the text box, specify a single gene whose expression values can be used to split the subjects into three categories: high, low and intermediate expression or click the icons to locate genes in the following databases. If entering more than one gene in the text box, separate 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 calIntegrator2 but does provide information about the gene(s).</p> <p>caBio – This link pulls identified genes into calIntegrator2 for analysis. Click the icon, enter Keyword(s) in the text box that opens and click Search.</p>  <p>Use the check boxes 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>
Over-expressed/ Under-expressed	Define the over- and under-expression criteria, expressed in terms of fold-change. Fold change is the ratio of the measured gene expression value for an experimental sample to the expression value for the control sample.
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 K-M plot.
Control Sample Set	One or more control sample sets are created by the study manager when the study is deployed. Select the control sample set you would like to use to calculate fold-change.

Table 5.2 Fields for selecting K-M gene expression plot values

- Click the **Create Plot** button. calIntegrator2 generates the plot which then displays below the K-M plot criteria ([Figure 5.4](#)).

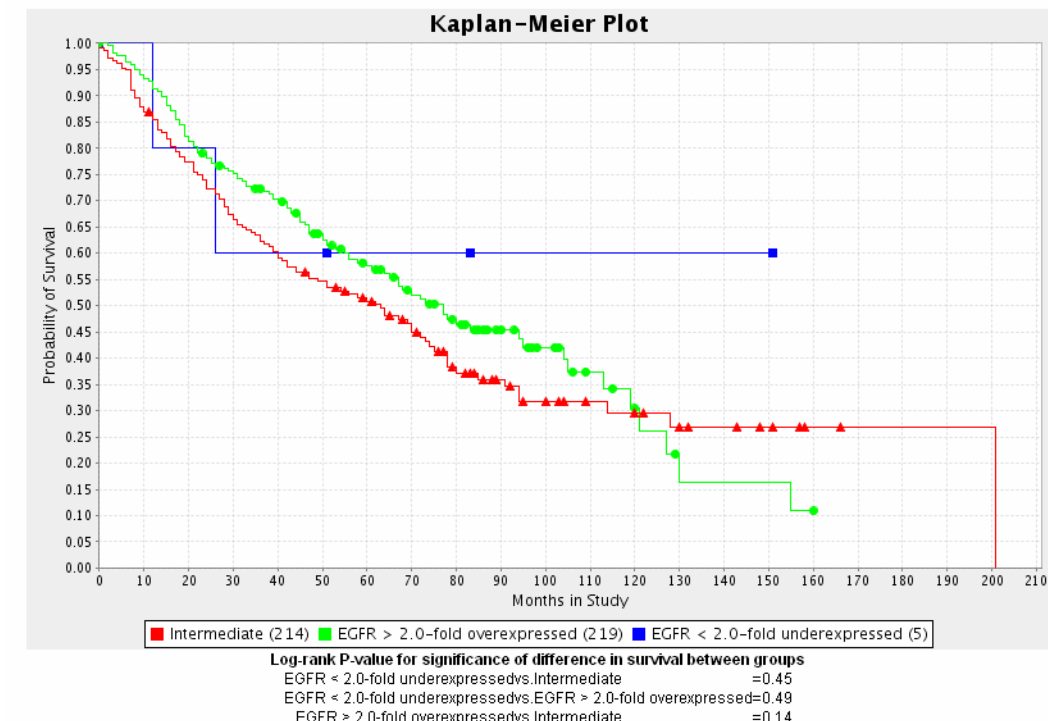


Figure 5.4 K-M plot generated from gene expression data.

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

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

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

K-M Plot for Queries

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

- 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 K-M plot have been saved.
- Under Analysis Tools on the left sidebar, select **K-M Plot**.

3. Select the **For Queries** tab (*Figure 5.5*).

Kaplan-Meier Survival Plots (draft)

For Annotation For Gene Expression **For Queries**

Query Based Kaplan-Meier Survival Plots

1.) Select Queries:

All Available Queries

- gender female
- equal to or > 60
- equal to or > 70
- Never smoke
- equal to or > 40
- equal to or > 30

Add >

< Remove

Selected Queries

▼ ▲

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

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

4.) Select Survival Value: Survival from enrollment ▼

Reset Create Plot

Figure 5.5 Fields for defining K-M plot parameters based on saved queries in caIntegrator2

4. Enter or select fields as described in *Table 5.3*.

	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 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...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 K-M plot.

Table 5.3 Fields for selecting K-M plot values based upon caIntegrator2 queries

5. Click the **Create Plot** button. caIntegrator2 generates the plot which then displays below the K-M plot criteria (*Figure 5.6*).

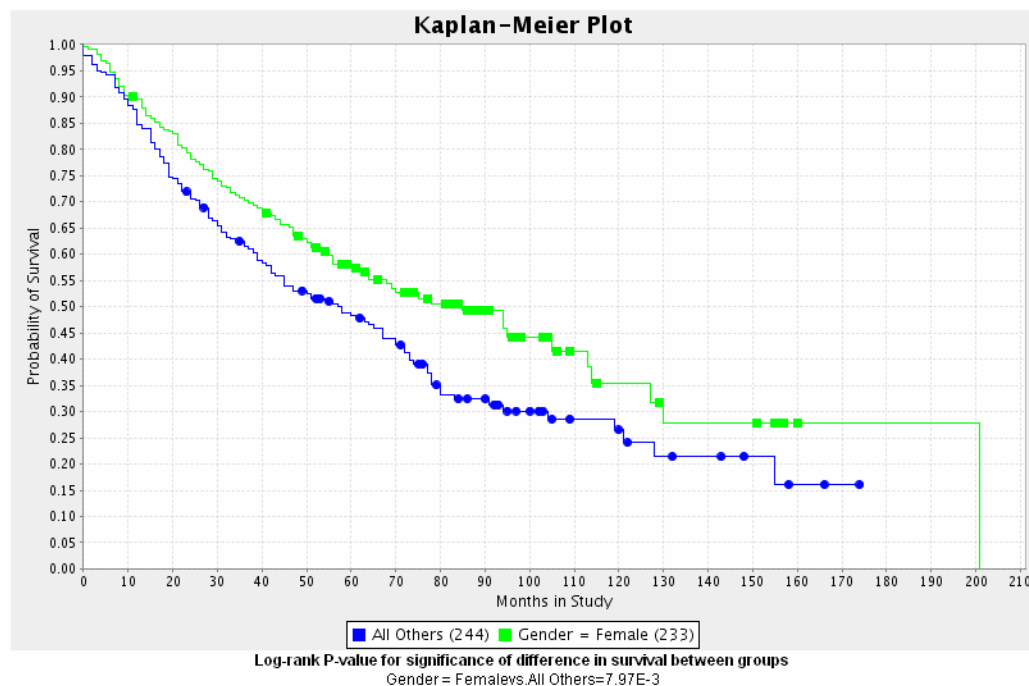


Figure 5.6 K-M Plot comparing statistics between subjects in two queries

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

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

Creating Gene Expression Plots

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

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

- [Gene Expression Value Plot for Annotation](#) – You can locate genes in the 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 70.

Gene Expression Value Plot for Annotation

To generate a gene expression plot, follow these steps:

1. Select the study whose data you want to analyze in the upper right portion of the 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 ([Figure 5.7](#)).

Gene Expression Value Plots

For Annotation For Genomic Queries For Clinical Queries

Annotation Based Gene Expression Plots

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

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

Annotation Type	Annotation	Values
Select Annotation Type	Select Annotation	

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](#).

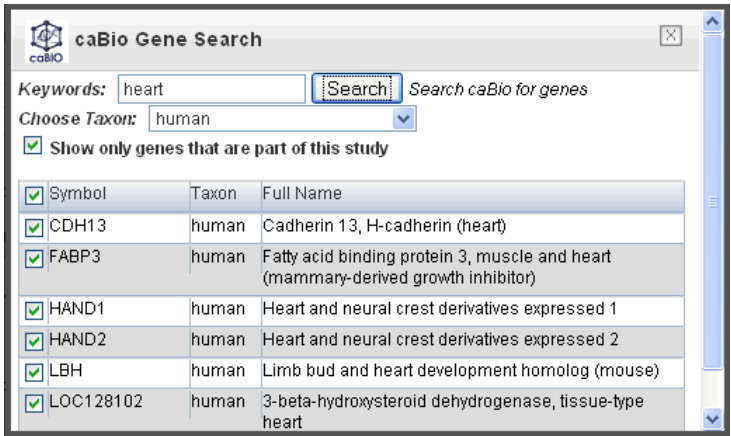


Field	Description
Gene Symbol	<p>Enter one or more gene symbols in the text box or click the icons to locate genes in the following databases. If entering more than one gene in the text box, separate entries by commas.</p> <p>CGAP – Use this directory to identify genes. Before clicking this link you 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 check boxes 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>
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 display an additional group of control samples for this study.

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

5. Click the **Create Plot** button. calIntegrator2 generates the plot which then displays below the Gene Expression Plot criteria in bar graph format (Figure 5.8). Legends below the plot indicate the plot input.

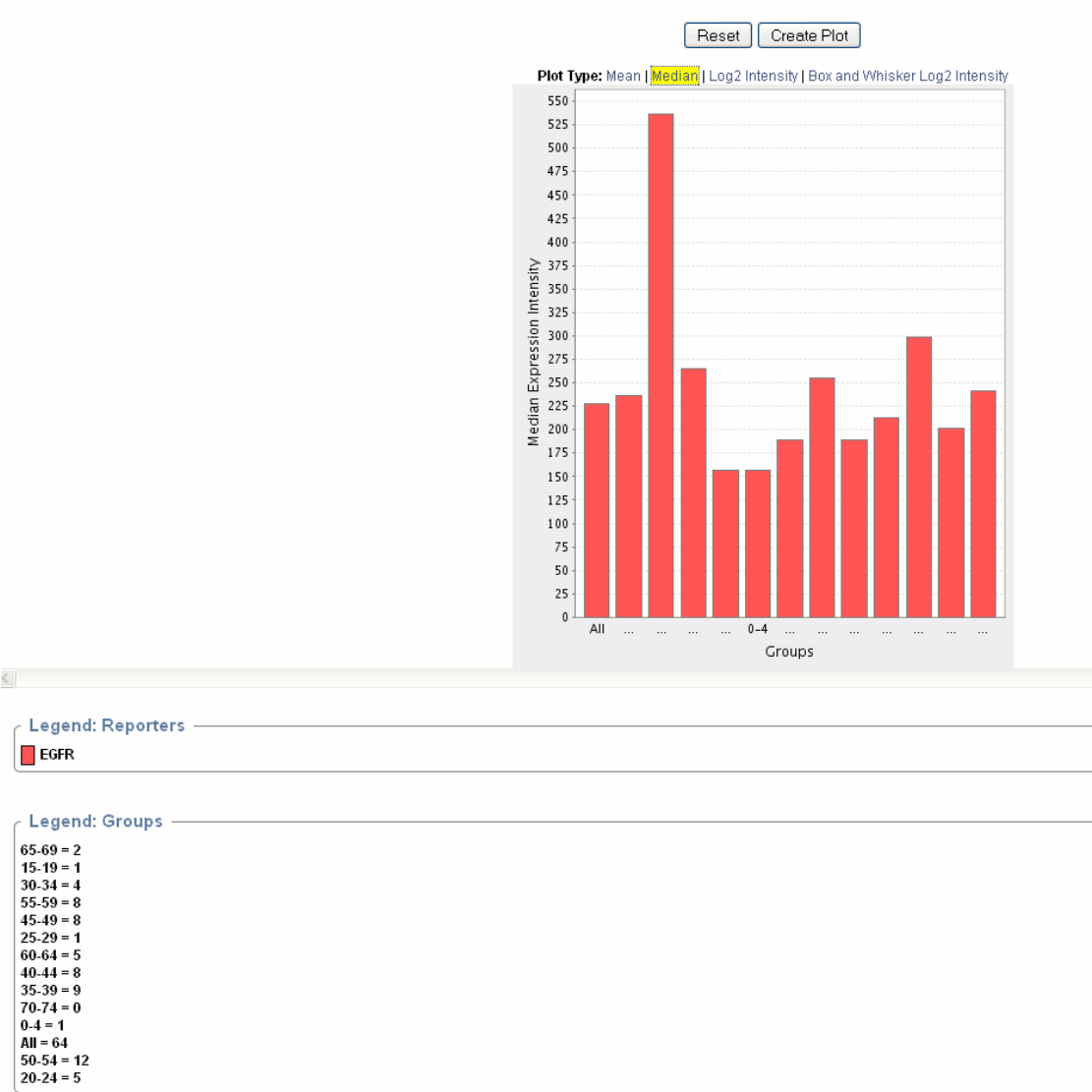


Figure 5.8 Gene expression plot based on selected annotations

6. 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 70

Figure 5.8 displays a plot with gene expression median calculation summaries. Legends below the plot indicate the plot input.

7. 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 42.

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 ([Figure 5.9](#)).

Gene Expression Value Plots

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

4. Enter or select fields as described in [Table 5.5](#).

	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. calIntegrator2 generates the plot which then displays below the Gene Expression Plot criteria. Legends below the plot indicate the plot input (*Figure 5.10*).

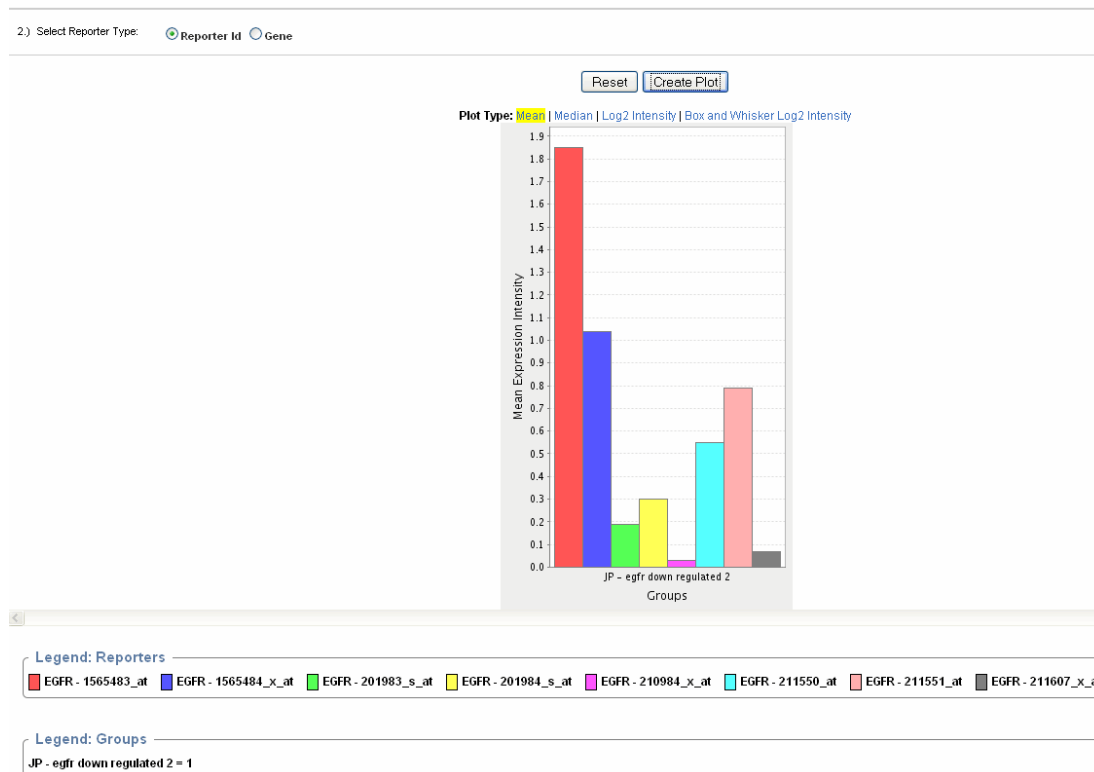


Figure 5.10 A gene expression plot (Mean) based on a genomic query.

- You can recalculate the data display by clicking the **Plot Type** above the graph. See *Understanding a Gene Expression Plot* on page 70. 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 calIntegrator page. You must select a study saved as a clinical study, but which has genomic data.
- Under Analysis Tools on the left sidebar, select **Gene Expression Plot**.

3. Select the **For Clinical Queries** tab (*Figure 5.11*).

Clinical Query Based Gene Expression Plots (draft)

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

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

3.) Select Queries:

All Available Queries

Add >

< Remove

Selected Queries

▼

▲

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. ▼

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

4. Enter or select fields as described in [Table 5.6](#).

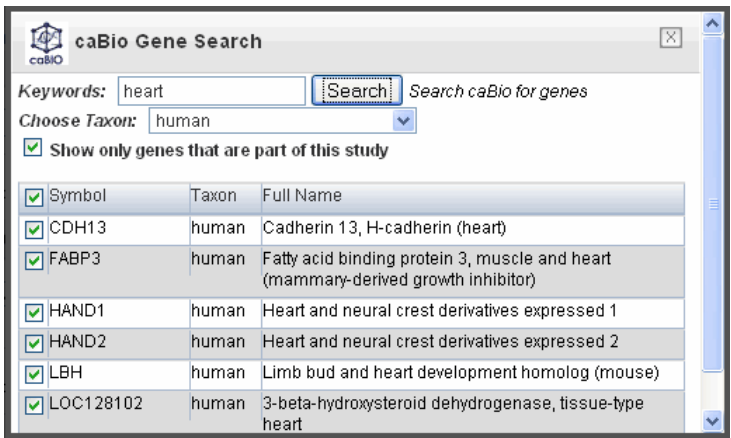


Field	Description
Gene Symbol	<p>Enter one or more gene symbols in the text box or click the icons to locate genes in the following databases. If entering more than one gene in the text box, separate entries by commas.</p> <p>CGAP – Use this directory to identify genes. Before clicking this link you 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 check boxes 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>
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 K-M 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.

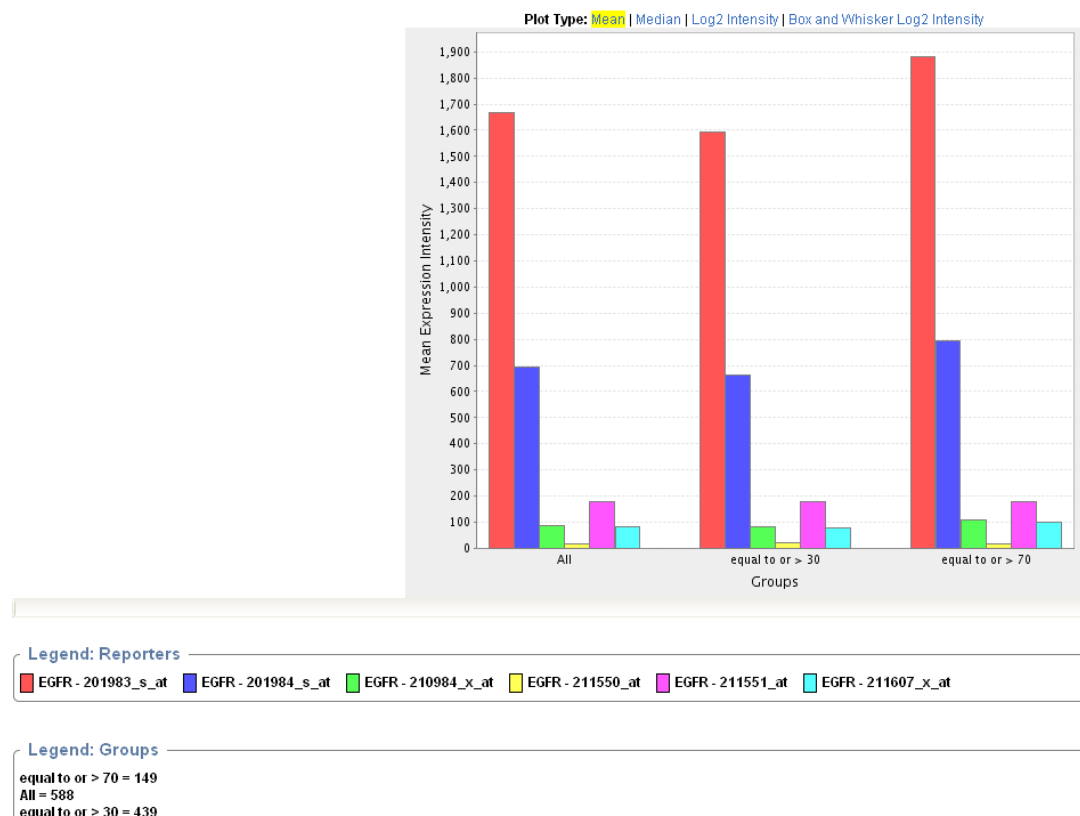


Figure 5.12 Gene expression plot based on clinical queries

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

Understanding a Gene Expression Plot

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

When you perform a Gene Expression simple search, by default the **Mean** Gene Expression Plot ([Figure 5.13](#)) appears.

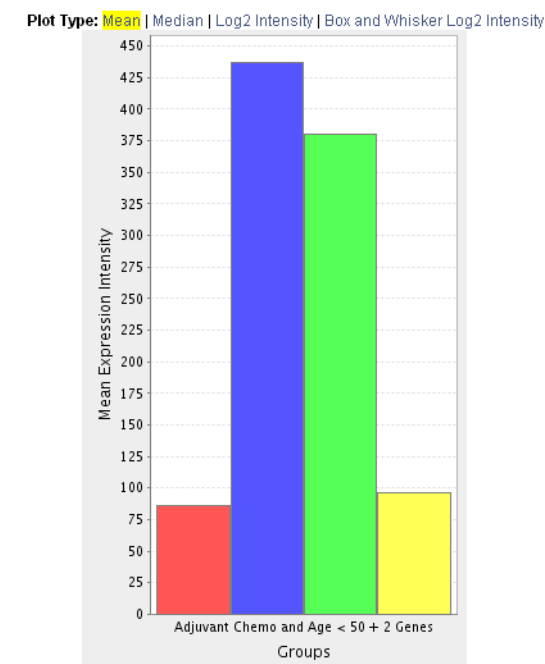


Figure 5.13 Gene expression plot calculating the mean

The **Mean** Gene Expression Plot ([Figure 5.13](#)) displays mean expression intensity (Geometric mean) versus Groups.

The **Median** Gene Expression Plot ([Figure 5.14](#)) displays the median expression intensity versus Groups..

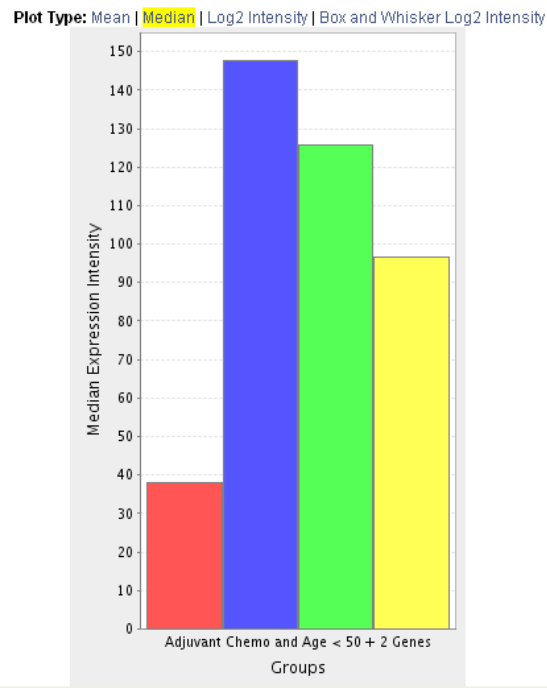


Figure 5.14 Gene expression plot calculating the median

The **Log2 Intensity** Gene Expression Plot ([Figure 5.15](#)) displays average expression intensities for the gene of interest based on Affymetrix GeneChip arrays (U133 Plus 2.0 arrays).

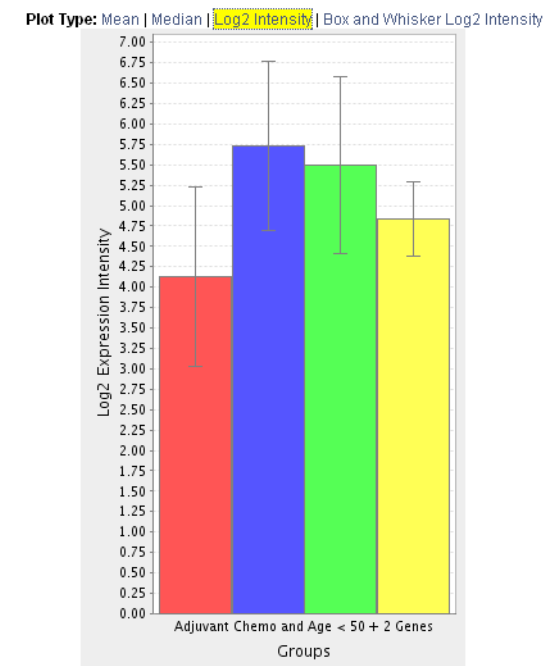


Figure 5.15 Gene expression plot displaying log2 intensity values

The box and whisker log2 expression intensity plot displays a box plot ([Figure 5.16](#), [Figure 5.17](#)). Example uses of box and whisker plots include the following:

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

Plot Type: Mean | Median | Log2 Intensity | **Box and Whisker Log2 Intensity**

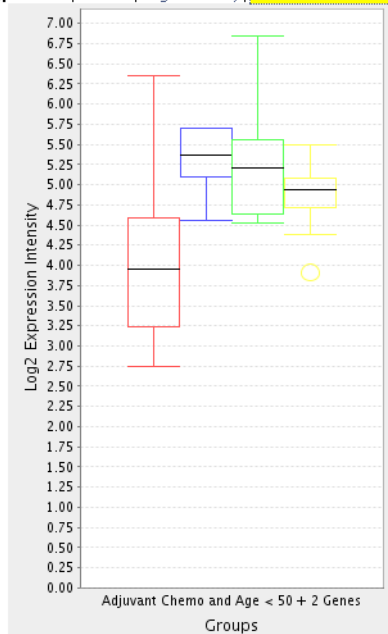


Figure 5.16 Box and whisker plot based on the same data set as represented in [Figure 5.13](#), [Figure 5.14](#), [Figure 5.15](#)

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

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

more than $1.5 \times \text{IQR}$ greater than Q3 and less than $1.5 \times \text{IQR}$ than Q1. Outliers, if present, are shown as open circles (*Figure 5.17*).

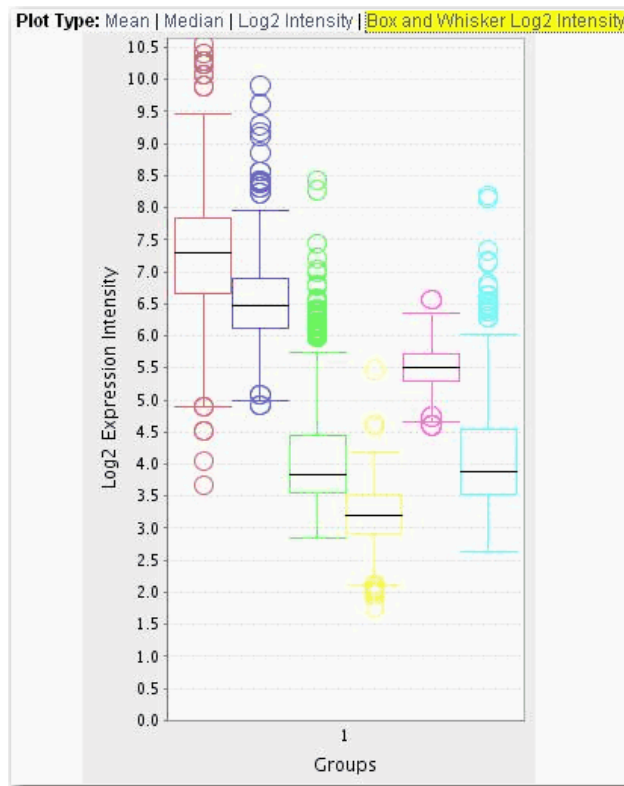


Figure 5.17 Box and whisker plot showing outliers

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

Analyzing Data with GenePattern

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

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

You have two options for using GenePattern from calIntegrator2:

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

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

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

1. Select the study whose data you want to analyze in the upper right portion of the calIntegrator2 page.
2. Click **GenePattern Analysis** in the left sidebar of calIntegrator2. This opens the GenePattern Analysis Status page (*Figure 5.18*).

GenePattern Analysis Status

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

Figure 5.18 GenePattern Analysis Status page

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

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

GenePattern Modules

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

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

- Select the study whose data you want to analyze in the upper right portion of the caIntegrator2 page.
- Click **GenePattern Analysis** in the left sidebar of caIntegrator2. This opens the GenePattern Analysis Status page.
- Make sure **GenePattern Modules** is selected in the drop down list. Click **New Analysis Job**.
- In the GenePattern Analysis dialog box (*Figure 5.19*), specify connection information, , described *Table 5.7*.

GenePattern Analysis

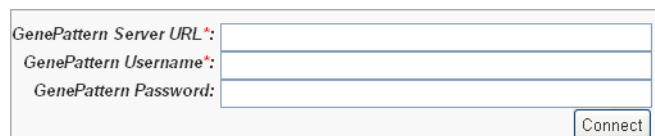


Figure 5.19 Dialog box for configuring the link to GenePattern

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

Table 5.7 Fields for selecting GenePattern configurations

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

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

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

Comparative Marker Selection (CMS) Analysis

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

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

To perform a CMS analysis, follow these steps:

1. Select the study whose data you want to analyze in the upper right portion of the calIntegrator2 page. You must select a study saved as a clinical study, but which has genomic data.
2. Click **GenePattern Analysis** in the left sidebar of calIntegrator2. This opens the GenePattern Analysis Status page.
3. In the GenePattern Analysis Status page, select **Comparative Marker Selection (Grid Service)** from the drop down list and click **New Analysis Job**. This opens the Comparative Marker Selection Analysis page (*Figure 5.20*).

Comparative Marker Selection Analysis

The screenshot shows the 'Comparative Marker Selection Analysis' web form. At the top, there are input fields for 'Job Name' (containing 'jhb1'), 'Preprocess Server' (a dropdown menu), 'Comparative Server' (a dropdown menu), and 'Clinical Queries' (a list box showing 'age 55-59'). Below these are 'Add >' and '< Remove' buttons. The main section contains various analysis parameters: 'Filter flag' (checkbox), 'Preprocessing Flag' (dropdown menu), 'Min Change' (text input: 3.0), 'Min Delta' (text input: 100.0), 'Threshold' (text input: 20.0), 'Ceiling' (text input: 2.1), 'Max Sigma Binning' (text input: 1), 'Probability Threshold' (text input: 1.0), 'Num Exclude' (text input: 0), 'Log Base Two' (checkbox), 'Number Of Columns Above Threshold' (text input: 1), 'Test Direction' (dropdown menu: two-sided), 'Test Statistic' (dropdown menu: T-test), 'Min Std' (text input: 1.0), 'Number Of Permutations' (text input: 1000), 'Complete' (checkbox), 'Balanced' (checkbox), 'Random Seed' (text input: 779948241), 'Smooth Pvalues' (checkbox), and 'Phenotype Test' (dropdown menu: one-versus-all). A 'Perform Analysis' button is located at the bottom right.

Figure 5.20 Comparative Marker Selection analysis parameters

4. Select or define CMS analysis parameters, described in [Table 5.8](#). An asterisk indicates required fields. The default settings are valid; they should provide valid results.

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

Table 5.8 Comparative Marker Selection analysis options

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

Table 5.8 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 ([Figure 5.21](#)).

GenePattern Analysis Status (draft)

Gene Pattern Modules


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

Figure 5.21 The progress of a GenePattern analysis that has been launched displays in the status column of page

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

Principal Component Analysis (PCA)

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

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

To perform a PCA analysis, follow these steps:

1. Select the study whose data you want to analyze in the upper right portion of the calIntegrator page. You must select a study with gene expression data.
2. Click **GenePattern Analysis** in the left sidebar of calIntegrator2. This opens the GenePattern Analysis Status page.
3. In the GenePattern Analysis Status page, select **Principle Component Analysis (Grid Service)** from the drop down list and click **New Analysis Job**. This opens the Principle Component Analysis page (*Figure 5.22*).

Principal Component Analysis

(draft)

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

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

* Job Name:

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

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

All Available Queries

- gender female
- Never smoke

Selected Queries

Add >

< Remove

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

Perform Analysis

Figure 5.22 Principal Component Analysis parameters

4. Select or define PCA analysis parameters, described in *Table 5.9*. You must enter a job name and select a clinical query, but you can accept the other default settings..

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

Table 5.9 PCA analysis options

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

Table 5.9 PCA analysis options

- If you want to preprocess the data set, click **Enable the Preprocess Dataset**. This opens an additional set of parameters ([Figure 5.23](#)), discussed in [Table 5.10](#). The preprocessing is executed prior to running the PCA.

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

* Preprocess Server: Default Broad service - <http://node255.broad.mit.edu:6060/wsrf/s>

Filter flag ☐

* Preprocessing Flag: no-disc-or-norm ▼

* Min Change: 3.0

* Min Delta: 100.0

* Threshold: 20.0

* Ceiling: 2.1

* Max Sigma Binning: 1

* Probability Threshold: 1.0

* Num Exclude: 0

Log Base Two ☐

* Number Of Columns Above Threshold: 1

Figure 5.23 Parameters for pre-processing parameters for PCA

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

Table 5.10 Parameters for preprocessing data sets for PCA

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

Table 5.10 Parameters for preprocessing data sets for PCA

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

GISTIC-Supported Analysis

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

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

- Select the study whose data you want to analyze in the upper right portion of the calIntegrator2 page. You must select a study with gene expression data.
- Click **GenePattern Analysis** in the left sidebar of calIntegrator2. This opens the GenePattern Analysis Status page.
- In the GenePattern Analysis Status page, select **GISTIC (Grid Service)** from the drop down list and click **New Analysis Job**. This opens the GISTIC Analysis page (*Figure 5.24*).

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:

Figure 5.24 GISTIC analysis criteria

4. Select or define GISTIC analysis parameters, as described in [Table 5.8](#). You must indicate a Job Name, but you can accept the other defaults settings, which are valid and should produce valid results.

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

GISTIC Parameters	Description
cnv File	<p>This selection is optional.</p> <p>Browse for the file. There are two options for the cnv file.</p> <p>Option #1 enables you to identify CNVs by marker name. Permissible file format is described as follows:</p> <p>A two column, tab-delimited file with an optional header row. The marker names given in this file must match the marker names given in the markers_file. The CNV identifiers are for user use and can be arbitrary. The column headers are:</p> <ol style="list-style-type: none"> 1. Marker Name 2. CNV Identifier <p>Option #2 enables you to identify CNVs by genomic location. Permissible file format is described as follows:</p> <p>A 6 column, tab-delimited file with an optional header row. The 'CNV Identifier', 'Narrow Region Start' and 'Narrow Region End' are for user use and can be arbitrary. The column headers are:</p> <ol style="list-style-type: none"> 1. CNV Identifier 2. Chromosome 3. Narrow Region Start 4. Narrow Region End 5. Wide Region Start 6. Wide Region End

Table 5.11 GISTIC analysis parameters

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

DATA IMPORT CONFIGURATIONS

This appendix describes configurations for importing data into a study.

Topics in this appendix include the following:

- [Subject Clinical Data Configuration](#) on this page
- [Delimited-Text Annotation Import](#) on page 85
- [Annotation Field Configuration](#) on page 86
- [Sample Data Configuration](#) on page 86
- [Genomic Data Configuration](#) on page 87
- [Imaging Data Configuration](#) on page 87

Subject Clinical Data Configuration

The following clinical data configuration information is collected:

- Clinical Data Source (delimited text)
- Protocol Id (of study to import)
- For delimited text, see [Delimited-Text Annotation Import](#). For subject annotation files, one field must be identified as the subject identifier.
- See [Annotation Field Configuration](#) for details on specification of visibility and browse configuration.

Delimited-Text Annotation Import

Delimited-text annotation files must be in standard comma-separated value format. The file must include a header line that specifies the name for each field. Each row of data must contain the same number of values as the header row. The file must include a column that will be designated as the identifier (e.g. subject identifier, sample identifier,

etc.) for each row. Optionally a file may include a single column that will be designated as a time-point indicator. Each row must contain a unique combination of identifier and time-point indicator of a unique identifier if no time-point is included. An example of the content of a file including a time-point is shown below.

```
"patientId", "timepoint", "bloodPressure", "weight"  
"1234", "T1", "120/80", "180"  
"1234", "T2", "125/80", "190"  
"5678", "T1", "120/85", "200"
```

After upload of the file, the Study Manager must indicate for each field:

- Field type (identifier, timepoint indicator, text, integer, float or boolean)
- After specification of these types, the file will be validated to ensure that the values are valid for the types selected and that the file conforms to the requirements given above.

Annotation Field Configuration

For each annotation field (regardless of the source), the Study Manager must specify the following information:

- Annotation semantics: each annotation field (whether associated with a subject, image series, image or sample) must either:
 - be associated with an existing annotation definition known to the system,
 - be associated to an existing CDE in caDSR or
 - have sufficient semantic metadata recorded so that the field may be submitted for registration as a CDE in caDSR.
- Field authorization: each field must be either declared publicly visible or a restricted to a list of groups. The default will be the visibility settings given at the study level.
- Whether the field is to be included in the results list for a given entity type (i.e. Subject, Sample, ImageSeries or Array Data) when browsing data (See Use Case: Browse Study Data).
- Whether the field is to be included in simple single-input searches when browsing data (See Use Case: Browse Study Data).

Sample Data Configuration

Sample data may be uploaded from either caArray 2 or from delimited-text import. Samples imported from caArray 2 may have annotation updated by use of the delimited-text import functionality if sample annotation is required. Import from caArray 2 requires specification of the following information:

- caArray server hostname
- caArray server JNDI port
- caArray username

- caArray password
- Either the experiment identifier (to import all samples in the experiment) or a file containing a comma-separated list of samples in the format “experiment identifier”, “sample name”.
- Mapping of samples to subjects. This may be specified by a comma-separated list in the format “subject identifier”, “sample identifier” or by a regular-expression based mapping formula.

When samples are imported via delimited-text import, the time-point is associated to the sample itself. This means that each sample may be associated with only one time-point (i.e. multiple time-points for the same sample are invalid).

Genomic Data Configuration

All genomic data (i.e. array data) is imported from caArray 2. First the Study Manager must specify sufficient information to map study samples to caArray 2 samples. If all samples were imported directly from caArray 2 as described in Special Requirement: Sample Data Configuration, no further information is required for this step. If samples were imported via delimited-text, the Study Manager must specify

- caArray server hostname
- caArray server JNDI port
- caArray username
- caArray password
- A mapping of calIntegrator2 sample identifiers to caArray 2 samples, specified as a comma-separated list in the format “calIntegrator2 sample identifier”, “caArray 2 experiment identifier”, “caArray 2 sample name”.

The system will enable the Study Manager to navigate easily to the selected caArray 2 instance.

Next, the system will indicate the available platforms and array data types available for the study samples. The Study Manager will indicate which platforms and data types to import and for each platform/data type combination will indicate:

- Whether to import the data
- The visibility of the data; either public or restricted to a set of groups. Low-level genotyping data (raw data and normalized) will always have restricted visibility.

Imaging Data Configuration

The following imaging data configuration information is collected:

- NBIA grid server hostname (defaults to NCICB instance)
- NBIA grid server port (defaults to NCICB instance port)
- Protocol Id

- Mapping of NBIA Patients to subjects imported from clinical data source. This may be specified by a comma-separated list in the format “subject identifier”, “NBIA patient identifier” or by a regular-expression based mapping formula.
- Which annotation fields to import from NBIA.
- The system will enable the Study Manager to navigate easily to the selected caArray 2 instance.

Additional annotation for either images or image series may be imported using the delimited-text import functionality.

INDEX

A

- account, requesting new user 6
- adding
 - clinical data 16
 - genomic data 24
 - imaging data 29
- annotation
 - assigning identifier 17
 - configuring field 86
 - importing delimited text 85
 - K-M plot for 54
 - searching for definitions 20
- Application Support i, 11
- assigning, annotation identifier 17

B

- box and whisker plot
 - interpretation 72
 - uses for 72

C

- caIntegrator2
 - logging in 8
 - logging out 10
 - requesting user account 6
 - using workspace 8
 - workspace 8
- caIntegrator2 User's Guide
 - introduction 1
 - organization 1
 - text conventions 2
- clinical study
 - adding data 16
 - data for 13
- columns, defining display 40
- Comparative Marker Selection (CMS)
 - data analysis 76
- configuring
 - annotation fields for import 86
 - copy number data 28
 - genomic data for import 87

- imaging data for import 87
- sample data for import 86
- subject clinical data for import 85
- control samples, uploading 27
- copy number data, configuring 28
- creating
 - K-M plot 54
 - study 14, 15

D

- data analysis, overview 53
- defining survival values 23
- delimited text annotation import 85
- DICOM, retrieving images 50

E

- editing a query 43
- exporting
 - data 52
 - query results 43

F

- fold change, control samples file 27

G

- gene expression, K-M plot for 56
- gene expression plot
 - description 53, 60, 70
 - for clinical queries 66
 - for genomic queries 64
 - interpreting 70
 - plot display, box & whisker 72
 - plot display, log2 intensity 71
 - plot display, mean 70
 - plot display, median 71
- GenePattern
 - analyses, description 73
 - analyses, in caIntegrator2 74
 - analyses, modules 75
 - CMS analysis 76

- GISTIC analysis 81
- PCA 78
- plot description 54
- genomic data
 - adding copy number data to 28
 - adding to study 24
 - configuring for import 87
 - for study 13
 - mapping to clinical data 26
- GISTIC-based data analysis 81

H

- hierarchy of objects, NBIA 37, 51

I

- imaging data
 - adding to study 29
 - configuring for import 87
 - for study 14
- importing delimited text annotations 85

K

- Kaplan_Meier plot see K-M plot
- K-M plot
 - creating 54
 - description 53, 54
 - for annotations 54
 - for gene expression 56
 - for queries 58

L

- logout link 10

M

- managing
 - platforms 32
 - queries 42
 - study 31
- mapping genomic to clinical data 26

N

- NBIA
 - forwarding imaging results to 49
 - viewing imaging results in 49
- NCICB Application Support i, 11

O

- objects in NBIA, hierarchy of 37, 51
- overview, chapters in guide 1

P

- patient
 - relationship to study, series, images 51
- platforms, managing 32
- plot
 - gene expression, description 53, 60
 - gene expression description 70
 - GenePattern description 54
 - K-M description 53
- Principal Component Analysis
 - data analysis 78

Q

- query
 - editing 43
 - exporting results 43
 - K-M plot for 58
 - managing 42
 - saving 42

R

- registering new user 6
- Results Type tab 40

S

- sample data, configuring for import 86
- saving query 42
- searching
 - annotation definitions 20
 - overview 35
 - study 36
- search results
 - browsing 46
 - exporting data 52
 - forwarding imaging results to NBIA 49
 - genomic data 46
 - imaging data 48
 - overview 45
 - retrieving DICOM images 50
 - viewing imaging data in NBIA 49
- sorting tab 40
- study
 - adding clinical data 16
 - adding genomic data 24
 - adding imaging data 29
 - clinical data, description 13
 - configuring copy number data 28
 - creating 13, 14, 15
 - deploying 31
 - editing 31
 - genomic data, description 13
 - imaging data, description 14

- managing [31](#)
- mapping genomic data to clinical [26](#)
- relationship patient, study, series, images [51](#)
- searching [36](#)
- uploading control samples to [27](#)
- subject clinical data, configuring for import [85](#)
- survival values, defining [23](#)

T

- Technical Support [i](#)
- text conventions in user guide [2](#)

U

- user's manual conventions [2](#)

