caINTEGRATOR v.1.3

User's Guide

NATIONAL® CANCER INSTITUTE

Center for Biomedical Informatics
and Information Technology

February 4, 2011

Credits and Resources

caIntegrator Development and Management Teams			
Development	***Quality Assurance***	***Documentation***	***Project and Product Management***
JP Marple[2]	Quy Phung[4]	JP Marple[2]	Shine Jacob[4]
Will Fitzhugh[2]	Henry Schaefer[4]	Jill Hadfield[1]	Anand Basu[1]
Eric Tavela[2]			Juli Klemm[1]
TJ Andrews[5]			Mervi Heiskanen[1]
Ngoc Nguyen[6]			
Matt Rehfuss[2]			
Huaitian Liu[7]			
Yuri Kotliarov[8]			
Karen Ketchum[4]			
Systems and Application Support		***Training***	
Cuong Nguyen[2]			
Deanna Siemaszko[3]			
[1] NCI Center for Biomedical Informatics and Information Technology (CBIIT)		[2] 5AM Solutions	[3]Terrapin Systems
[4]Enterprise Solutions And Consulting (ESAC)	[5]ScenPro	[6]Claris LLC	[7] Science Application International Corporation (SAIC)
[8]National Cancer Institute (NCI)			

Contacts and Support	
NCICB Application Support	http://ncicb.nci.nih.gov/NCICB/support Telephone: 301-451-4384 Toll free: 888-478-4423

TABLE OF CONTENTS

USING THE CAINTEGRATOR V.1.3 USER'S GUIDE

This chapter introduces you to the *caIntegrator v.1.3 User's Guide* and suggests ways you can maximize its use.

Topics in this chapter include:

Introduction to the caIntegrator User's Guide

The *caIntegrator v.1.3 User's Guide* is the companion documentation to the caIntegrator software application. The user's guide includes information and instructions for the end user about using caIntegrator.

Organization of this Guide

The *caIntegrator v.1.3 User's Guide* contains the following chapters and appendices:

Using the caIntegrator User's Guide — This chapter introduces you to the *caIntegrator v.1.3 User's Guide* and suggests ways you can maximize its use.

Chapter 1 Getting Started in caIntegrator — This chapter introduces general caIntegrator procedures and how to obtain help to use caIntegrator.

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

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

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

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

Chapter 6 Administering User Accounts —This chapter describes the process for creating and managing user accounts in caIntegrator.

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

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.
Italics	Used to reference other documents, sections, figures, and tables.	*caCORE Software Development Kit 1.0 Programmer's Guide*
`Italic boldface monospace type`	Text that you type	In the New Subset text box, enter ***`Proprietary Proteins.`***
Note:	Highlights a concept of particular interest	**Note:** This concept is used throughout the installation manual.

Table 2.1 caIntegrator User's Guide Text Conventions

Convention	*Description*	*Example*
Warning!	Highlights information of which you should be particularly aware.	**Warning!** Deleting an object will permanently delete it from the database.
{ }	Curly brackets are used for replaceable items.	Replace `{root directory}` with its proper value, such as `c:\cabio`

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

CHAPTER

1

GETTING STARTED WITH CAINTEGRATOR

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

Topics in this chapter include:

About caIntegrator

NCI, Center for Biomedical informatics and Information Technology (CBIIT) is developing a novel translational informatics platform called caIntegrator that allows researchers and bioinformaticians to access and analyze subject annotation and experimental data across multiple subject annotation trials and studies. The caIntegrator 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, subject annotation trials data, etc.) in a cohesive fashion.

caIntegrator is a web based or locally installed portal that allows researchers and study managers to access the biomedical informatics infrastructure and data analysis tools established by caIntegrator from one common software platform. As a caIntegrator user, you can perform the following tasks:

- Integrate subject annotation data with genomic and/or imaging data

- Import data of various types in a predefined flat format, and create new studies with multiple study data
- Update an existing study to add new attributes or to add/modify data
- Perform analyses on study data

Registering as a New caIntegrator User

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

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

Figure 1.1 New user account registration form

3. In the Register form, enter the appropriate information[1].
   - **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 caIntegrator without an account to browse and

1. Items with an asterisk or highlight are required.

search available experiments and download data while your account is verified and activated.

–**Username***

–**LDAP Password***

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

If your LDAP profile is not validated, caIntegrator 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 caIntegrator, and continue entering the appropriate information in the **Account Details** section.

Role	*Description*	*Permissible 1.0 Actions*
Study Manager	Creates, owns and manages studies	Create studies Assign annotations to 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

Table 1.1 caIntegrator 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 caIntegrator system administrator and an account request confirmation email is also sent to you, the prospective user, at your specified email address. In response, the caIntegrator 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 caIntegrator.

When your account is registered, the user ID and password you are assigned determine your access rights for the software.

Welcome to the caIntegrator Workspace

The caIntegrator Welcome workspace enables quick access to all caIntegrator functions and information before you login. The Welcome page also displays after you log in, before you open any studies (*Figure 1.2*).

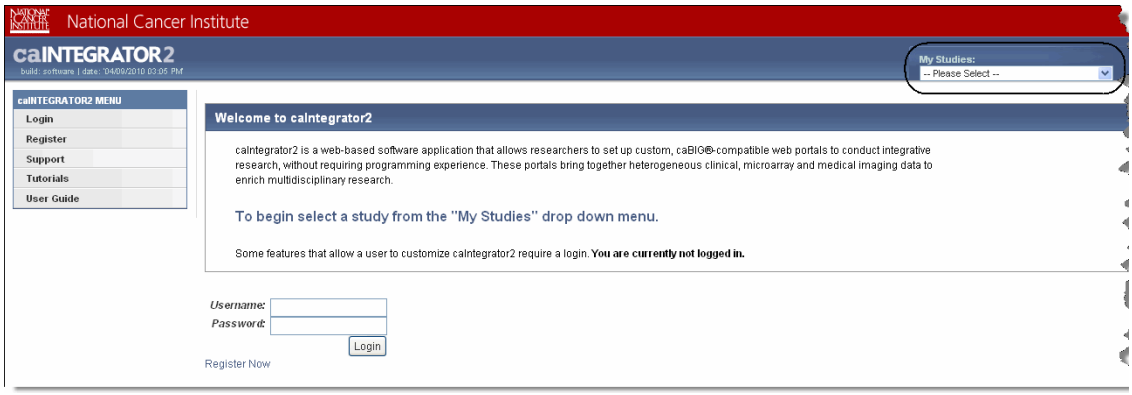


Figure 1.2 Welcome page that displays before and after login

Without logging in, you can browse any public studies. To do so, select from the drop-down list of public studies in the upper right-hand corner of your browser (*Figure 1.2*).

To log into caIntegrator, 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

To access caIntegrator functions, use the options listed on the left sidebar of the workspace.

caIntegrator Functions

When you log into caIntegrator, 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 caIntegrator option in the workspace (*Figure 1.2*).

Sidebar Option	*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 caIntegrator Study*.
Study Data	Click **Saved 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 61. Click **Saved Lists** > **Global Lists or** > **My Lists** to open gene lists that have been saved for a study. From any page in caIntegrator that shows such a group, you can save a such a list of genes to be used for searches or analyses. See *Creating a Gene or Subject List* on page 67. Click **Saved Copy Number Analysis** to open an Edit GISTIC Analysis page. You can review genes retrived from the analysis or, with appropriate permissions, you can edit the metadata for the analysis. See *Editing a GISTIC Analysis* on page 72.
Analysis Tools	Click any of the listed options to open a page where you can launch an analysis of the data in the selected study. • Generate a K-M Plot. See *Creating Kaplan-Meier Plots* on page 80. • Generate a Gene Expression Plots. See *Creating Gene Expression Plots* on page 86. • Launch GenePattern Analysis. *Analyzing Data with GenePattern* on page 99.
Study Management	Click either of the listed options to manage the selected study through editing or deleting it or by creating a new study. • Click **Manage Studies.** See *Managing a Study* on page 45. • Click **Create a New Study.** See *Configuring and Deploying a Study* on page 16.
Application Management	Click **Manage Platforms** to identify, add or remove platforms that caIntegrator supports. For more information, see *Managing Platforms* on page 47.

Table 1.2 caIntegrator tabs

Sidebar Option	Function
caIntegrator Menu	• Click **Support** to view contact information for Application Support. • Click **Tutorials** to view a tutorial to help you get started using caIntegrator. • Click **User Guide** to open the caIntegrator v.1.0 User's Guide in PDF format.

Table 1.2 caIntegrator 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 well as public studies.) 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 and/or saved lists for that study

Viewing Existing Studies

If you have not logged into caIntegrator, you can view any public studies in your browser. After logging in, you can view existing studies for which you have been granted permission . In the upper right corner of the page, in the My Studies drop down list, select the study you want to review or work in (*Figure 1.3*).

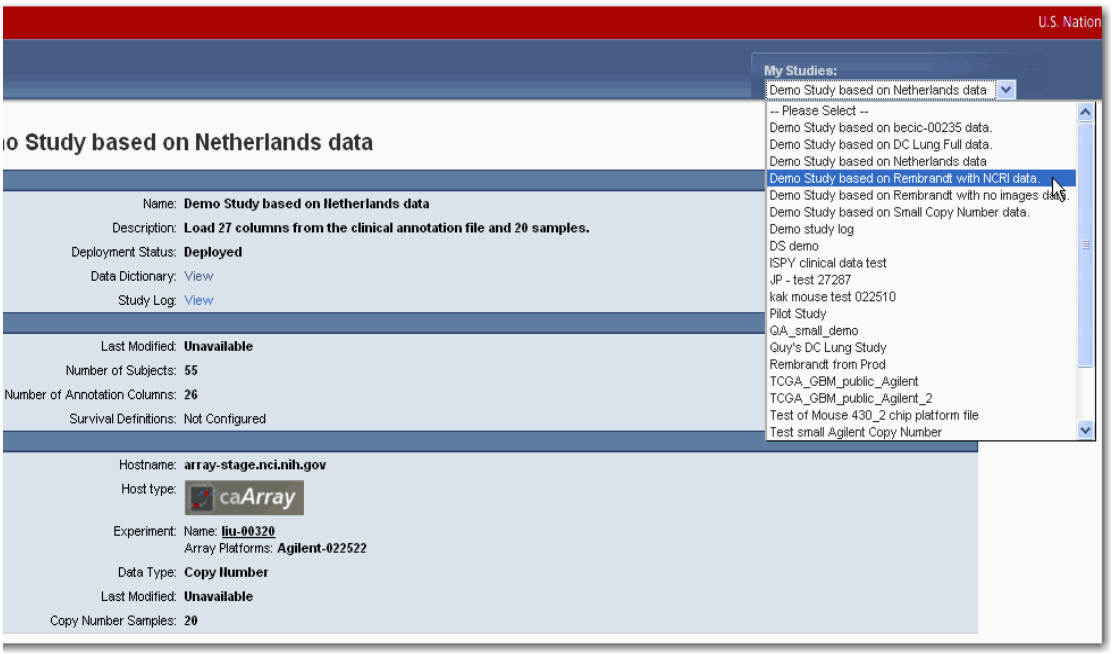


Figure 1.3 Drop-down list for selecting existing studies

The study you select opens in the browser. You can review the study data for which you have been granted permission.

After selecting the study name, in the **My Studies** drop-down list, 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.

When the annotations are uploaded during the creation of the study each field is defined by the study manager.

- Because in looking at the study, you may not know the meaning of all the annotations, you can open a reference page with a summary of the annotations. Click the **Data Dictionary: View** link on the study home page (*Figure 1.4*).
- From the study summary page, you can also open a log for the study. Click the **Study Log: View** link on the page to see all log entries with descriptions.

Data Dictionary

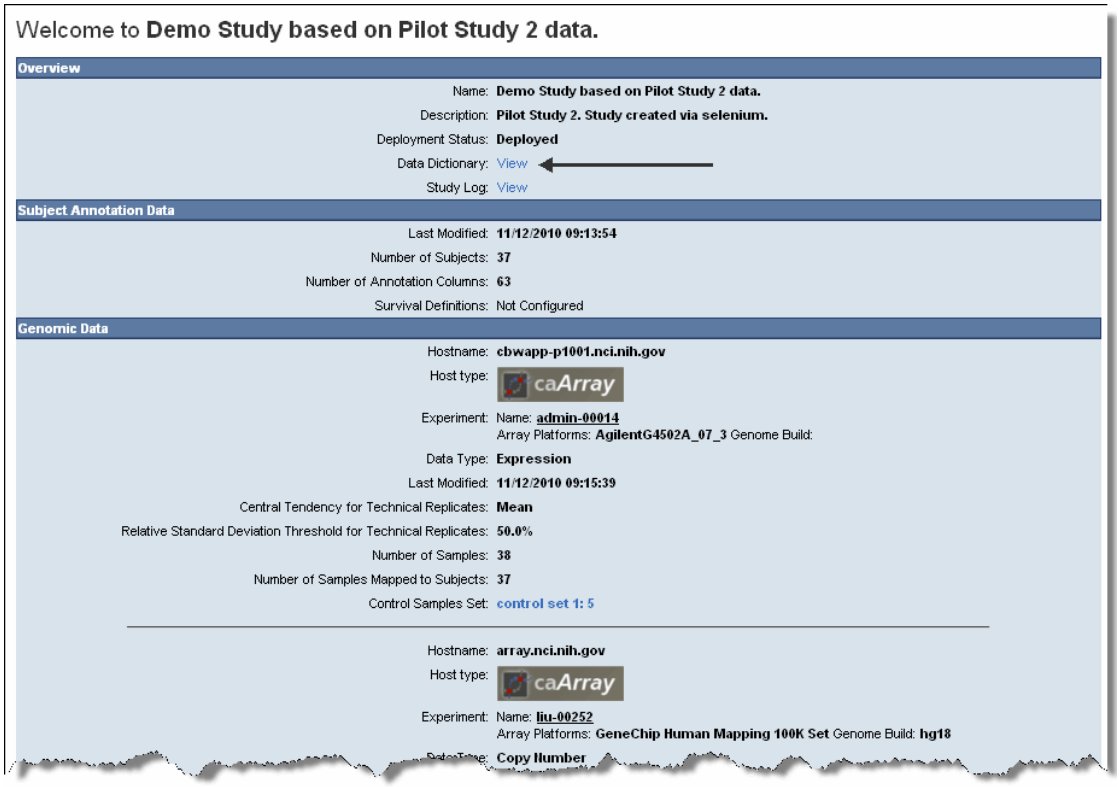


Figure 1.4 A link on a study home page opens a data dictionary summary

The Data Dictionary consists of a table that clarifies all annotations used in the study. It displays their field descriptors, descriptions, caDSR identifiers (if used), caDSR IDs and definitions, data type, and permissible settings (*Figure 1.5*). The **Restrictions** column

indicates whether or not masks have been applied to numeric data in the study. For more information, see *Assigning An Identifier or Annotation* on page 23..

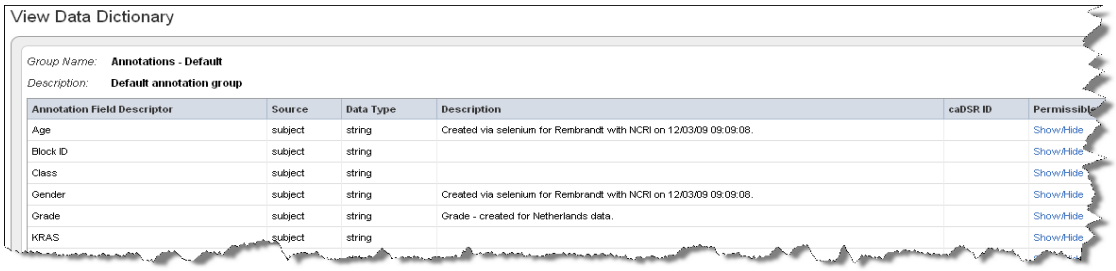


Figure 1.5 Page for viewing data dictionary details

For more information about study details, see *Creating/Editing a Study* on page 17.

Study Log

The study log which you can open by clicking the **Study Log > View** link on the study summary page lists step used to create a study. For more information, see *Viewing/Editing a Log* on page 18.

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.

When you open online help, the table of contents displays in the left panel.

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

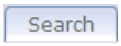
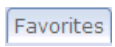
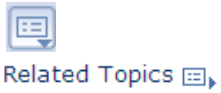
Icon or Button	***Description***
	Locates and highlights your current topic in the table of contents.
	Select a topic from the complete online help table of contents.
	Select a topic from the online help index.
	Perform word searches of Help by entering query text in the search text box.
	Create a list of your frequently-accessed topics.
	Opens other closely related topics.
	Prints the current topic.

Table 1.3 Online help tips

Icon or Button	Description
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 caIntegrator, 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: • 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 caIntegrator.

Topics in this chapter include:

Creating a Study – Overview

You can create a caIntegrator study by importing subject annotation study data, genomics data and imaging data, using a combination of spreadsheet/files and existing caGrid applications as source data. Each instance of caIntegrator 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 caIntegrator.

- **Subject Annotation** – Subject annotation data refers to pre-subject annotation, phenotypic, subject annotation, pathology or any other annotations associated with a subject.The subject annotation 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 caIntegrator. You, as the study creator, must have access to the subject annotation data file, as the file does not come from a caBIG® repository.
- **Genomic** – To use caIntegrator to integrate array data, the data should be imported into caArray, either locally or in 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 subjects in the subject annotation data files. See

- **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 (required) and annotation IDs in the second column. You must also have an image mapping file in CSV format. This file indicates correlations between subject annotation subjects or images in NBIA and subjects in the subject annotation 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 a 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 (subject annotation, 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*).

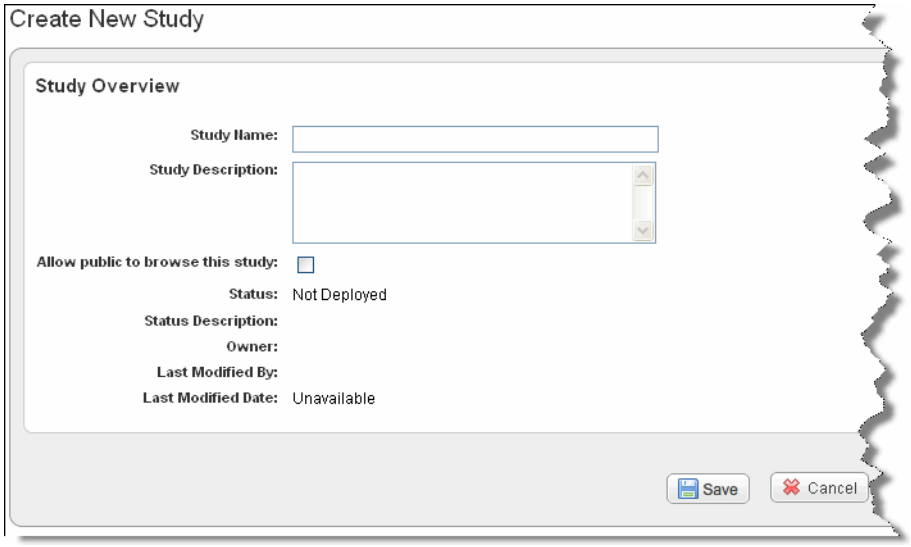


Figure 2.1 Create Study page

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*).

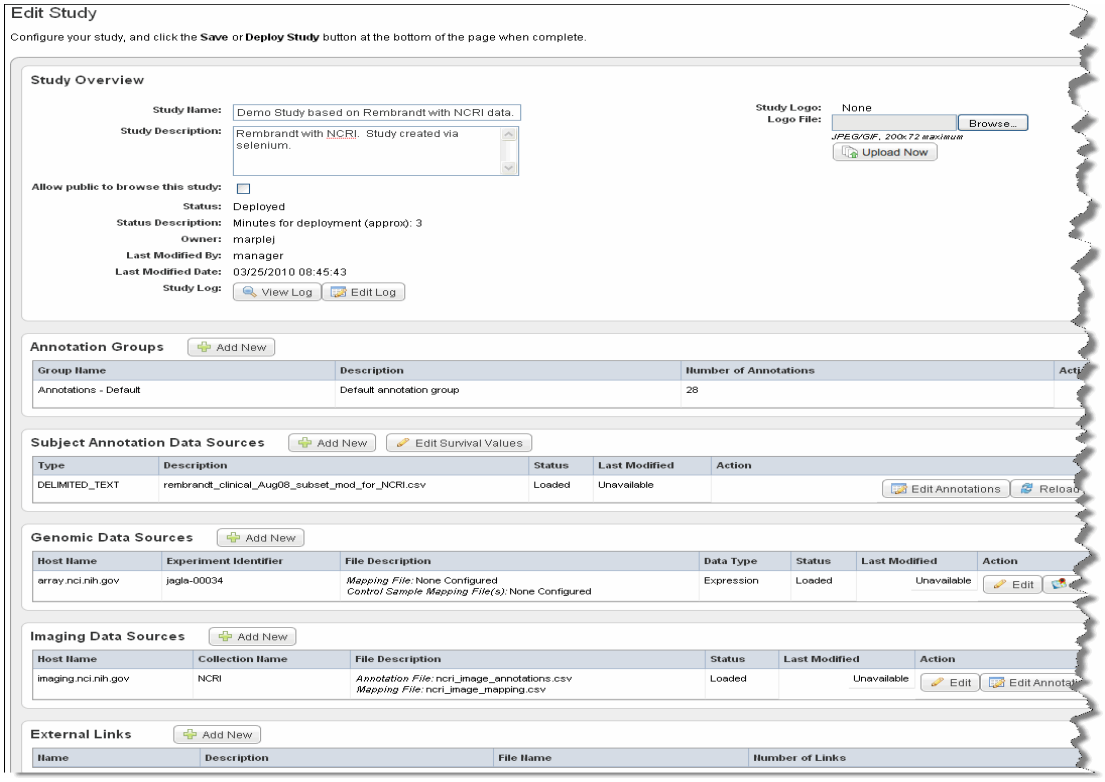


Figure 2.2 Edit Study page

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

1. Enter or change(if editing) the name and/or description, if you choose.
2. Check the checkbox to make the study publicly available, if appropriate.
3. For the study log feature, click **View Log** or **Edit Log.** See *Viewing/Editing a Log* for details about the log.
4. 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.

5. 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.

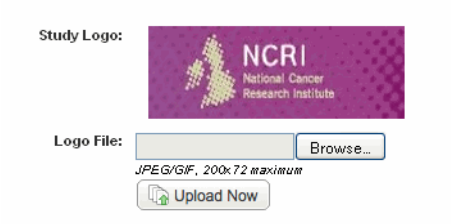


Figure 2.3 Example of a logo added to the caIntegrator browser on the Edit Study page

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

Viewing/Editing a Log

On the Edit Study page, as a study manager you can open a detailed log for the study.

1. Click **View Log** on the Edit Study page to simply review an existing log. The log records all steps comprising activity in the study, with the most recent displaying at the top of the log.
2. To edit a log, click **Edit Log** on the Edit Study page.

   Add an appropriate description/annotations to the individual log entries.
3. Check the **Update** box next to the description, then click **Save** to save the edits. The descriptions will now be available when any user views the log.

See also *Study Log* on page 12.

Working with Annotations – An Overview

One of the most important factors in creating a study in caIntegrator is in properly annotating the data. Because the process can be relatively complex, you might want to review the steps for working with annotations.

Annotation workflow summary:

1. Add an annotation group. This optional step is for users who have a rigid data dictionary of all annotations relevant to the study. This step can also be helpful in cases where a study has many annotations. For more information, see *Adding An Annotation Group* on page 19.
2. Add subject annotation data. This consists of multiple sub-steps.
   a. Add a new subject annotation data sources file. This step uploads the file and starts the workflow for assigning uploaded data definitions. See *Editing an Annotation Group* on page 20, step 1.
   b. Edit the annotations. This step opens the Define Fields for Subject Data page. See *Editing an Annotation Group* on page 20, step 2.

c. In the Define Fields for Subject Data page, review possible definitions in the annotation group associated with this study. See *Define Fields Page for Editing Annotations* on page 21.

d. Assign the visibility of each annotation definition. See *Editing an Annotation Group* on page 20, step 1.

e. Locate and verify the assignment as “identifier” for one annotation. See *Assigning An Identifier or Annotation* on page 23.

f. Review, verify and assign definitions for each annotation. You can do this in one of four ways:

— Accept existing default definitions as described in the associated annotation group. See *Assigning An Identifier or Annotation* on page 23.

— Create or manage definitions manually. See *Assigning An Identifier or Annotation* on page 23.

— Search for and use definitions existing in other caIntegrator studies. see *Searching for Annotation Definitions* on page 26.

— Search for and use definitions from caDSR. see *Searching for Annotation Definitions* on page 26.

3. Load the Subject Annotation Source. Up until this point, you can periodically save your work with the annotations, but before you can deploy the study, you must complete this step.

4. Deploy the study. See *Deploying the Study* on page 45.

Adding An Annotation Group

An annotation group is a group of annotation definitions configured in a CSV file. This feature is primarily meant for the Study Manager who knows that they have tightly restricted vocabulary definitions that are relevant to a study. In this optional step, you can review the uploaded Group Definition Source file before assigning the appropriate definition for your study.

To add an annotation group, follow these steps:

1. On the Edit Study page for a study, Annotation Groups section, click the **Add New** button.

2. On the Edit Annotation Group page that opens, enter a name for the annotation group.

3. Enter a description (optional).

4. Browse for the Group Definition Source CSV file.

   The CSV file must include columns with these column headers in the first row: File Column Name, Field Type, Entity Type, CDE ID, CDE Version, Annotation Def Name, Data Type, Permissible, and Visible. Subsequent rows in the file define each subject annotation column in the subject annotation file.

a. If a subject annotation is defined by a CDE Public ID, values for the following columns are required: File Column Name, Field Type, Entity Type, CDE ID, and Visible; a value for CDE Version is optional.

– OR –

b. If a subject annotation definition is not defined by a CDE Public ID, values for the following columns are required: File Column Name, Field Type, Entity Type, Annotation Def Name, Data Type (String, Date, Numeric), Permissible (Yes or No), and Visible (Yes or No).

5. Click **Save**. This uploads the file, whose name now displays on the Edit Study page under Annotation Groups.

When you open the Define Fields for Subject Data page, the annotation definitions in the file you uploaded display on the page, available for assignment in the study. Additionally, you can view the definitions by viewing the annotation group listed in the first column of the matrix.

Note: Annotation definitions by default are visible only to the Study Manager's group. They are not visible to all caIntegrator users, unless you change the visibility for each.

Editing an Annotation Group

To edit an annotation group, on the Edit Study page for a study with an existing annotation group, click the **Edit Group** button.

1. You can change the Name and Description for the group.
2. A list of annotation definitions applied to the original annotation group displays on the Edit Annotation Group page.
   - In the drop-down list, you can select a different annotation group for the annotation definition.
   - You can change the visibility for the annotation definition.
   - Click **Change Assignment** to modify the properties of the annotation definition.
3. Click **Update Annotations** to confirm your edits for the group.

Adding Subject Annotation Data

The Edit Study page, described in *Creating/Editing a Study* on page 17, opens after you save a new study or click to edit an existing study.

To add subject annotation metadata on this page, follow these steps:

1. In the Subject Annotation Data Sources section of the page, click the **Add New** button. The page expands to reveal new fields for you to identify information about the annotation data sources.
2. Navigate to locate a subject annotation data file which is required for a study. Files must be in CSV file format.

3. Click the appropriate box if you want caIntegrator to **Create an annotation definition if one is not found**.
4. Click **Upload Now** to load the annotation source data.

After the data file is uploaded to this study, it will be listed in the Subject Annotation Data Sources section of the Edit Study page.

From this page you can initiate editing the annotations. In the Subject Annotation Data Sources section, click **Edit Annotations** corresponding to the subject annotations that have been uploaded for the study. This open the *Define Fields Page for Editing Annotations*.

Define Fields Page for Editing Annotations

The Define Fields for Subject Data page opens when you click **Edit Annotations** in the Subject Annotation Data Sources or the Image Data Sources section of the Edit Study page (*Figure 2.4*). The exception to this is if you have not yet imported annotations for the imaging data for the study, In that case, when you click the **Edit Annotations** button in the Imaging Data Sources section, a page opens where you can identify and upload image annotation data (*Adding or Editing Image Annotations* on page 42).

If this Define Fields page opens after clicking the Edit Annotations button, working with this page is identical for both subject and image annotations

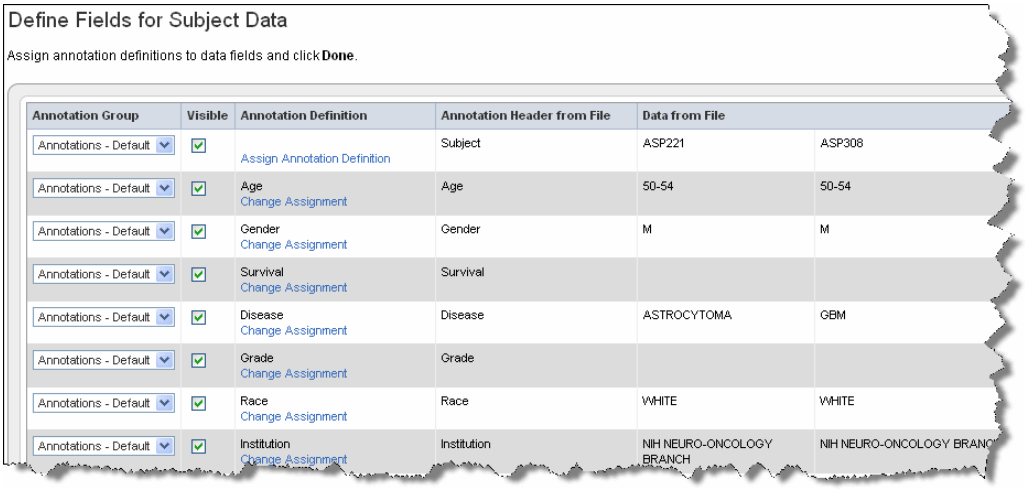


Figure 2.4 Define Fields for Subject Data page

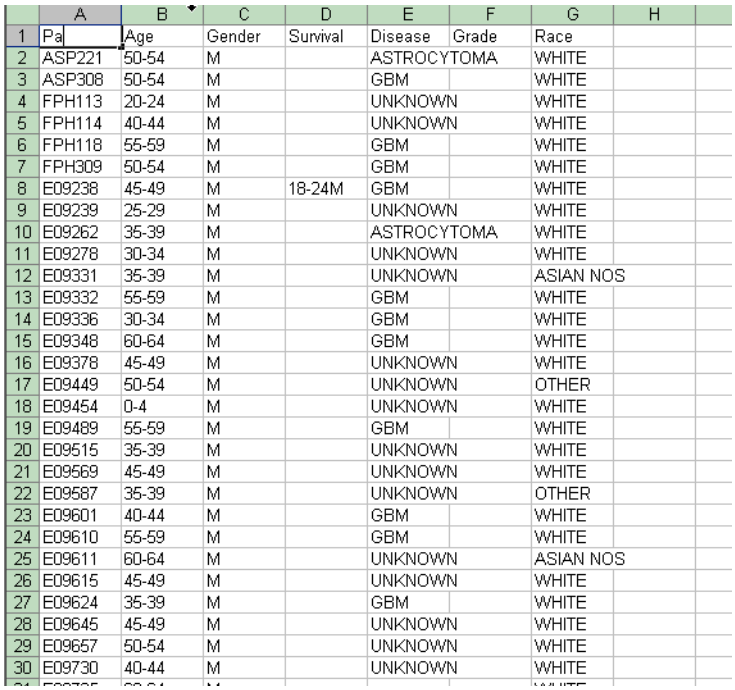
The first column of the table on this page displays annotation groups that have been created for this study. For more information, see *Adding An Annotation Group* on page 19.

To add subject or image annotation metadata in this page, follow these steps:

1. You can specify visibility of specified annotation data in the **Visible** column.
    - Select a checkbox for a row to make the corresponding data visible to all subscribers of the study or anonymous users if the study is made available to the public.

- Clear a checkbox to hide the corresponding annotation from any subscriber or anonymous user of the study. Data continues to exist but does not show up in query fields nor query results.

2. The Annotation Header from File column on the Define Fields for Subject (or Image) Data 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 caIntegrator 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 caIntegrator

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

- 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.

Note the following regarding the list of annotations on this page:

— 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 **Annotation Definition** column above the blue **Change Assignment** link.

— When the annotation definition has not been assigned, and the area above the blue **Assign Annotation Definition** link is blank, no correlating term exists in the database. In this case, you must specify the field type, and then the term will populate the space. See *Assigning An Identifier or Annotation* for more information.

— A field name that displays in red indicates an error in the annotation. Click the **Change Assignment** button for more information about the error.

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 Field Descriptor 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.

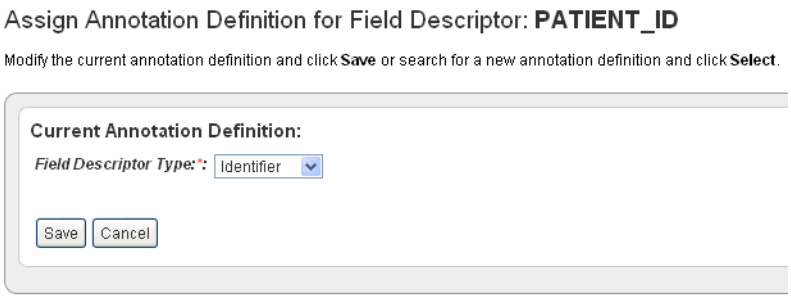


Figure 2.6 The Assign Annotation Definition dialog box

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**.
2. Click **Save** to save the identifier. This returns you to the Define Fields for Subject 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 Annotation Definition you want to review, click **Change Assignment**.
4. In the Assign Annotation Definition for Field Descriptor 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 and/or manage your own definitions manually. See Step 6.
- You can search for and use definitions in other caIntegrator studies. See *Searching for Annotation Definitions* on page 26.

- You can search for and use definitions found in caDSR. See *Searching for Annotation Definitions* on page 26.

5. Review the current annotation definition in the Assign Definition page, Current Annotation Definition section. Click **Cancel** to return to the Define Fields... page.

   You can still initiate a search for another annotation definition in the Search for an Annotation Definition section if you choose to change the definition (*Figure 2.7*). See *Searching for Annotation Definitions* on page 26. Click **Save** to retain any changes.

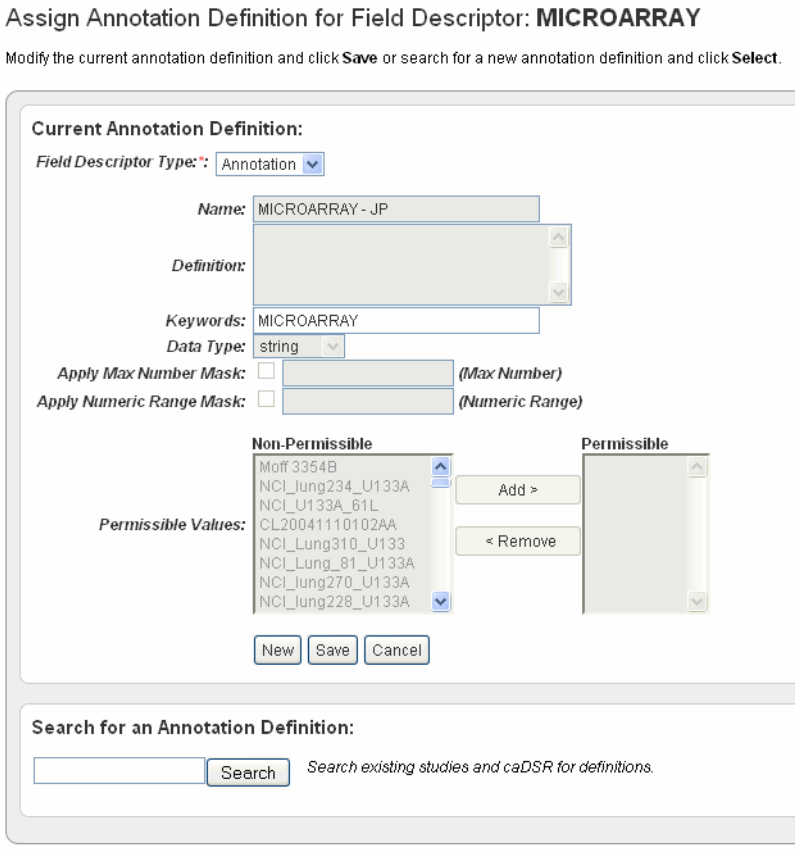


Figure 2.7 Current Annotation Definition

6. 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	Select a string (default), numeric, or date.

Table 2.1 Annotation fields for new definitions

Annotation Field	*Field Description*
Apply Max Number Mask	This field is available only for numeric-type annotations, or when a new definition is created. This feature is unavailable when permissible values are present. Select the box and enter a maximum number for the mask, such as “80” for age. When you query results above the value of the mask, then the system displays the mask and not the actual age. **Note:** If you enter masks of both “max number” and “range”, caIntegrator applies both masks at the same time. The Data Dictionary page now has a Restrictions column that shows restrictions whenever a mask has been applied.
Apply Numeric Range Mask	This field is available only for numeric-type annotations, or when a new definition is created. This feature is unavailable when permissible values are present. Select the box and enter a width of range for the mask, such as “5” representing blocks of 5 years. For example, if you enter a width of 5, the query only allows age blocks of 0-5, 6-10, 11-15, etc. When you query results above the value of the mask, then the system displays the mask and not the actual age ranges. **Note:** If you enter masks of both “max number” and “range”, caIntegrator applies both masks at the same time. The Data Dictionary page now has a Restrictions column that shows restrictions whenever a mask has been applied.

Table 2.1 Annotation fields for new definitions

Annotation Field	*Field Description*
Permissible/Non-permissible Values	**Note:** The first time you load a file, before you assign annotation definitions (step 3 on page 23), these panels may be blank. If the column header for the data is already "recognizable" by caIntegrator, 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. When you select or change annotation definitions by selecting matching definitions (described in *Searching for Annotation Definitions* on page 26), this may add (or change) the list of non-permissible values in this section. 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. 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. To edit a field's permissible values, you must change the annotation definition. You can do this even after a study has been deployed. **Note:** You cannot edit permissible values in an existing annotation definition. To change permissible values, you must create a new annotation.

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 caIntegrator studies or in caDSR.

1. Enter search keyword(s) in the **Search** text box on the Assign Annotation Definition page. Click **Search** or click **Enter** to launch the search. After a few moments, the search results display on the page (*Figure 2.8*).

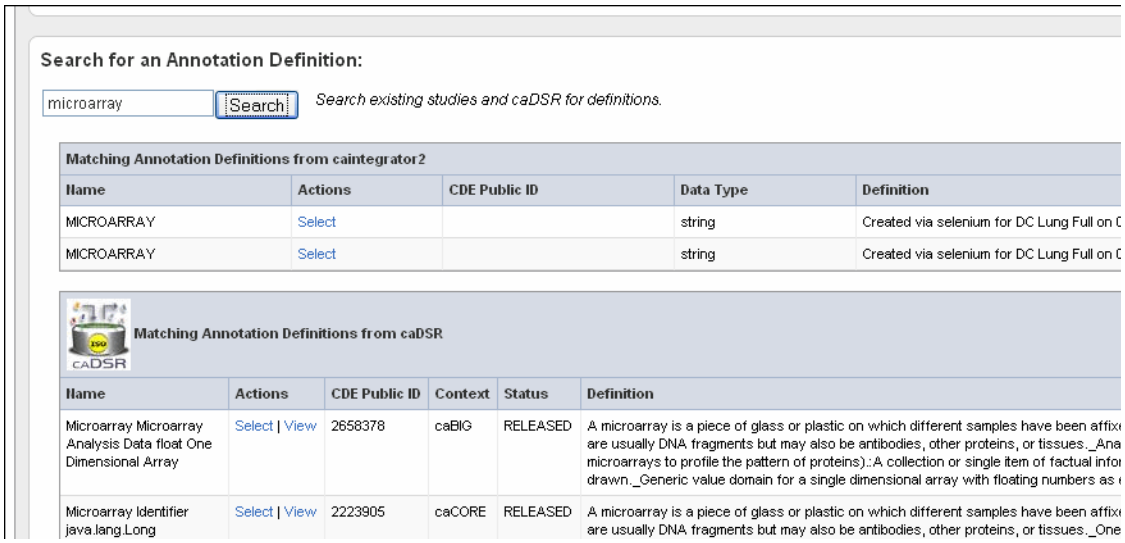


Figure 2.8 Results for annotation definition search

2. To view the definitions corresponding to any of the “Matching Annotation Definitions”, which are those currently found in other caIntegrator 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 Subject 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 24.

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 Fields for Subject 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 Subject Data page, be sure and designate the data types 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 Subject Data screen has an entry, one as the unique 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".

Tip: Saving your entries in this way saves the study by name and description, but does not deploy the study. See *Deploying the Study* on page 45.

The Edit Study page now displays a "Not Loaded" status for the file whose annotations (column headers) you have defined (*Figure 2.9*).

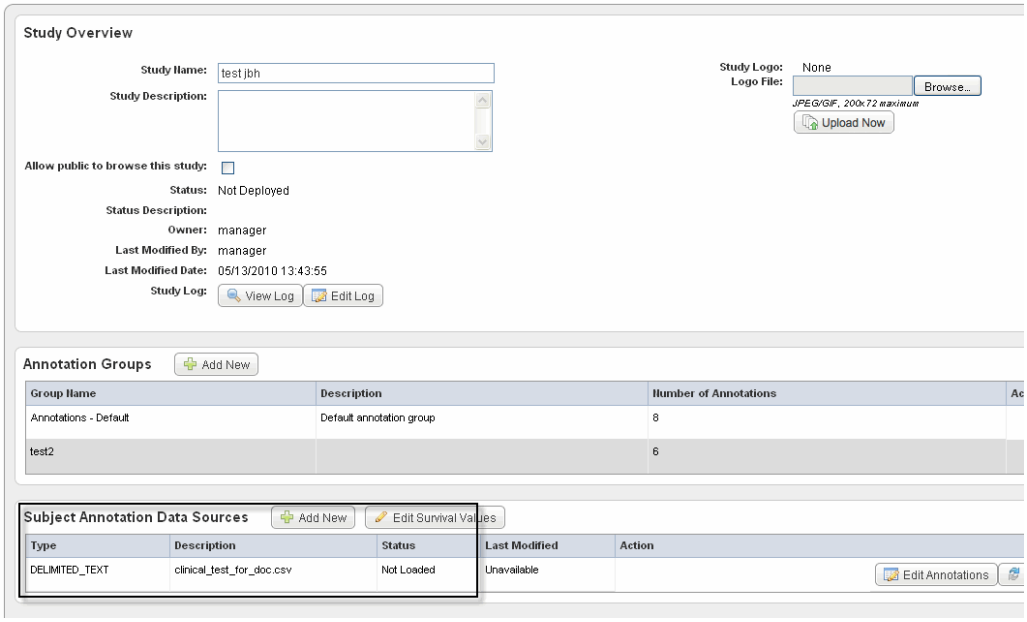


Figure 2.9 Example file whose annotations have been defined but not yet loaded

Status definitions:

- **Definition Incomplete** – An annotation definition or definitions must be modified on the Define Fields for Subject Data page. This status may be displayed because an identifier has not been selected. See *Define Fields Page for Editing Annotations* on page 21.
- **Not Loaded** – The annotation definitions must be loaded before a study can be deployed. If an error appears after attempting to load a subject annotation source, cick the **Edit Annotations** button which takes you to the Define Fields for Subject Data page where the problematic annotations will appear in red. See *Define Fields Page for Editing Annotations* on page 21.
- **Loaded** – The annotation definitions are properly loaded.

6. Click the **Load Subject Annotation Source** button in the Action section to load the data file you have configured, The **Deploy Study** button, to this point has been unavailable, but this step activates the button.

   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.

7. Click **Deploy Study**. caIntegrator now loads data from the file to the caIntegrator database, and the file status changes to “Loaded”.

   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 17.

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 45.

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

See also *Deploying the Study* on page 45.

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 caIntegrator data 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 caIntegrator, 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 2.10*).

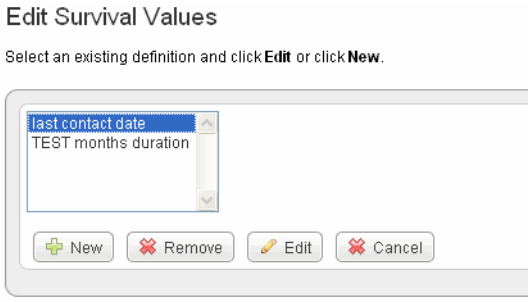


Figure 2.10 Survival Value Definition dialog box

2. Click **New** to enter new survival value definitions.

   - OR -

   Click **Edit** to edit existing survival value definitions.

3. The dialog box extends, now displaying radio buttons and three drop-down lists that show column headers for date metadata in the spreadsheet you have uploaded. *Figure 2.11* displays survival value ranges that have already been added to a study.

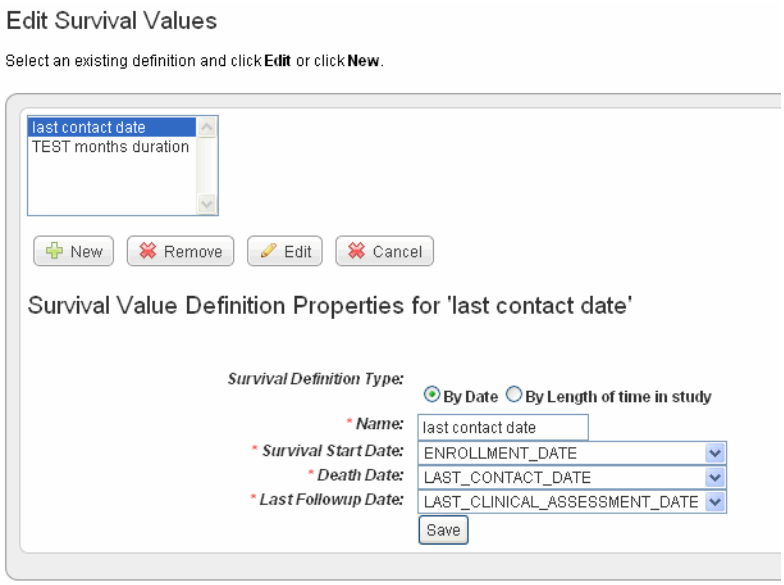


Figure 2.11 Survival Definitions example

Survival values can be defined by Date or by Length of time in study. Select the radio button for the category that defines your survival data.

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.

- **Survival Definition Type** – Select whether the survival time is defined by dates or length of time subject was in the study.
- **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 Length Units** – Select the appropriate units for this data.
- **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 80.

Updated the Edit Survival Value Definitions page, now has a radio button and 2 different types of ways to define survival values.

Adding/Editing Genomic Data

Note: Genomic data must be parsed and stored in caArray to be able to analyze it in caIntegrator.

Once you have loaded subject annotation data and identified patient IDs, you can add either one or more sets of array genomic sample data from caArray, which caIntegrator maps by sample IDs to the patient IDs in the subject annotation data, covered in this section, or you can load imaging files from NBIA, also mapped by IDs to the patient data, covered in *Working with Imaging Data* on page 40. 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 subject annotation data, and the corresponding array data in caArray.

caIntegrator supports a limited number of array platforms. For more information, see *Managing Platforms* on page 47.

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

1. On the Edit Study page where you have selected and added the subject annotation data, click the **Add New** 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 2.12*). This fields are described below.

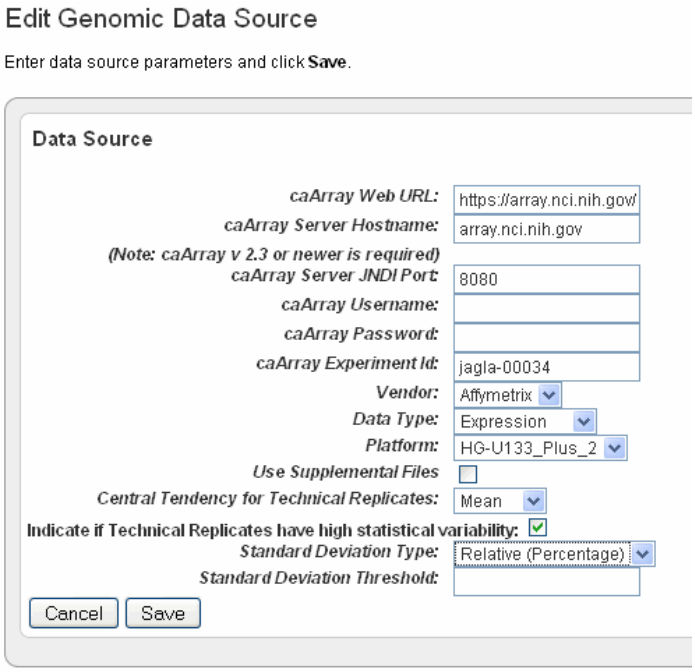


Figure 2.12 Edit Genomic Source dialog box

- **caArray Web URL** – Enter the URL for the caArray to be used for the genomic data sources. This will enable a user to link to the referenced caArray experiment from the study summary page.
- **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 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 subject annotation 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**.
- **Platform** **(needed only for Agilent)** – If appropriate, select the **Agilent** platform.

Note: Because you can add more than one set of genomic data to a study, a study can also have multiple platforms, one for each set of genomic data.

- **Central Tendency for Technical Replicates** – If more than one hybridization is found for the reporter, the hybridizations will be represented by this method.
- **Indicate if technical replicates have high statistical variability** – If more than one hybridization is found, checking this box will display a ** in the genomic search results when a reporter value has high statistical variability.
- **Standard Deviation Type** - When the checkbox for indicating if technical replicates have high statistical variability is checked, this parameter becomes available. Select in the drop-down the calculation to be used to determine whether or not to display a ** (see previous bullet point).
  - **Relative**, which calculates the Relative Standard Deviation in percentage value
  - **Normal**, which calculates the Standard Deviation in numeric value
- **Standard Deviation Threshold** – When the checkbox for indicating if technical replicates have high statistical variability is checked, this parameter becomes available. This is the threshold at which the Standard Deviation Type is exceeded and the reporter is marked with a **.

2. Click **Save**.

   caIntegrator 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 2.13*).

Figure 2.13 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 Subject Annotation Data

Because the goal of caIntegrator is to integrate data from subject annotation, genomic and imaging data sources, data from uploaded source files must be mapped to each other. Mapping files can map to caArray genomic data of two types: “imported and parsed” and that stored in supplemental files.

Creating a Mapping File

You, as the caIntegrator study manager, must create a Subject to Sample mapping file before following the actual mapping steps. This file provides caIntegrator with the information for mapping patients to caArray samples.

1. Start with the 6-column mapping file template, described as follows:
   - *All platforms* – Raw (level 1) data cannot be mapped; only normalized, processed (level 2) data is acceptable.
   - The required six-column file format uses the following columns:
     - **Subject ID**
     - **Sample ID**
     - **Name of supplemental file** (if appropriate, as attached to the experiment in caArray)
     - **Probe Header** – Name of column header (in the supplemental file) which contains the probe IDs.
     - **Value Header** – Name of column header (in the supplemental file) which holds the level 2 data.
     - **Sample Header** – Name of column header (in the supplemental file) which holds the level 2 data.

   Note: Only one of the last 2 columns is used: a single sample per file uses the Value Header column; multiple samples per file used Sample Header column. Unused columns are blank.

   Figure 2.14 shows an example multiple sample mapping file in CSV format.

	A	B	C	D	E	F	G
1	ASP221	Sample_Exp_1	Affy_FC_Test.TXT	Probe Set ID	na	Hybriddization Ref	
2	ASP308	Sample_Exp_2	Affy_FC_Test.TXT	Probe Set ID	na	Hybriddization Ref	
3	FPH113	Sample_Exp_3	Affy_FC_Test.TXT	Probe Set ID	na	Hybriddization Ref	
4	FPH114	Sample_Exp_4	Affy_FC_Test.TXT	Probe Set ID	na	Hybriddization Ref	
5	FPH118	Sample_Exp_5	Affy_FC_Test.TXT	Probe Set ID	na	Hybriddization Ref	
6	FPH309	Sample_Control_1	Affy_FC_Test.TXT	Probe Set ID	na	Hybriddization Ref	
7	E07733	Sample_Control_2	Affy_FC_Test.TXT	Probe Set ID	na	Hybriddization Ref	
8	E09176	Sample_Control_3	Affy_FC_Test.TXT	Probe Set ID	na	Hybriddization Ref	
9	E09214	Sample_Control_4	Affy_FC_Test.TXT	Probe Set ID	na	Hybriddization Ref	
10	E09262	Sample_Control_5	Affy_FC_Test.TXT	Probe Set ID	na	Hybriddization Ref	

Figure 2.14 Mapping file in CSV format, showing multiple samples

2. When you use the mapping file, make sure you use the patient ID for mapping.
3. Determine whether your data in caArray is "imported and parsed" or "supplemental". Fill in the 6-column mapping file according to the following standard:
   - Imported and parsed – Complete only the first two columns of the 6-column mapping file as described above. You can ignore the remaining columns.
   - Supplemental – Supplemental data comes in two flavors: "single sample per file" and "multiple samples per file". Only one of the last two columns is used. If the supplemental data format is:

— Single sample per file – the column named "Sample_Header" can be left empty.

— Multiple samples per file – the column named "Value_Header" can be left empty.

Note: Supplemental files from caArray for mapping data must be configured appropriately. For information, see *Supplemental Files Configuration* on page 133.

The following steps use data of either type.

Steps for Mapping Genomic Data

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

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

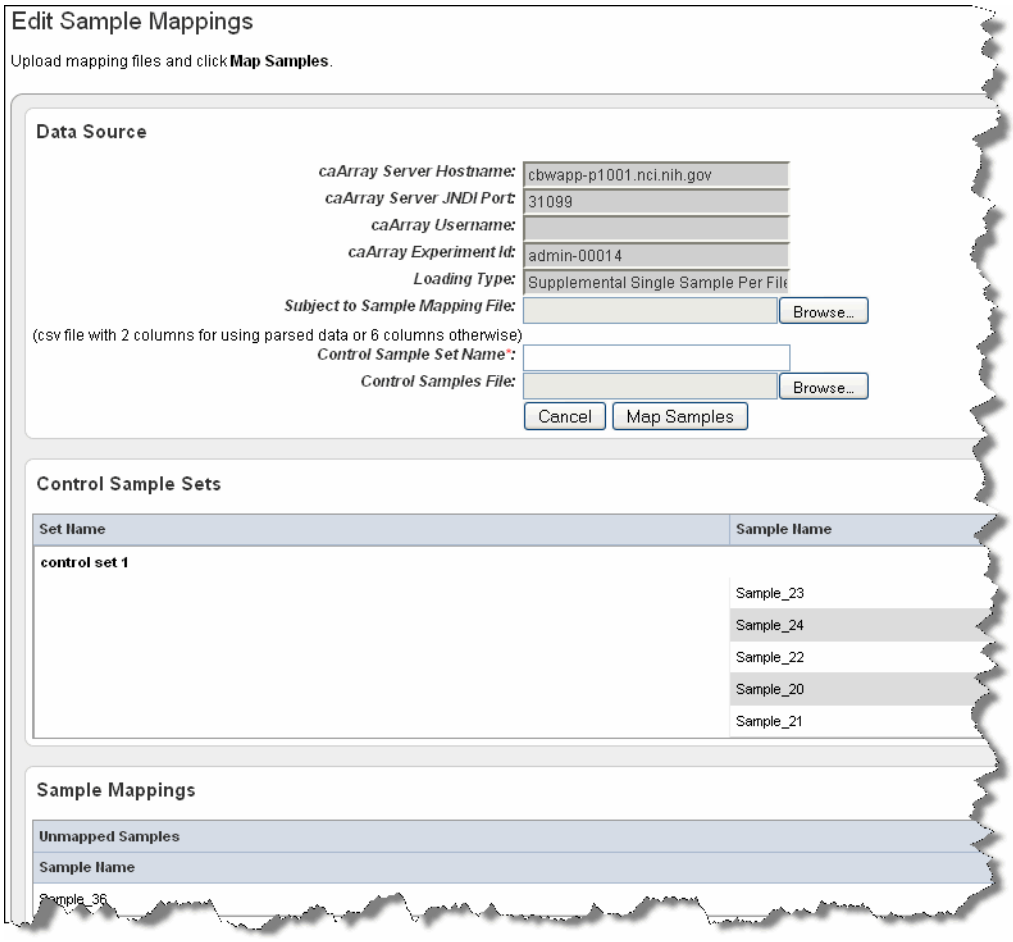


Figure 2.15 Edit Sample Mappings page showing some already mapped samples

2. The first two caArray fields may be populated with the information for the instance of caArray to which you have access. You can, however, enter the following caArray information, if appropriate.

- **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** – Enter your caArray account user name and password; you must have permissions in caArray for the experiment if it is private. If the data is public, you can leave this field blank.

3. **caArray Experiment ID** – Enter the caArray Experiment ID which you know corresponds with the subject annotation 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.
4. Enter the **Loading Type** of the data file you plan to map. (File types are described in *Creating a Mapping File* on page 33).
5. In the Subject to Sample Mapping File section, click **Browse** to navigate for the Sample Mapping CSV file that you created (described in *Creating a Mapping File* on page 33). This provides caIntegrator with the information for mapping patients to caArray samples.
6. Click the **Map Samples** button.

   If the caArray data you have identified is imported and parsed, when you click the Map Samples button, the mapping takes place as the date is uploaded into caIntegrator. If the caArray data is supplemental, the mapping does not occur until the study is deployed.

   Mapped samples will be listed in the Samples Mapped to Subjects section. Unmapped samples show at the top of the caIntegrator page. They were loaded from caArray, but they are not in the mapping file. These are not used for integration.

   Note: If you have already mapped samples, when you first open this page they are listed in the Samples Mapped to Subjects section. If you have not already mapped samples, all of the samples in the caArray experiment you selected are listed as unmapped, because caIntegrator does not know how these sample names correlate to the patient data in the subject annotation file until you upload the subject to sample mapping file.

7. Scroll down the page to see samples that are mapped to the patients in the subject annotation data (*Figure 2.16*).

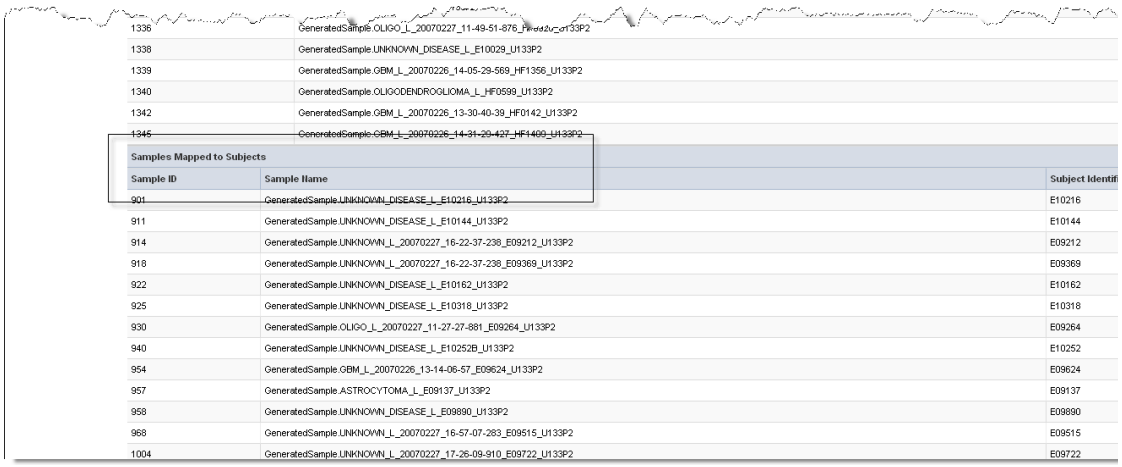


Figure 2.16 Example of samples mapped to patients' data

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. The control sets display at the top of the page once they have been uploaded (*Figure 2.17*).

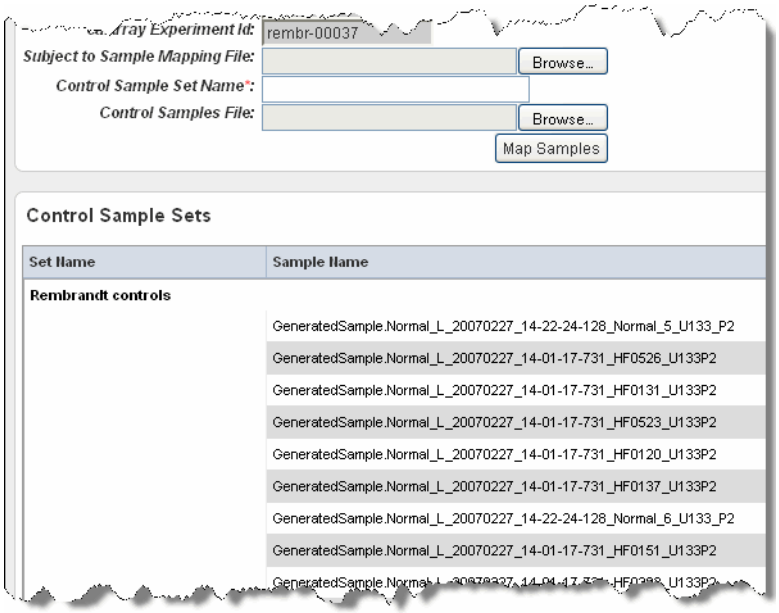


Figure 2.17 Example list of control samples

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

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

Note: If a Control Set is to be used in Gene Expression For Annotation, or Gene Expression plots for Annotation Query, then the control set should be composed of only samples which are mapped to subjects.

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.

The name specified in the third column of the mapping file is specific for each array manufacturer as follows:

- Affymetrix – The third column of the mapping file must contain filenames that end in .cnchp. The corresponding experiment in caArray must have these files and the extensions must match .cnchp.
- Agilent – The third column must name a file which contains level 2 copy number data. Level one copy number will not work. This file name is repeated for each line in the mapping file.

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** button.

   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 2.18*).

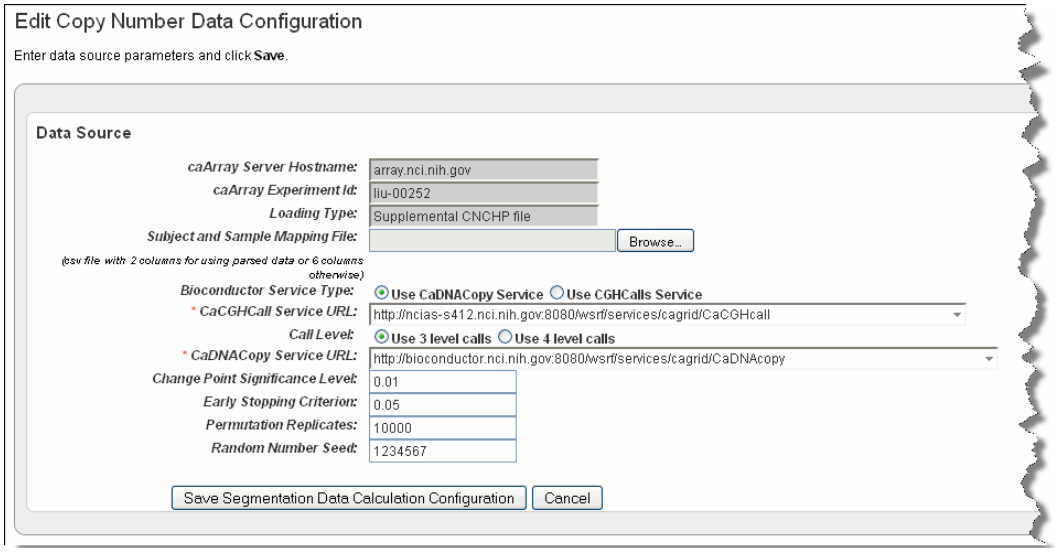


Figure 2.18 Edit Copy Number page

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.

Field	*Description*
caArray Service 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 Experiment ID	Enter the caArray Experiment ID which you know corresponds with the copy number data.
Loading Type	Enter the **Loading Type** of the data file you plan to map.
Subject and Sample Mapping File	Browse for the appropriate CN mapping file. The file must be a CSV file with 3 column format for mapping single data file and 5 column format for mapping 1 data file per sample.
Bioconductor Service Type	This is the type of bioconductor module that will be used for segmentation. Select between the two options: **DNAcopy** or **CGHcall**.
caCGHcall Service URL	Enter the URL for the grid service used to access the caCGHcall service. For more information, see http://www.bioconductor.org/help/bioc-views/release/bioc/html/CGHcall.html
Call Level	An input parameter to CGHcall. This is the number of discrete values used to represent the copy number level. Select between two options: **3** (consisting of discrete values of -1, 0, 1) or **4** (consisting of discrete values -1, 0, 1, 2)
caDNACopy Service URL*	Control for selecting the URL which hosts the caDNACopy grid service For more information, see http://www.bioconductor.org/packages/2.6/bioc/html/DNAcopy.html.
Change Point Significance Level	Significance levels for the test to accept change-points
Early Stopping Criterion	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 **Save Segmentation Data Calculation Configuration** 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.

Caution: After a study has been deployed and the genomic source has been loaded, you cannot change these copy number parameters without reloading the data from caArray first.

Remapping Copy Number Data in a Deployed Study

Occasionally you may need to remap copy number data in a deployed study. To do so, follow these steps:

1. Select the **Manage Studies** button and select **Edit** for the study you wish to remap.
2. On the Edit Study page, select **Edit** under the Genomic Data Sources header.
3. Without altering any information, select **Save**. When the warning box appears, select **OK**.
4. Select **ConfigureCopyNumberData**.
5. Enter the new mapping file in the Subject and Sample Mapping File field. Select **Save Segmentation Data Calculation** configuration.
6. Select **Deploy Study**.

Working with Imaging Data

Once you have loaded subject annotation data and identified patient IDs, you can add either array genomic sample data from caArray which caIntegrator maps by sample IDs to the patient IDs in the subject annotation data, or you can upload image data from NBIA, also mapped by IDs to the subject data. Once you have configured an NBIA image data source for adding images, then you can import image annotation data for the images. 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 subject annotation data, and the corresponding imaging data in NBIA.

Adding or Editing Imaging Data Files from NBIA

To add images from NBIA to the study you are creating, follow these steps:

1. On the Edit Study page, under the Imaging Data Sources section click the **Add New** button.

   Note: If you have already provided an imaging data source, it is listed in this section of the Edit Study page. To edit the imaging data source, click the **Edit** button which opens the same dialog box described in the following steps.

2. In the Edit Imaging Data Source dialog box, configure the imaging data source in the fields (*Figure 2.19*). Asterisks indicate required fields..

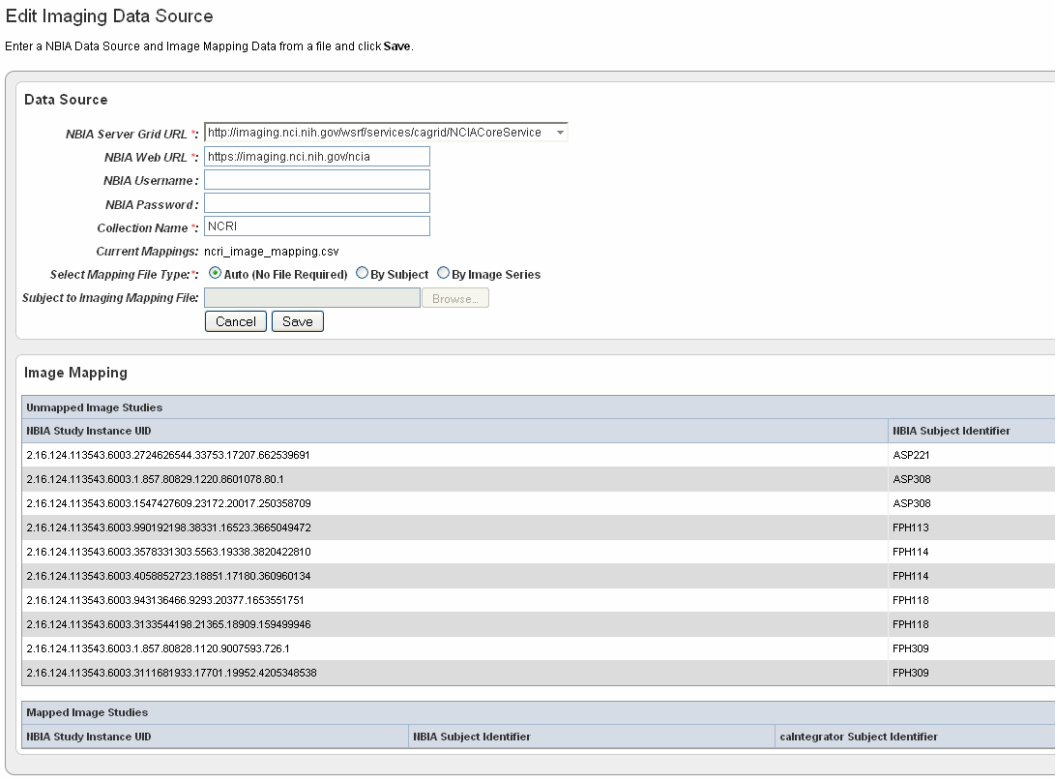


Figure 2.19 Edit Image Data Source dialog box

- **NBIA Server Grid URL*** – Enter the URL for the grid connection to NBIA.
- **NBIA Web URL** *– Enter the URL of the web interface of the NBIA installation.
- **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 you want to retrieve.
- **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 is required. Selecting this takes all subject annotation 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 mapping file to be uploaded. The “subject annotation to imaging mapping file” must be in CSV format with two columns that map the caIntegrator subject annotation subject ID to the NBIA subject ID.

— **By Image Series** – Requires a file to be uploaded. The subject annotation to imaging mapping file needs to be a two column mapping (CSV) from the caIntegrator subject annotation subject ID to the NBIA study instance UID.

- **Subject to Imaging Mapping File** – Click **Browse** to navigate to the appropriate subject annotation to imaging mapping file. See **Select Mapping File Type*** field description.

Note: If mapping files have already been uploaded for the data sources you are editing, the Image Mapping tables of the dialog box show the mapping from NBIA Image Series Identifier to caIntegrator Subject Identifier.

3. Click **Save** to upload the data from NBIA to caIntegrator.

   The imaging data displays on the Edit Study page under the Imaging Data Sources section (*Figure 2.20*).

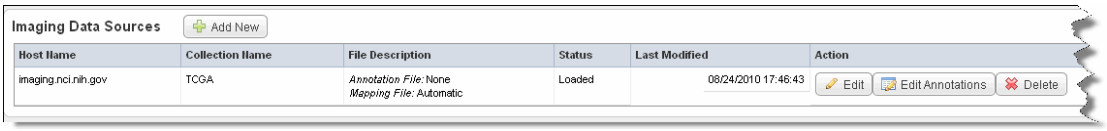


Figure 2.20 Imaging Data Sources section of the Edit Study page.

4. Once the data is uploaded, you can add image annotations. For more information, see *Adding or Editing Image Annotations*.

Adding or Editing Image Annotations

After you have configured an image data source with an NBIA Grid service and uploaded the image data, described in *Adding or Editing Imaging Data Files from NBIA* on page 40, you can load image annotations into caIntegrator from a file in CSV format or through an Annotations and Image Markup (AIM) service.

Tip: The image data shown in the Imaging Data Sources section indicate whether or not annotations have already be imported from a file for these sources. See marked area in *Figure 2.20*.

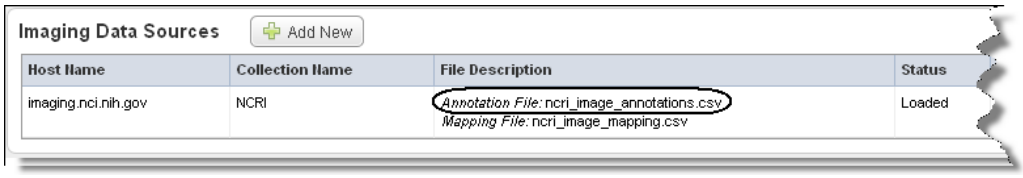


Figure 2.21 Imaging Data Sources section of the Edit Study page. The circled section in this screen shot indicates that annotations have been uploaded for this image collection.

To add image annotations from a file, follow these steps:

1. On the Edit Study page, click the **Edit Annotations** button under the Image Data Sources section.

   Note: If you have not yet imported annotations, clicking this button opens the page from which you can import image annotations (*Figure 2.22*).

Continue with the steps in this section. If you are editing annotations, clicking this button opens the Define Fields for Image Annotations dialog box where you can edit annotations. See *Define Fields Page for Editing Annotations* on page 21.

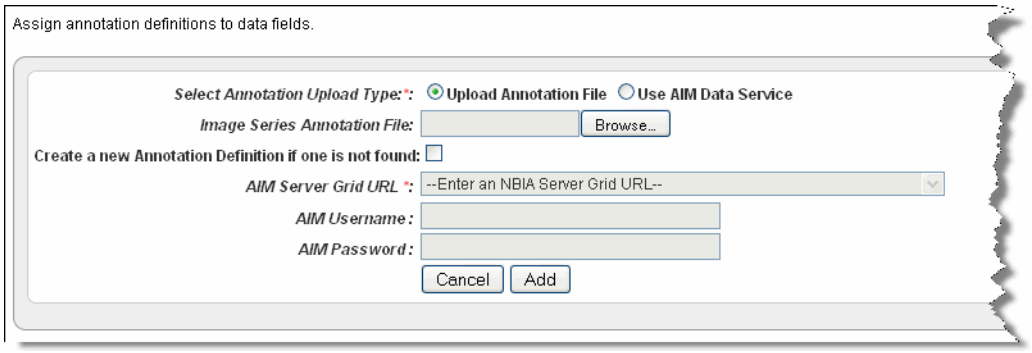
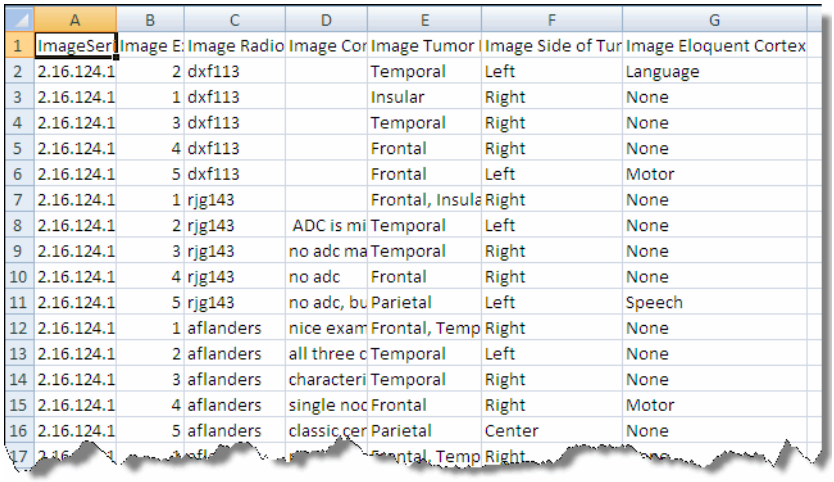


Figure 2.22 Page for adding imaging data annotations

2. Select the radio button **Upload Annotation File**.
3. Click **Browse** to select an annotation CSV file for upload.

   Note: An image annotation CSV file must include an Image Series ID column. See the highlighted column in *Figure 2.23*.



	A	B	C	D	E	F	G
1	ImageSer	Image E	Image Radio	Image Cor	Image Tumor	Image Side of Tur	Image Eloquent Cortex
2	2.16.124.1	2	dxf113		Temporal	Left	Language
3	2.16.124.1	1	dxf113		Insular	Right	None
4	2.16.124.1	3	dxf113		Temporal	Right	None
5	2.16.124.1	4	dxf113		Frontal	Right	None
6	2.16.124.1	5	dxf113		Frontal	Left	Motor
7	2.16.124.1	1	rjg143		Frontal, Insula	Right	None
8	2.16.124.1	2	rjg143	ADC is mi	Temporal	Left	None
9	2.16.124.1	3	rjg143	no adc ma	Temporal	Right	None
10	2.16.124.1	4	rjg143	no adc	Frontal	Right	None
11	2.16.124.1	5	rjg143	no adc, bu	Parietal	Left	Speech
12	2.16.124.1	1	aflanders	nice exam	Frontal, Temp	Right	None
13	2.16.124.1	2	aflanders	all three c	Temporal	Left	None
14	2.16.124.1	3	aflanders	characteri	Temporal	Right	None
15	2.16.124.1	4	aflanders	single noc	Frontal	Right	Motor
16	2.16.124.1	5	aflanders	classic cer	Parietal	Center	None
17	2.16		afl		ntal, Temp	Right	

Figure 2.23 Image annotations file

4. Check the box for **Create a new Annotation Definition if one is not found** (if appropriate).
5. Click **Add**.

To load image annotations through an AIM service, follow these steps:

1. On the Edit Study page, click the **Edit Annotations** link under the Image Data Sources section.
2. Select the radio button **Use AIM Data Service**.
3. Select an **AIM Server Grid URL**.

4. Click **Add**.

Using either method, the image annotations are uploaded to caIntegrator. After this occurs, when you click the **Edit Annotations** button, the system opens to the Define Fields for Imaging Data page where you can edit the annotations. For more information, see *Define Fields Page for Editing Annotations* on page 21. You must assign identifiers and annotations to the data in the same way you did with the subject annotation data. For more information, see *Assigning An Identifier or Annotation* on page 23 and *Searching for Annotation Definitions* on page 26.

Adding External Links

This feature on the Edit Study page allows you to configure a CSV file with URLs to be used as external links relevant to a study. This allows you to easily share or configure references.

To add an external link, follow these steps:

1. As a study manager, you can configure a CSV file with URLs to be used as external links.
2. On the Edit Study page, click the **Add** button under External Links section. External links can be any URL(s) to resources that are hosted external to caIntegrator but are relevant to the study being deployed.
3. Assign a name to the external link.
4. Add a description for the link, if appropriate.
5. Browse for the CSV file containing URLs (HTTP linked) to resources outside of caIntegrator.
6. Click **Upload Now**. caIntegrator does not validate any links in the file being uploaded.

Once you have created external links for a study, when the study is open, an External Links section showing the link(s) displays on the left sidebar of the page (*Figure 2.24*).

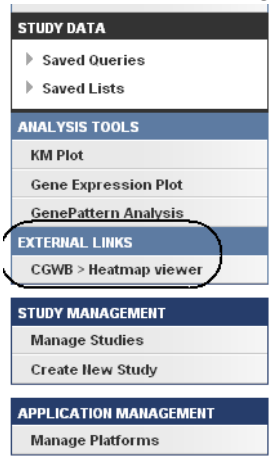


Figure 2.24 Left sidebar displaying external links

Click the link to open a page that displays appropriately formatted web page links (*Figure 2.25*).

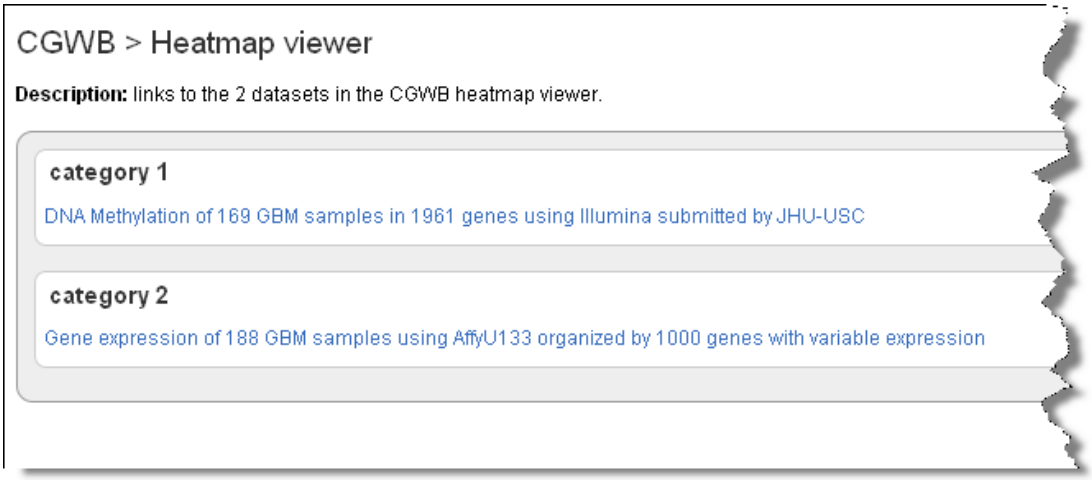


Figure 2.25 An example of exernal links

Deploying the Study

When you are ready to deploy the study, click the **Deploy Study** button on the Edit Study page. caIntegrator 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 caIntegrator while study is being deployed.

Managing a Study

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

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 2.26*).

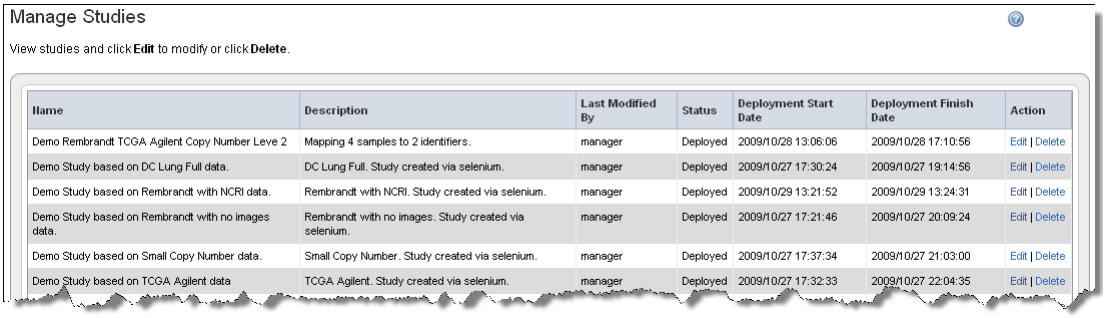


Figure 2.26 Manage Studies page

All of the "in process" or "completed" studies display on this page, with associated metadata. Note that whoever edited or updated the study last is shown in the Last Modified Column, indicated as the Study Manager.

2. Click the **Edit** link corresponding to your study of choice to open the Edit Studies page (*Figure 2.27*).

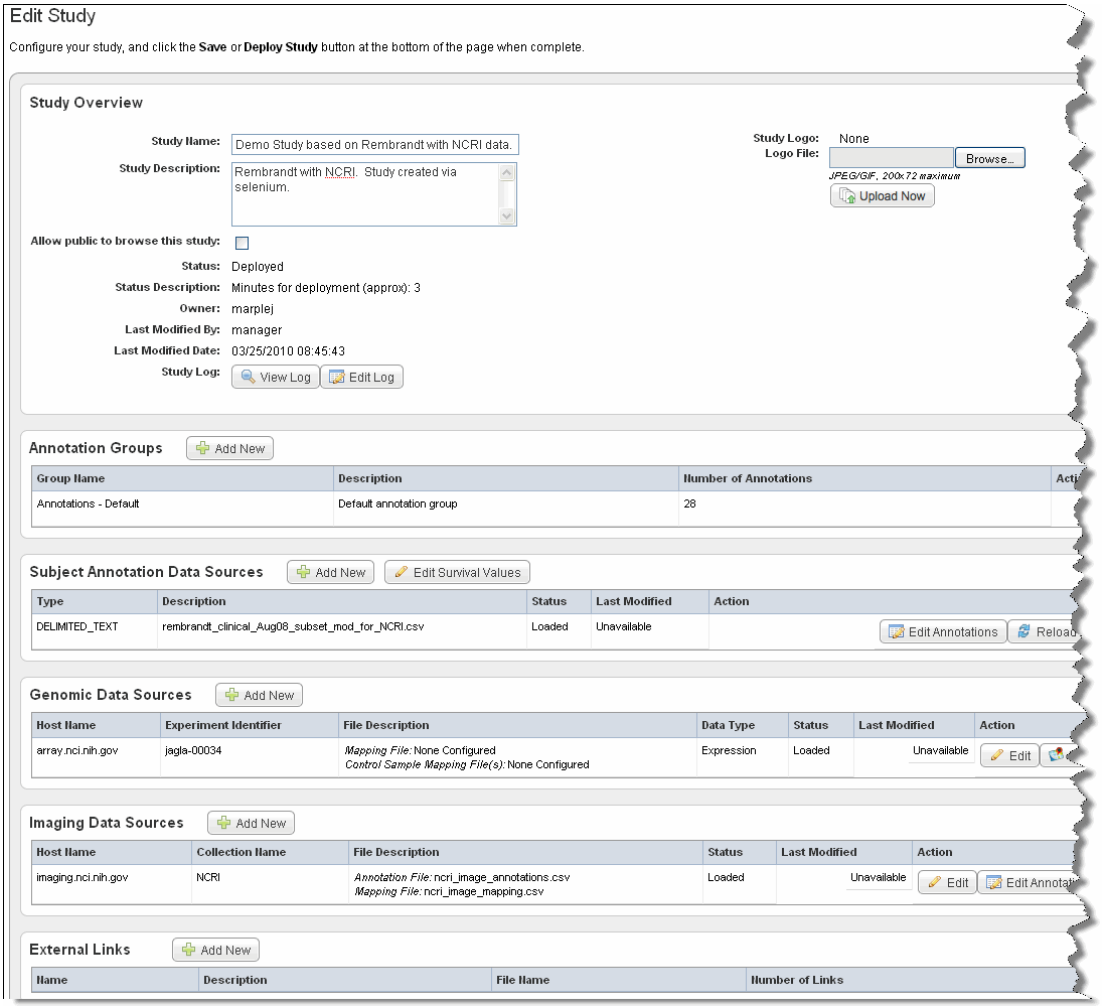


Figure 2.27 Edit Studies page where you can edit any details for an existing study

On this page you can edit any details such as adding or deleting files, survival values, and so forth. For information about working with the Edit Study feature, see *Creating/Editing a Study* on page 17.

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

Managing Platforms

caIntegrator 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, caIntegrator 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 caIntegrator study without a corresponding supported array platform. If you add more than one set of genomic data to the study, you can specify more than one platform for the study.

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

To manage platforms in caIntegrator, follow these steps:

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

   The Manage Platforms page that opens lists the platforms caIntegrator currently supports, those that the system can pull from caArray (*Figure 2.28*). You can also add a new platform by entering information in the fields in the Create a New Platform section.

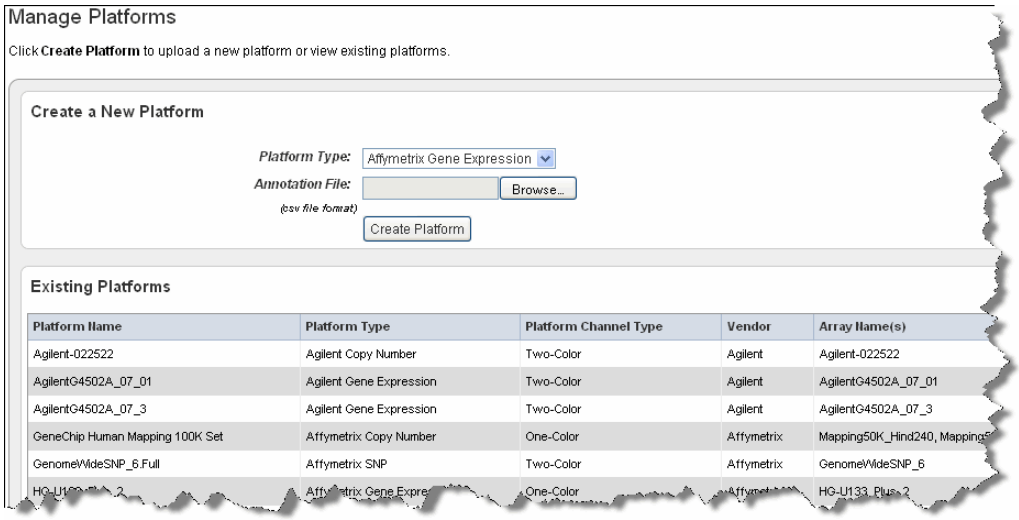


Figure 2.28 Manage Platforms page

2. To add a platform, in the Platform Type field, select the appropriate platform type from the drop down list.
3. Click **Browse** to navigate for the Affymetrix or Agilent file you want to add.

   Note: Tab-delimited .txt or .txv Agilent platform annotation files must contain the following column headers: ProbeId, GeneSymbol, GeneName and Accessions.

4. Enter a **Platform Name** if the file is a NON-GEML.xml file.

   Depending on the Platform Type you select, there may be other parameters to provide here as well, such as **Platform Channel Type** for an Agilent platform.

5. Click the **Browse** button to browse for the appropriate annotation file. When you have located it, click **Open** in the Upload File dialog box. The system displays the annotation file you select in the Annotation File box.
6. Once all parameters have been entered, click **Create Platform**.

The platform deployment can be time-consuming. If the platform takes more than 12 hours to deploy, caIntegrator displays a "timed out" message. At that point, you can delete the platform, even if it has not loaded to the system.

Note: Platform loading can fail if the manufacturer's platform annotation file is missing data.

CHAPTER

3

SEARCHING A CAINTEGRATOR STUDY

This chapter describes the processes for searching studies within caIntegrator.

Topics in this chapter include:

Search Overview

The search and browse functions in caIntegrator allow you to search for subject annotation data, genomic data or imaging data that were uploaded into the application as part of a study. When gene expression and imaging data are uploaded into a caIntegrator study, mapping files that correlate sample IDs in those files to subject IDs (patient IDs) in the subject annotation data file must also be uploaded. When you launch a search, caIntegrator finds and integrates the subject annotation, 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:
    - **[Annotations]** – Annotation data can be labeled ‘default’ or given the annotation ‘group’ name when annotation groups are specified by the manager, for example, chronologic, therapy, diagnosis, patient, or other annotation group types. This selection searches one or more uploaded CSV

files for data identifiers or annotations (column headers) specified during study creation.

- **[Genomic]** – Genomic data can be gene expression or copy number data. This selection searches caArray experiments samples uploaded in the study for gene expression or copy number data by gene name, reporter ID, chromosome number, chromosome coordinates and/or segmentation values representing amplification or deletion.
- **Image Data** – Searches NBIA imaging 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 subject/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. Follow 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 caIntegrator, 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*).

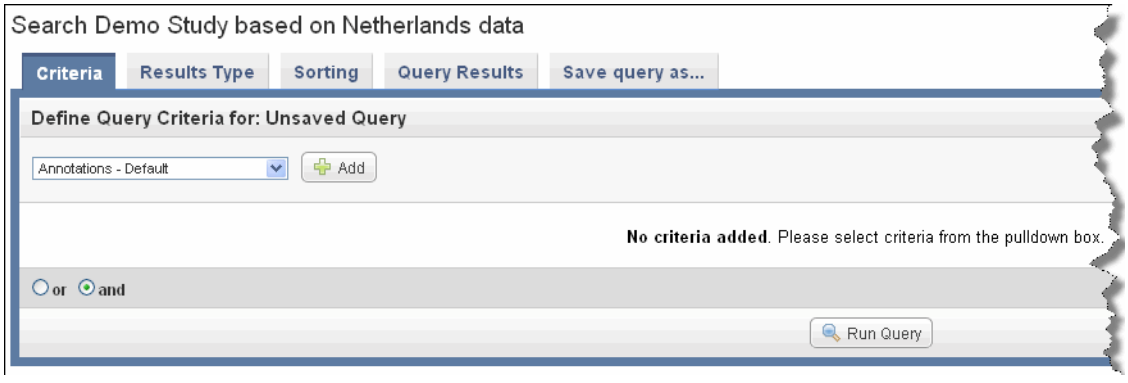


Figure 3.1 Search page

3. On the Criteria tab, in the drop-down list, select the type of data you want to search (*Figure 3.2*).

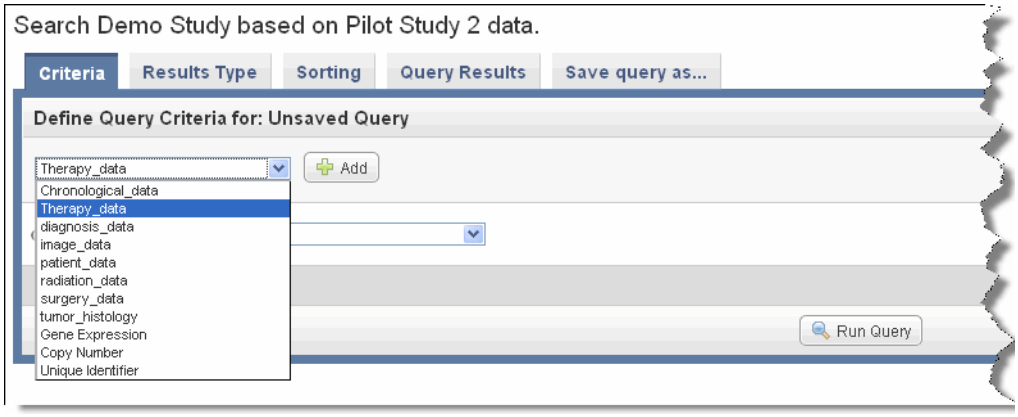


Figure 3.2 Default or defined annotation data types are available in the search criteria drop-down list

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.

- **Annotations** (listed as 'default' or by annotation group name when specified when the study was created)
- **Gene Expression** or **Copy Number**
- **Image Series**

4. Click **Add** to further define criteria for the search.

Continue with:

Annotation and Image Data Searches on page 52

Gene Expression Data Searches on page 54

Copy Number Searches on page 55

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 Tab* on page 58 and *Sorting Tab* on page 60.

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.

Annotation and Image Data Searches

Note: If the study manager defined the study's own annotation groups, then those group names are listed in the criteria drop-down list. If the study manager did not define the study's annotation groups when the study was created, then all annotations are placed, by default, in a group called "Annotations default".

- Once you select an annotation group data type, an additional drop-down list displays data elements that are annotation definitions specified when the data was uploaded into the study (*Figure 3.3*).

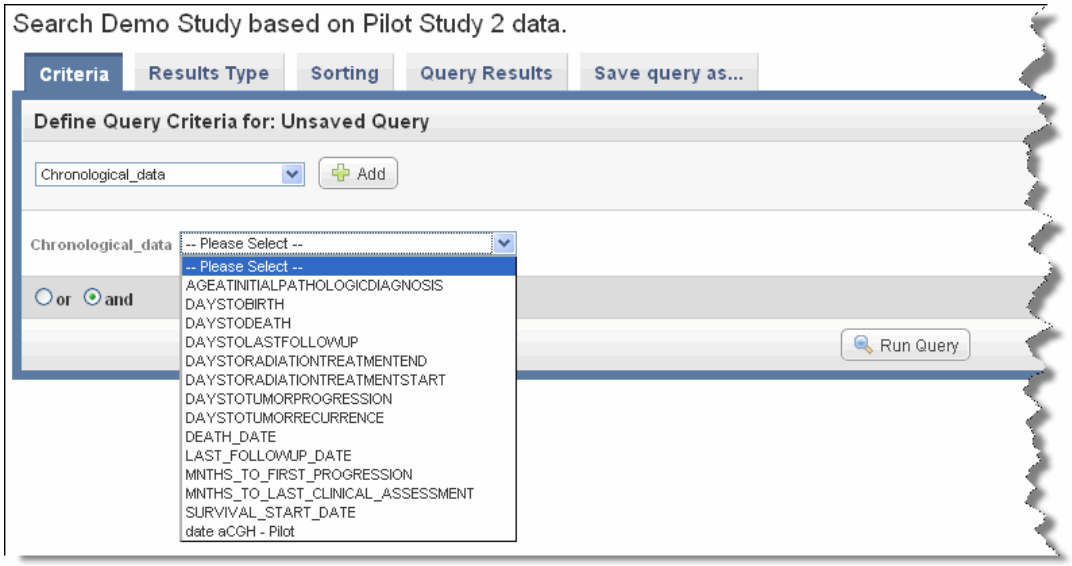


Figure 3.3 Annotation data elements available in the search criteria drop-down list reflect definitions specified in the corresponding study

- Select a search criterion from among the options. You can make only one selection at a time.

Note: If the study includes imaging data, imaging annotations should be available in the Annotations list.

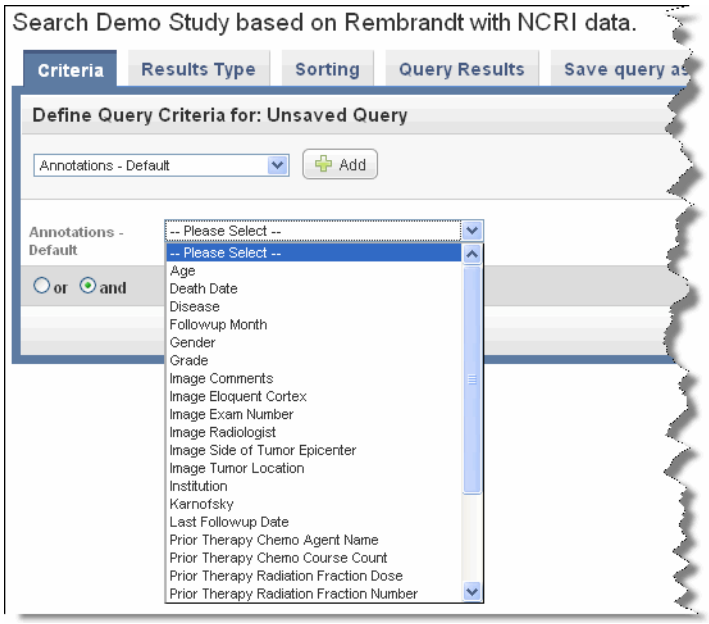


Figure 3.4 Annotation search criteria, including criteria for imaging

- 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 3.5*).

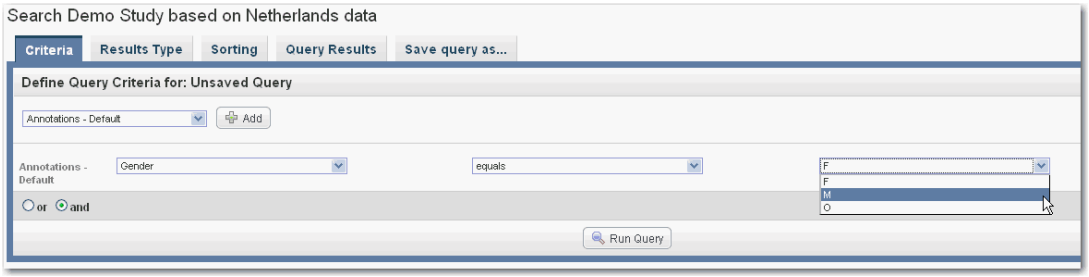


Figure 3.5 You may be able to further define search criteria when you select a specific subject annotation or imaging annotation element

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 50.

Gene Expression Data Searches

1. For the Gene Expression selection, select **Gene Name, Expression Level** or **Fold Change**. If the study includes multiple platforms, a **Platform** option is also visible.

2. **Gene Name, Expression Level** or **Fold Change** – Enter one or more gene symbols in the text box or click the icons to locate genes in the following databases. If you enter more than one gene in the text box, separate the entries by commas. If multiple platforms are part of the study, your platform selection in the Fold Change query criteria determines the control samples that are available.

   Note: If you leave the gene symbols field blank, caIntegrator searches all gene symbols for a match to the other criteria you specify.

caIntegrator provides three methods whereby you can obtain gene names for a gene expression search. For information about selecting genes, see *Choosing Genes* on page 56.

Additional fields display for the Expression Level selection.

Range Type –

a. Select the range type from the following options:

   >=: Greater than the entered value

   <=: Less than the entered value

   Inside Range: Looks for all matching values that occur between the two levels you enter

   Outside Range: Looks for all matching values that are not in between the two levels that you enter.

b. **Expression Level**: This criteria changes slightly based the range type you select. Enter the level appropriate for your query.

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 caIntegrator. You can enter query values in greater/lesser-than-or-equal-to arguments.

3. Select or enter data for the Fold change fields shown in *Figure 3.6*:

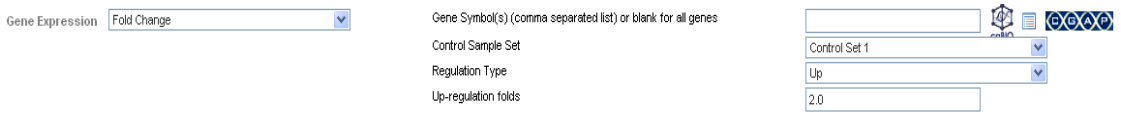


Figure 3.6 Fields for identifying fold change search criteria

- **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** – Enter a numerical value representing fold change. The number you enter here is 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.

  For example, if you enter 2.0 in this field, after selecting **Up** in the previous field, the search will locate genes whose expression is 2 times (2-fold up regulation) the base value.

Continue with step 5 in *Searching a Study* on page 50.

Copy Number Searches

In some diseases, like cancer, cells that are abnormal can exhibit a change in the chromosomal structure in that parts of a chromosome can be amplified or deleted. ‘Copy number’ experiments that measure variation in genomic structure use molecular markers to detect amplification or deletion of chromosomal segments. Typically, copy number alteration experiments compare a genomic sample from a diseased tissue (for example, a tumor) to a control sample (for example, blood).

The Copy Number query option, as described in *Searching a Study*, appears only if copy number data have been uploaded to the study. A copy number search identifies patients or samples that have a copy number amplification or deletion in the genome range specified. Searches can be constructed with gene names, chromosome number and/or chromosome coordinates. You can enter query values in greater/lesser-than-or-equal-to arguments.

1. For the Copy Number selection, select **Gene Name** or **Segmentation**.
2. **Gene Name** – Enter one or more gene symbols in the text box, separated by commas, or click the icons to locate genes in the following databases.

   Note: If you leave the gene symbols field blank, caIntegrator searches all gene symbols for a match to the other criteria you specify.

caIntegrator provides three methods whereby you can obtain gene names for a copy number search. For information about selecting genes, see *Choosing Genes* on page 56.

Additional fields display for the Segmentation selection.

3. Select or enter data for the copy number query fields shown in *Figure 3.7*.

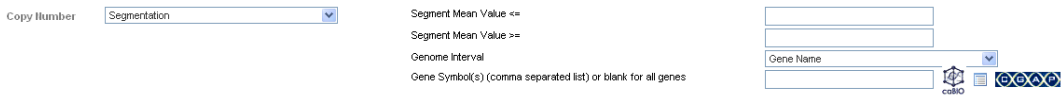


Figure 3.7 Fields for identifying copy number search criteria

Segmentation is the process of defining the chromosomal boundaries (coordinates) of the region deleted or amplified in the sample.

- **Segment Mean <=** – Enter the value equal to or less than the higher limit of change.
- **Segment Mean >=** – Enter the value equal to or greater than the lower limit of change.
- **Genome Interval > Gene Name** – Enter one or more gene symbols in the text box, separated by commas, or click the icons to locate genes in the following databases.

  Note: If you leave the gene symbols field blank, caIntegrator searches all gene symbols for a match to the other criteria you specify.

  caIntegrator provides three methods whereby you can obtain gene names for a copy number search. For information about selecting genes, see *Choosing Genes* on page 56.
- **Genome Interval > Chromosome Number** – In the text box that opens, enter the chromosome number you want the query to search against.
- **Genome Interval > Chromosome Coordinates** – In the **From** and **To** text boxes that open, enter the range on the chromosome you want to search. This defines the chromosomal boundaries of the region with the suspected copy number variations.

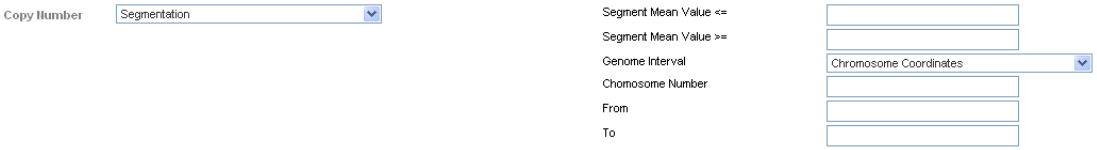


Figure 3.8 Fields for identifying copy number chromosome coordinates values

The Bioconductor DNAcopy algorithm (see *Copy Number Data* on page 66) identifies the location of the amplification or deletion and then reports it as the base pair at the start and stop of the segment. Each segment is then catalogued with chromosome number, start coordinate, stop coordinate, genes in the segment, and the segment mean value.

Continue with step 5 in *Searching a Study* on page 50.

Choosing Genes

caIntegrator provides three methods whereby you can obtain gene names for a gene expression search.

- **caBIO** – This link searches caBIO, then pulls identified genes into caIntegrator for analysis.

  a. Click the **caBIO** icon ( ).

  b. Enter **Search Terms**. Note that caIntegrator can perform a search on a partial HUGO symbol. For example, as search using ***ACH*** would find matches with 'achalasia' and 'arachidonate'.

  c. Select if you want to search in **Gene Keywords**, **Gene Symbols, Gene Alias, Database Cross Reference Identifier** or **Pathways** (from the drop-down list).

  — **Gene Keywords** searches the description field in caBIO; the result displays in the Full Name Column.

  — **Gene Symbols** searches only the Unigene and HUGO gene symbols in caBIO.

  — **Gene Alias** searches for one or more gene symbols which are synonymous for the current gene symbol.

  — **Database Cross Reference Identifier** searches for the symbol for this gene as it appears in other databases.

  — **Pathways** searches only the pathway names in caBIO. Note that searching in Pathways is a two step process. First, the initial Pathway search produces search results which are pathways. Second, from the pathway search results screen, you must select pathways of interest, then click **Search Pathways for Genes** to obtain a list of genes related to the selected pathways.

  d. Select the **Any** or **All** choice to determine how your search terms will be matched. **Any** finds any match for any search term you entered. **All** finds only results that match all of the search terms.

  e. Choose the **Taxon** from the drop-down list and click **Search.** The search results display (*Figure 3.9*).

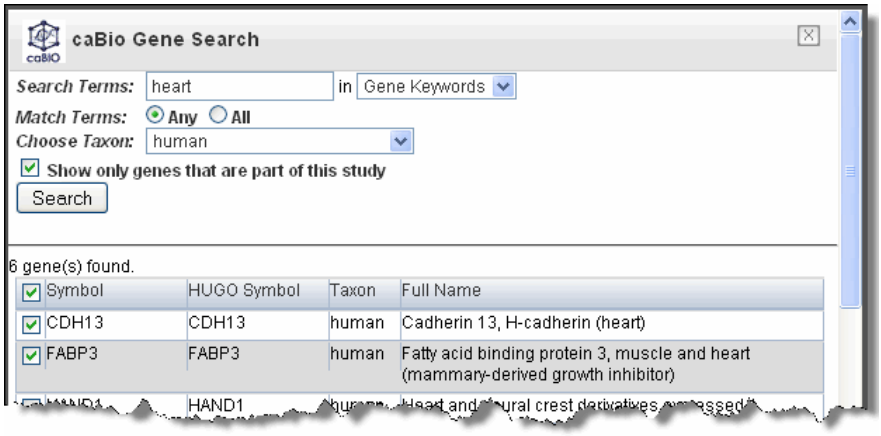


Figure 3.9 Example caBIO gene search criteria and search results

f. In the search results, use the check boxes to identify the genes whose symbols you want to use in the gene expression analysis.

g. Click **Use Genes** at the bottom of the page. This pulls the checked genes into the Criteria tab (*Figure 3.10*).

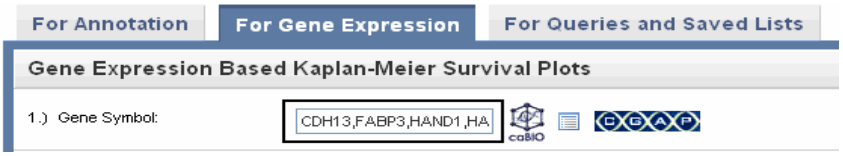


Figure 3.10 Genes pulled in from caBIO display on the Criteria tab

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

  a. Click the Genes List icon ( ) to open a Gene List Picker dialog. For more information, see *Creating a Gene or Subject List* on page 67.

  — GISTIC Amplified genes is a list of gene symbols in which the corresponding regions of the genome are significantly amplified.

  — GISTIC Deleted genes is a list of gene symbols in which the corresponding regions of the genome are significantly deleted.

  b. In the drop-down menu that lists previously saved gene lists, select a gene list. In the list that appears, use the check boxes to identify the genes whose symbols you want to use in the gene expression analysis.

  c. Click **Use Genes** at the bottom of the dialog. This pulls the checked genes into the Search Criteria tab.

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

Query Results

You can specify columns 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 Tab* and *Sorting Tab* to expand the display options and re-run the search, having set the display parameters.

For more information, see *Viewing Query Results* on page 63.

Results Type Tab

The selection you make on the Results Type tab determines whether caIntegrator displays search results for subject annotation or genomic data. It filters the search based on the criteria you set on the Criteria tab, whether it is annotation, gene expression or image series data type(s). In other words, if you select annotation 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 annotation criteria you defined on the Criteria tab.

1. On the Results Type tab, select the **Annotation, Copy Number** or **Genomic** radio button to search annotation data (*Figure 3.11*).

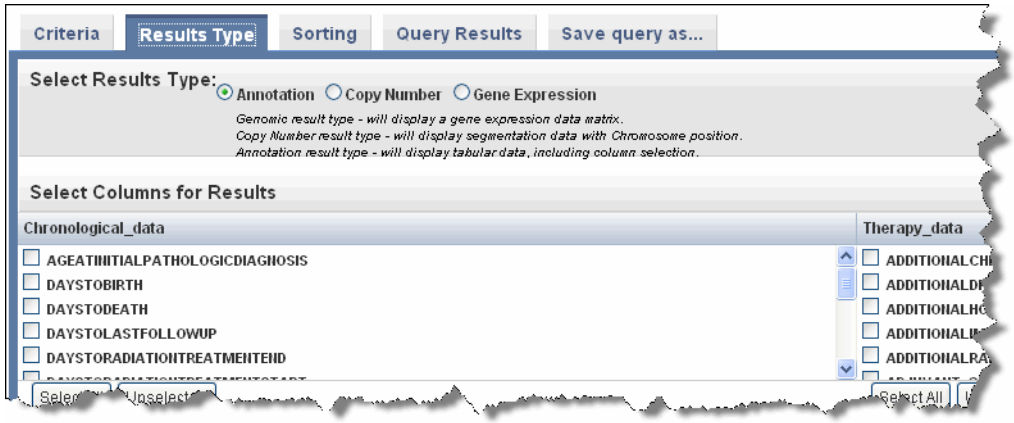


Figure 3.11 Results Type tab, annotation options

Annotation – 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 subject annotations, the Patient or Subject Identifier displays by default in the search results.

Results display as tabular data.

Copy Number – This option appears only if the open study includes copy number data. If you select this option, the annotation elements initially displayed on this tab disappear, and you are asked to run the query again. Based on the criteria you defined, the Query Results tab shows a data matrix containing samples against the genomic region you specified. For more information, see *Copy Number Searches* on page 55 and *Copy Number Data* on page 66.

Gene Expression – Select the **Reporter Type** and **Results Orientation**.

- **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
- **Genes in rows/Subjects in columns** or **Genes in columns/Subjects in rows** – Determines query results matrix format

Results display in a gene expression data matrix. For more information, see *Gene Expression Data* on page 64.

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 **Annotation**. 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 77.

Results display as tabular data. For more information, see *Subject Annotation and Imaging Data* on page 64.

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 61.

Sorting Tab

On the Sorting tab, you can set the sort order for data columns in the query results. You can also indicate whether column contents are sorted in ascending or descending order.

The columns that display on the Sorting tab are those criteria that you selected on the *Results Type Tab* for an Annotation Results type search.

Note: Sorting is not applicable to copy number search results. For those results, no options are available on the Sorting tab.

1. Select the Sorting tab and indicate the left to right column order of the Search Results by changing one or more numbers in the Column Order column in this table (*Figure 3.12*).

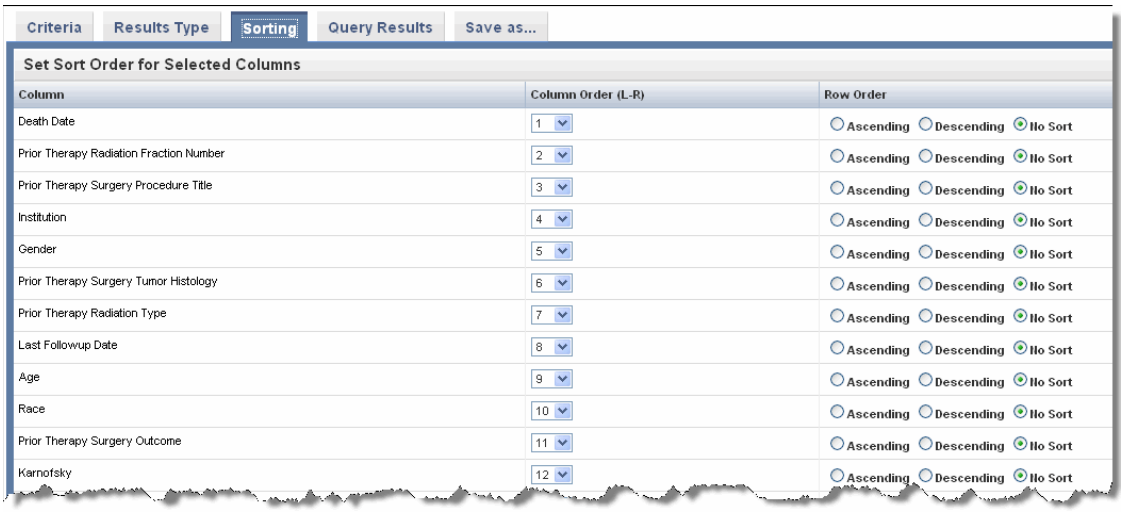


Figure 3.12 Sorting tab

2. In the **Row Order** column, indicate how you want columns sorted, **Ascending** or **Descending**, or leave the default, **No Sort**, if you choose.

3. Click **Run Query** at the bottom of the page to execute your sorting changes in the search results. When you do so, the change in column order is visible on the Query Results tab, as well as on the Sorting tab. For example, any column that you have indicated to be number “1” now appears in Query Results immediately after the Subject Identifier column and at the top of the Set Sort Order table on the Sorting tab.

   Sorting parameters are saved as part of the query if you choose to save it using the Save Query feature. See *Saving a Query* on page 61.

4. If you click the **Reset** button before running the query from the Sorting tab, the original column settings are restored.

For information about the search results, see *Chapter 4 Viewing Query Results*.

Managing Queries

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

For more information, see these topics:

Saving a Query on page 61

Editing a Query on page 61

Exporting Query Results on page 62

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 pop-up 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 78.

CHAPTER

4

VIEWING QUERY RESULTS

This chapter describes search results that caIntegrator returns after queries.

Topics in this chapter include the following:

Query Results Overview

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

If you have not configured the 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 Tab* on page 58.

caIntegrator 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 62.

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.

Subject Annotation and Imaging Data on page 64

Gene Expression Data on page 64

Expanding Imaging Data Results on page 73

Subject Annotation 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*). .

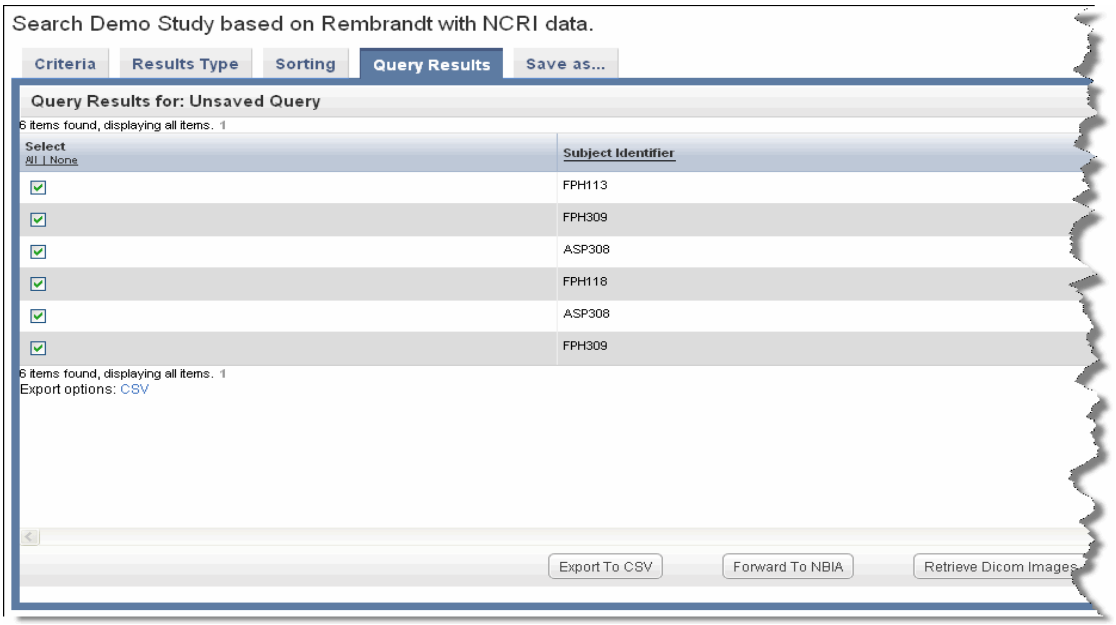


Figure 4.1 Query Results page

You can add details for 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 Tab* on page 58.

Gene Expression Data

If after defining gene expression criteria on the Criteria tab, you select the **Gene Expression** 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 caIntegrator.

You can select on the Results Type tab a preferred orientation for displaying the results: genes in rows and subjects in columns, or genes in columns and subjects in rows.

For Gene criteria, the cells display the median gene expression value for each gene. By each gene symbol, caIntegrator displays an icon ( ) which you can click to open the Cancer Genome Anatomy Project (CGAP) showing data for the gene (*Figure 4.2*).

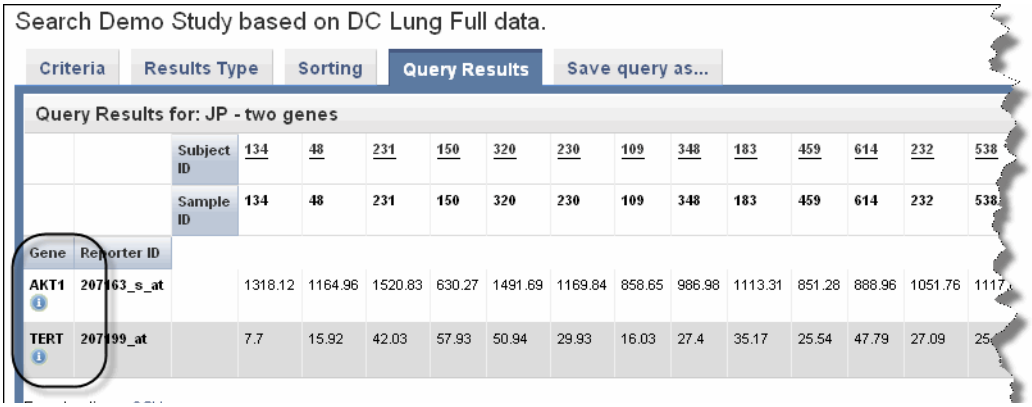


Figure 4.2 Genomic query result matrix after gene criterion has been specified

If you have selected **Gene Expression** on the Results Type tab, then the column headers are a clickable label which sorts the entire table on that column. If you selected Reporter ID on the Results Type tab, the Reporter ID is clickable (and the gene is not clickable).

For fold-change criteria, the cells display the normalized signal-based value for a given reporter for a given sample. In the results matrix, caIntegrator highlights matrix values for fold change results that meet fold change criteria. Red represents upregulated values and blue indicates downregulated values (*Figure 4.3*, *Figure 4.4*).

Figure 4.3 Gene Name search 6 genes, Reporter Type: Gene. Genes display in rows and subjects appear in columns.

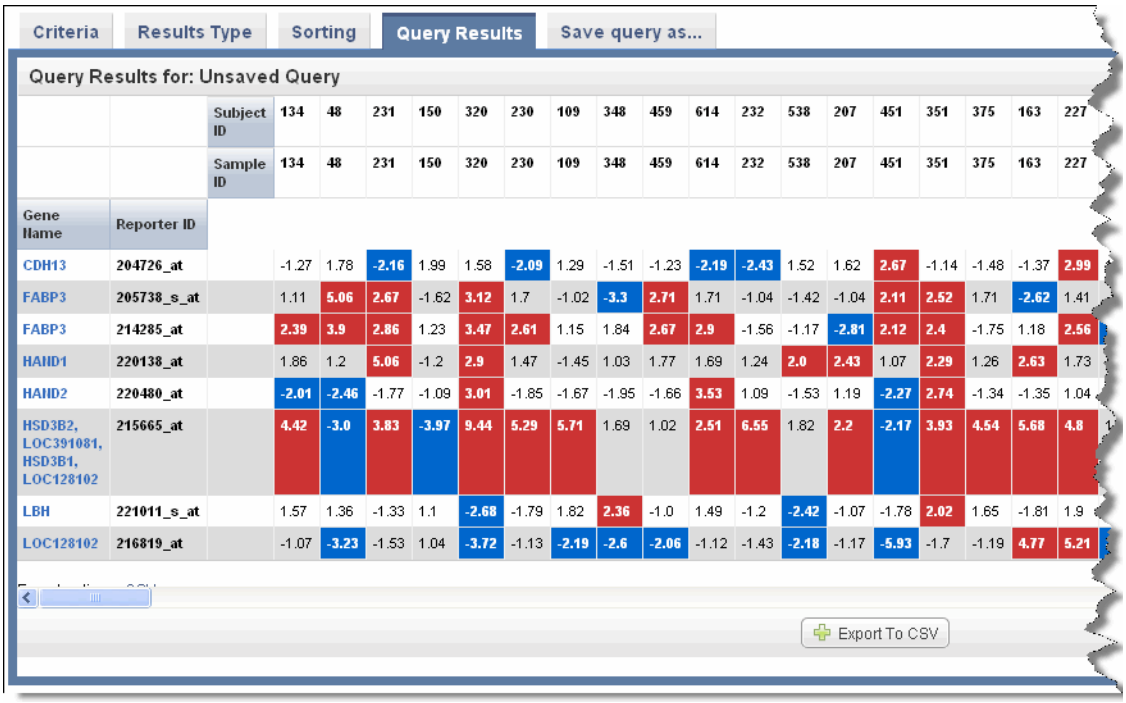


Figure 4.4 Gene Name search 6 genes, Reporter Type: Reporter ID. Genes display in rows and subjects appear in columns.

- Genomic data does not display in tandem with subject annotation and imaging data; it only displays when you select the **Gene Expression** result type on the Results Type tab. Genomic data is however, filtered by subject annotation 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 caIntegrator on the appropriate tabs.

You can save genes identified in the search results as a gene list.

Copy Number Data

If after defining copy number criteria on the Criteria tab and running a copy number query, (see *Copy Number Searches* on page 55), you should select the **Copy Number** result type on the *Results Type Tab*, and rerun the query. Copy number data search results display in a data matrix containing samples vs. genomic regions.

- Gene symbols display parallel to chromosome regions on the matrix.
- Sample ID column headings display the Subject ID/Sample ID (for example, E09262/E09262) because each calculation is based on a comparison of a tumor and matched blood sample from the same subject.
- The values in the Sample ID columns are mean segment values as calculated by the DNAcopy algorithm (*Figure 4.5*). These are expressed as log2(test/

reference, as in tumor/normal). For more information about the algorithm, see http://www.bioconductor.org/packages/2.6/bioc/html/DNAcopy.html.

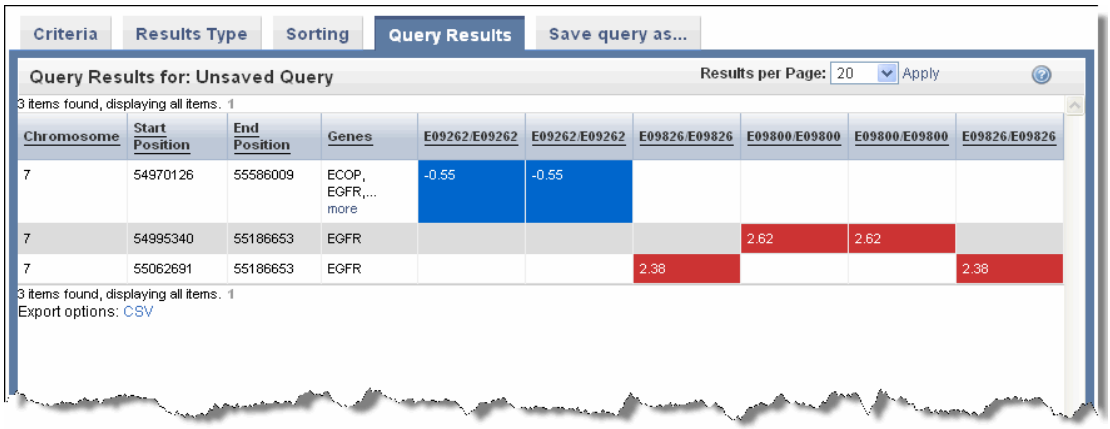


Figure 4.5 Data matrix displaying copy number search results

DNAcopy ouput values can be negative. If the test and the reference genomic samples both have two copies of a chromosomal region, the ratio of test/reference is '1', and the log2(1) = 0. That is, if there is no change in the chromosomal structure, then the value is 0. If there are more copies in the test sample (amplification of the chromosomal segment), the ratio of test to reference is greater than 1, and the log2(test/reference) is greater than 0. For example, if the test sample has 6, the ratio or test/reference is 6/2 = 3; log2(3) = 1.58. In a deletion, the test is less than the reference, for example 1. The DNAcopy output value would be log2(1/2) = log2(0.5) = -1.0. Values below -0.6 are often considered a deletion.

Creating a Gene or Subject List

From any page in caIntegrator that shows such a group, you can save a list of genes or subjects so you can use it for searches or analyses. This functionality can also be used where a gene or subject list was created outside of caIntegrator, for example, a list of subjects with validated mutation such as from TCGA projects, or a list of subjects with high EGFR expression or any lists of subjects with genomic or clinical characteristics determined with other tools.

To create a list, follow these steps:

1. Click the **Create New List** link in the left sidebar. This opens the Manage List page (*Figure 4.6*):

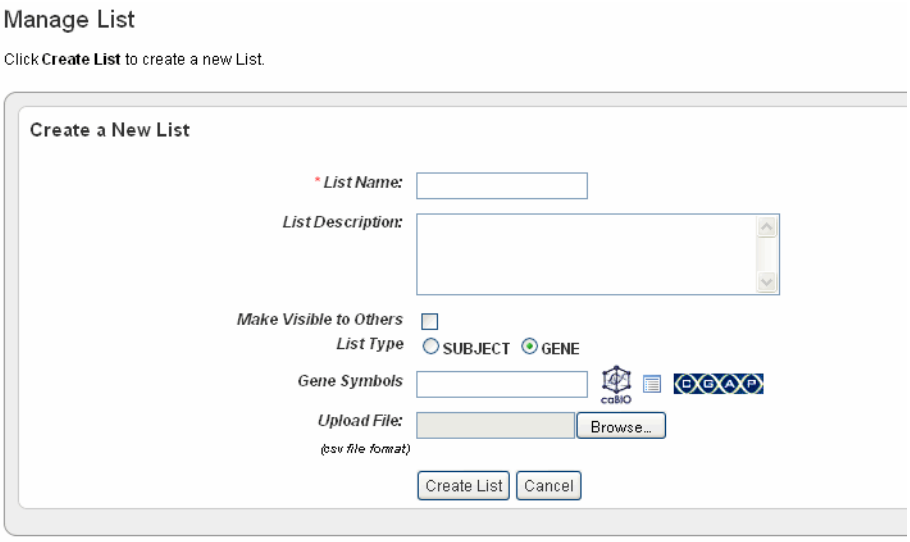


Figure 4.6 Manage Gene List page

2. Enter a name for the gene or subject list.
3. Enter a description (optional).
4. Select **Make Visible to Others** if you want the list to be visible to anyone who views the study. This selection places the list in the Global List folder in the left sidebar under Saved Lists. In any box where you can select lists, the term '**Global**' will identify any list so identified when the list is created.
5. Select the List Type, **Subject** or **Gene**.
    - If you select Subject, enter the **Subject IDs** in the text box that appears. Proceed with step 7.
    - If you select Gene, proceed with step 6.
6. For **Gene Symbol**, enter one or more gene symbols in the text box or click the icons to locate genes in the following databases. If you enter more than one gene in the text box, separate the entries by commas.

caIntegrator provides three methods whereby you can obtain gene symbols for creating a gene list: For more information, see *Choosing Genes* on page 69.

7. If you so choose, you can upload a gene or subject ID list. For the Upload File field, click the **Browse** button to navigate to a .csv file made up of gene symbols. caIntegrator converts the comma-separated content to a gene list.

8. Click **Create List** at the bottom of the page. caIntegrator now opens the Edit [Subject or Gene] List page which shows the name and symbols of the newest gene list (*Figure 4.7*).

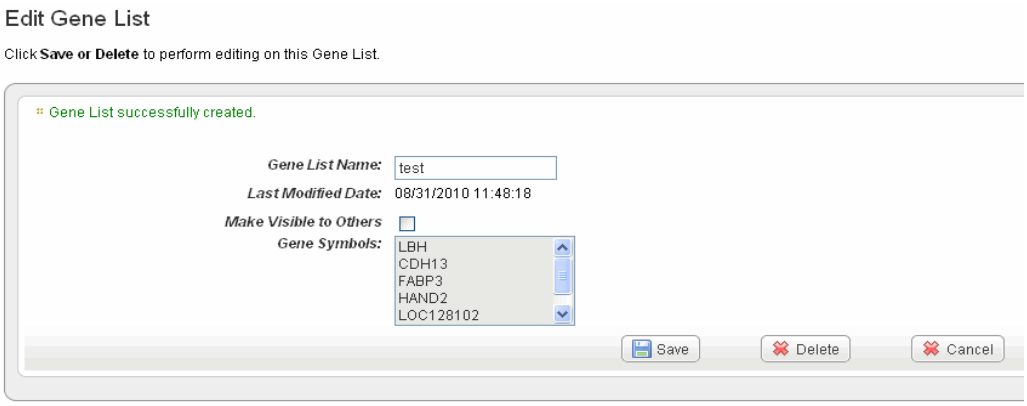


Figure 4.7 The Edit Gene List for reviewing, editing the name or deleting a gene list. The Edit Subject List page is comparable.

See Editing a List for information about the edit feature.

Note: When you perform a GISTIC analysis, caIntegrator automatically saves the retrieved genes in the Saved Copy Number analysis in the left sidebar. For a query or plot analysis, they also appear in the Gene Picker dialog box described in *Choosing Genes* on page 69.

Choosing Genes

caIntegrator provides three methods whereby you can obtain gene names for a gene expression search.

- **caBIO** – This link searches caBIO, then pulls identified genes into caIntegrator for analysis.

  a. Click the **caBIO** icon ( ).

  b. Enter **Search Terms**. Note that caIntegrator can perform a search on a partial HUGO symbol. For example, as search using ***ACH*** would find matches with ‘achalasia’ and ‘arachidonate’.

  c. Select if you want to search in **Gene Keywords**, **Gene Symbols, Gene Alias, Database Cross Reference Identifier** or **Pathways** (from the drop-down list).

  — **Gene Keywords** searches the description field in caBIO; the result displays in the Full Name Column.

  — **Gene Symbols** searches only the Unigene and HUGO gene symbols in caBIO.

  — **Gene Alias** searches for one or more gene symbols which are synonymous for the current gene symbol.

— **Database Cross Reference Identifier** searches for the symbol for this gene as it appears in other databases.

— **Pathways** searches only the pathway names in caBIO. Note that searching in Pathways is a two step process. First, the initial Pathway search produces search results which are pathways. Second, from the pathway search results screen, you must select pathways of interest, then click **Search Pathways for Genes** to obtain a list of genes related to the selected pathways.

d. Select the **Any** or **All** choice to determine how your search terms will be matched. **Any** finds any match for any search term you entered. **All** finds only results that match all of the search terms.

e. Choose the **Taxon** from the drop-down list and click **Search.** The search results display (*Figure 4.8*).

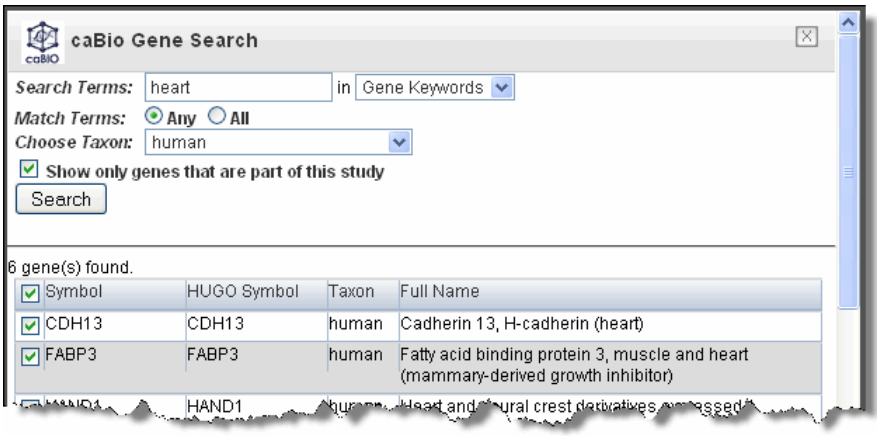


Figure 4.8 Example caBIO gene search criteria and search results

f. In the search results, use the check boxes to identify the genes whose symbols you want to use in the gene expression analysis.

g. Click **Use Genes** at the bottom of the page. This pulls the checked genes into the Criteria tab (*Figure 4.9*).

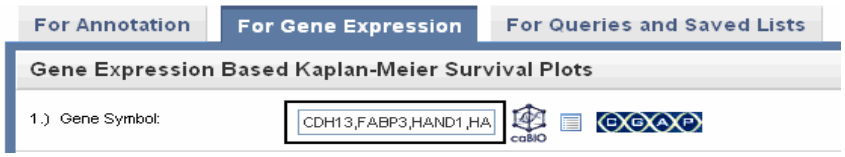


Figure 4.9 Genes pulled in from caBIO display on the Criteria tab

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

  a. Click the Genes List icon ( ) to open the Gene List Picker dialog box. For more information, see *Creating a Gene or Subject List* on page 67.

  b. In the drop-down menu that lists previously saved gene lists, select a gene list. In the list that appears, use the check boxes to identify the genes whose symbols you want to use in the gene expression analysis.

— GISTIC Amplified genes is a list of gene symbols in which the corresponding regions of the genome are significantly amplified.

— GISTIC Deleted genes is a list of gene symbols in which the corresponding regions of the genome are significantly deleted.

c. Click **Use Genes** at the bottom of the dialog. This pulls the checked genes into the Search Criteria tab.

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

Editing a Gene or Subject List

To view a gene list or subject list in caIntegrator, under **Study Data** in the left sidebar, click **Saved Lists** > **Global Lists**, or **My Lists**. Select the list/analysis you want to open. The system displays gene or subject lists that have been saved for the open study.

You can initiate the following functions on this page:

- Click on any of the list names or the list icon ( ) to rerun the query from which the gene or subject list was first created. If the list is a gene list, in the query results, you can click on the gene icon ( ) to open the Cancer Genome Anatomy Project (CGAP) showing metadata for the gene.
- Click the edit icon ( ) to open an Edit Gene/Subject List dialog box. On this page you can review the list of gene symbols or subject IDs included in the list (*Figure 4.10*).

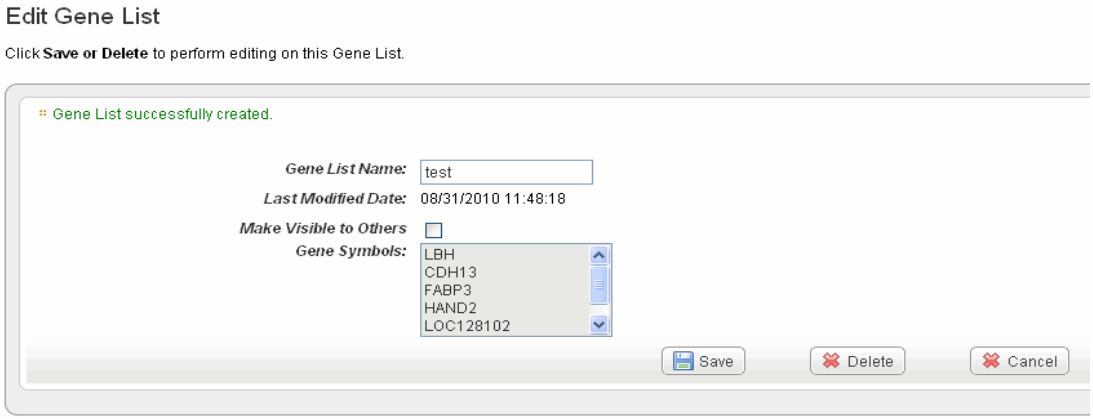


Figure 4.10 Edit Gene List allows you to edit gene lists for a study

In the Edit [List Type] dialog box, you can perform the following tasks:

- To rename the list in the **[List Type] List Name** text box, enter the new list name.
- You can change the visibility of the list in the appropriate check box.

- To delete the list, click the **Delete** button.
- Click **Save** to save your changes or **Cancel** to leave the page without making changes.

Once a list is created, you cannot edit the list contents.

Editing a GISTIC Analysis

To view a GISTIC analysis page in caIntegrator where you can review or edit analysis parameters and results, under **Study Data** in the left sidebar, click **Saved Copy Number Analysis**. Select the analysis you want to open. The system displays analysis parameters and gene lists that that were retrieved from the analysis (*Figure 4.11*).

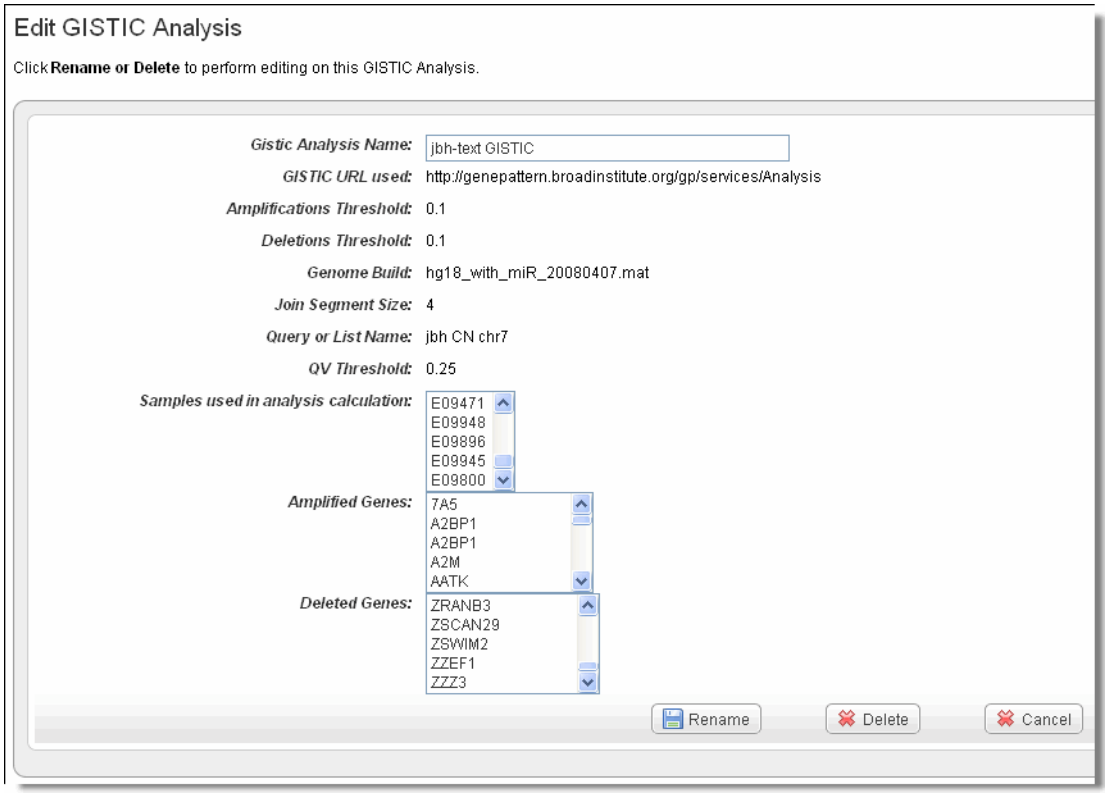


Figure 4.11 Edit GISTIC allows you to view and edit analysis parameters. From this page you can rename or delete the analysis.

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

From this page you can rename or delete the analysis.

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

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

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

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 77.

If you run a 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 4.12*).

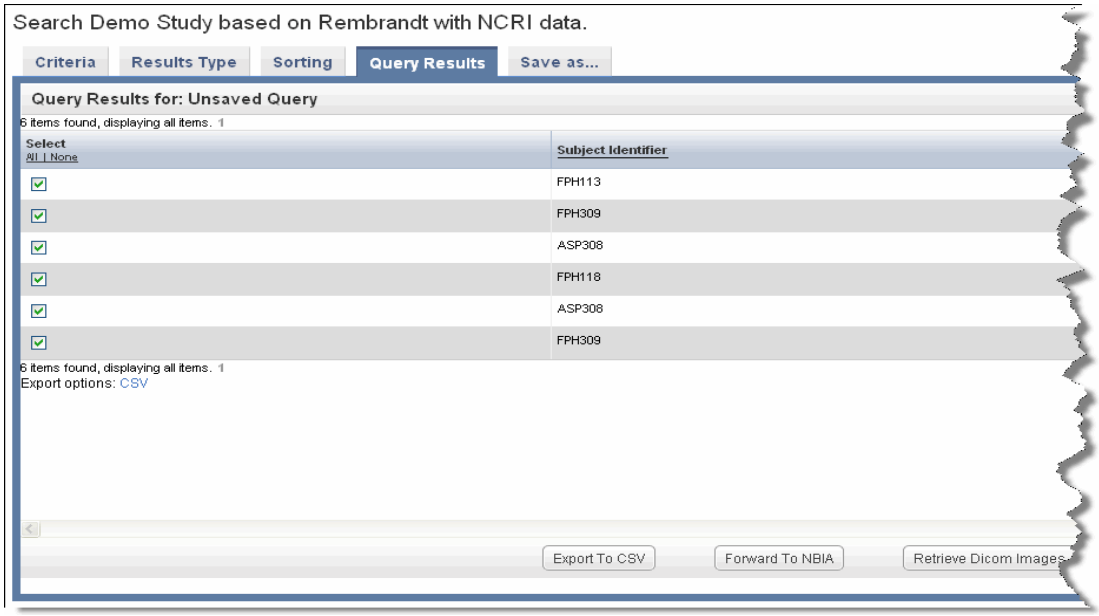


Figure 4.12 With imaging criteria only and no column definition, only Subject IDs display

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, or any annotations you check on the Results Type page. The check boxes work in conjunction with buttons at the bottom of the results page (*Figure 4.13*).

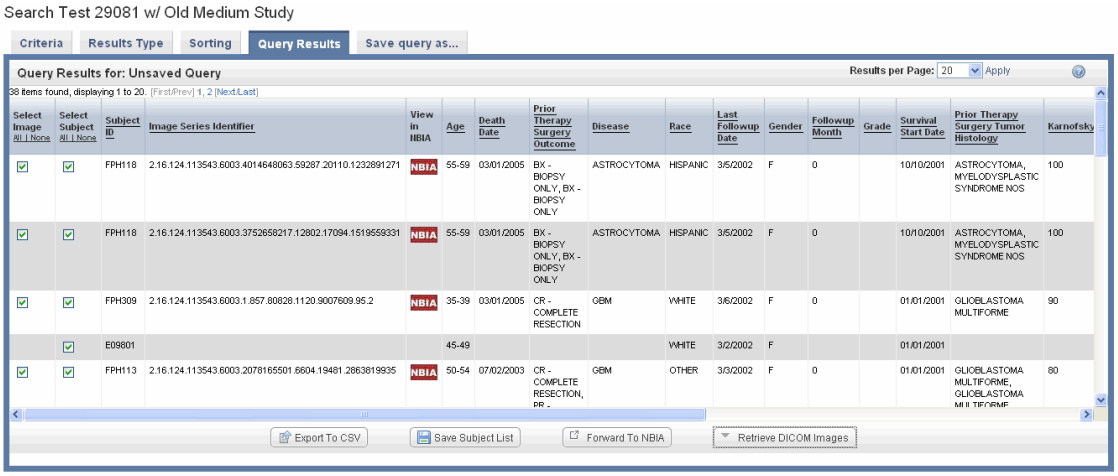


Figure 4.13 By expanding display parameters on the Results Type tab, you can view complete details for image search results

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 to which the image belongs
- Image feature attributes
- The subject ID. Click the subject ID under 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 Tab* on page 58.

See also *caIntegrator and NBIA*, *Retrieving DICOM Images* and *Example of Retrieving Images:*

caIntegrator and NBIA

Images can be accessed in NBIA if you see buttons on the Search Results page. See the Imaging Note in *Results Type Tab* on page 58. 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 4.14*). 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 4.14 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 caIntegrator study was NOT configured with image annotation for an image series, caIntegrator 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 caIntegrator 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. caIntegrator retrieves the corresponding image(s) from NBIA through the grid. NBIA organizes the download file by patient ID, StudyInstance UID, and ImageSeries UID, and compresses it into a zip file. When caIntegrator 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 4.15*). For more information, see the note below.

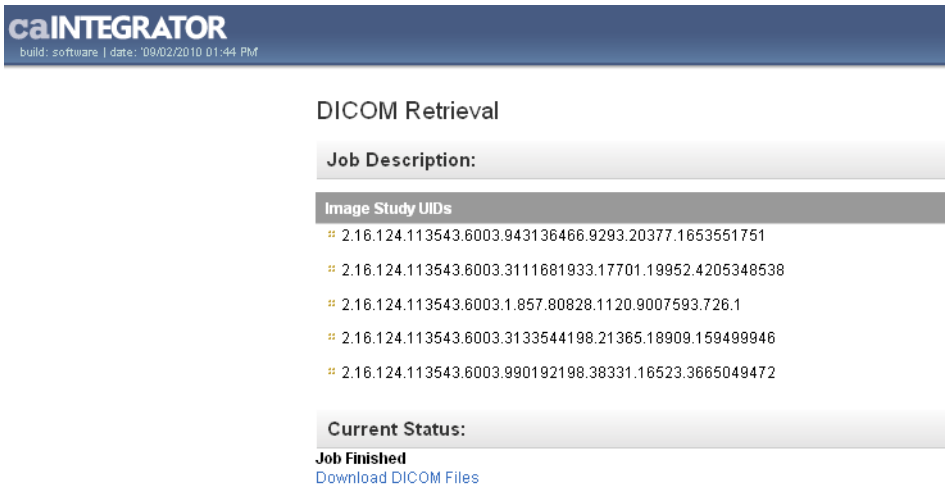


Figure 4.15 DICOM Retrieval result

Click the **Download DICOM** link to download and save the file. caIntegrator 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, caIntegrator aggregates to the *image series*. If results are a mixture and you select multiple rows, caIntegrator 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 77.

Example of Retrieving Images:

If you are searching a study that has image data and image annotation(s) for at least one image series, you would follow these steps:

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.

   caIntegrator 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.

subject annotation 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 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 gene expression data displays as query results.

CHAPTER

5

ANALYZING STUDIES

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

Topics in this chapter include the following:

Data Analysis Overview

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

You can verify that the study has “Deployed” status by selecting the study name in the My Studies dropdown selector. After selecting the study name, click **Home** in the left sidebar of the caIntegrator menu. A study summary should appear, including a status field. If the status is not deployed, or if the study summary does not appear, then the study is not deployed nor available for analysis.

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

- **K-M Plot**: This tool analyzes subject annotation data, generating a Kaplan-Meier (K-M) plot based on survival data sets. See *Creating Kaplan-Meier Plots* on page 80.

- **Gene Expression Plot**: This tool analyzes annotation, subject annotation or genomic data based on gene expression values. See *Creating Gene Expression Plots* on page 86.
- **GenePattern**: This feature provides an express link to GenePattern where you can perform analyses on selected caIntegrator studies, or it enables you to perform several GenePattern analyses on the grid. See *Analyzing Data with GenePattern* on page 99 .

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-Meier method analyzes comparative groups of patients or samples. In caIntegrator, 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, 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.

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

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 29.

K-M Plot for Annotations

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

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

3. Select the **For Annotation** tab at the top of the page (*Figure 5.1*).

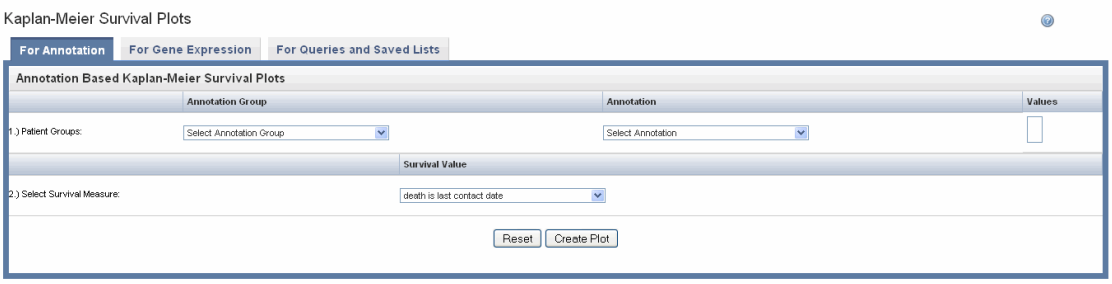


Figure 5.1 Fields for defining annotation data for a K-M plot

4. The groups to be compared in the K-M plot originate from one patient group. Varying data sets are based upon multiple values corresponding to the selected annotation. Define Patient Groups using these options:

   - **Annotation Type** – Select the annotation type that identifies the patient group. Selections are based on the data in the chosen study.
   - **Annotation** – Select an annotation. Fields are based on the annotation type you select. For example, if you choose **Subject**, then you could select **Gender** or **Radiation Type** or any field that would distinguish the patients into groups based upon their values.

     Note: Only annotations that are defined with permissible values display in the dropdown list.
   - **Values** – Using conventional selection techniques, select two or more values which will be the basis for the K-M plot. Permissible (available) values or “No Values” correspond to the selected annotation.

5. **Survival value** is the length of time the patient lived. caIntegrator displays valid survival values entered for this study. Select the survival measure which is the unit of measurement for the survival value to be used for the plot.

6. Click the **Create Plot** button.

caIntegrator generates the plot which then displays below the plot criteria (*Figure 5.2*).

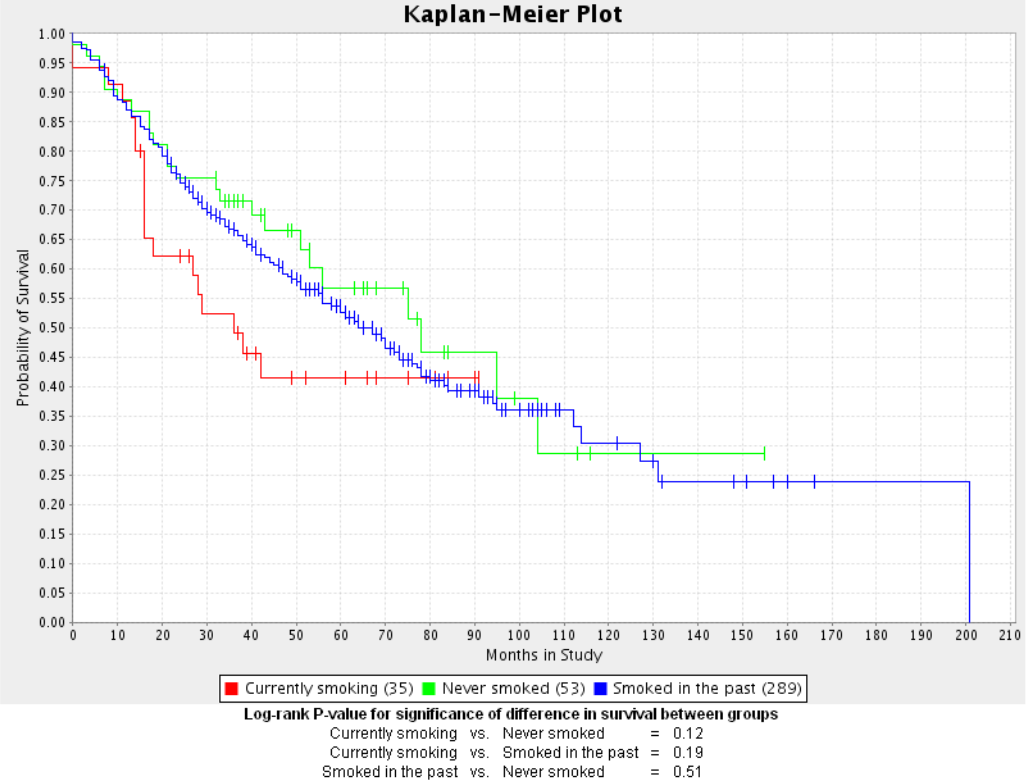


Figure 5.2 A K-M plot generated for groups based on annotations

- The number of subjects for each group appears embedded in the legend of the graph below the plot.
- caIntegrator generates a P-value for the selected groups; it displays at the bottom of the page. A low P-value generally has more significance than a high P-value.

  Note: For information regarding the P-value calculation, see *Creating Kaplan-Meier Plots* on page 80.

K-M Plot for Gene Expression

caIntegrator allows you to compare expression levels for one given gene in different representative groups. The relative expression level is referred to as "fold change". Fold change is the ratio of the measured gene expression value in an experimental sample as determined by a reporter to a reference value calculated for that reporter against all control samples. The reference value is calculated by taking the mean of the $\log_2$ of the expression values for all control samples for the reporter in question. The $\log_2$ mean value (n) is then converted back to a comparable expression signal by returning 2 to the exponent *n*.

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

1. Select the study whose data you want to analyze in the upper right portion of the caIntegrator page. You must select a study with gene expression data.

2. Under Analysis Tools on the left sidebar, select **K-M Plot**.
3. Select the **For Gene Expression** tab (*Figure 5.3*).

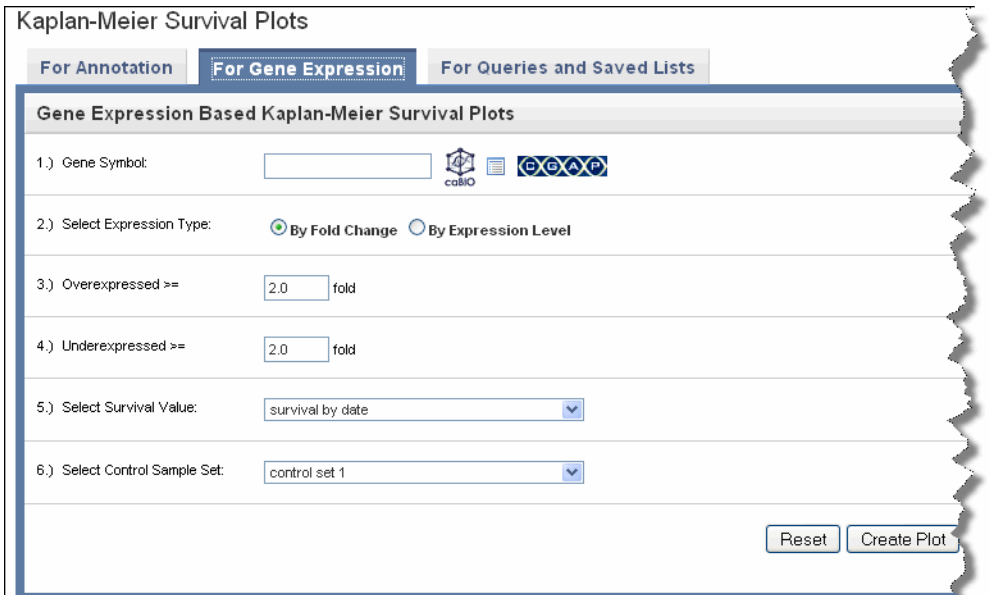


Figure 5.3 Fields for defining gene expression data for a K-M plot

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

   caIntegrator provides three methods whereby you can obtain gene symbols for calculating a KM plot for gene expression. For more information, see *Choosing Genes* on page 97.

5. Select **By Fold Change** or **By Expression Level**.

   - **Fold Change: 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.
   - **Expression Level:** This option allows you to run a KM gene expression plot when there is no control group nor reference data set. Enter values **Above** or **Below Expression Level**.

6. **Survival value** – This field is required for both expression types. The length of time the patient lived. For **Survival Value**, select the survival measure which is the unit of measurement for the survival value to be used for the plot.
7. **Control Sample Sets** – This field is required only for fold change data. One or more control sets are created by the study manager when a study is deployed. Select the **Control Sample Set** you would like to use to calculate fold-change.

   Note: If the study has more than one platform associated with it, the platform is inherently selected when you select the control set. Control sets are comprised of samples from only one platform.

8. Click the **Create Plot** button. caIntegrator generates the plot which then displays below the plot criteria (*Figure 5.4*)..

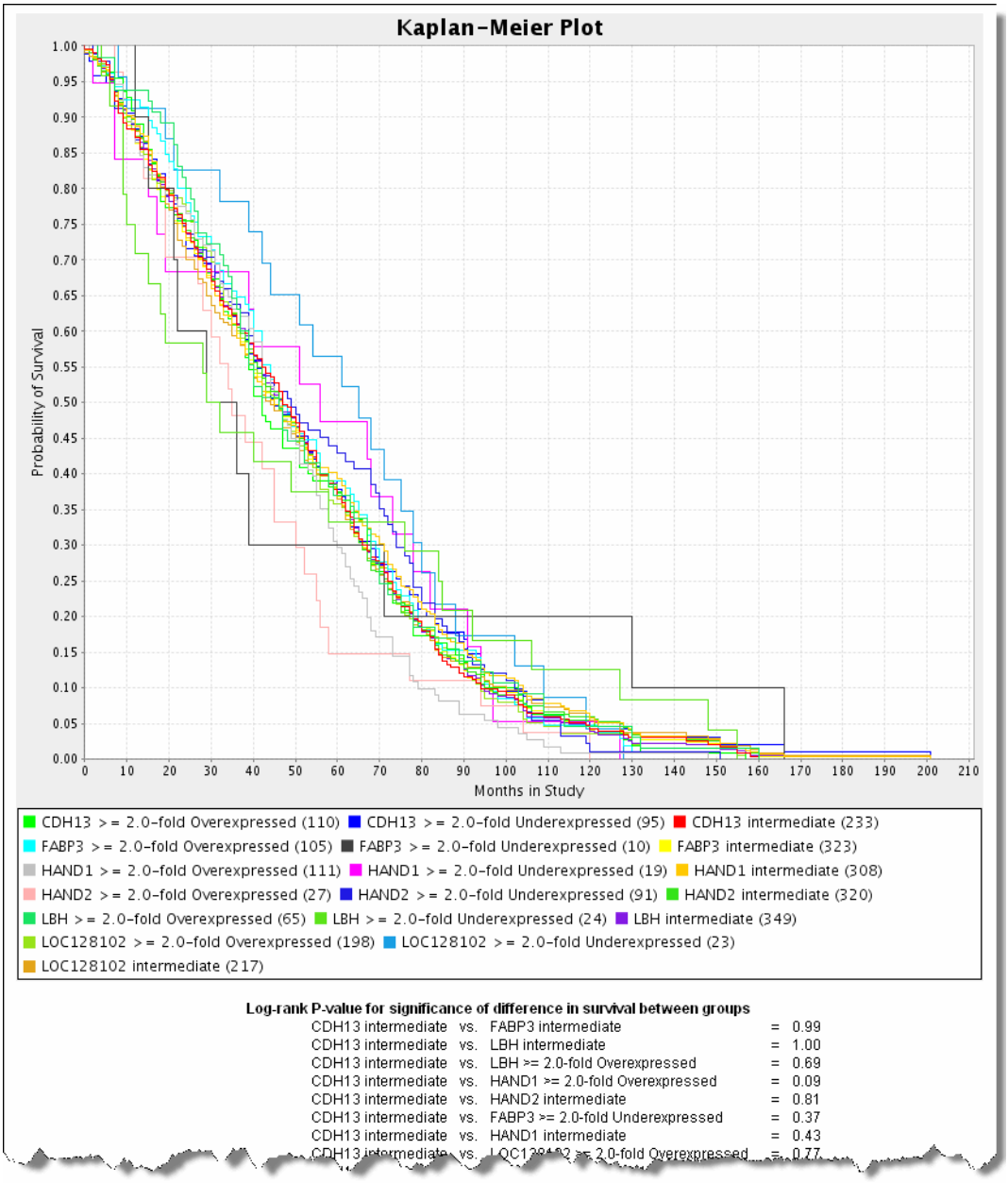


Figure 5.4 K-M plot generated from gene expression data.

- The gene symbol for each group represented in the data appears with its color correlation to the plot embedded in the legend of the graph below the plot. Three lines on the plot represent each gene symbol entered for the plot. Each line of the three represents a subgroup of people carrying the gene--one line for overexpressed values, one line for under expressed values and one line for intermediate values which represents gene values that are not up-regulated nor down-regulated.

- In queries that include a fold change criterion and that are configured to return genomic data, raw expression values are replaced with calculated fold change values.
- A P-value is also generated for the selected groups; it displays at the bottom of the page. A low P-value generally has more significance than a high P-value.

Note: For information regarding the P-value calculation, see *Creating Kaplan-Meier Plots* on page 80.

K-M Plot for Queries and Saved Lists

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

1. Select the study whose data you want to analyze in the upper right portion of the caIntegrator page. The queries you identify for the K-M plot must have been saved previously in caIntegrator.
2. Under Analysis Tools on the left sidebar, select **K-M Plot**.
3. Select the **For Queries and Saved Lists** tab (*Figure 5.5*).

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

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

   Note: Genomic queries do not appear in the lists; they cannot be selected for this type of K-M plot.
5. **Exclusive Subject in Queries** – Check the box if you want to exclude any subjects that appear in both (or all) queries selected for the plot, thus eliminating overlap.
6. **Add Additional Group...all other subjects** – Check the box to create an additional group of all other subjects that are not in selected query groups.

7. **Survival value** – The length of time the patient lived. Select the survival measure which is the unit of measurement for the survival value to be used for the plot.
8. Click the **Create Plot** button. caIntegrator generates the plot which then displays below the plot criteria (*Figure 5.6*).

.

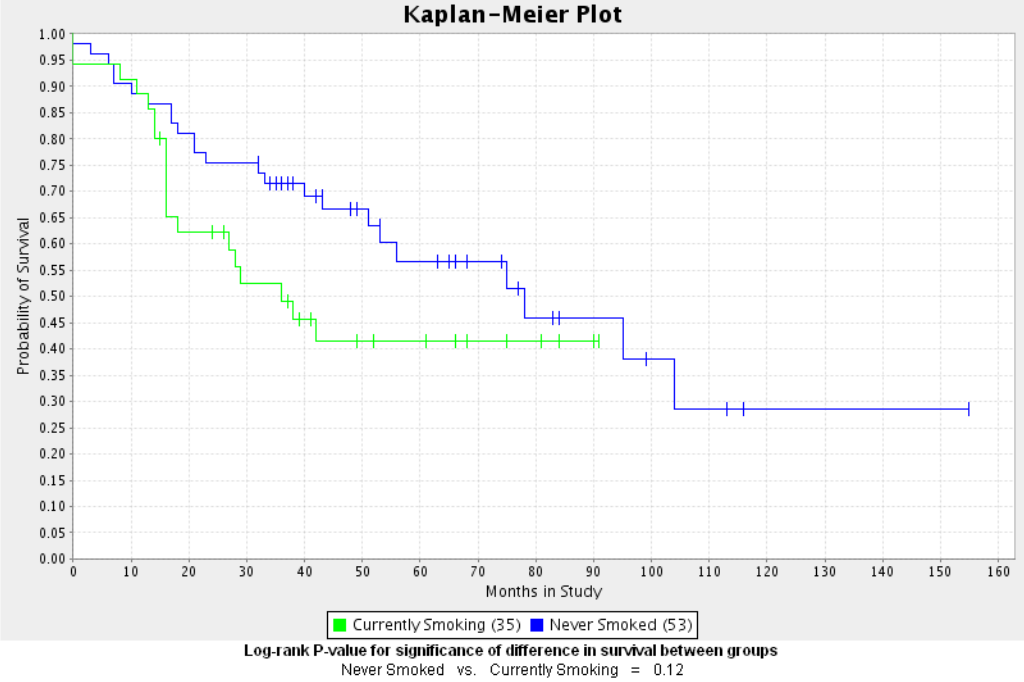


Figure 5.6 K-M Plot comparing statistics between subjects in two queries

- 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.

  Note: For information regarding the P-value calculation, see *Creating Kaplan-Meier Plots* on page 80.

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.

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

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

See also *Understanding a Gene Expression Plot* on page 93.

Gene Expression Value Plot for Annotation

To generate a gene expression plot, follow these steps:

1. Select the study whose data you want to analyze in the upper right portion of the caIntegrator page. (You must select a study which has genomic data.)
2. Under Analysis Tools on the left sidebar, select **Gene Expression Plot**. This opens a page with three tabs
3. Select the **For Annotation** tab (*Figure 5.7*).

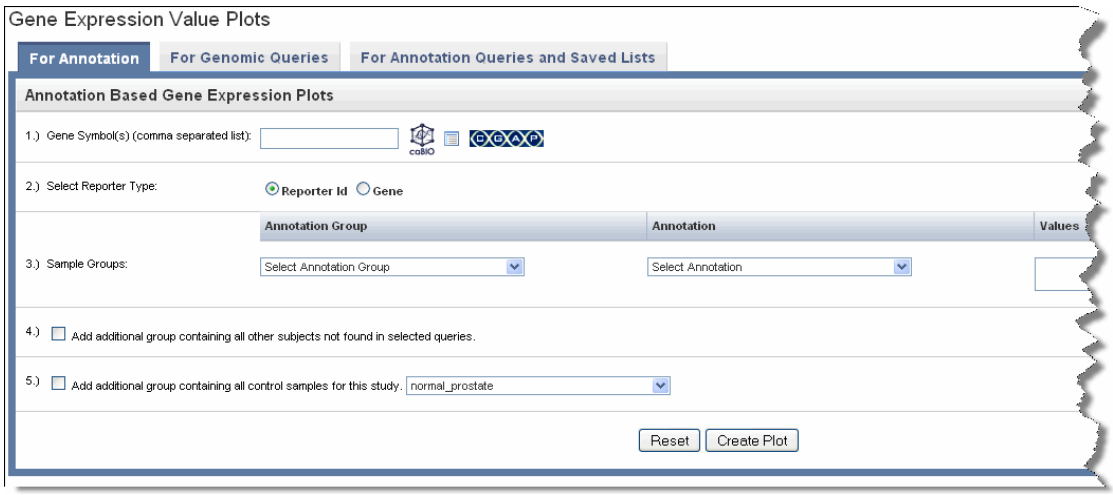


Figure 5.7 Gene expression value tab for configuring gene expression annotation value plot

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

   caIntegrator provides three methods whereby you can obtain gene symbols for calculating a gene expression plot. For more information, see *Choosing Genes* on page 97.

5. **Reporter Type** – Select the radio button that describes the reporter type:
   - **Reporter ID** – Summarizes expression levels for all reporters you specify.
   - **Gene Name** – Summarizes expression levels at the gene level.

- **Platform** – This field displays only if the study has multiple platforms. Select the appropriate platform for the plot. The platform you select determines the genes used for the plot.

6. **Sample Groups** – Choose among the following options:
   - **Annotation Type** – Select the annotation type. Selections are based on the data in the chosen study
   - **Annotation** – Select an annotation. Fields are based on the annotation type you select. For example, if you choose Subject, then you could select Gender or Radiation Type or any field that would distinguish the patients into groups based upon study values.
   - **Values** – Using conventional selection techniques, select one or more values which will be the basis for the plot. Permissible (available) values or "No Values" correspond to the selected annotation.
7. **Add Additional Group...** – Define as follows:
   - **...all other subjects** – Check the box to create an additional group of all other subjects that are not in selected query groups.
   - **...control group** – Check the box to display an additional group of control samples for this study. The control set should be composed of only samples which are mapped to subjects. See *Uploading Control Samples* on page 37.
8. Click the **Create Plot** button. caIntegrator generates the plot which then displays below the plot criteria in bar graph format (*Figure 5.4*).

Legends below the plot indicate the plot input. By default, the plot shows the mean of the data. *Figure 5.8* displays a plot with gene expression median calculation summaries.

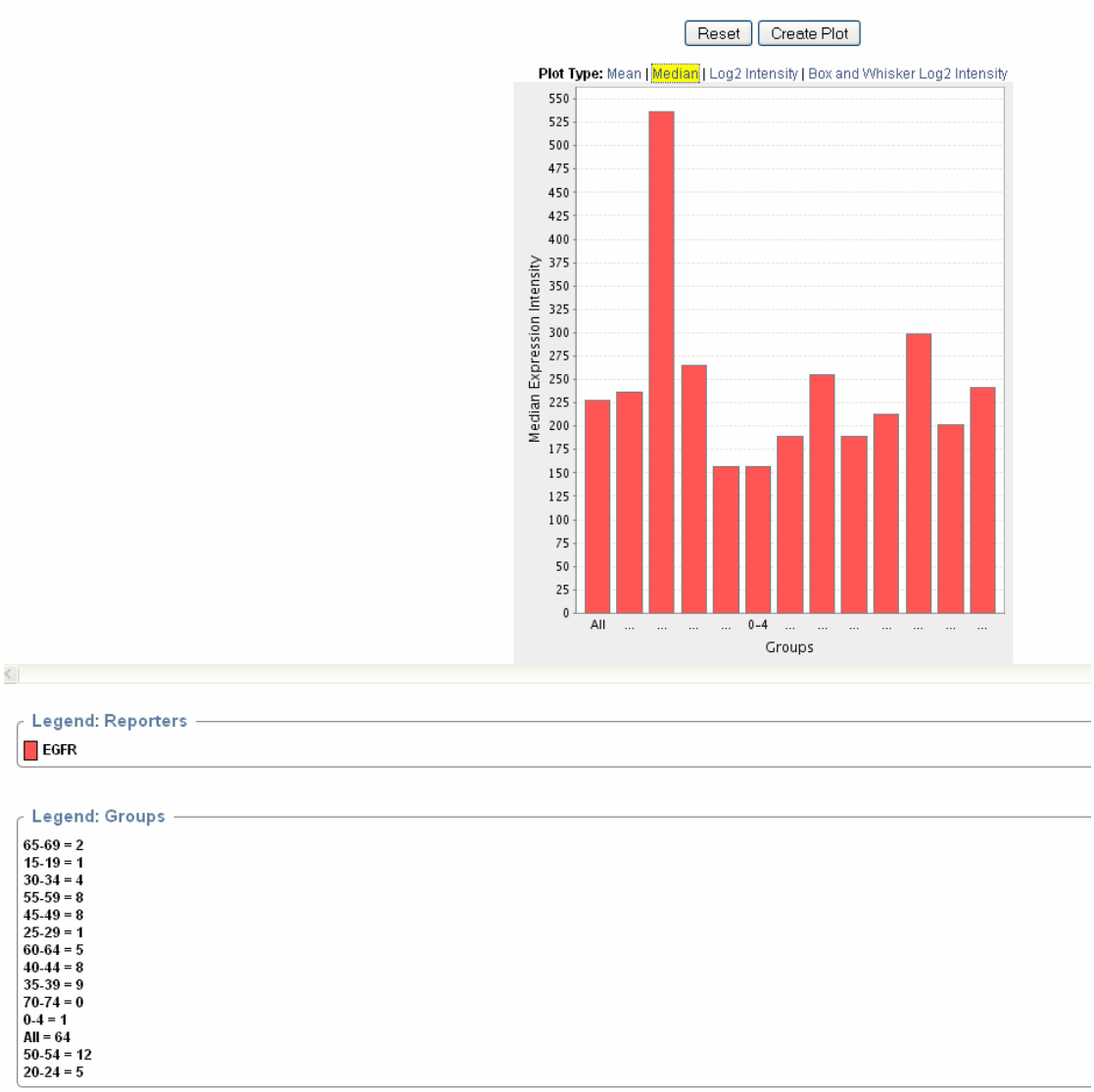


Figure 5.8 Gene expression plot based on selected annotations

- You can recalculate the data display by clicking the **Plot Type** above the graph. See *Understanding a Gene Expression Plot* on page 93.
- 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 61.

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*).

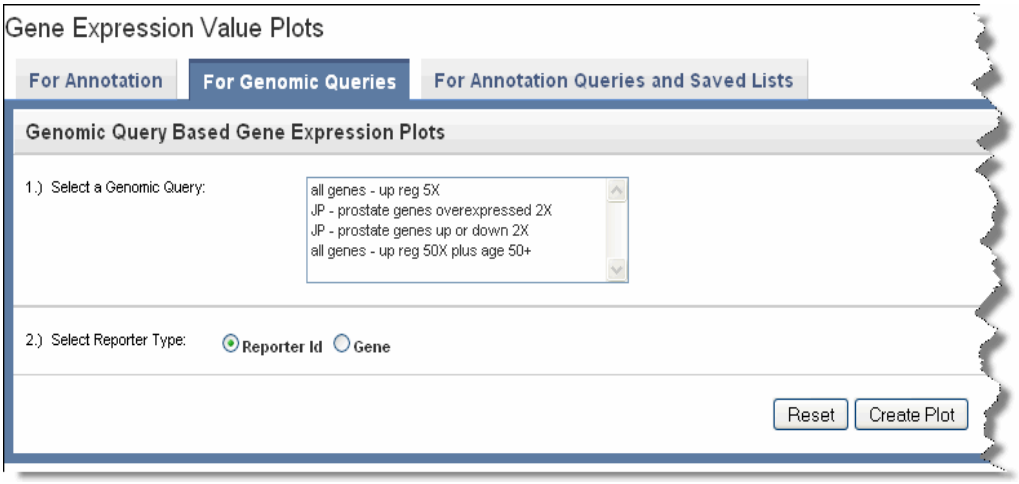


Figure 5.9 Gene expression value tab for configuring gene expression genomic queries plot

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

6. Click the **Create Plot** button. caIntegrator generates the plot which then displays below the 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.

7. You can recalculate the data display by clicking the **Plot Type** above the graph. See *Understanding a Gene Expression Plot* on page 93.

8. You can modify the plot parameters and click the **Reset** button to recalculate the plot.

Gene Expression Value Plot for Annotation and Saved List Queries

Data to be analyzed on this tab must have been saved as a subject annotation query, but it must have genomic data identified in the query. For more information, see *Adding/Editing Genomic Data* on page 31. For the genomic data, you must identify genes whose expression values are used to calculate the plot.

To generate a gene expression plot using an annotation 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 saved as a subject annotation study, but which has genomic data.

2. Under Analysis Tools on the left sidebar, select **Gene Expression Plot**.

3. Select the **For Annotation Queries and Saved Lists** tab (*Figure 5.11*).

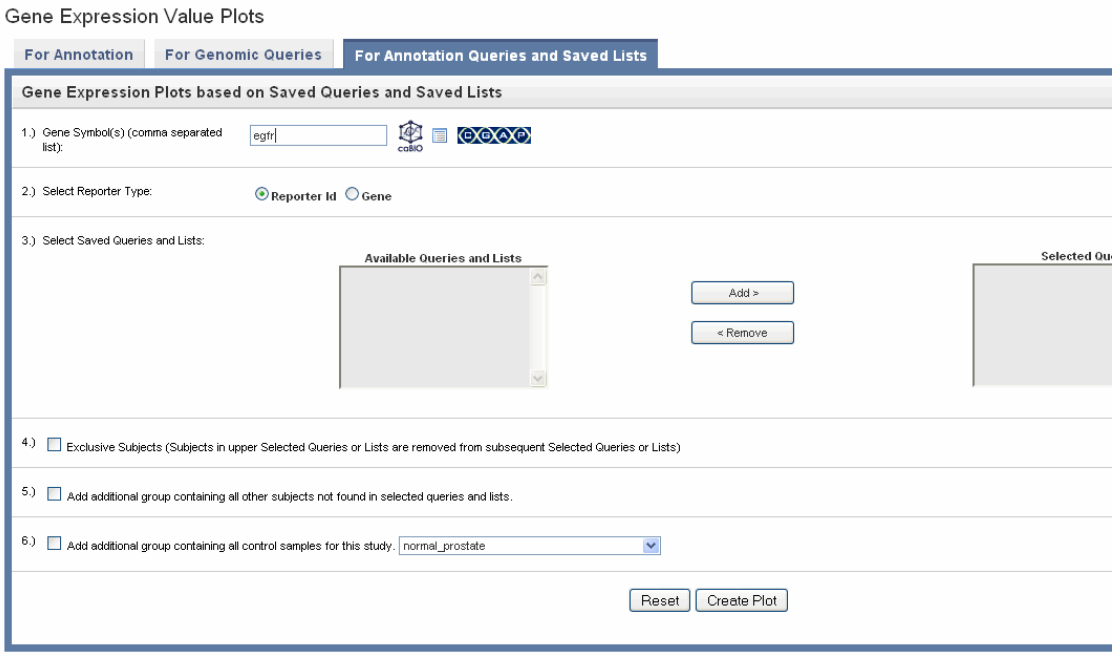


Figure 5.11 Gene expression value tab for configuring gene expression annotation queries plot

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

   caIntegrator provides three methods whereby you can obtain gene symbols for calculating a gene expression plot. For more information, see *Choosing Genes* on page 97.

5. For **Reporter Type**, select the radio button that describes the reporter type:

   - **Reporter ID** – Summarizes expression levels for all reporters you specify.
   - **Gene Name** – Summarizes expression levels at the gene level.
   - **Platform** – This field displays only if the study has multiple platforms. Select the appropriate platform for the plot. The platform you select determines the genes used for the plot.

6. For **Saved Queries,** choose among the available saved queries and lists. Build your selections in the right panel by using the **Add >** and **Remove <** buttons.

   Note: The [SL] and [Q] prefixes to list names indicate "Subject Lists" or "Saved Queries". A "G" in the prefix indicates the list is Global. For more information, see *Creating a Gene or Subject List* on page 67.

7. Check the **Exclusive Subjects...** option to remove subjects in your queries and lists selection from queries or lists you use subsequently for analysis, using them exclusively for the current analysis.

8. For the **Add Additional Group...** options, define as follows:

— **...all other subjects** – Check the box to create an additional group of all other subjects that are not in selected query groups.

— **...control group** – Check the box to display an additional group of control samples for this study. The control set should be composed of only samples which are mapped to subjects. See *Uploading Control Samples* on page 37.

9. Click the **Create Plot** button. caIntegrator generates the plot which then displays below the plot criteria in bar graph format (*Figure 5.4*).

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

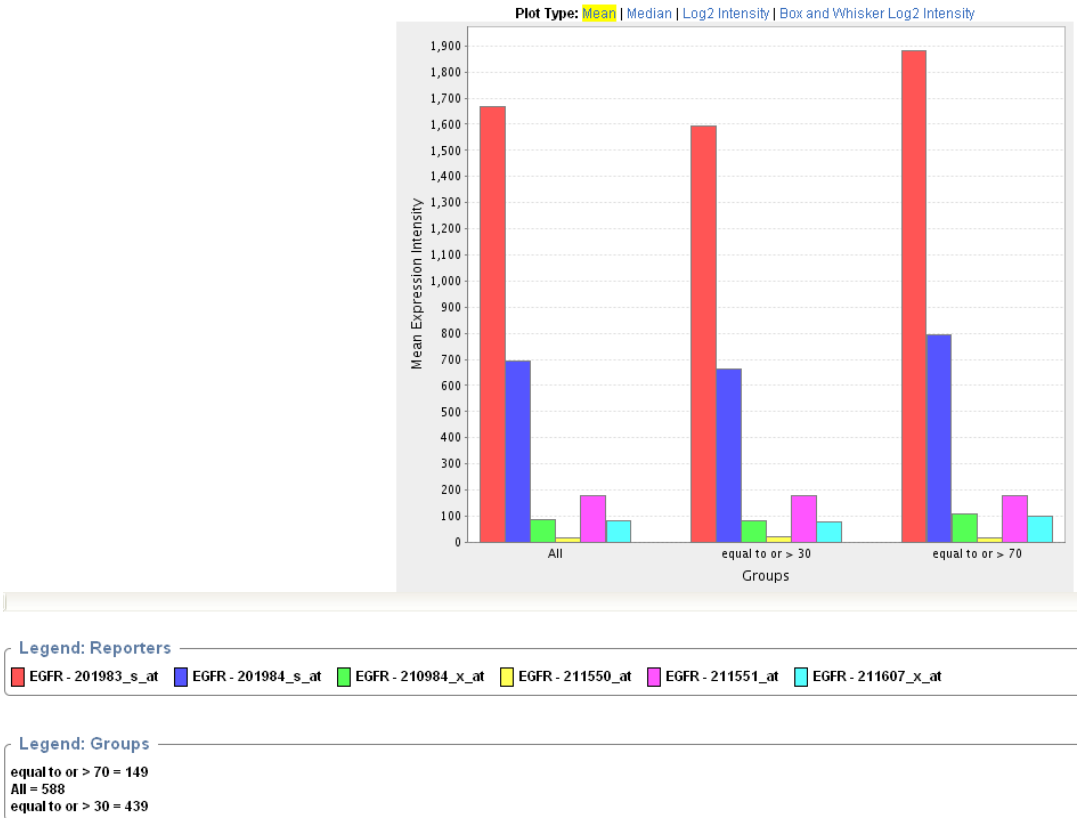


Figure 5.12 Gene expression plot based on annotation queries gene expression values

- You can recalculate the data display by clicking the **Plot Type** above the graph. See *Understanding a Gene Expression Plot* on page 93.
- 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.

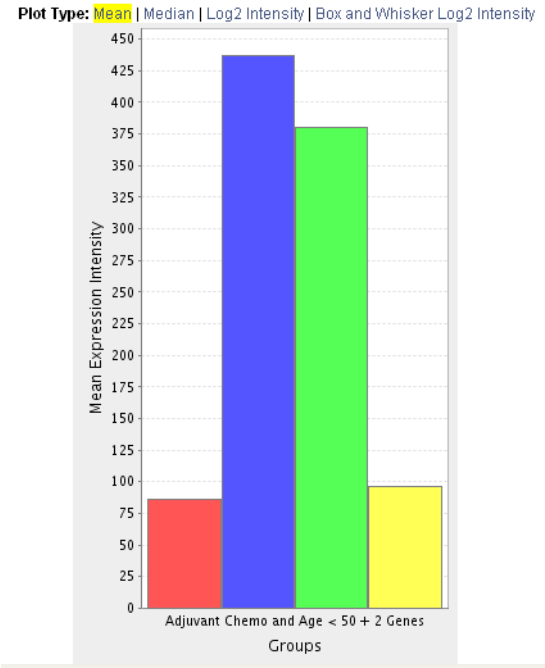


Figure 5.13 Gene expression plot calculating the mean

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.

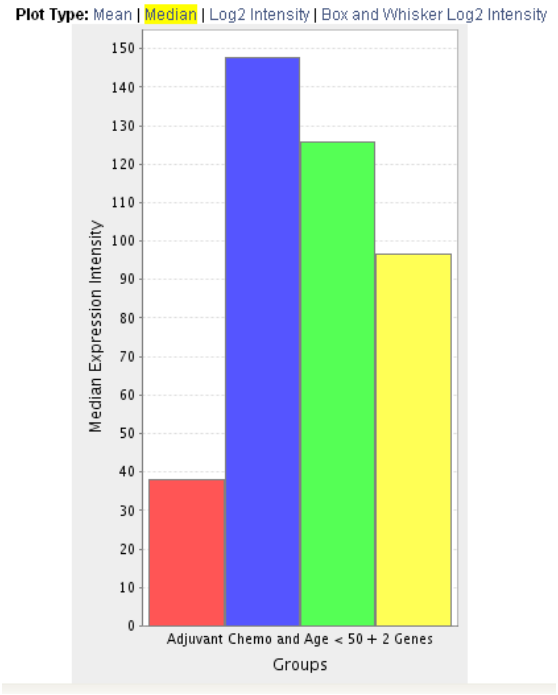


Figure 5.14 Gene expression plot calculating the median

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).

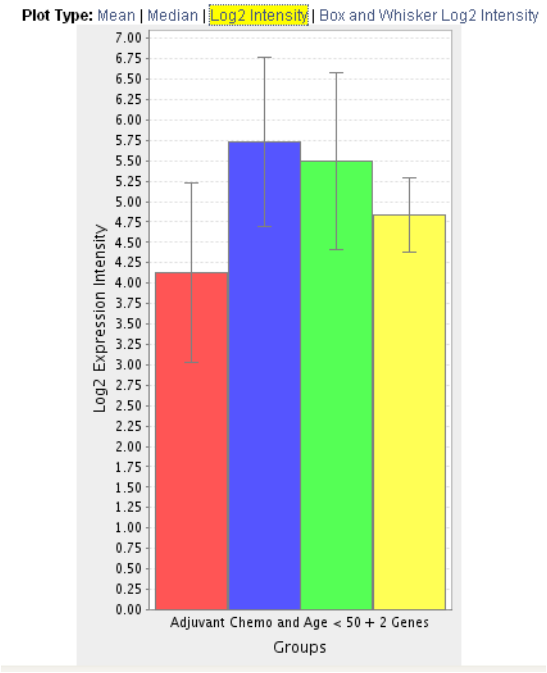


Figure 5.15 Gene expression plot displaying log2 intensity values

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 center, spread, and overall range are immediately apparent.

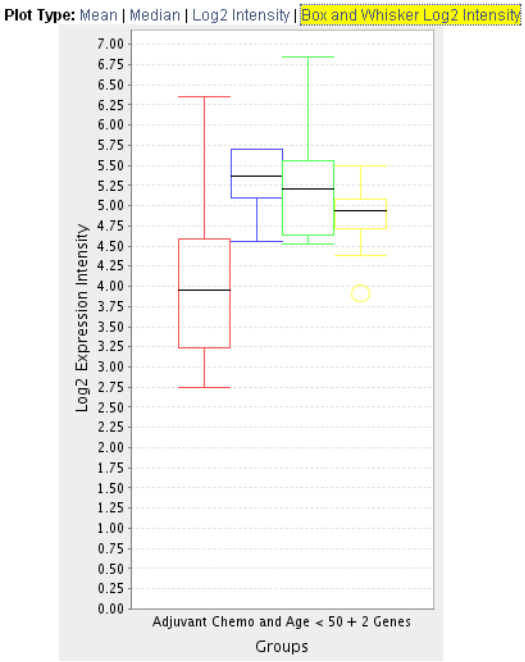


Figure 5.16 Box and whisker plot based on the same data set as represented in Figure 5.13, Figure 5.14, Figure 5.15

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 * IQR greater than Q3 and less than 1.5 * IQR than Q1. Outliers, if present, are shown as open circles (*Figure 5.17*).

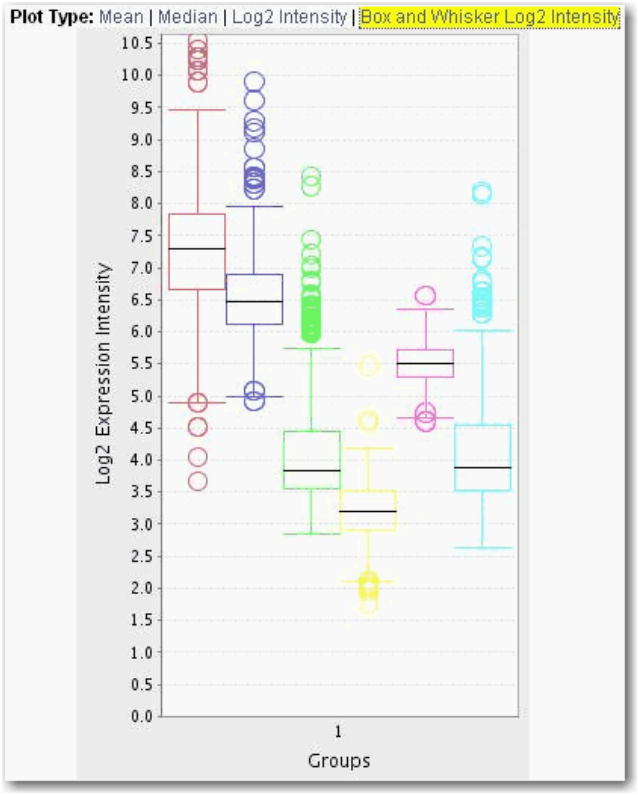


Figure 5.17 Box and whisker plot showing outliers

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.

Choosing Genes

caIntegrator provides three methods whereby you can obtain gene names for data analysis.

- **caBIO** – This link searches caBIO, then pulls identified genes into caIntegrator for analysis.

  a. Click the **caBIO** icon ( ).

  b. Enter **Search Terms**. Note that caIntegrator can perform a search on a partial HUGO symbol. For example, as search using ***ACH*** would find matches with ‘achalasia’ and ‘arachidonate’.

  c. Select if you want to search in **Gene Keywords**, **Gene Symbols, Gene Alias, Database Cross Reference Identifier** or **Pathways** (from the drop-down list).

  — **Gene Keywords** searches the description field in caBIO; the result displays in the Full Name Column.

— **Gene Symbols** searches only the Unigene and HUGO gene symbols in caBIO.

— **Gene Alias** searches for one or more gene symbols which are synonymous for the current gene symbol.

— **Database Cross Reference Identifier** searches for the symbol for this gene as it appears in other databases.

— **Pathways** searches only the pathway names in caBIO. Note that searching in Pathways is a two step process. First, the initial Pathway search produces search results which are pathways. Second, from the pathway search results screen, you must select pathways of interest, then click **Search Pathways for Genes** to obtain a list of genes related to the selected pathways.

d. Select the **Any** or **All** choice to determine how your search terms will be matched. **Any** finds any match for any search term you entered. **All** finds only results that match all of the search terms.

e. Choose the **Taxon** from the drop-down list and click **Search.** The search results display (*Figure 5.18*).

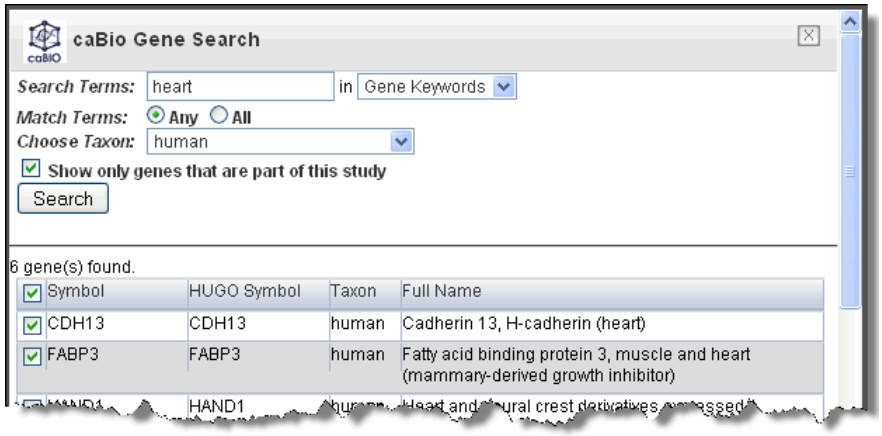


Figure 5.18 Example caBIO gene search criteria and search results

f. In the search results, use the check boxes to identify the genes whose symbols you want to use in the gene expression analysis.

g. Click **Use Genes** at the bottom of the page. This pulls the checked genes into the Criteria tab (*Figure 5.19*).

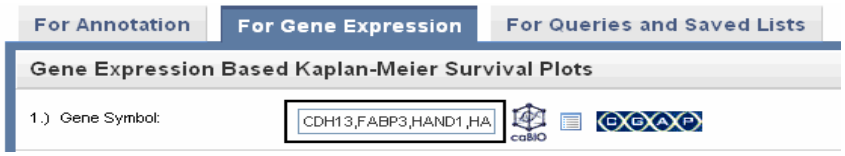


Figure 5.19 Genes pulled in from caBIO display on the Criteria tab

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

a. Click the Genes List icon ( ) to open the Gene List Picker dialog. For more information, see *Creating a Gene or Subject List* on page 67.

— GISTIC Amplified genes is a list of gene symbols in which the corresponding regions of the genome are significantly amplified.

— GISTIC Deleted genes is a list of gene symbols in which the corresponding regions of the genome are significantly deleted.

b. In the drop-down menu that lists previously saved gene lists, select a gene list. In the list that appears, use the check boxes to identify the genes whose symbols you want to use in the gene expression analysis.

c. Click **Use Genes** at the bottom of the dialog. This pulls the checked genes into the Search Criteria tab.

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

Analyzing Data with GenePattern

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

Information is included in this section for connecting to GenePattern from caIntegrator. Specifics for launching GenePattern tools from caIntegrator 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 caIntegrator:

- 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.broadinstitute.org/gp/pages/login.jsf.
  In caIntegrator, enter the URL for connecting: http://genepattern.broadinstitute.org/gp/services/Analysis, then enter your user ID and password.

- Option 2 – Use GenePattern on the grid.

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

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

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

2. Click **GenePattern Analysis** in the left sidebar of caIntegrator. This opens the GenePattern Analysis Status page (*Figure 5.20*).

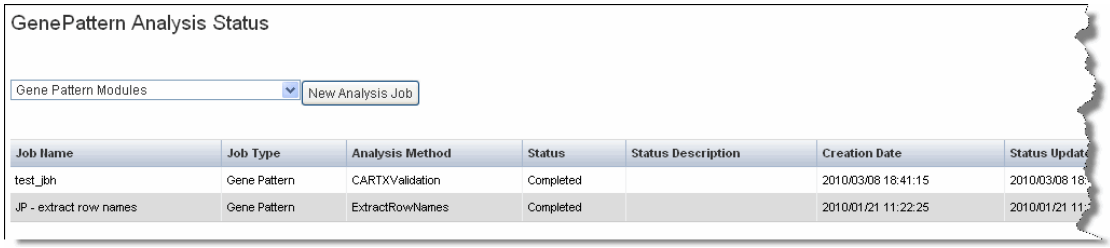


Figure 5.20 GenePattern Analysis Status page

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 100.
   - **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 103.
   - **Principal Component Analysis (Grid Service)**. This option enables you to run this GenePattern analysis on the grid. See *Principal Component Analysis (PCA)* on page 105.
   - **GISTIC (Grid Service)**. This option enables you to run this GenePattern analysis on the grid. See *GISTIC-Supported Analysis* on page 108.
4. Click the **New Analysis Job** button to open a corresponding page where you can configure the analysis parameters.

GenePattern Modules

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

To configure the link for accessing GenePattern from caIntegrator, open the appropriate page as described in *Analyzing Data with GenePattern* on page 99.

1. Select the study whose data you want to analyze in the upper right portion of the caIntegrator page.
2. Click **GenePattern Analysis** in the left sidebar of caIntegrator. This opens the GenePattern Analysis Status page.
3. Make sure **GenePattern Modules** is selected in the drop down list. Click **New Analysis** Job.

4. In the GenePattern Analysis dialog box (*Figure 5.21*), specify connection information, described *Table 5.1* and click **Connect**.

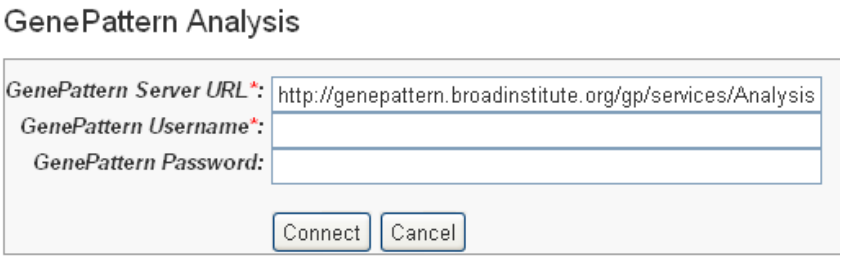


Figure 5.21 Dialog box for configuring the link to GenePattern

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

Table 5.1 Fields for selecting GenePattern configurations

5. After logging in with the GenePattern profile, the dialog box expands to includes fields for defining your GenePattern analysis..

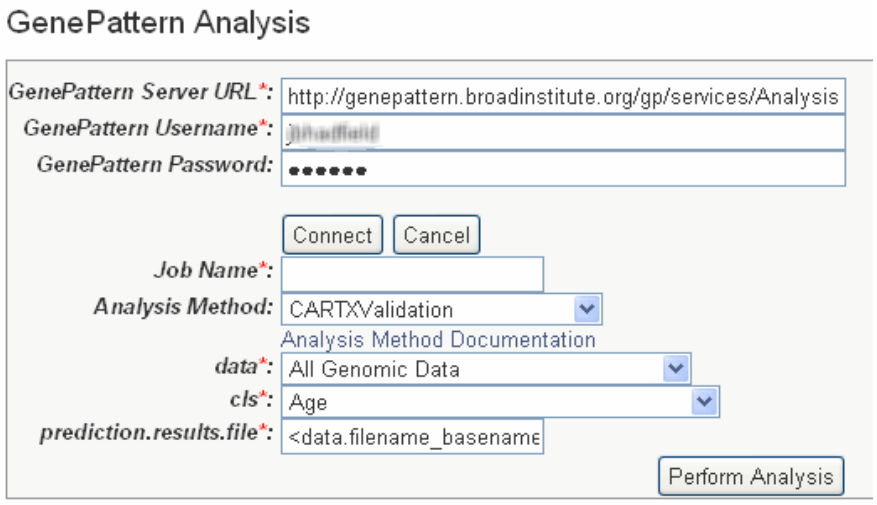


Figure 5.22 GenePattern module options

6. Enter information for the following fields. Fields with a red asterisk are required:
   a. **Job Name*** – Enter a unique name for the analysis
   b. **Analysis Method** – Select any method from the drop down list. Click Analysis Method Documentation for descriptions of the different analysis methods.

c. **Data*** – All genomic data is selected by default. Select from the list any list that has been created for this study.

d. **cls*** – Select any annotation field

The CLS file format defines phenotype (class or template) labels and associates each sample in the expression data with a label. It uses spaces or tabs to separate the fields. The CLS file format differs somewhat depending on whether you are defining categorical or continuous phenotypes:

— Categorical labels define discrete phenotypes; for example, normal vs tumor).

— Continuous phenotypes are used for time series experiments or to define the profile of a gene of interest (gene neighbors).

Note: Most GenePattern modules are intended for use with categorical phenotypes. Therefore, unless the module documentation explicitly states otherwise, a CLS file should define categorical labels.

e. **prediction.results.file** – Enter the name of this file which is part of the output from a GenePattern module.

7. Click **Perform Analysis**. Based on the analysis method you select, you may be asked to add more information for the analysis.For more information, refer to the GenePattern Help site: http://genepattern.broadinstitute.org/gp/getTaskDocCatalog.jsp

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

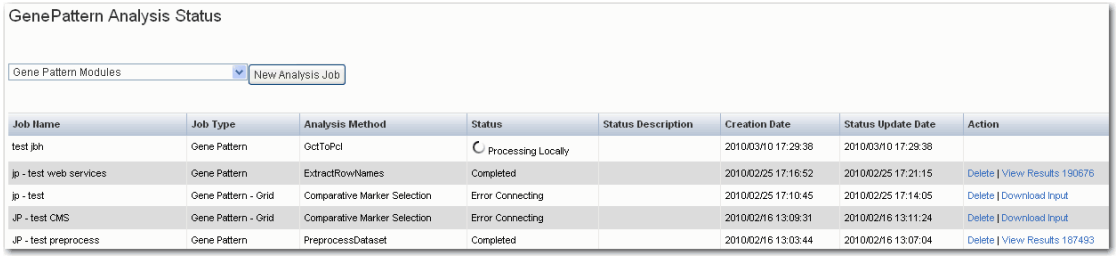


Figure 5.23 GenePattern Analysis Status page displays a list of GenePattern analysis performed on the current study

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

Tip: If you run these analyses within GenePattern itself, you may be able to view results in the GenePattern visualization module. Click **View Results** on the row where the results are listed. If you run them on the grid from caIntegrator, your results will be available only in spreadsheet and XML format.

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

Comparative Marker Selection (CMS) Analysis

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

For more information, see the GenePattern website: http://www.broadinstitute.org/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 caIntegrator page. You must select a study saved as a subject annotation study, but which has genomic data.
2. Click **GenePattern Analysis** in the left sidebar of caIntegrator. This opens the GenePattern Analysis Status page.
3. In the GenePattern Analysis Status page, select **Comparative Marker Selection (Grid Service)** from the drop down list and click **New Analysis Job**. This opens the Comparative Marker Selection Analysis page (*Figure 5.24*).

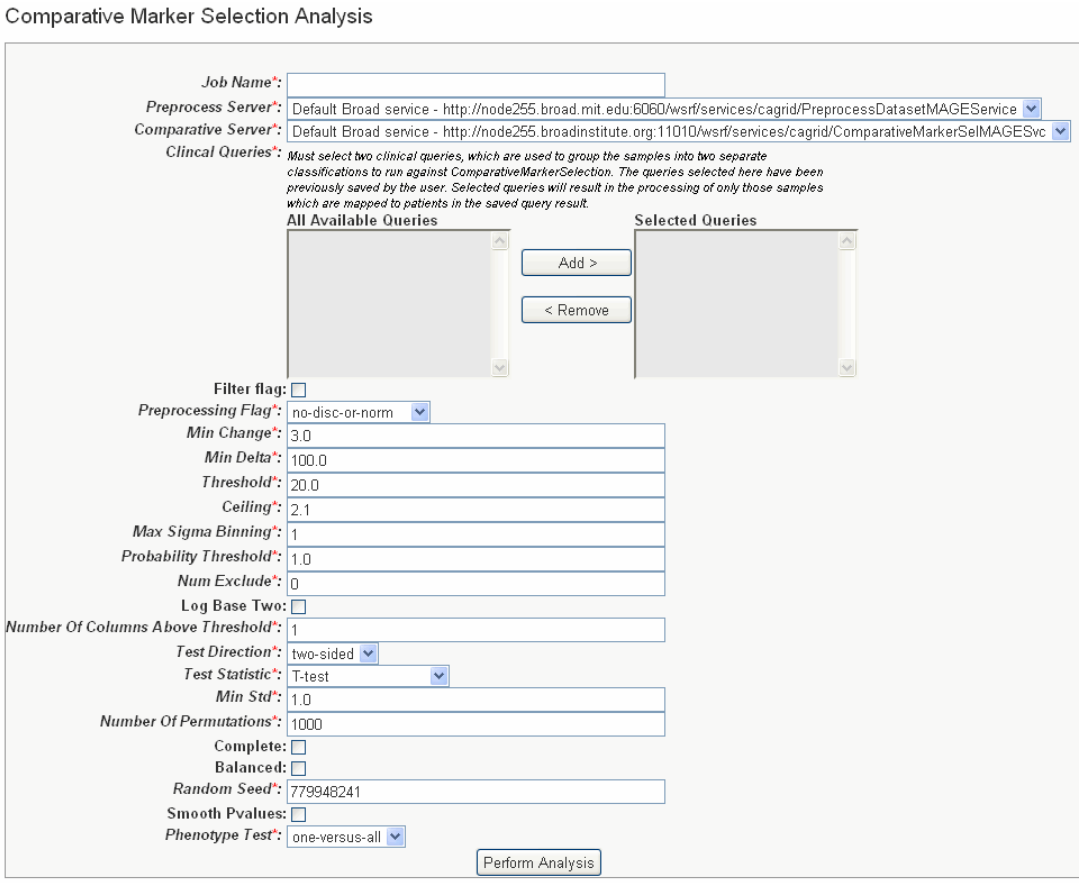


Figure 5.24 Comparative Marker Selection analysis parameters

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

CMS Parameter	*Description*
Job Name*	Assign a unique name to the analysis you are configuring.
Preprocess Server*	A server which hosts the grid-enabled data GenePattern PreProcess Dataset module. Select one from the list and caIntegrator will use the selected server for this portion of the processing.
Comparative Server*	A server which hosts the grid-enabled data GenePattern Comparative Marker Selection module. Select one from the list and caIntegrator will use the selected server for this portion of the processing.
Annotation Queries and Lists*	All subject annotation queries and gene lists with appropriate data for the analysis are listed. Select and move two or more queries from the **All Available Queries** panel to the **Selected Queries** panel using the **Add >** and **Remove <** buttons. **Note:** The [SL] and [Q] prefixes to list names indicate "Subject Lists" or "Saved Queries". A "G" in the prefix indicates the list is Global. For more information, see *Creating a Gene or Subject List* on page 67.
Filter Flag	Variation filter and thresholding flag
Preprocessing Flag*	Discretization and normalization flag
Min Change*	Minimum fold change for filter
Min Delta*	Minimum delta for filter
Threshold*	Value for threshold
Ceiling*	Value for ceiling
Max Sigma Binning*	Maximum sigma for binning
Probability Threshold*	Value for uniform probability threshold filter
Num Exclude*	Number of experiments to exclude (max & min) before applying variation filter
Log Base Two	Whether to take the log base two after thresholding; default setting is "Yes".
Number of Columns Above Threshold*	Remove row if n columns are not >= than the given threshold In other words, the module can remove rows in which the given number of columns does not contain a value greater or equal to a user defined threshold.
Test Direction*	The test to perform (up-regulated for class0; up-regulated for class1, two sided). By default, Comparative Marker Selection performs the two-sided test.
Test Statistic*	Select the statistic to use.
Min Std*	The minimum standard deviation if test statistic includes the min std option. Used only if test statistic includes the min std option.

Table 5.2 Comparative Marker Selection analysis options

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

Table 5.2 Comparative Marker Selection analysis options

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

   caIntegrator takes you to the JobStatus/Launch page where you will see the job and its status in the Status column of the list (*Figure 5.25*).

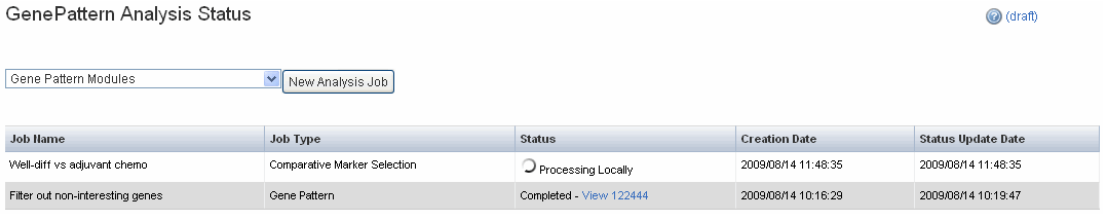


Figure 5.25 The progress of a GenePattern analysis that has been launched displays in the status column of page

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

Principal Component Analysis (PCA)

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

You can configure GenePattern grid parameters for preprocessing the dataset in addition to PCA module parameters. For more information, see the GenePattern website: http://www.broadinstitute.org/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 caIntegrator page. You must select a study with gene expression data.
2. Click **GenePattern Analysis** in the left sidebar of caIntegrator. This opens the GenePattern Analysis Status page.
3. In the GenePattern Analysis Status page, select **Principal Component Analysis (Grid Service)** from the drop down list and click **New Analysis Job**. This opens the Principal Component Analysis page (*Figure 5.26*).

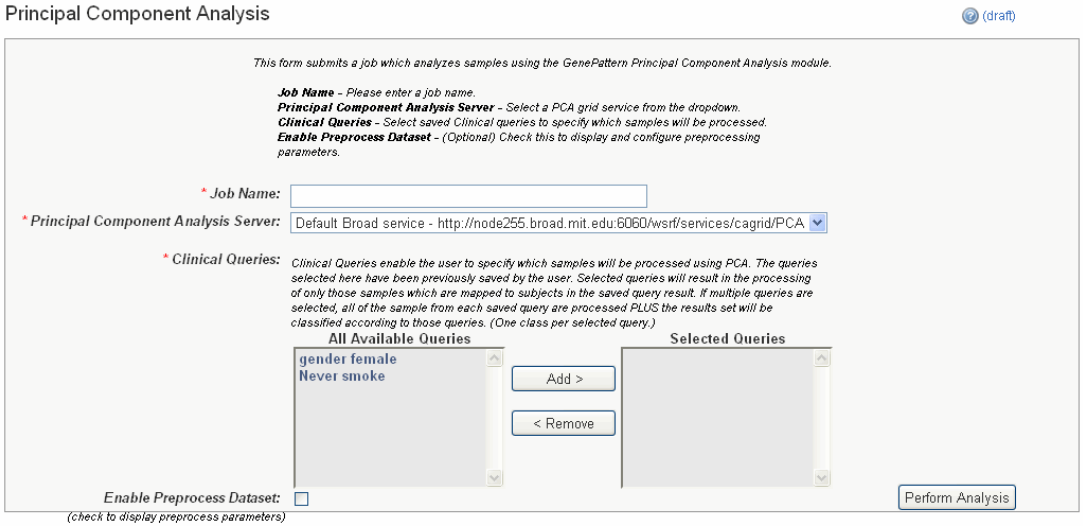


Figure 5.26 Principal Component Analysis parameters

4. Select or define PCA analysis parameters, described in *Table 5.3*. You must enter a job name and select an annotation 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 caIntegrator will use the selected server for this portion of the processing.
Annotation Queries*	All annotation queries display in this list. Select one or more of these queries to define which samples are analyzed using PCA. If you select more than one query, then the union of the samples returned by the multiple queries is analyzed.

Table 5.3 PCA analysis options

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

Table 5.3 PCA analysis options

5. If you want to preprocess the data set, click **Enable the Preprocess Dataset**. This opens an additional set of parameters (*Figure 5.27*), discussed in *Table 5.4*. The preprocessing is executed prior to running the PCA.

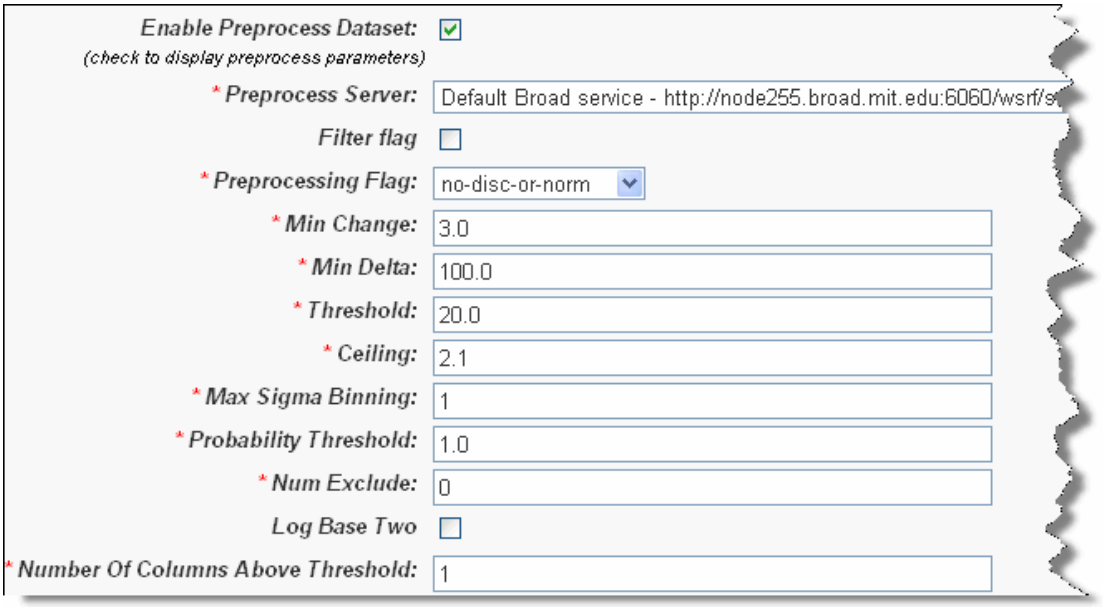


Figure 5.27 Parameters for pre-processing parameters for PCA

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

Table 5.4 Parameters for preprocessing data sets for PCA

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

Table 5.4 Parameters for preprocessing data sets for PCA

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

GISTIC-Supported Analysis

Note: The GISTIC test option displays only if the study contains copy number or SNP data. For more information, see *Configuring Copy Number Data* on page 38.

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.broadinstitute.org/cgi-bin/cancer/software/genepattern/modules/gp_modules.cgi.

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

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

3. In the GenePattern Analysis Status page, select **GISTIC (Grid Service)** from the drop down list and click **New Analysis Job**. This opens the GISTIC Analysis page (*Figure 5.28*).

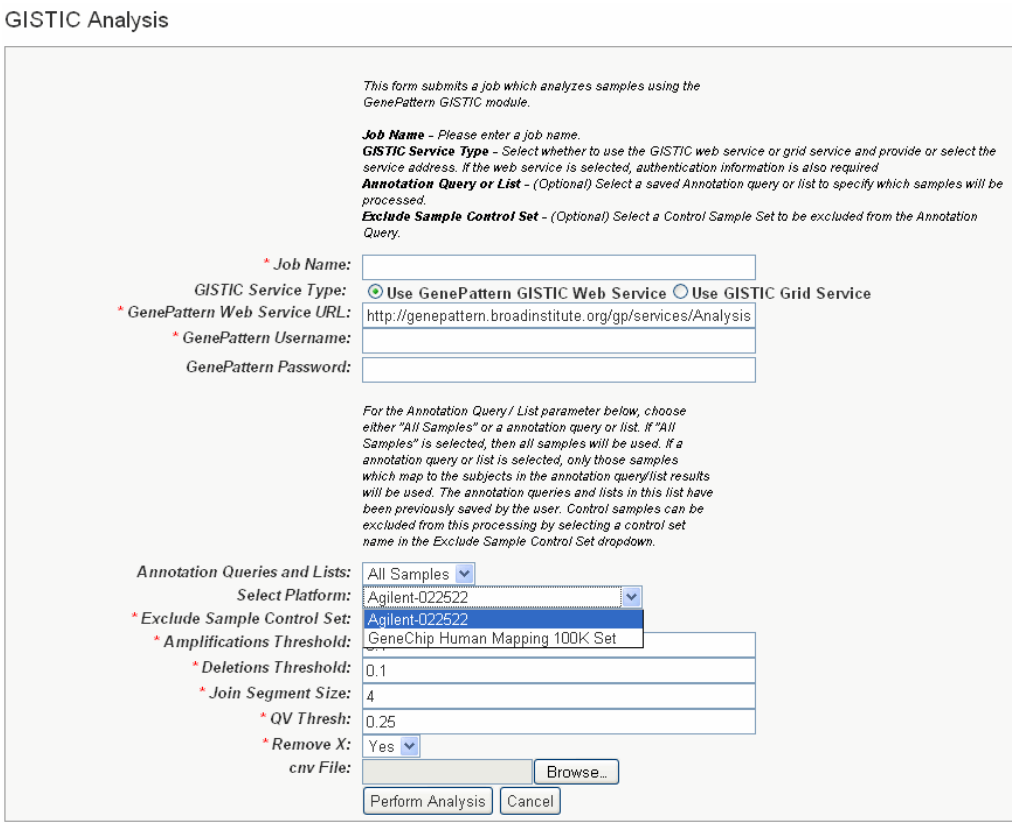


Figure 5.28 GISTIC analysis criteria

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

GISTIC Parameters	*Description*
Job Name*	Assign a unique name to the analysis you are configuring.
GISTIC Service Type*	Select whether to use the GISTIC web service or grid service and provide or select the service address. If the web service is selected, authentication information is also required
GenePattern User Name/ Password	Include these to log into GenePattern for the analysis.
Annotation Queries and Lists	All annotation queries display in this list as well as an option to select all non-control samples. Select an annotation query if you wish to run GISTIC on a subset of the data and select all non-control samples if wish to include all samples.
Select Platform	This option appears only if more than one copy number platform exists in the study. Select the appropriate platform from the drop-down list (*Figure 5.28*).

Table 5.5 GISTIC analysis parameters

GISTIC Parameters	*Description*
Exclude Sample Control Set	From the drop-down list, select the name of the control set you want to exclude from the analysis. Click **None** if that is applicable.
Amplifications Threshold*	Threshold for copy number amplifications. Regions with a log2 ratio above this value are considered amplified. Default = 0.1.
Deletions Threshold*	Threshold for copy number deletions. Regions with a log2 ratio below the negative of this value are considered deletions. Default = 0.1.
Join Segment Size*	Smallest number of markers to allow in segments from the segmented data. Segments that contain fewer than this number of markers are joined to the neighboring segment that is closest in copy number. Default = 4.
QV Thresh[hold]*	Threshold for q-values. Regions with q-values below this number are considered significant. Default = 0.25.
Remove X*	Flag indicating whether to remove data from the X-chromosome before analysis. Allowed values = {1,0}. Default = 1(yes).
cnv File	This selection is optional. Browse for the file. There are two options for the CNV file. **Option #1** enables you to identify CNVs by marker name. Permissible file format is described as follows: A two column, tab-delimited file with an optional header row. The marker names given in this file must match the marker names given in the markers_file. The CNV identifiers are for user use and can be arbitrary. The column headers are: 1. Marker Name 2. CNV Identifier **Option #2** enables you to identify CNVs by genomic location. Permissible file format is described as follows: A 6 column, tab-delimited file with an optional header row. The 'CNV Identifier', 'Narrow Region Start' and 'Narrow Region End' are for user use and can be arbitrary. The column headers are: 1. CNV Identifier 2. Chromosome 3. Narrow Region Start 4. Narrow Region End 5. Wide Region Start 6. Wide Region End

Table 5.5 GISTIC analysis parameters

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

6. When the job is complete, the system displays a completion date on the GenePattern Analysis status page. Click the **Download** link. This downloads zipped result files to your local work station. The number of files and their file type will vary according to the processing. The results format is compatible with GenePattern visualizers and can be uploaded within GenePattern.
7. Additionally, upon completion of a successful GISTIC anaylsis, caIntegrator automatically displays the two gene lists that it generates in the Gene List Picker so that you can use them in a caIntegrator query or plot calculation. The lists are visible only to your userID. For more information, see *Choosing Genes* on page 97. The genes will also display in **Saved Copy Number Analyses** in the left sidebar. See *Editing a GISTIC Analysis* on page 72.

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

Viewing Data with the Integrative Genomics Viewer

Once you have run a query for gene expression, *Gene Expression Data Searches* on page 54, or copy number data, *Copy Number Searches* on page 55, you can view results in the Integrative Genomics Viewer (IGV).

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

Notes: For more information about the Integrative Genomics Viewer or to connect independently to the IGV home page, click this link: http://www.broadinstitute.org/igv/log-in.
The IGV viewer and the NCI Heat Map viewer both require you to install a version of Java containing Java Web Start. For more information, see *Java for IGV and Heat Map Viewer* on page 116.

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

Method 1

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

3. Follow the instructions through the intermediate dialog boxes. After clicking **Open** with the Java program listed, the `IGV.jnlp` opens, displaying the dataset in the computer screen (*Figure 5.29*).

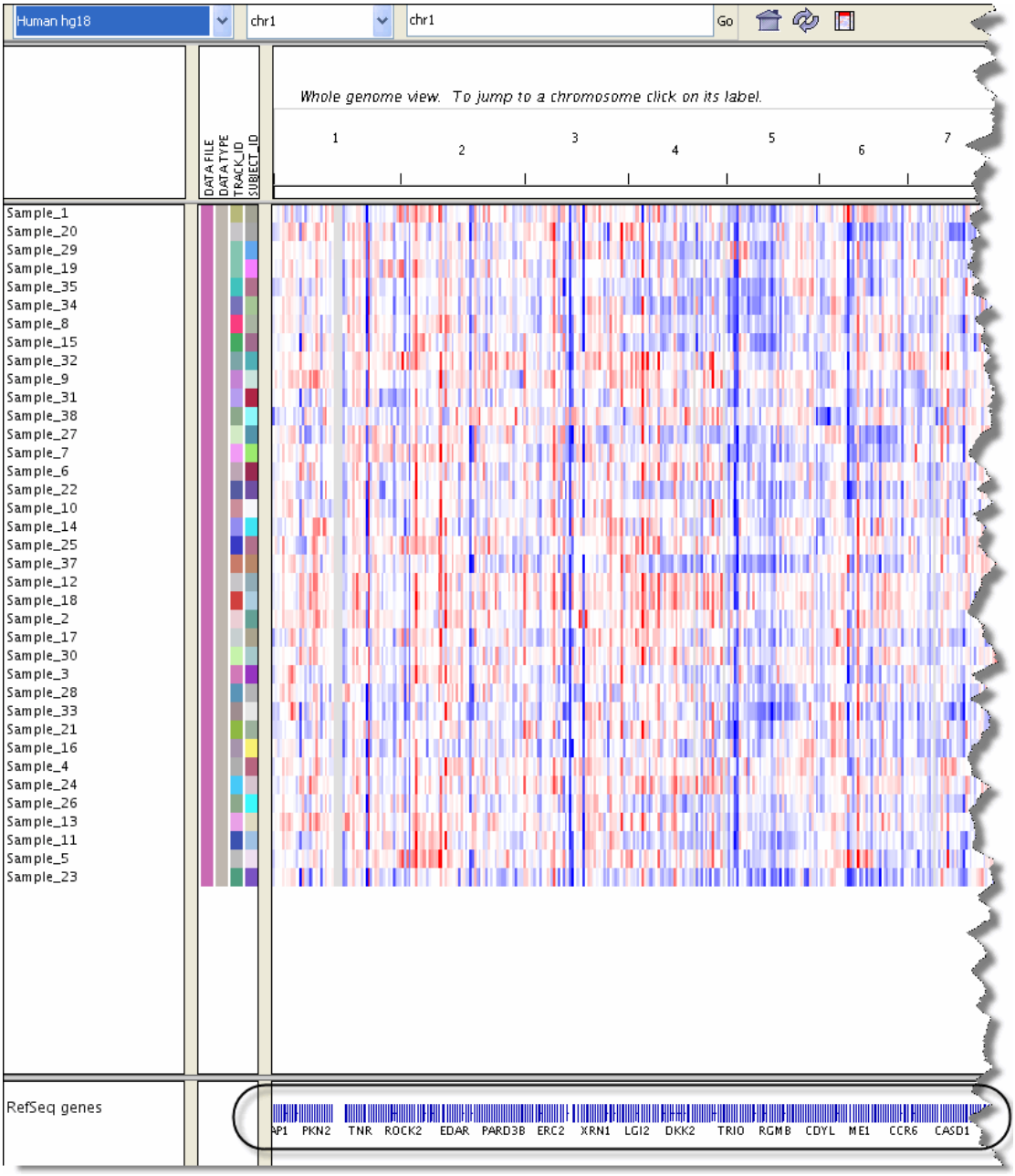


Figure 5.29 IGV Viewer displaying expression results from data isolated in caIntegrator

4. Move your mouse to hover over the genes graphic at the bottom of the page, indicated in *Figure 5.29*.

5. Click the mouse when you've identified a gene of interest.

This opens the genome site at UCSC http://genome.ucsc.edu/index.html?org=Human&db=hg18&hgsid=174031247, where you can learn more about the gene (*Figure 5.30*).

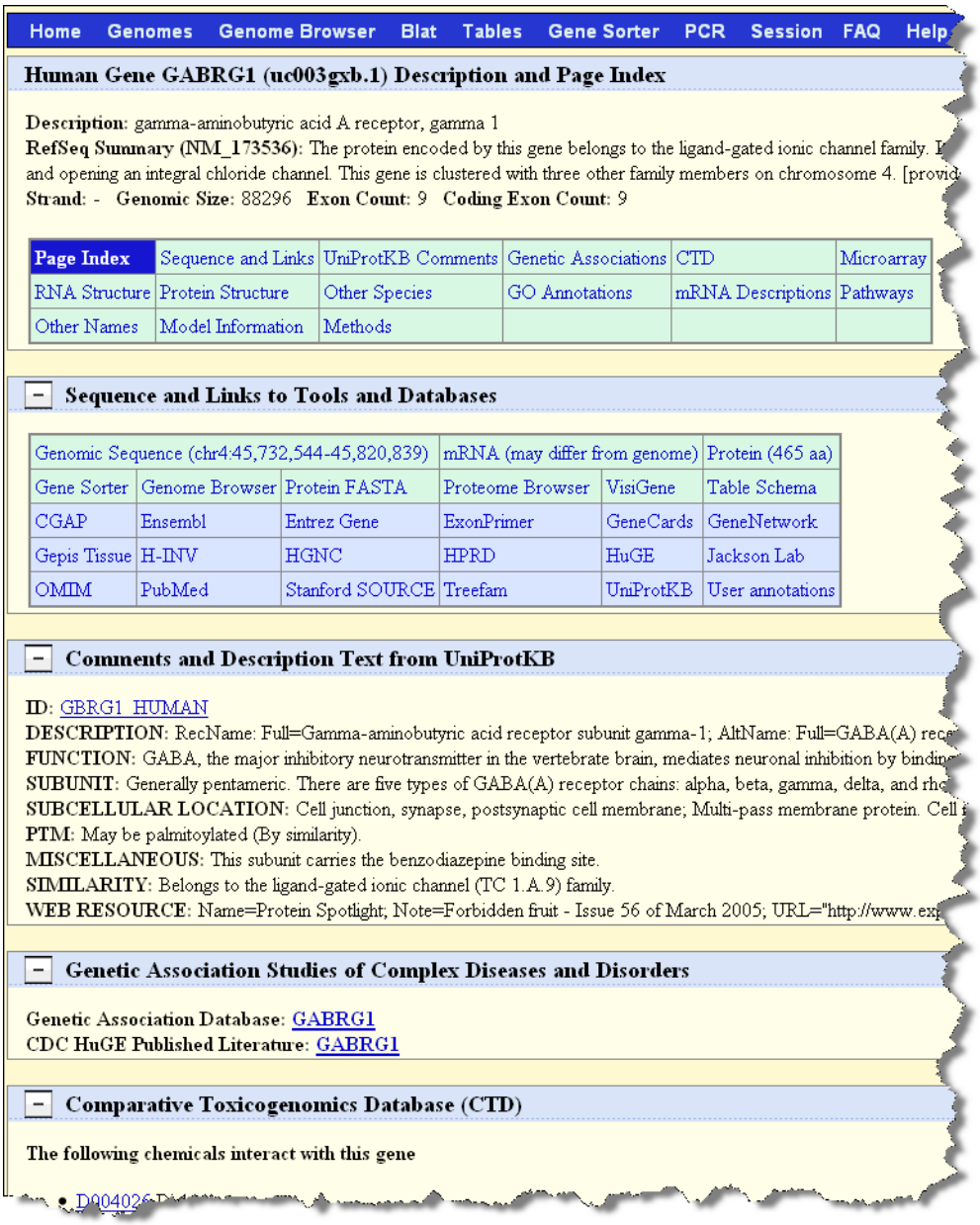


Figure 5.30 Example of the kind of metadata you can learn about a gene at the UCSC genome website

6. Go to the following website for a user guide for IGV: http://www.broadinstitute.org/software/igv/UserGuide.

Method 2

1. With the appropriate study open, click **Integrative Genomics Viewer** on the left sidebar.

2. This opens the View IGV Selector page (*Figure 5.31*).

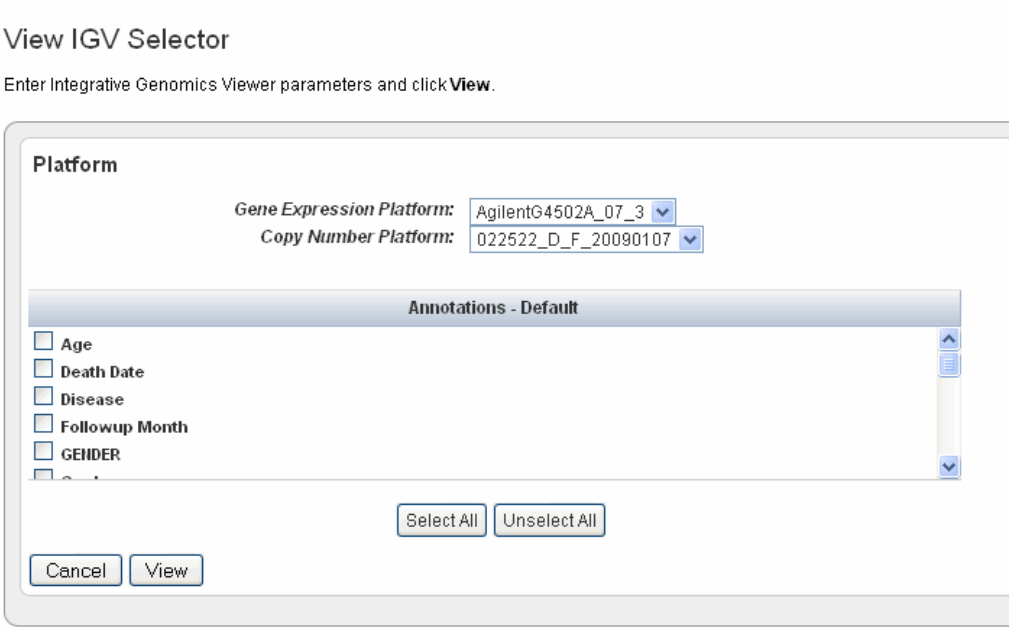


Figure 5.31 The page for configuring the connection to the IGV

3. In the drop-down list, select the **Gene Expression Platform** for the data you want to view.
4. Select the **Copy Number Platform** ID.
5. The Annotations - Default panel displays existing annotation fields for the gene expression data in the open study. Select those fields you want to view when you open the IGV. Use the buttons for convenience if you want to **Select All** or **Unselect All**, when all are checked.
6. Click **View** to see the data in the Integrative Genomic Viewer. caIntegrator creates IGV files of the data.
7. After the files are created, click the **Launch Integrative Viewer** hypertext link that appears.
8. Continue with step 3 on page 112.

Viewing Data with Heat Map Viewer

Once you have run a query for gene expression, *Gene Expression Data Searches* on page 54, or copy number data, *Copy Number Searches* on page 55, you can view results in the Heat Map Viewer (HMV).

Notes: For more information about the Heat Map Viewer or to connect independently to the HMV home page, click this link: https://cgwb.nci.nih.gov/cgi-bin/heatmap.
The IGV viewer and the NCI Heat Map viewer both require you to install a version of Java containing Java Web Start. For more information, see *Java for IGV and Heat Map Viewer* on page 116.

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

Method 1

1. With the appropriate study open, at the bottom of the Query Results page, click the **View in Heat Map Viewer** button.
2. If you click the button at the bottom of the page with any of the query results line items selected, caIntegrator creates HMV files, with a monitor informing you of this. After the files are created, click the **Launch Heat Map Viewer** hypertext link.
3. Follow the instructions through the intermediate dialog boxes. After clicking **Open** with the Java program listed, the `retrieveFile.jnlp` runs, displaying the dataset in the computer screen (*Figure 5.32*).

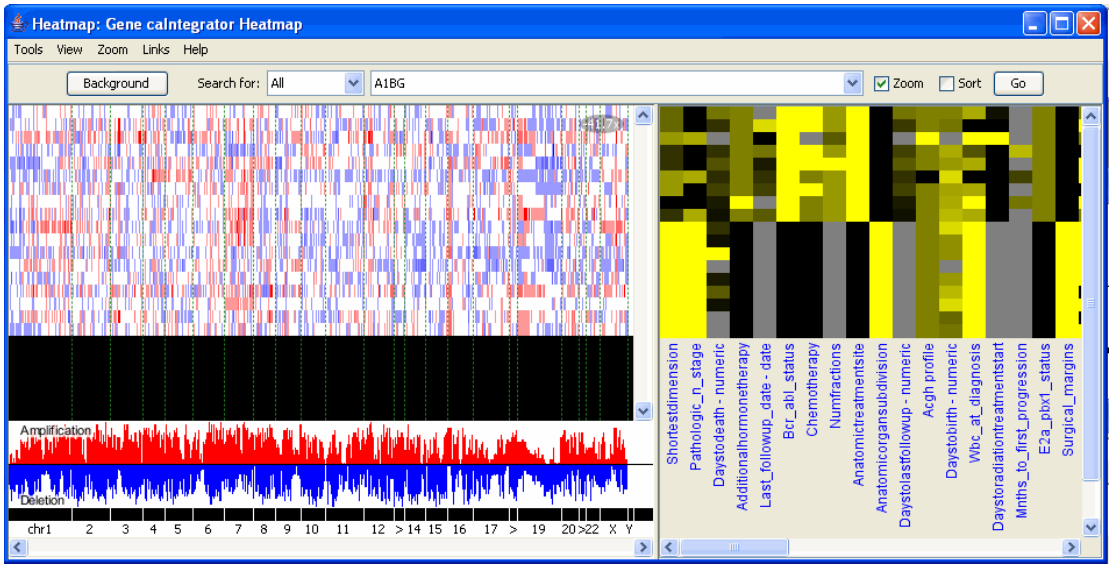


Figure 5.32 Data display in Heat Map Viewer

4. Go to the following website for Heat Map Viewer documentation: https://cgwb.nci.nih.gov/goldenPath/heatmap/documentation/index.html.

Method 2

1. With the appropriate study open, click **Heat Map Viewer** on the left sidebar.

2. This opens the View Heat Map Viewer Selector page (*Figure 5.33*).

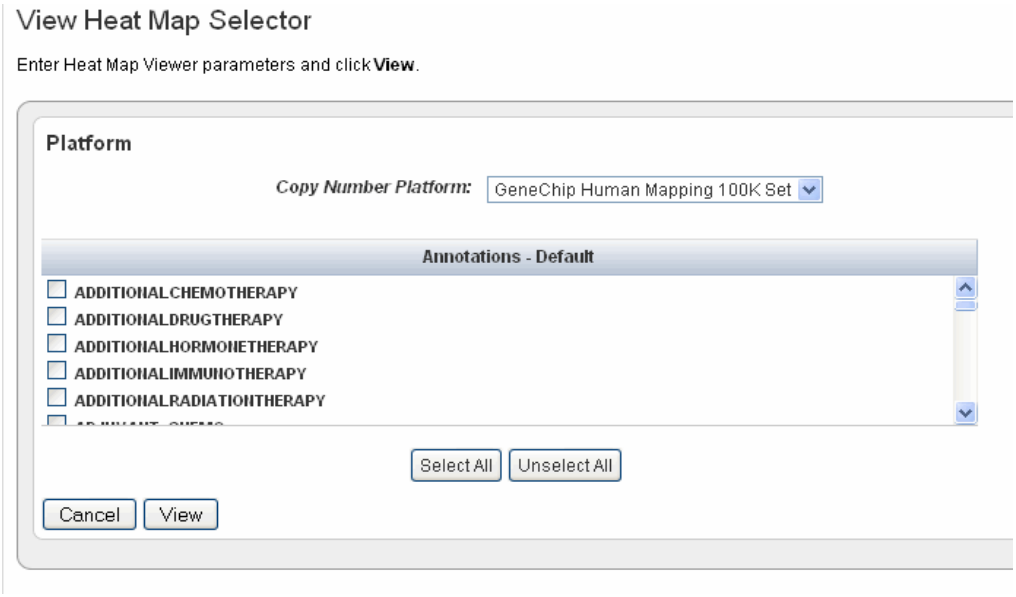


Figure 5.33 View Heat Map Selector page

3. Select the appropriate **Copy Number Platform** in the drop down list.
4. The Annotations - Default panel displays existing annotation fields for the gene expression data in the open study. Select one or more annotations in the annotation list. For convenience, you can use the **Select All** or **Unselect All buttons**.
5. Click **View** to view the data you select in Heat Map Viewer. caIntegrator creates Heat Map Viewer files of the data.
6. After the files are created, click the **Launch Heat Map Viewer** hypertext link that appears.
7. Continue with step 3 on page 115.
8. For interpretation of the results and using HMV features, see the help files opened from HMV.

Java for IGV and Heat Map Viewer

To use the IGV and the NCI Heat Map viewer, you must install a version of Java containing Java Web Start. You must install recent versions of the Java Development Kit (JDK 1.5.0 aka JDK 5.0 or newer) or Java Runtime Environment (JRE 1.5.0 aka JRE 5.0 or newer). The easiest option is to install JRE 5.0. For more information, see: http://www.java.com/en/download/faq/java_webstart.xml.

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

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

CHAPTER

6

ADMINISTERING USER ACCOUNTS

This chapter describes the process for creating and managing user accounts in caIntegrator. It also discusses the processes for managing ownership and access to studies in caIntegrator.

Note: The options for performing user management tasks are visible in caIntegrator on the left sidebar of the browser only if you have these Admin privileges.

Administering caIntegrator User Accounts Using UPT

Note: If you are interested in registering an account in caIntegrator, see *Registering as a New caIntegrator User* on page 6.

In caIntegrator, all tasks related to creating and managing user accounts can be performed only by a caIntegrator administrator using the CBIIT User Provisioning Tool (UPT) v. 4.2. The following sections discuss the use of the UPT for performing these tasks. For further information about UPT, see Chapter 3 of the CSM 4.2 Programmer's Guide located here: https://gforge.nci.nih.gov/docman/view.php/12/18945/caCORE_CSM_v42_ProgrammersGuide.pdf

The UPT is a separately installed application which serves as the user management interface for all National Cancer Institute CBIIT Life Sciences Distribution (LSD) applications, including caIntegrator. The UPT application is the central point for all user management functionality within caIntegrator. You can use UPT to add new users and to apply user group assignments to the caIntegrator database directly. The UPT groups can refer to predefined groups such as Study Manager or Study Investigator, which determine what roles the user has.

The following terms are used both in this chapter and in the UPT to define user-related roles:

- **User** – a person who is accessing caIntegrator. The user has an associated account and user ID.

- **User Group** – a group of users, typically grouped by organization and role, for example, "Columbia University Study Managers"
- **Protection Group** – a group of studies given a secure status and typically grouped by organization, for example, "Columbia University Protected Studies".

Steps for Creating User Access to caIntegrator

The following steps summarize the process for establishing user access to caIntegrator:

1. A potential user requests a user account in caIntegrator. See *Registering as a New caIntegrator User* on page 6.
1. You, as a caIntegrator administrator, check if the **User** already exists in caIntegrator. If not, create the new user. See *Creating a New caIntegrator User* on page 120.
2. Check if the requestor's **User Group** already exists in caIntegrator. If not, create a new **User Group**. See *Creating a New User Group* on page 122.
3. Check if the **Protection Group** (e.g. "Columbia University Protected Studies"), containing the studies to which this user wants access currently exists. If not, create a new **Protection Group**. See *Creating a New Protection Group* on page 123.

   Note: If the Protection Group already exists, contact the Organizational Contact person to confirm that it is OK to give this person access to this Protection Group.
4. Give the requestor's **User Group** access to the **Protection Group**. See *Assigning a User Group to a Protection Group* on page 124.
5. Add the **User** to the **User Group**. See *Adding a User to a User Group* on page 127

Creating a New caIntegrator User

To create a new User in caIntegrator, follow these steps:

1. Login to UPT as a caIntegrator Admin.
2. First, search to see if the user already exists. Click the **User** menu option.
3. On the User page that opens, click **Select an Existing User**.

4. Use the form and search for the user. If you define no criteria, UPT returns a list of all caIntegrator users currently in the system (*Figure 6.1*).

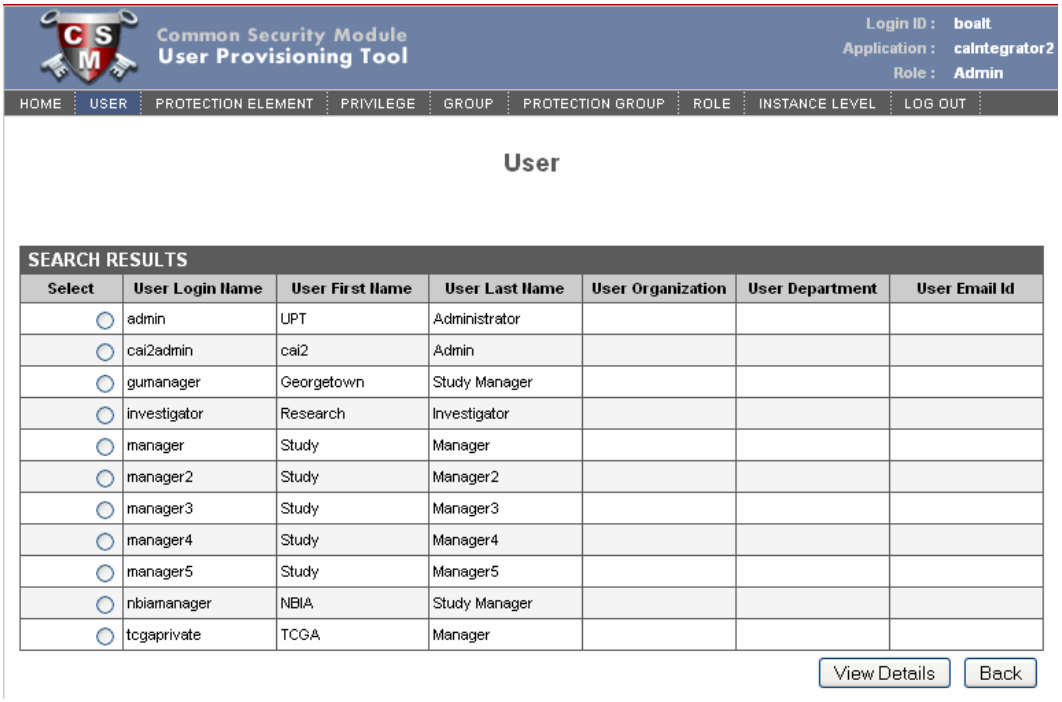


Figure 6.1 A list of current caIntegrator users displays in UPT after a user search

5. If the user does not already exist (is not listed in the search results), then create a new user. To do so, select the **User** menu option again, then click **Create a New User**.

This opens the page for creating a new caIntegrator user (*Figure 6.2*).

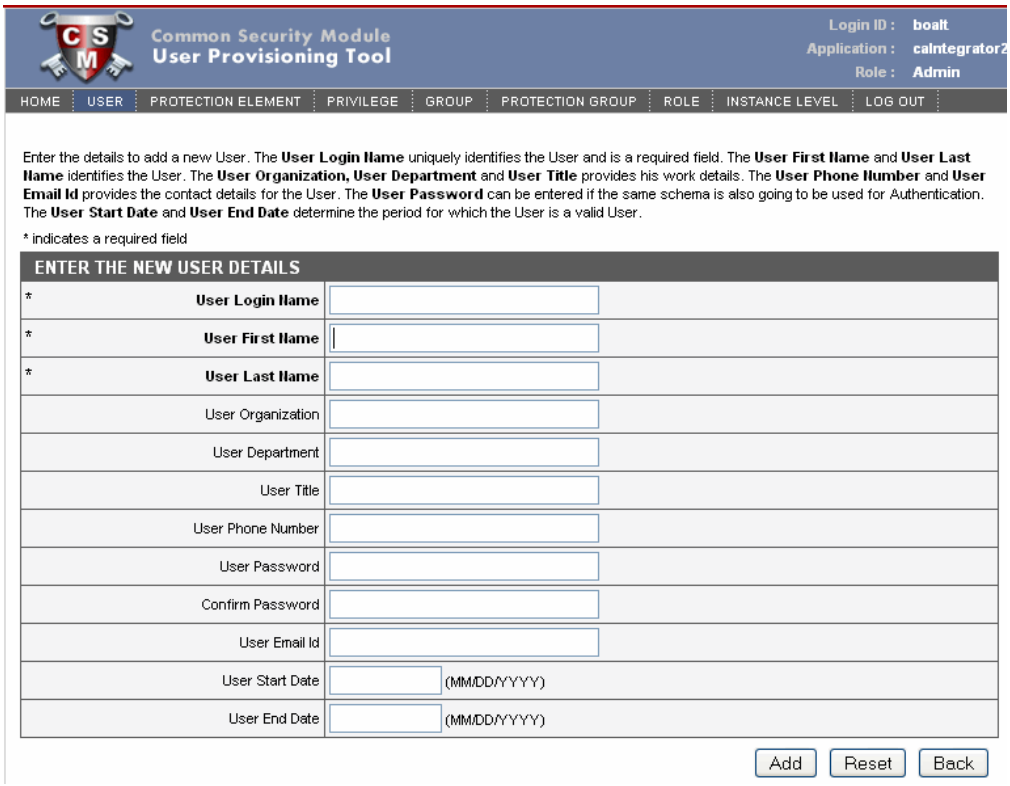


Figure 6.2 UPT page for creating new user details

6. Enter details for the following required fields:
   - **User Login Name**
   - **User First Name**
   - **User Last Name**
   - **User Password**

   Caution: If the requestor is an LDAP user, then the User Login Name must match the LDAP login ID AND the User Password field must be left blank. If the requestor is not an LDAP user, then provide a password.

   - **User Organization**
   - **User Department**
7. Click **Add** to confirm the new user.

Creating a New User Group

You can assign a user group to a protection group. The advantage of working with a user group is that you do not have to assign roles to each user individually. You can assign users to a user group to which you assign a role, and then assign that user group to the protection group, or you can assign a role collectively to a protection group after it is created.

To create a new User Group in caIntegrator, follow these steps:

1. Login to UPT as caIntegrator Admin.
2. First search for an existing group that the user wishes to join. Click the **Group** menu option.
3. On the Group page that opens, click **Select an Existing Group**.
4. Use the form and search for the group. If you define no criteria, UPT returns a list of all caIntegrator groups currently in the system
5. If a user group does not already exist, then create a new user group. Click the **Group** menu option, then click **Create a new Group**.
6. On the form that opens (*Figure 6.3*), enter a unique **Group Name** and a description, if appropriate. Click **Add**.

   Note: The recommended naming convention for a new User Group is ***[insert organization name] Study [insert role]s***. Example: "Columbia University Study Managers".

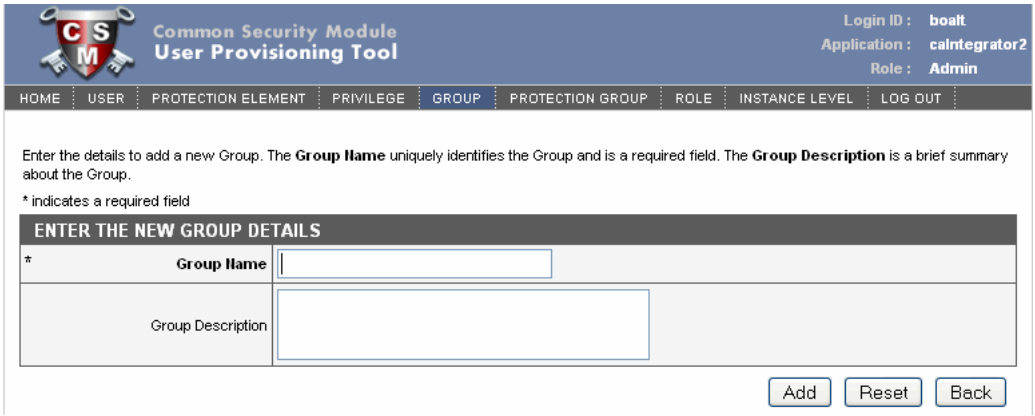


Figure 6.3 UPT page for creating a new group

Creating a New Protection Group

If you prefer that a study or group of studies have limited access, you can assign a user to a particular protection group and assign roles which allow the users in the protection group study access. A protection group provides security or limited access for studies listed there.

To create a new Protection Group in caIntegrator, follow these steps:

1. Login to UPT as caIntegrator Admin.
2. Click the **Protection Group** menu option.

3. On the page that opens, click **Create a New Protection Group**. The page opens for defining PG Group details (*Figure 6.4*).

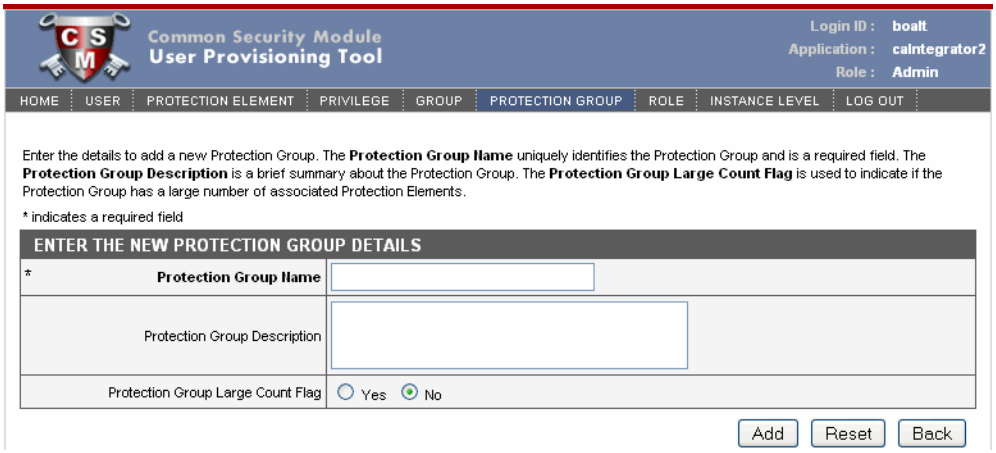


Figure 6.4 UPT page for creating a new protection group

4. Enter a unique **Protection Group Name** and Description, if appropriate. Click **Add**.

   Note: The recommended naming convention is ***[insert organization name here] Protected Studies***. Example: "Columbia University Protected Studies".

Assigning a User Group to a Protection Group

To give a User Group access to a Protection Group (a group of protected studies), follow these steps:

1. Login to UPT as caIntegrator Admin.
2. Find the user group that you want to assign. Click the **Group** menu option and click **Select an Existing Group**. In the page that opens, click **Search**. If you define no criteria, UPT returns a list of all caIntegrator groups currently in the system (*Figure 6.5*).

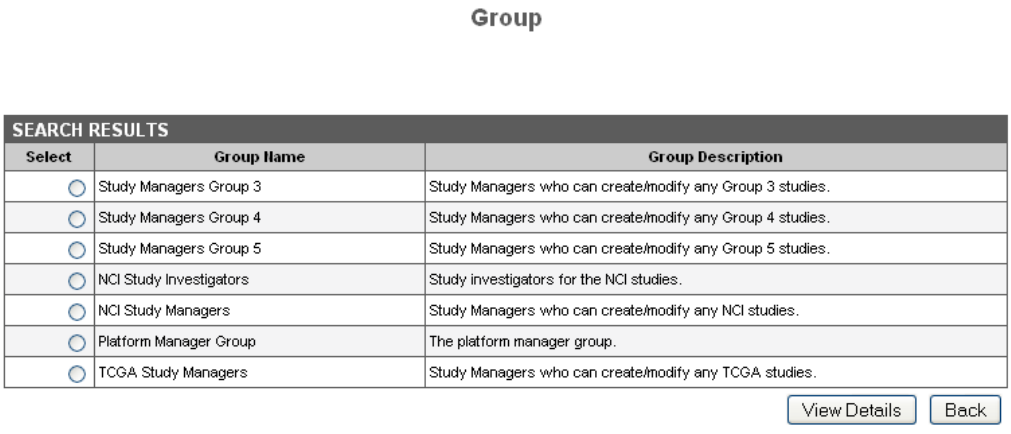


Figure 6.5 UPT page showing Group search results

3. Select the radio button next to the group name you want to assign to the Protection Group. Click **View Details**. This opens the Group Details page (*Figure 6.6*).

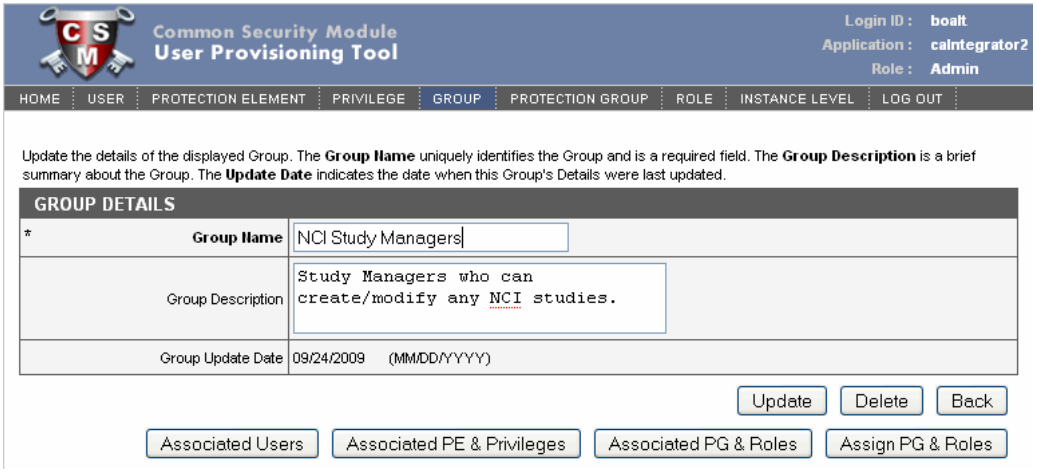


Figure 6.6 UPT page showing details for a selected group

4. Below the group details, click **Associated PG & Roles**. The page that opens displays any PG to which the user group is already assigned (*Figure 6.7*).

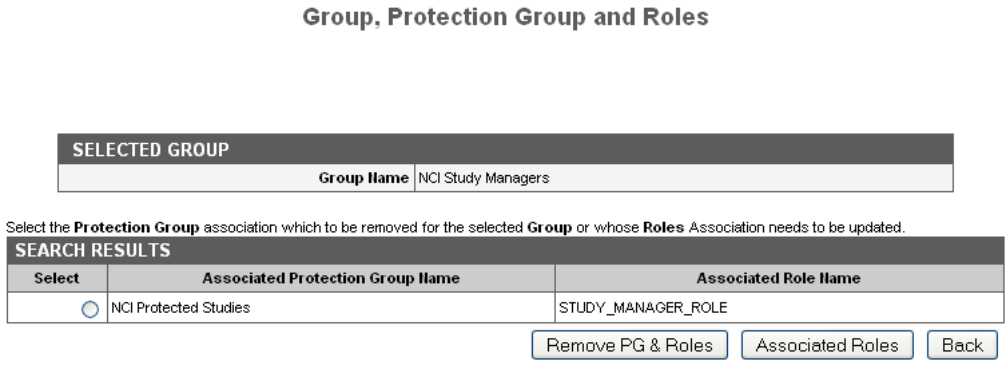


Figure 6.7 UPT page that shows any PGs to which the select user group is assigned

5. Below the group name, examine if the Protection Group of your choice is already listed there. If so, this means your user group is already assigned to the protection group of choice, and you can skip the remainder of the steps in this section. If the Protection Group is not listed there, then click **Back**.

6. Back on the User Group details page, click **Assign PG & Roles**. This opens the Group, Protection Group and Roles Association page (*Figure 6.8*).

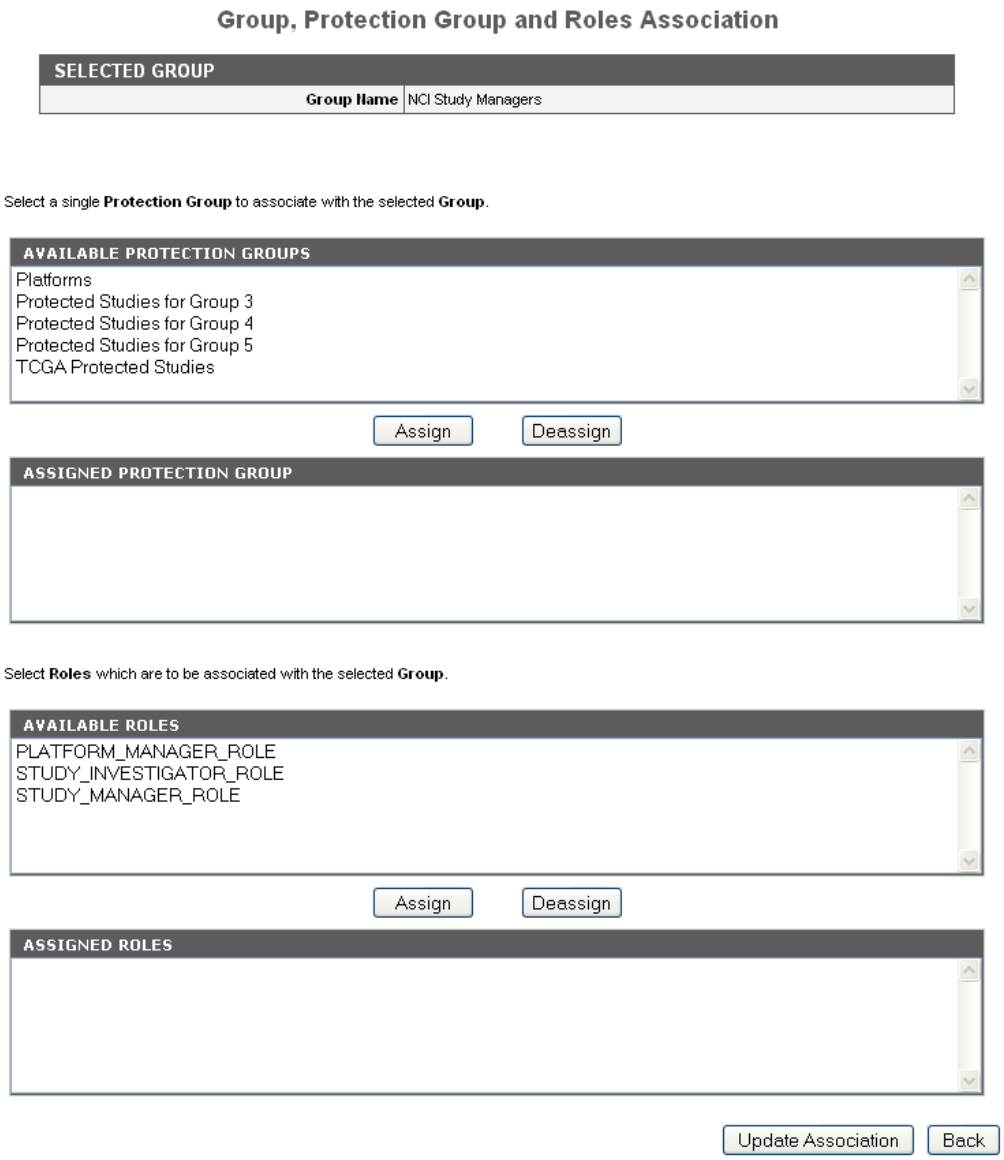


Figure 6.8 UPT page for assigning user group to a protection group and selected roles

7. From the list of Available Protection Groups, highlight your PG of choice and click **Assign**.

Now you can assign a role to the user. The caIntegrator Roles are defined in *Table 6.1*:

Role Name	***Role Definition***
STUDY_MANAGER_ ROLE	Assigning this role allows the user to modify existing studies, create new studies, and deploy existing studies.

Table 6.1 Names and definitions for caIntegrator roles

Role Name	*Role Definition*
STUDY_INVESTIGATOR_ROLE	Assigning this role allows the user to search the study, save queries about the study and perform analyses.
PLATFORM_MANAGER_ROLE	Assigning this role allows the user to create and delete array platforms for the entire caIntegrator installation. **Caution**: Array platforms are shared by all users and studies in the caIntegrator installation. A user with this role can affect the platforms that are used by by all users and studies in the caIntegrator installation.

Table 6.1 Names and definitions for caIntegrator roles

8. If this user group is a group of study managers, then select STUDY_MANAGER_ROLE. If this user group is a group of study investigators, then select STUDY_INVESTIGATOR_ROLE. Click **Assign**.
9. Click **Update Association** at the bottom of the page. This completes the assigning of the user group to the protection group you chose.

   Note: If a **User** has the STUDY_MANAGER_ROLE role for more than one **Protection Group**, then any study that the **User** creates will be assign to each of those **Protection Groups**.

Adding a User to a User Group

To add a user to an existing user group, follow these steps:

1. Login to UPT as caIntegrator Admin.
2. Find the user that you want to assign to a user group. Click the **User** menu option, then click **Select an Existing User**.

3. Enter the name of the user you are looking for and click **Search**. If you define no criteria, UPT returns a list of all caIntegrator users currently in the system (*Figure 6.10*).

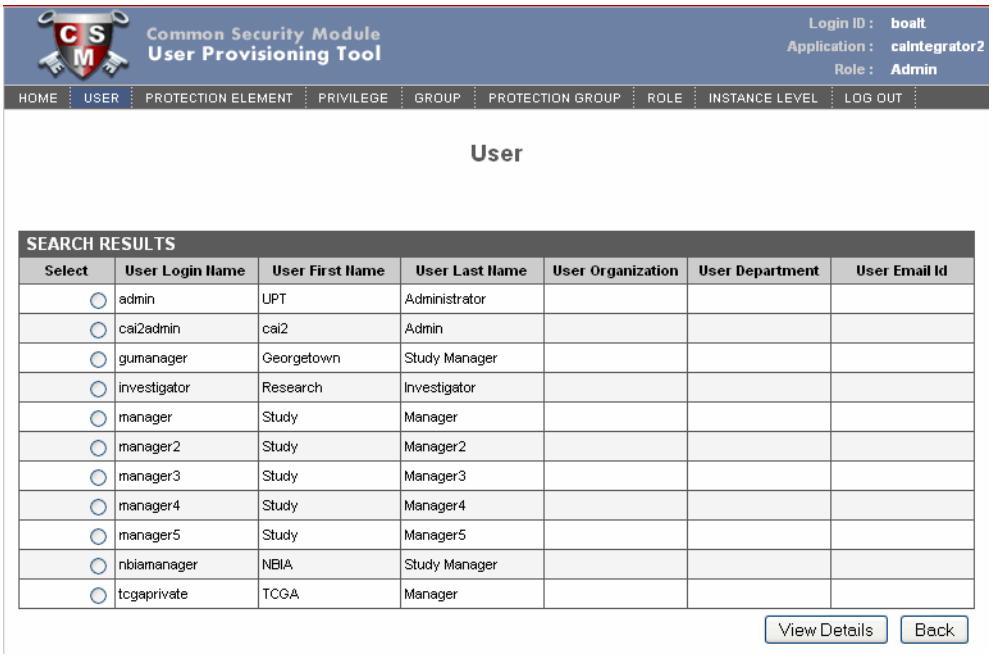


Figure 6.9 UPT page showing a list of caIntegrator users

4. Select the radio button next to the name and click **View Details** (*Figure 6.10*).

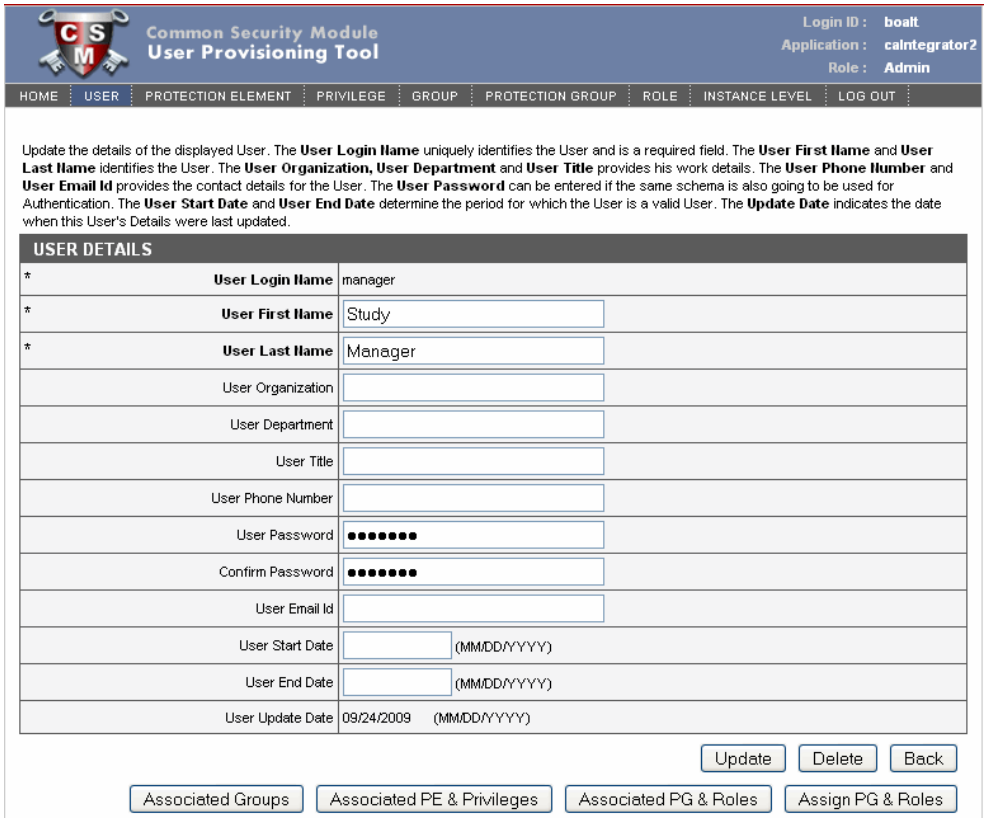


Figure 6.10 UPT page showing details for a selected user

5. Click the **Associated Groups** button at the bottom of the page. This opens the page where you can assign a user to a group (*Figure 6.11*).

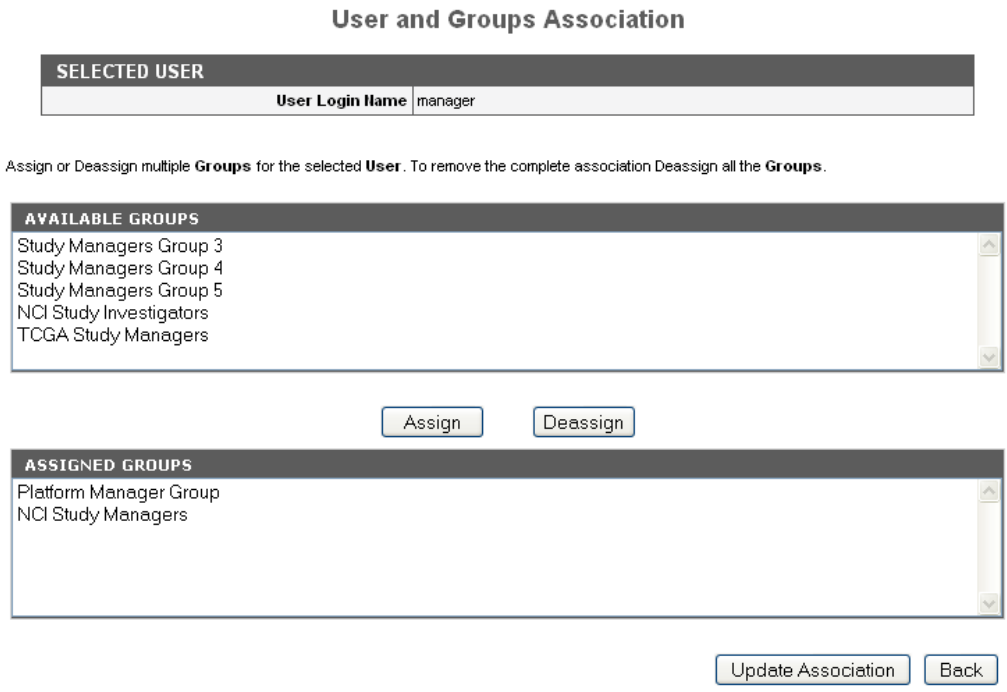


Figure 6.11 UPT page for assigning a user to user groups

6. Select the group(s) that you want the user to be in and click **Assign**.
7. At the bottom of the page click **Update Association**. This completes the assigning of the user to the user group. Now the user will have access to any studies to which the user group has been given access.

   Tip: You can add a user to more than one user group. For example, a user could be assigned to "Columbia University Study Managers" as well as to "Columbia University Study Investigators".

Changing a User Password

To change a password for a User, follow these steps:

1. Confirm if the User is an LDAP user or not. If the User is an LDAP user, then this person must change their password using the NCI password change utility. Skip the rest of these steps.
2. If the User is not an LDAP user, then continue with the rest of these steps.
3. Login to UPT as caIntegrator Admin.
4. Find the User that you want to change. Click the **User** menu option, then **Select an Existing User**.
5. Enter the name of the user you are looking for and click **Search**. If you define not criteria, UPT returns a list of all caIntegrator users.
6. Select the radio button next to the name and click **View Details**

7. Replace the **User Password** and **Confirm Password** fields with the new password (*Figure 6.12*).

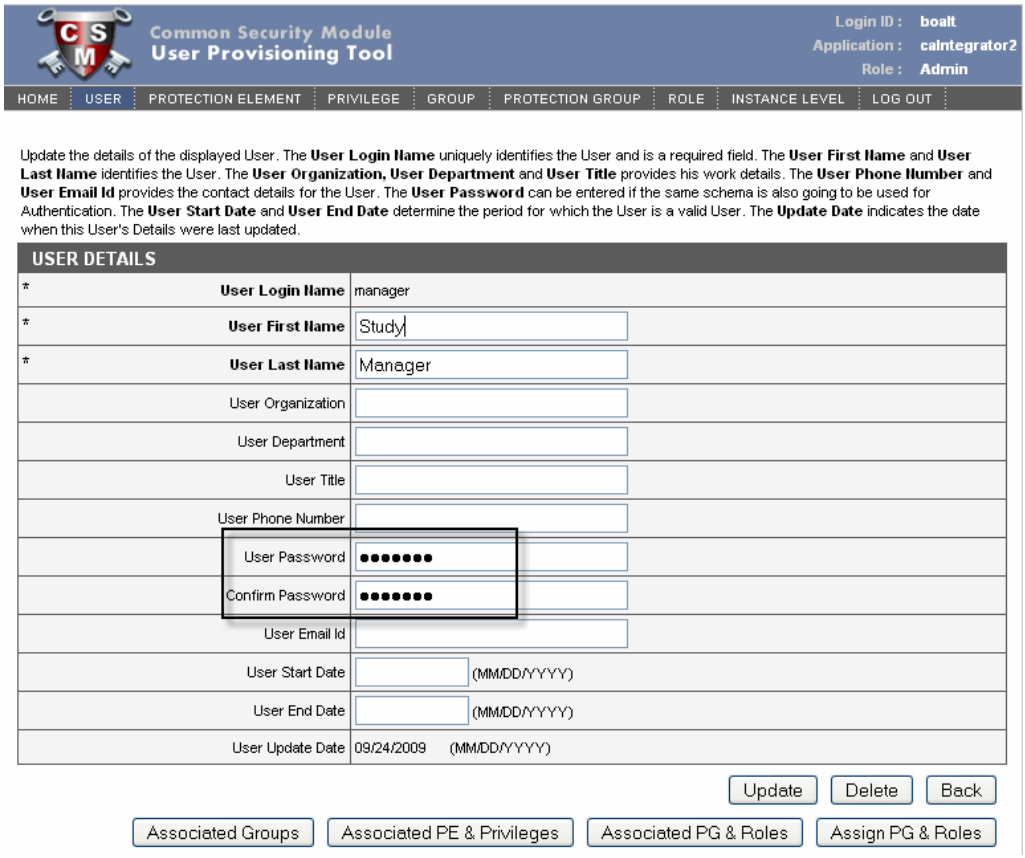


Figure 6.12 UPT page where you can edit user details, such as a password

8. At the bottom of the page click **Update**.

APPENDIX

A

DATA IMPORT CONFIGURATIONS

This appendix describes configurations for importing data into a study.

Topics in this appendix include the following:

Subject Annotation Data Configuration

The following subject annotation data configuration information is collected:

- Subject annotation 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 restricted to a list of groups. The default will be the visibility settings given at the study level. For more information, see *Define Fields Page for Editing Annotations* on page 21.
- Whether the field is to be included in the results list for a given entity type (i.e. Subject, Sample, Image Series or Array Data) when browsing data.
- Whether the field is to be included in simple single-input searches when browsing 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 caIntegrator sample identifiers to caArray 2 samples, specified as a comma-separated list in the format “caIntegrator sample identifier”, “caArray 2 experiment identifier”, “caArray 2 sample name”.

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

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

- 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.

See also *Supplemental Files Configuration*.

Supplemental Files Configuration

This section describes the format that must be used when creating supplemental files for use by caIntegrator. The supplemental files described here are to be added to an experiment in caArray prior to configuring a study in caIntegrator.

The file itself is a tab-delimited text file. The file extension can be anything, though users typically use `.txt`. The name of each supplemental file must be unique within a caArray experiment.

Inside the file, each row in the file contains the data from one reporter. Each column in the file must have a unique header name, that is, you cannot give two different columns the same column name.

There are two supported formats for these files: *Single Sample Format* and *Multiple Sample Format*.

Single Sample Format

- Minimum of two required columns
- One column must contain the reporter/probe name
- One column must contain the value be reported by the reporter
- The file can have additional columns, though other than reporter/probe name and value mentioned above, the rest will be ignored
- One single sample file for each sample in the experiment

An example of single sample format file is shown in *Figure A.2*

ProbeID	signal:Log2
A_14_P112718	0.01
A_16_P15000916	0.5166
A_16_P15001074	0.4965
A_16_P00000012	0.1553
A_16_P00000014	-0.5684
A_16_P00000017	NA
A_16_P00000021	-0.6415
A_16_P00000023	-0.6041
A_16_P00000027	-0.374
A_16_P00000033	-0.493
A_16_P35001586	-0.465
A_16_P15001533	0.1939
A_16_P00000060	-0.0576
A_16_P15001594	-0.0674
A_16_P00000082	0.1754
A_16_P00000090	0.0878
A_16_P00000099	-0.4532
A_16_P15001666	-0.1321
A_16_P00000112	0.8277
A_16_P00000114	0.1214
A_16_P00000127	0.8765
A_16_P00000136	0.1701
A_16_P00000140	-0.3096

Figure A.1 Example of single sample format file

Multiple Sample Format

- One column must contain the reporter/probe name.
- Each additional columns are the reporter values such that there is one column per sample.
- One multiple sample file for the whole experiment.

Note: Currently the multiple sample format is slower to load than the single sample format for platforms other than Agilent Copy Number. Future releases should show improvements in this performance.

An example of multiple sample format file is shown in *Figure A.2*.

Hybridization Ref			TCGA-09-0364-01A-02D-0357-04	TCGA-13-0723-01A-02D-0357-04	TCGA-13-0757-01A-01D-0357-04
ProbeID	Chr	Pos	signal:Log2	signal:Log2	signal:Log2
A_14_P112718	1	554268	0.01	0.2111	-0.2282
A_16_P15000916	1	554287	0.5166	1.2929	0.8352
A_16_P15001074	1	639581	0.4965	0.4769	0.8294
A_16_P00000012	1	736483	0.1553	0.3047	-0.0251
A_16_P00000014	1	742533	-0.5684	0.3057	-0.7369
A_16_P00000017	1	746956	NA	0.3308	-1.1949
A_16_P00000021	1	757922	-0.6415	-0.2729	-1.3004
A_16_P00000023	1	769590	-0.6041	0.4084	0.0886
A_16_P00000027	1	784458	-0.374	-0.2919	-0.7865
A_16_P00000033	1	792413	-0.493	-0.3545	-1.0858
A_16_P35001586	1	800905	-0.465	0.0274	-0.2939
A_16_P15001533	1	823964	0.1939	0.4682	-0.1703
A_16_P00000060	1	836543	-0.0576	0.3648	0.1564
A_16_P15001594	1	842726	-0.0674	0.4352	-0.0918
A_16_P00000082	1	847646	0.1754	0.4735	-0.2007
A_16_P00000090	1	853295	0.0878	0.3196	0.6272
A_16_P00000099	1	857406	-0.4532	-0.2435	-1.0838
A_16_P15001666	1	862519	-0.1321	0.0441	0.0934
A_16_P00000112	1	865691	0.8277	0.9984	0.5806
A_16_P00000114	1	868794	0.1214	0.1444	-0.0517
A_16_P00000127	1	875165	0.8765	1.0351	1.1245

Figure A.2 Example of a multiple sample format file

The following software programs create the supplemental data format used by caArray:

- Affymetrix Expression Console – This software produces supplemental files. In Expression Console, use the “Export Result” function to create these files. Note that when you use an algorithm other than MAS5 to normalize the data (for example using RMA or Plier), Expression Console automatically creates a […`summary.txt`] file that contains extra lines on top of the derived data results. The extra lines all start with a “#” to signify that it is a remark. These lines are ignored by caIntegrator parsing.
- Agilent GeneSpring GX – This software can export a results table in .txt format.

Imaging Data Configuration

The following imaging data configuration information is collected:

- NBIA grid server hostname (defaults to CBIIT instance)
- NBIA grid server port (defaults to CBIIT instance port)
- Protocol Id
- Mapping of NBIA Patients to subjects imported from subject annotation 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 enables the Study Manager to navigate easily to the selected caArray 2 instance.

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

GLOSSARY

This glossary defines acronyms, abbreviations, and terminology used in this guide.

Term	*Definition*
caBIG	cancer Biomedical Informatics Grid
caBIO	Cancer Bioinformatics Infrastructure Objects
caCORE	cancer Common Ontologic Representation Environment
caDSR	Cancer Data Standards Repository
caMOD	Cancer Models Database
CGH	Comparative Genomic Hybridization
EBI	European Bioinformatics Institute
EVS	Enterprise Vocabulary Services
MAGE 1.1	MAGE 1.1 is a widely-used microarray data standard or guideline
MAGE-ML software format	XML-based standard for representation of microarray data
MIAME 1.1	MIAME1.1. is a standard or guideline for the minimum amount of information required to make a microarray record useful to others.
MGED Ontology	MGED Ontology is a controlled vocabulary standard that concisely defines terms as they relate to Microarrays and caArray as a whole
MGED	Microarray Gene Expression Data Society
MMHCC	Mouse Models of Human Cancers Consortium
NCI	National Cancer Institute
NCICB	National Cancer Institute Center for Bioinformatics
URI	Uniform Resource Identifier
URL	Uniform Resource Locators
XML	Extensible Markup Language (http://www.w3.org/TR/REC-xml/) - XML is a subset of Standard Generalized Markup Language (SGML). Its goal is to enable generic SGML to be served, received, and processed on the Web in the way that is now possible with HTML. XML has been designed for ease of implementation and for interoperability with both SGML and HTML

INDEX

T

U

V