

User's Guide



June 30, 2011 v.1



Contents

Chapter 1: Getting Started with tranSMART 1
Feature Overview 1
Search Tool 2
Dataset Explorer 3
Sample Explorer 3
Gene Signature Wizard 3
Logging In 4
Tools 5
Opening a Particular Tool at Login 6

Chapter 2: Search Tool 7
Search Tasks 7
Defining a Search Filter 8
Building a Search String 13
Saving a Search Filter or Search String 16
Working With Search Results 18
TEA Analyses 46
TEA Indicators Applied to Individual Biomarkers 46
TEA Indicators Applied to an Analysis 47
TEA Indicators Applied to an Individual Gene 49

Chapter 3: Dataset Explorer 51
Overview of the UI 51
Using Dataset Explorer 53
Public and Private Studies 54
Selecting the Study 54
Populating the Study Groups 59
Generating Summary Statistics 65
Defining Points of Comparison 68
Printing or Saving the Contents of the Results/Analysis View 69
Copying Individual Charts in the Results/Analysis View 69
Generating Heat Maps 70
Types of Heat Maps 70
Interactive Heat Maps 74
Requirements for Generating Heat Maps 75
Instructions for Generating Heat Maps 79
Export Heat Map Data Points 93

Generating a Principal Component Analysis 93
Multi-Dimensional Projections 95
Generating a Survival Analysis 97
Hazard Ratio and Relative Risk 99
Generating a Haploview 100
Running the SNP Viewer 102
Examples 104
Viewing SNP Data 110
Running the Integrated Genomics Viewer 113
Examples 115
Viewing IGV Data 119
Change the default display 122
Hypothesis Testing Across Multiple Studies 125
Across-Trials Comparisons of Clinical Data 125
Across-Trials Comparisons of RBM Data 127
Viewing a Study 129
Asynchronous Operations 130
Jobs Tab 131
Viewing a Logged Job 132

Chapter 4: Sample Explorer 133
Select a Primary Search Filter 133
View and Refine Sample Search Results 136
Select and Remove Search Filters 137
Find Samples in the Biobank 137
Locate the Source of the Samples in Dataset Explorer 138
Display Sample Data in a Heat Map 141
Manage the Sample Search Result List 142

Chapter 5: Gene Signatures and Gene Lists 145
Creating a Gene Signature 145
Step 1. Adding the Genes to a Text File 145
Step 2. Creating the Gene Signature 149
Performing Actions on Your Gene Signatures 153
Performing Actions on Other Users' Signatures 154
Viewing a Gene Signature Definition 155

Chapter 6: Other Tasks 157
Other Tasks 157

Appendix A: How TEA Scores Are Calculated 159
Data Inputs to the TEA algorithm 159
Operations 159
Result 161
Appendix B: Rules for Loading OmicSoft Data 163
Rules for Loading OmicSoft Data 163

Chapter 1

Getting Started with tranSMART

The tranSMART application reflects the efforts of Johnson & Johnson's informatics group to integrate data from internal and external data sources within a single data warehouse, and to provide scientific end users the tools to search for, view, and analyze the data in the warehouse.

The core internal data is a historical base of immunology trials with associated biomarker data from gene expression, RBM, and SNP experiments, including both raw and analyzed data, as well as oncology datasets from internal collaborations. Additionally, tools are provided to search through Johnson & Johnson's other data repositories, including Hydra, SharePoint, and some of Beerse's data analysis pipeline.

External data sources include publicly available resources such as the Gene Expression Omnibus repository, Dana-Farber Cancer Institute GeneChip Oncology Database, and MeSH Ontology.

The tranSMART application presents scientists with a search tool to query this vast ocean of disparate data through a Google-like user interface.

As users become more sophisticated in developing patterns of search terms, they can save and share their efforts with fellow researchers. A second tool, called the Dataset Explorer, allows the properly authorized user to create and study cohorts of patients that have been involved in completed clinical research efforts. Dataset Explorer also provides the user with tools to compare an individual (or group) in one study against a person or cohort in another study.

> **Note:** There may be some minor differences between the UI objects illustrated in this guide and the ones you see on your screen.

Feature Overview

tranSMART contains the following major features:

- Search tool
- Dataset Explorer
- Sample Explorer
- Gene Signature Wizard

Search Tool

tranSMART provides a Google-like search tool that lets you search across multiple data sources for information related to items of interest, such as biomarkers, compounds, diseases, genes, gene signatures, and clinical trials.

The scope of a search includes Johnson & Johnson clinical studies, externally conducted experiments, in vivo/in vitro studies, Dana Farber's published GCOD dataset, Johnson & Johnson document repositories, curated published literature, and some Johnson & Johnson internal tools.

Search tool functionality includes:

- Searching within a particular category, such as diseases, genes, or compounds, or searching across all categories.
- Building complex search criteria that let you precisely define what to search for.
- Saving search criteria for easy recall and re-execution.
- Emailing search criteria to colleagues.

Search Results

In searches of clinical trials and experiments, tranSMART displays complete listings of all analyses related to the trials and experiments that are found.

tranSMART flags "meaningful" results in the analysis lists. Meaningful analyses are those where the signature genes are differentially modulated in a statistically significant way, indicating that the associated target or pathway is probably affected by the treatment, disease or other topic examined in the experiment.

Search result functionality includes:

- Displaying details of a particular clinical trial or experiment by clicking the name of the trial or experiment in the results list.
- Accessing a number of gene-related sites – Entrez Gene, Entrez Global, Pictor, Hydra, GeneCards, and Google Scholar – by clicking the name of a gene in the results list.
- Viewing the technical report or protocol used for an analysis.
- Viewing clinical trial search results in a heat map.
- Exporting the complete results list to a Microsoft Excel file.
- Exporting details of a particular study, experiment, or other result to a Microsoft Excel file.
- Exporting results of gene, gene pathway, or gene signature searches to Johnson & Johnson's Pictor site.
- Exporting results of gene, gene pathway, gene signature, and curated literature searches to Ariadne Genomics Pathway Studio.

Dataset Explorer

Dataset Explorer is an i2b2-based tool that lets you compare two sets of study groups based on one or more points of comparison. You define both the criteria that populate the study groups and the points of comparison between the study groups.

Dataset Explorer leverages the familiar navigation tree interface of Microsoft Windows Explorer to display data from clinical trials, and also leverages intuitive drag-and-drop functionality to help you build the criteria for populating the study groups and to add the points of comparison.

Dataset Explorer functionality includes:

- Saving the criteria used to populate the study groups.
- Emailing the study group criteria to colleagues.
- Using a heat map to visualize the change in the expression of a specific protein from one sample to another.
- Using principal component analysis (PCA) to reduce the dimensionality of the dataset and to identify new, meaningful variables in the dataset.
- Using a haploview tool to analyze the differences in allele frequency in two or more loci from one sample to the next.

Sample Explorer

Sample Explorer lets you search for datasets of tested tissue and blood samples, within categories such as tissue type, pathology, and test type (such as gene expression or SNP). You can also search only for samples that are included in the Biobank.

Once you find samples of interest, you can perform tasks such as:

- Displaying data from one or more sample datasets in a visualization such as a heat map.
- Linking back to the Dataset Explorer study for which the samples were collected.
- Viewing reference information to help you locate samples in the Biobank.

Gene Signature Wizard

tranSMART provides a wizard to help you create and define gene signatures and gene lists.

You can use your gene signature or gene list in tranSMART searches to find studies where the differentially regulated genes match those in your gene signature or list. This can help you develop hypotheses about diseases or treatments that may have similar genes disregulated.

Gene signature functionality includes:

- Keeping the gene signature or list private so that only you can access it and use it in searches, or making it publicly available to all tranSMART users.
- Cloning an existing gene signature or list – either yours or a public one – as the starting point for creating and defining a new gene signature or list.
- Exporting all details of a gene signature or list to a Microsoft Excel file.

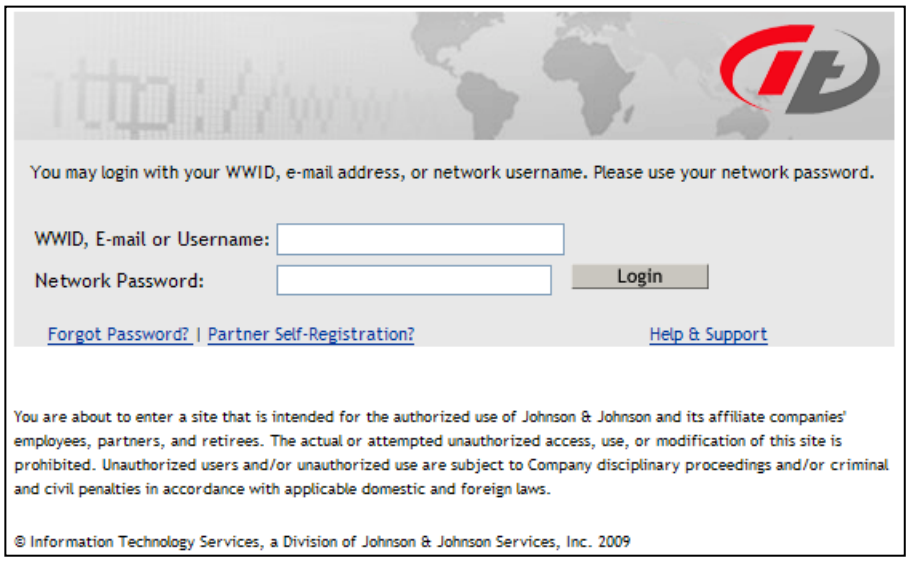
Logging In

To log into tranSMART:

1. Type the address of the tranSMART software into your browser's URL field:

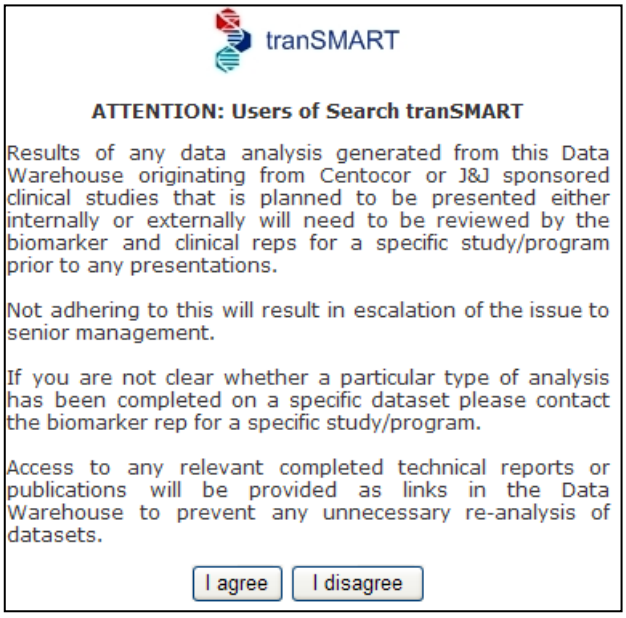
   https://transmart.jnj.com/transmart

   The login screen appears:



2. Type your tranSMART login credentials, then click **Login**.

The following security note appears:



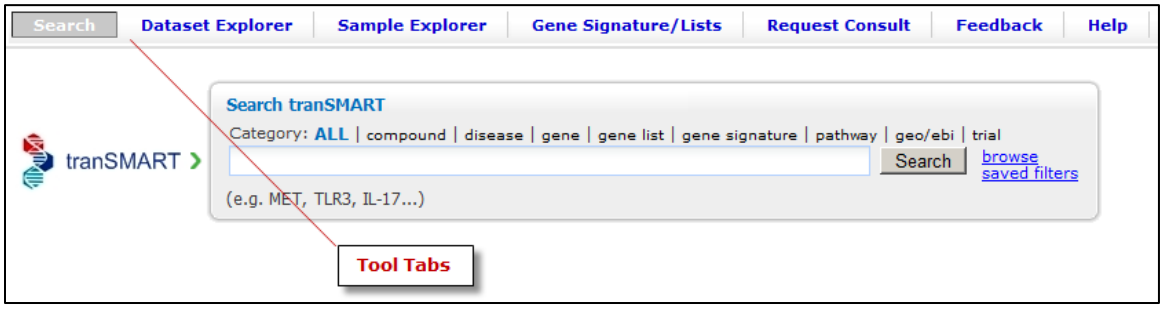
3. Read the data security note. If you agree, click **I agree**.

Tools

tranSMART provides the following tools:

- **Search** – Search across internal and external data sources for research data and literature related to search terms that you provide.
- **Dataset Explorer –** View study data for subjects that you select, based on criteria that you specify. Also, compare data generated for subjects in two different study groups, based on criteria and points of comparison that you specify.
- **Sample Explorer –** Search for test samples using pre-defined search filters such as tissue, pathology, and dataset.
- **Gene Signature/Lists –** View definitions of existing gene signatures and add new gene signature definitions.
- **Request Consult –** Email a request for information that cannot be found in the data warehouse.
- **Feedback –** Email questions, problem reports, enhancement requests, or any other feedback about the tranSMART application.
- **Help –** Display links to the tranSMART documentation set.

Select the tranSMART tool to use by clicking one of the tool tabs at the top of the tranSMART window:



Opening a Particular Tool at Login

By default, tranSMART opens the Search tool after you log in. However, you can specify the tool for tranSMART to open immediately after login by including the tool name in the address you type into your browser's URL field.

To automatically open a particular tranSMART tool immediately after login, use an address listed below:

> **Note:** The addresses below are case sensitive.

- Search tool – either of the following:

  https://transmart.jnj.com/transmart

  https://transmart.jnj.com/transmart/search

- Dataset Explorer tool

  https://transmart.jnj.com/transmart/datasetExplorer

- Sample Explorer tool

  https://transmart.jnj.com/transmart/sampleExplorer

- Gene Signature/Lists tool

  https://transmart.jnj.com/transmart/geneSignature

Chapter 2
Search Tool

A tranSMART search returns information in one or more of the following result categories:

- **Clinical Trials** – Internal Johnson & Johnson clinical trials.
- **mRNA Analysis** – Public or internal mRNA experiments.
- **mRNA Profiles** – Gene expression profiles from public cancer datasets curated by the Dana Farber Cancer Institute GCOD database.
- **Literature** – Curated documents within broad subject matter categories such as oncology and asthma.
- **Documents** – Documents from Johnson & Johnson's document repositories.
- **Pictor** – Johnson & Johnson's interface to pathway databases such as KEGG, GeneGO, and Ingenuity.
- **ResNet** – Ariadne Genomics functionality for performing search and pathway operations.
- **GeneGo** – Direct access to the EZ Search functionality of GeneGo's MetaCore.

Search Tasks

tranSMART provides a Google-like interface for searching across internal Johnson & Johnson data sources as well as external data sources with a single query, based on one or more search filters that you define.

A search filter is one of the following:

- The name of a biomedical concept such as a gene, pathway, compound, disease, or other item of medical interest. These filter names are pre-defined in tranSMART. You can browse lists of these filter names and select the filter you want, or type part of a filter name in the **Search** field, causing tranSMART to display a list of filters that begin with the text you type.
- Any text that you type into the **Search** field and then click the **Search** button or press the **Enter** key. By clicking **Search** or pressing **Enter** rather than selecting from a list of pre-defined filter names, you instruct tranSMART to limit its search to internal Johnson & Johnson document repositories, looking for the exact text you typed (not case-sensitive).

You can base your search on a single search filter or on a multi-filter search string.

Defining a Search Filter

There are several ways to define a search filter:

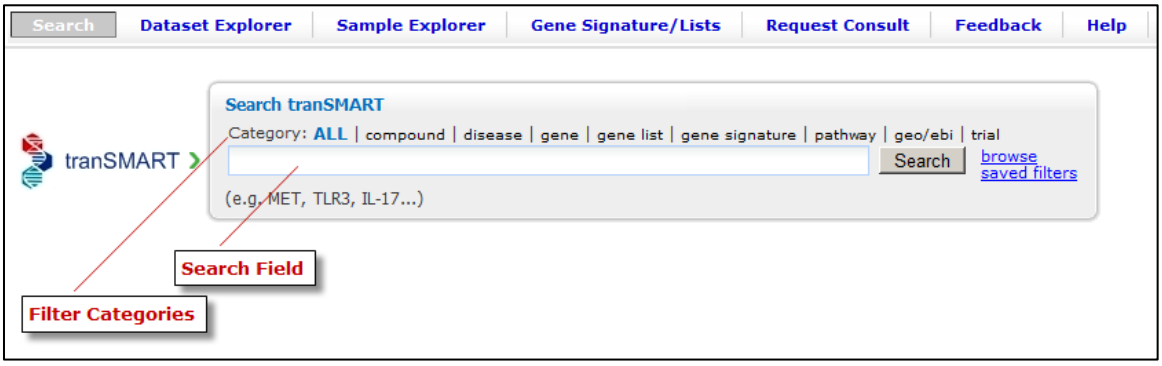
- Type part or all of the filter name directly into the **Search** field.
- Browse all the pre-defined filters within filter categories (such as clinical trials or diseases).
- Use a saved search filter or search string.

Type the Filter Name

To search the internal and external data sources for information related to a filter name:

1. Click the tab for the *Search* tool at the top of the tranSMART window.
2. Click the search filter category to search within (for example, search only compounds, or search only genes).

   The search engine first filters by the filter category you select, and then filters by the name you type. To search across all filter categories, click **all**.



> **Note:** You can only specify one search filter in the **Search** field shown above. For instructions on creating a multi-filter search string, see Building a Search String on page 13.

3. Type part or all of the filter name into the **Search** field.

Up to 20 matches that begin with the text you type are displayed in a dropdown list below the **Search** field. For example, the following list appears for the search filter **nev** when searching across all filter categories:



Note: You can also search for aliases. For example, to find the gene PTK7, you can type part or all of the name PTK7 or its alias, CCK4.

4. Do one of the following:

   - If the name of the filter you want appears in the list, click the filter name. The search begins immediately.
   - If the filter name you want does not appear in the list, type a more complete name in the **Search** field. For example, if you typed only **ne** in the **Search** field, no entries for "nevus" or "nevi" appear in the list. Typing a **v** after the text you already typed displays a list like the one shown above.
   - If no list appears after you type a complete filter name, click the **Search** button. It is possible that documentation related to the name you typed exists in Johnson & Johnson's document repositories.

   Note: When you click the **Search** button, tranSMART searches Johnson & Johnson's document repositories for the exact text you typed in the **Search** field. Wildcard characters are not supported, nor will tranSMART search for words and phrases that begin with the text you typed (for example, typing **nev** does not return **nevus**, **nevi**, or any other words and phrases beginning with **nev**.

5. To start another search using a new search filter, click **clear all** above the search result:



Alternatively, you can click the tranSMART logo, or simply type a new filter in the **Search** field.

See Working With Search Results on page 18 for information on viewing and refining search results.

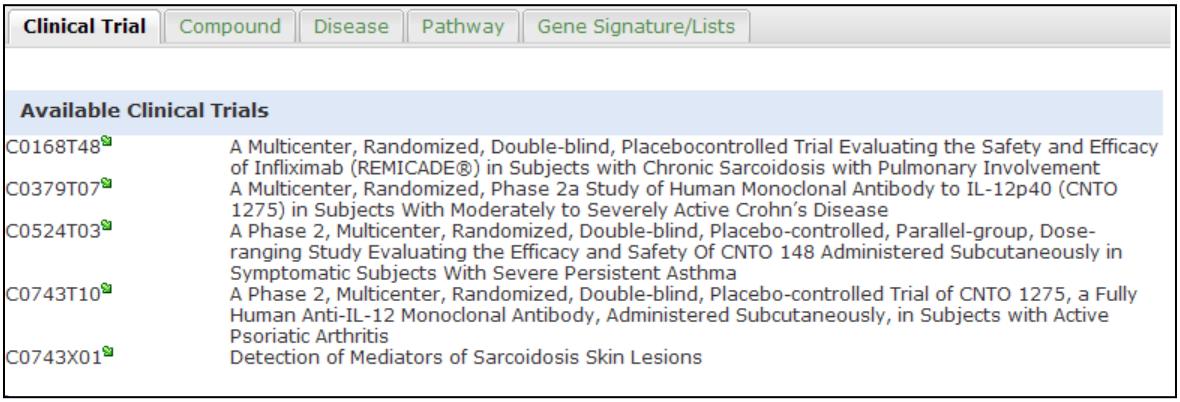
Browse for a Filter Name

You can browse through all the pre-defined filters in each of the following areas:

- Clinical Trial
- Compound
- Disease
- Pathway
- Gene Signature/Lists

To browse the pre-defined filters:

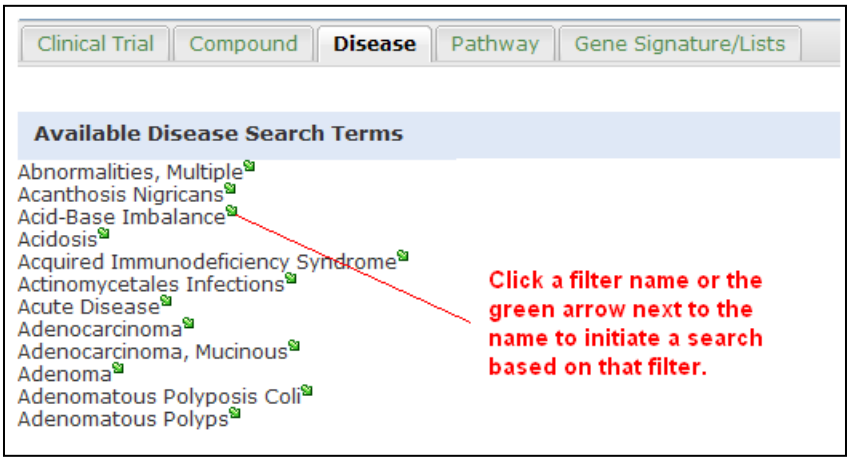
1. Click the tab for the Search tool at the top of the tranSMART window.
2. Click the **browse** link to the right of the **Search** button. A window similar to the following appears:



Note: The search engine ignores any filter category you may have selected and any filter text you may have entered in the **Search** field.

3. Click the tab for the area in which you want to browse for filters.

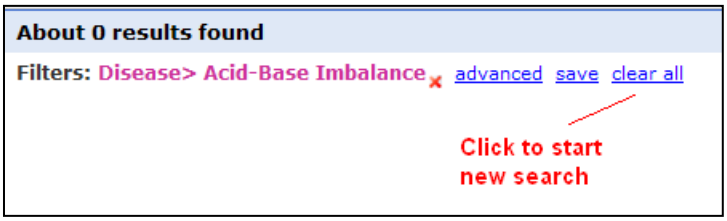
4. To initiate a search for information related to a filter, click the filter name or the green arrow after the name:



   After you click a filter, the search begins immediately.

5. To browse for another filter. click **browse** again. There is no need to clear the previous result before clicking **browse**.

6. To start another search using a new search filter, click **clear all** above the search result:



   Alternatively, you can click the tranSMART logo, or simply type a new filter in the **Search** field.

See Working With Search Results on page 18 for information on viewing and refining search results.

Use a Saved Search Filter

There are two ways to access a saved search filter:

- Retrieve the saved filter from a list of filters that you created and saved. The instructions in this section describe this method.
- Click a link to a saved filter that someone else has created, saved, and emailed to you.

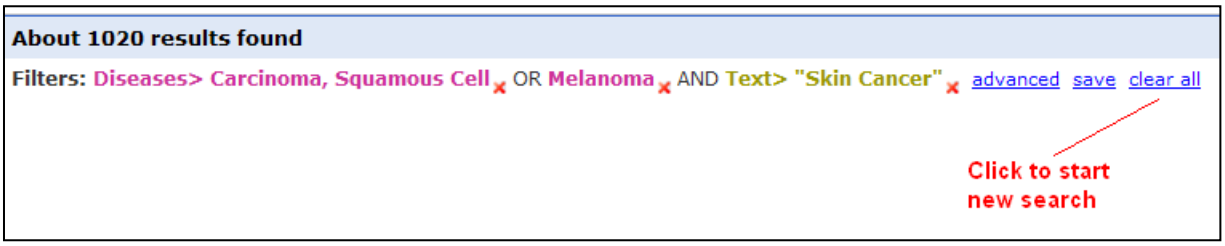
See Saving a Search Filter or Search String on page 16 for more information, including instructions on saving search filters and search strings.

To search against a filter that you created and saved:

1. Click the tab for the Search tool at the top of the tranSMART window.
2. Click the **saved filters** link to the right of the **Search** button. A list of filters that you created and saved appears:



3. To search against a saved filter in the list, click the **select** link to the right of the saved filter name. The search begins immediately.
4. To start another search using a new search filter, click **clear all** above the search result.



Alternatively, you can click the tranSMART logo, or simply type a new filter in the **Search** field.

See Working With Search Results on page 18 for information on viewing and refining search results.

Building a Search String

You can make the scope of your search more precise by building a multi-filter search string. The filters in a search string are joined by the logical operators `AND` and `OR`.

Rules for Building a Search String

The following rules apply to building a multi-filter search string:

- Filters within the same filter category (such as compounds, diseases, and genes) are joined by the logical operator `OR`.

  For example, if you add the filters `Diseases> Melanoma` and `Diseases> Melanoma, Experimental` to a search string, the search engine evaluates them as in the following expression:

  ```
  (Diseases> Melanoma OR Diseases> Melanoma, Experimental)
  ```

- Filters within different filter categories are joined by the logical operator `AND`.

  For example, if you add the filters `Diseases> Melanoma`, `Diseases> Melanoma, Experimental`, and `Pathway> KEGG-Melanoma` to a search string, the search engine evaluates them as in the following expression:

  ```
  (Diseases> Melanoma OR Diseases> Melanoma, Experimental) AND
  Pathway> KEGG-Melanoma
  ```

- Filters that are not among the pre-defined filters are assigned to the filter category **Text>**.

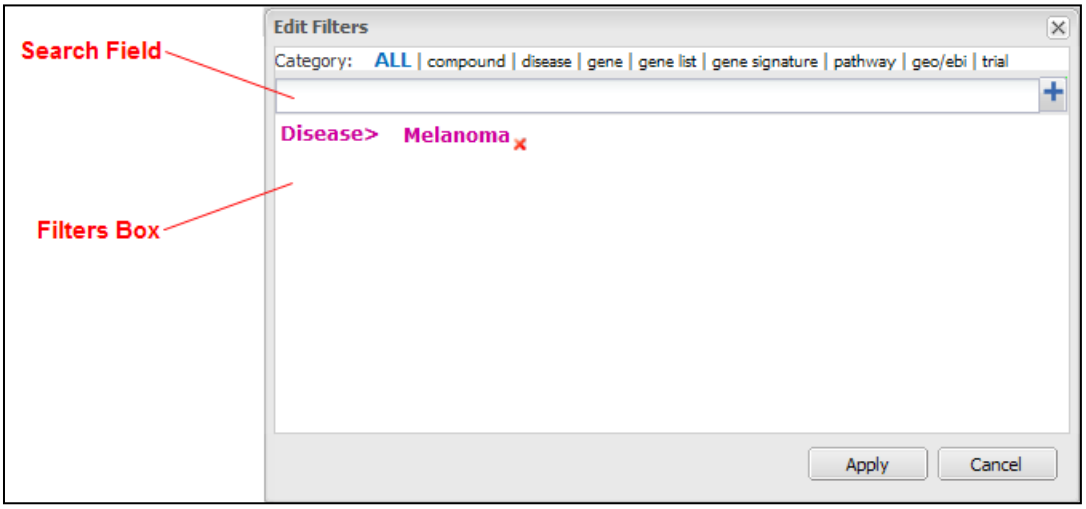
Instructions for Building a Search String

To build a multi-filter search string:

1. Define a search filter using any of the methods described in Defining a Search Filter on page 8.
2. When the results window appears, click **advanced**:

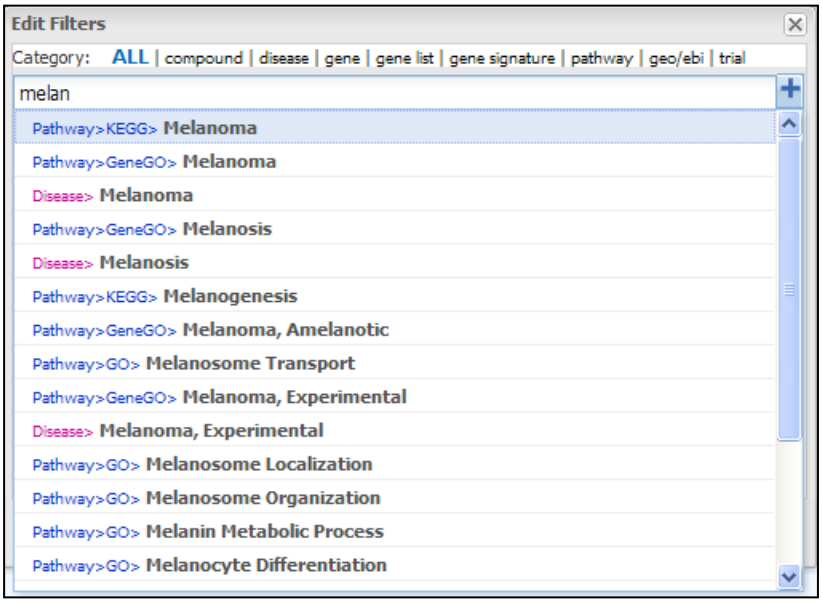


The Edit Filters dialog appears, displaying the filter you just created:



3. To add another filter, type part or all of a filter name into the **Search** field.

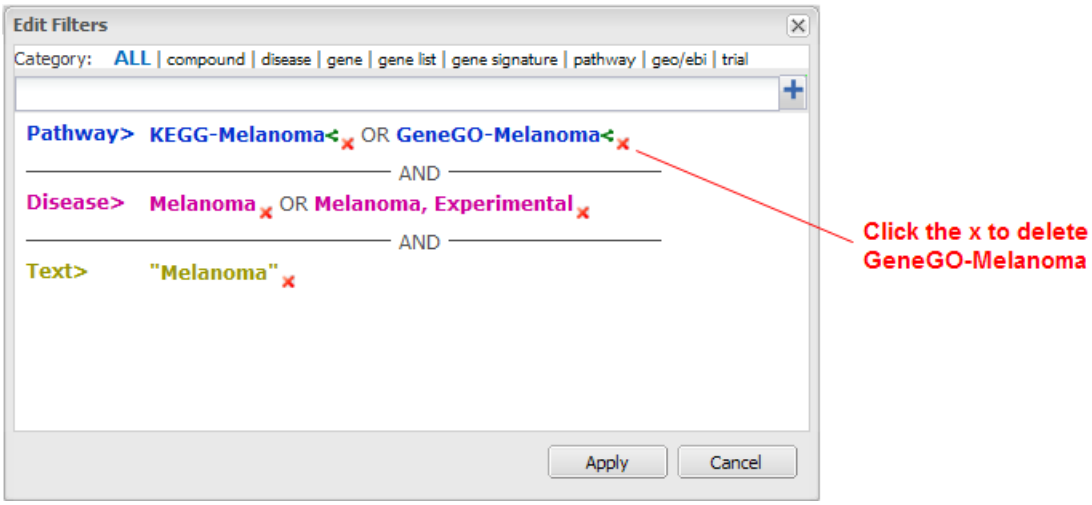
   Up to 20 matches for the text you type are displayed in a dropdown list below the **Search** field. For example, the following list appears for the search filter **melan**:



   Do one of the following:

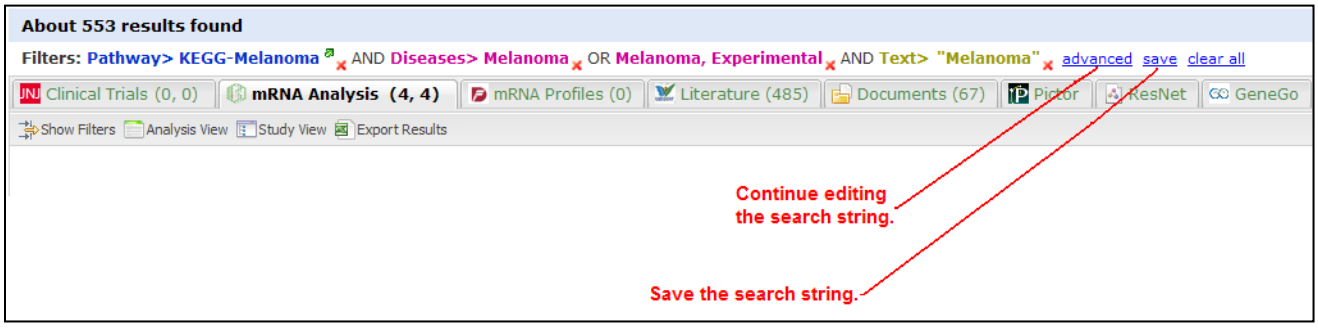
   - If the name of the filter you want appears in the list, click the filter name. The tranSMART software inserts the filter into the **Filters Box**.
   - If the filter you want does not appear in the list, type a more complete name in the **Search** field.
   - If no list appears after you type a complete filter name, or if you want to search Johnson & Johnson's documentation repositories for the text you typed, click the plus-sign button (+) to the right of the **Search** field.

4. Repeat the previous step for each new filter to add to the search string.
5. Optionally, to delete a filter from the search string, click the red x (✘) to the right of the filter name:



Note: If you click the green Split icon (⋖) after a pathway name, the pathway name is replaced by a list of all the genes in the pathway, each with a red x (✘) after the gene symbol. This allows you to delete one or more of the pathway's genes from the search filter.

6. When finished defining the search string, click **Apply** to begin the search.
7. When the results window appears, you can continue editing the search string or save it, as follows:
   - To continue editing the search string, click **advanced**.
   - To save the search string, click **save**.



The search engine evaluates this search string as in the following expression:

```
Pathway> KEGG-Melanoma AND (Diseases> Melanoma OR Diseases>
Melanoma, Experimental) AND Text> Melanoma
```

See Saving a Search Filter or Search String on page 16 for more information about saving search filters and search strings.

Saving a Search Filter or Search String

To save a search filter or search string:

1. After defining the search filter or search string, run the search and click **save** in the results window:

   The Create Filter window appears:



2. In the **Name** field, type a name for the search filter or search string.

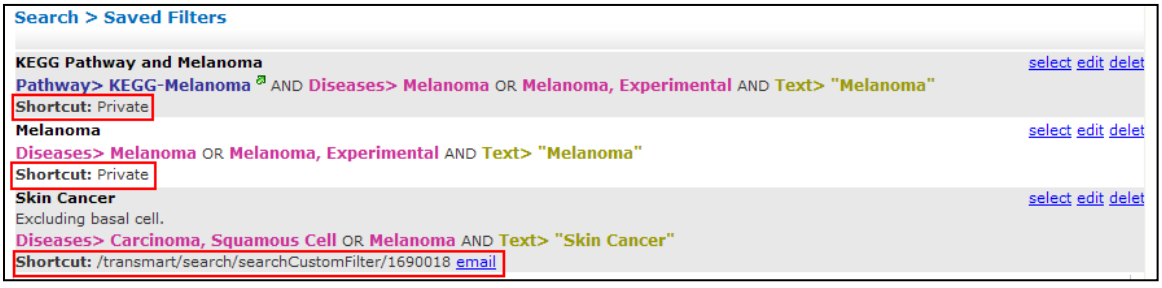
3. Optionally, in the **Description** field, type a description of the search filter or search string. In the saved filters list, the description appears immediately below the name of the search filter or search string.

4. Check the **Private Flag** checkbox to prevent others from using this search filter or search string, or clear the checkbox to allow others to use the search filter or search string.

   If a filter is public, a shortcut (link) to the filter is displayed in the **saved filters** list, and an **email** link is provided, allowing you to email the shortcut to others. If a filter is private, the saved filter is marked "Private," and the filter shortcut and **email** link are not displayed.

   Note: Only the person who created and saved a search filter can see that filter in the saved filter list. To let a colleague use a search filter you saved, you must (1) mark the filter as Public, and (2) click the **email** link to send the shortcut for the search filter to the colleague.

   In the following **Saved Filters** list, the first two entries are private and the third is public:

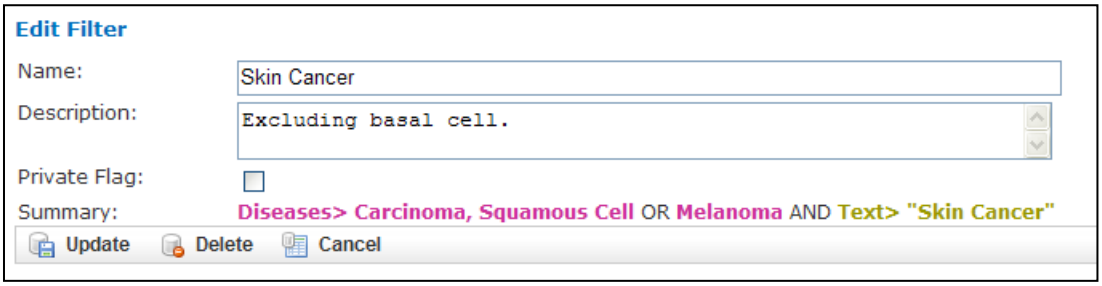


5. When finished, click **Create** to save the new search filter or search string, or click **Cancel** to abandon it.

Editing and Deleting Saved Filters

To edit a saved filter:

1. Click the tab for the Search tool at the top of the tranSMART window.
2. Click the **saved filters** link to the right of the **Search** button. A list of your saved search filters appears.
3. Click the **edit** link to the right of the saved filter name. The Edit Filter window appears:



4. Make one or more of the following changes:
   - In the **Name** field, modify the name of the saved filter.
   - In the **Description** field, add or modify an optional description of the saved filter. In the **saved filters** list, the description appears immediately below the saved filter name.
   - Check the **Private Flag** checkbox to prevent others from using this saved filter, or clear the checkbox to allow others to use the saved filter.

     Another user can use a filter you created and saved only (1) if the filter is public, and (2) you email the user the shortcut (link) to the filter.
   - To delete the filter you are editing, click the **Delete** button (Delete).

     Note: These are the only changes you can make to a saved filter. To make changes to the filter itself, run the search against the filter, then click **advanced** to define a new search filter based on the existing one. For details, see Instructions for Building a Search String on page 13.
5. When finished making changes, click the **Update** button to save your changes, or click the **Cancel** button to abandon them.

To delete a saved filter from the saved filters list:

1. Click the tab for the Search tool at the top of the tranSMART window.
2. Click the **saved filters** link to the right of the **Search** button. A list of saved search filters appears.
3. Click the **delete** link to the right of the saved filter name.

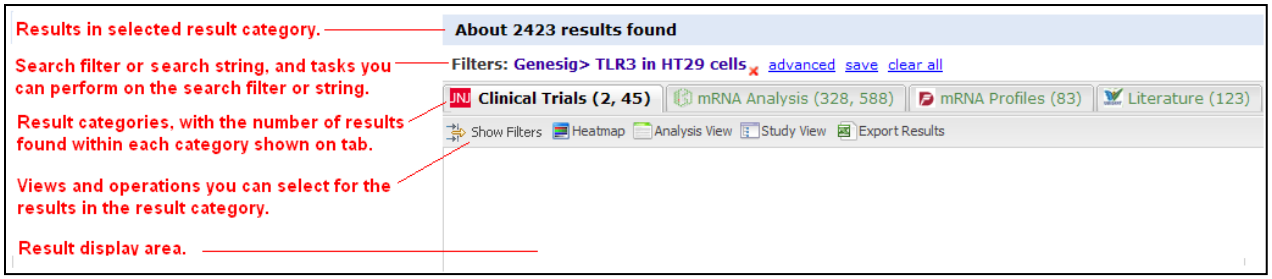
Working With Search Results

The results window displays all the clinical, documentary, and other information found in the data warehouse that relates to the search filter or search string.

The content of the results window varies, depending on the result category you select (for example, Clinical Trials or mRNA Analysis) and the type of view you want to use to display the results (for example, Heatmap, Analysis View, Study View). Some result categories also let you further refine the results by adding more filters to the search.

To select a result category to view, click the tab that contains the result category name.

The following figure shows the sections of the results window:



The number of results in a given result category is displayed on the result category's tab. For example, in the preceding figure, 83 mRNA profiles were returned.

The tabs for the result categories Clinical Trials and mRNA Analysis display pairs of numbers. These numbers represent the following results:

- **Clinical Trials (x, y)**
  - x = the number of statistically significant analyses. These hits can be viewed in the Analysis View.
  - y = the total number of analyses. These hits can be viewed in the Study View.

  For example, in the preceding figure, 2 statistically significant analyses were returned, and a total of 45 analyses were returned.
- **mRNA Analysis (x, y)**
  - x = the number of statistically significant analyses. These hits can be viewed in the Analysis View.
  - y = the total number of analyses. These hits can be viewed in the Study View.

  For example, in the preceding figure, 328 statistically significant analyses were returned, and a total of 588 analyses were returned.

A *statistically significant analysis* is one in which the genes in a gene signature, gene list, or pathway are differentially modulated in a statistically significant way, indicating that the associated target or pathway is probably affected by the treatment, disease or other topic examined in the study.

To qualify as a statistically significant analysis, certain data points (such as p-value) must be evaluated and attain an aggregate score that meets or exceeds a particular threshold, based on an internal Johnson & Johnson algorithm. For information on the rules that determine how analysis results are ranked, see TEA Analyses on page 46.

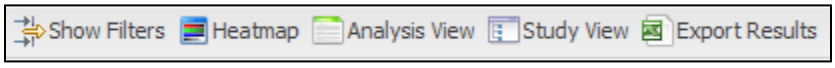
The following sections describe the views and operations available for each result category:

- Clinical Trials Tab (page 19)
- mRNA Analysis Tab (page 28)
- mRNA Profiles Tab (page 33)
- Literature Tab (page 35)
- Documents Tab (page 40)
- Pictor Tab (page 42)
- ResNet Tab (page 43)
- GeneGo Tab (page 44)

Clinical Trials Tab

This result category contains data from internal Johnson & Johnson clinical trials.

Click the **Clinical Trials** tab to display the results in this category. The buttons in the following figure appear at the top of the results list. You may see fewer buttons, depending on the results of your particular search:



These buttons give you access to the following views and operations:

- **Show Filters** – Define additional filters to further refine the search results.
- **Heatmap** – View the results as a heat map.
- **Analysis View** – View a list of the statistically significant analyses of the clinical trials.

  Results are sorted from the highest-scoring analysis down to the lowest.
- **Study View** – View a list of the clinical trials and, optionally, *all* the analyses for each clinical trial – that is, those analyses that are considered statistically significant and those that are not.

  Results are sorted from the clinical trial with the most matches with the search filter or search string, down to the one with the least matches.
- **Export Results** – Export descriptions of each clinical trial, and also all the analysis data from each of the clinical trials, to a Microsoft Excel file. All clinical trial descriptions are written to one worksheet in the file, and all analysis data is written to a second worksheet in the file.

The following sections describe clinical trial results for the gene signature **TLR3 in HT29 cells**.

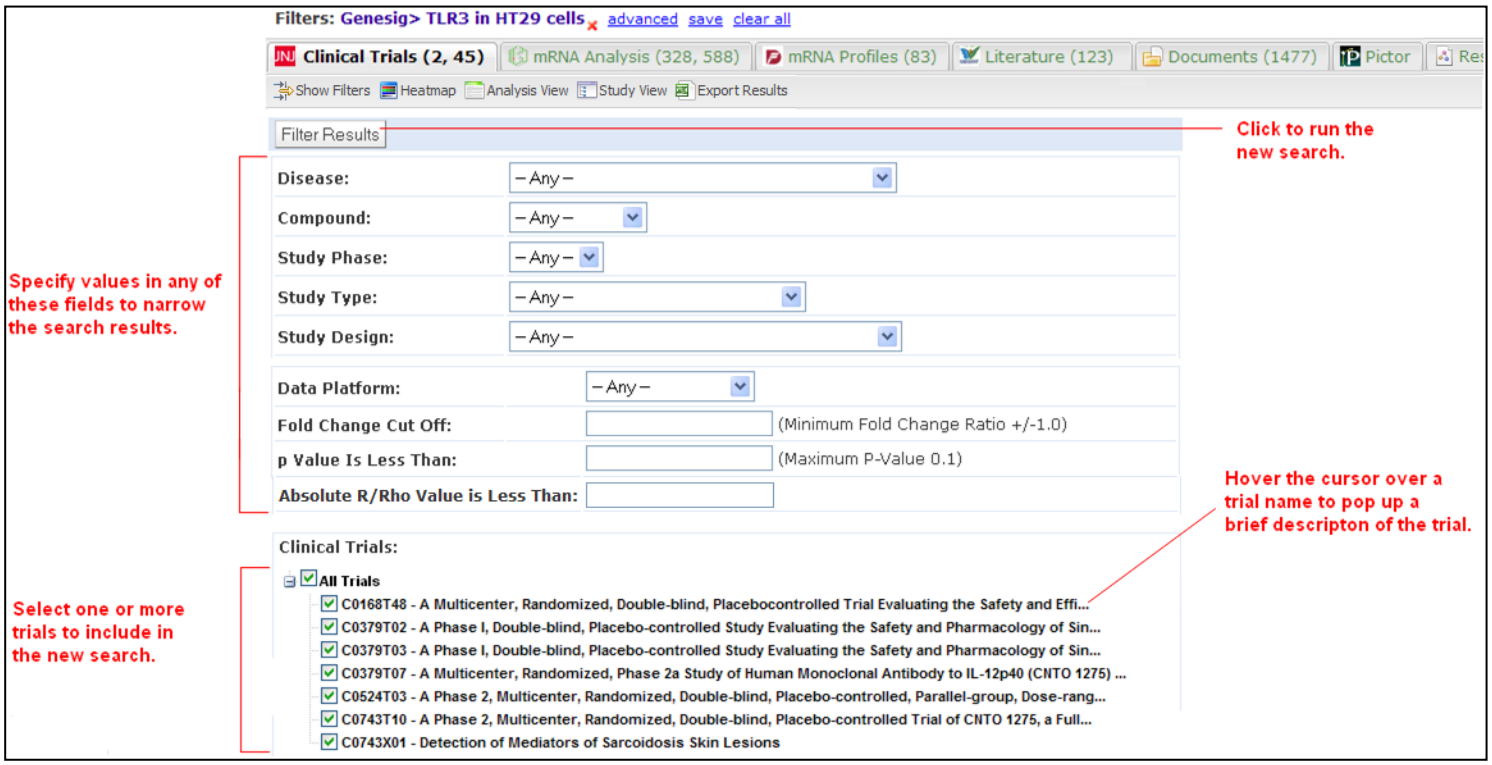
Show Filters

Click the **Show Filters** button to further refine the search results. When you click the button, a window containing filter fields appears (shown below), and the **Show Filters** button is replaced by the **Hide Filters** button.

In the figure below, filter selections are set for the broadest possible search.

To narrow the search:

1. Specify one or more filters – for example, specify a particular p-value to search against, and/or select a particular disease from the dropdown list.
2. Exclude one or more trials from the search by clearing the checkbox next to the trial name.
3. Click **Filter Results** to start the search.



Heatmap

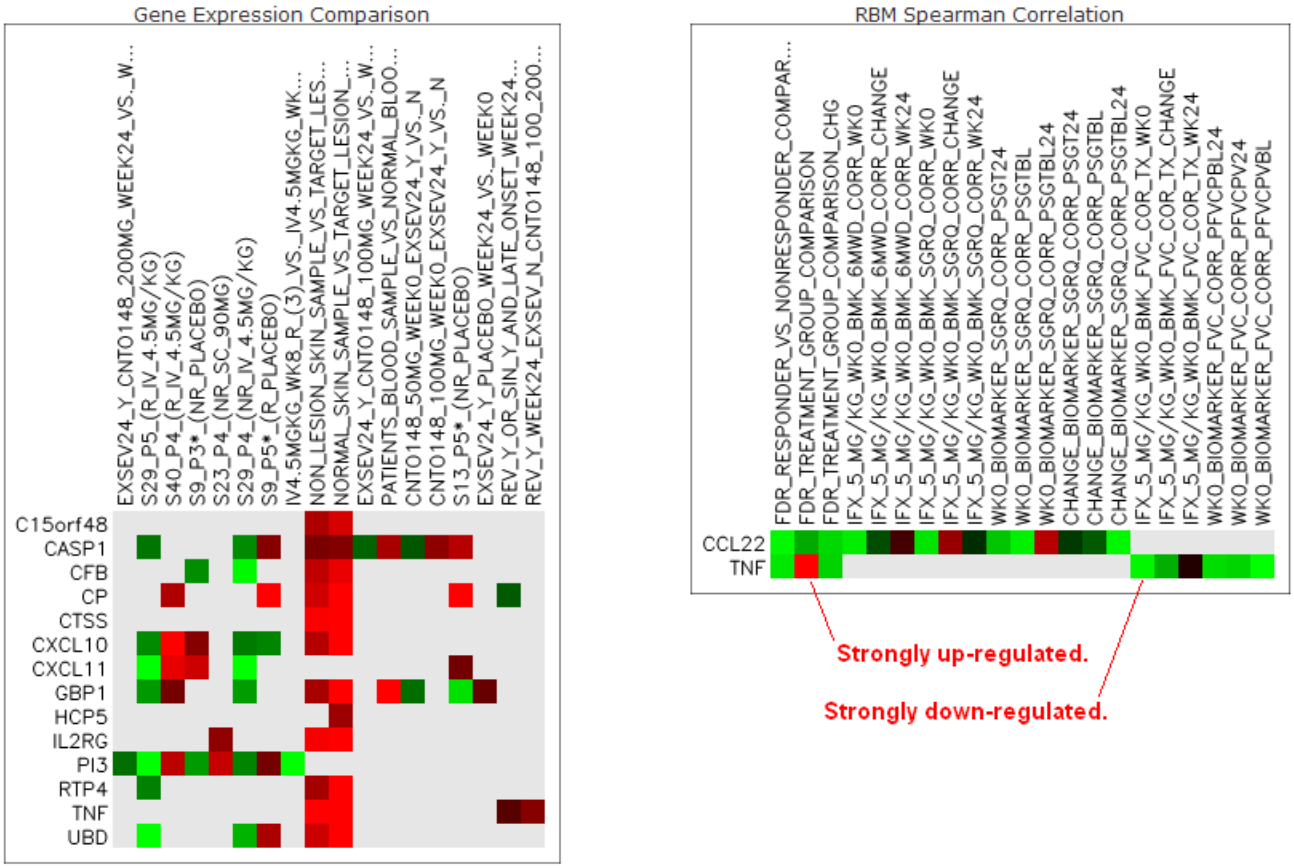
A heat map is a visualization of biomarker hits across multiple analyses of clinical trials. Click the **Heatmap** button to view clinical trial results as a heat map.

The heat map is a matrix of data points where relative values are indicated by color. In heat maps for clinical trials, the intersection of the vertical axis (genes) and the horizontal axis (analyses of clinical trial experiments) indicates the relative degree of protein production for the given gene during the particular experiment.

The degree of protein production for a gene is indicated as follows:

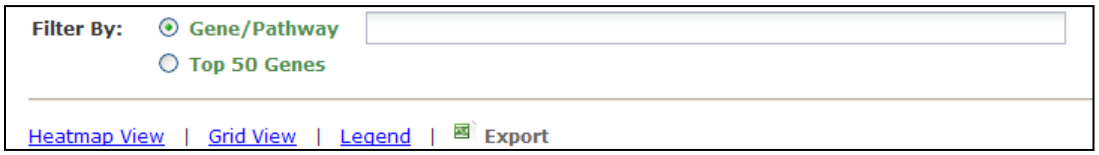
- **Red** – Up-regulated gene. The brighter the shade of red, greater the degree of protein production.
- **Green** – Down-regulated gene. The brighter the shade of green, the more pronounced is the underproduction of protein.
- **Black** – The gene is neither up-regulated nor down-regulated.
- **Gray** – There is no relevant data for this gene in the given clinical trial experiment.

In the rightmost matrix in the figure below, the TNF gene was strongly up-regulated in the **FDR_treatment_group_comparison** experiment, and strongly down-regulated in the **IFX 5 mg/kg wk0 bmk_FVC_cor_Tx wk0** experiment:



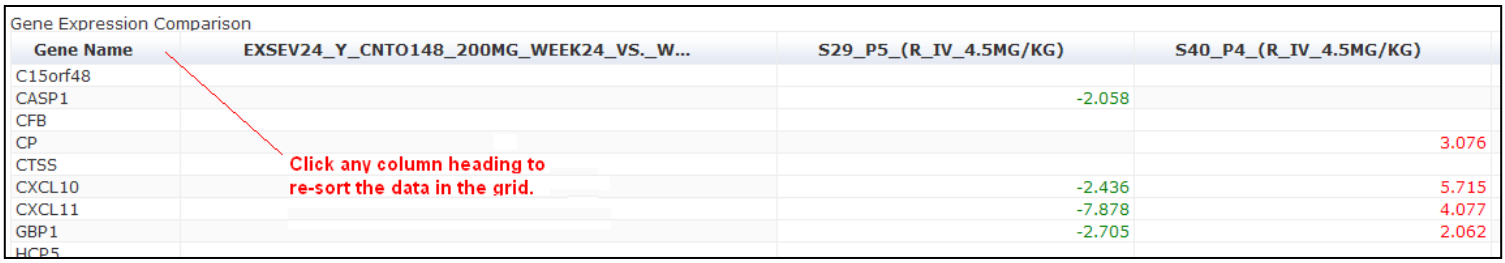
Operations You Can Perform in Heat Map View

The following controls appear at the top of the heat map:



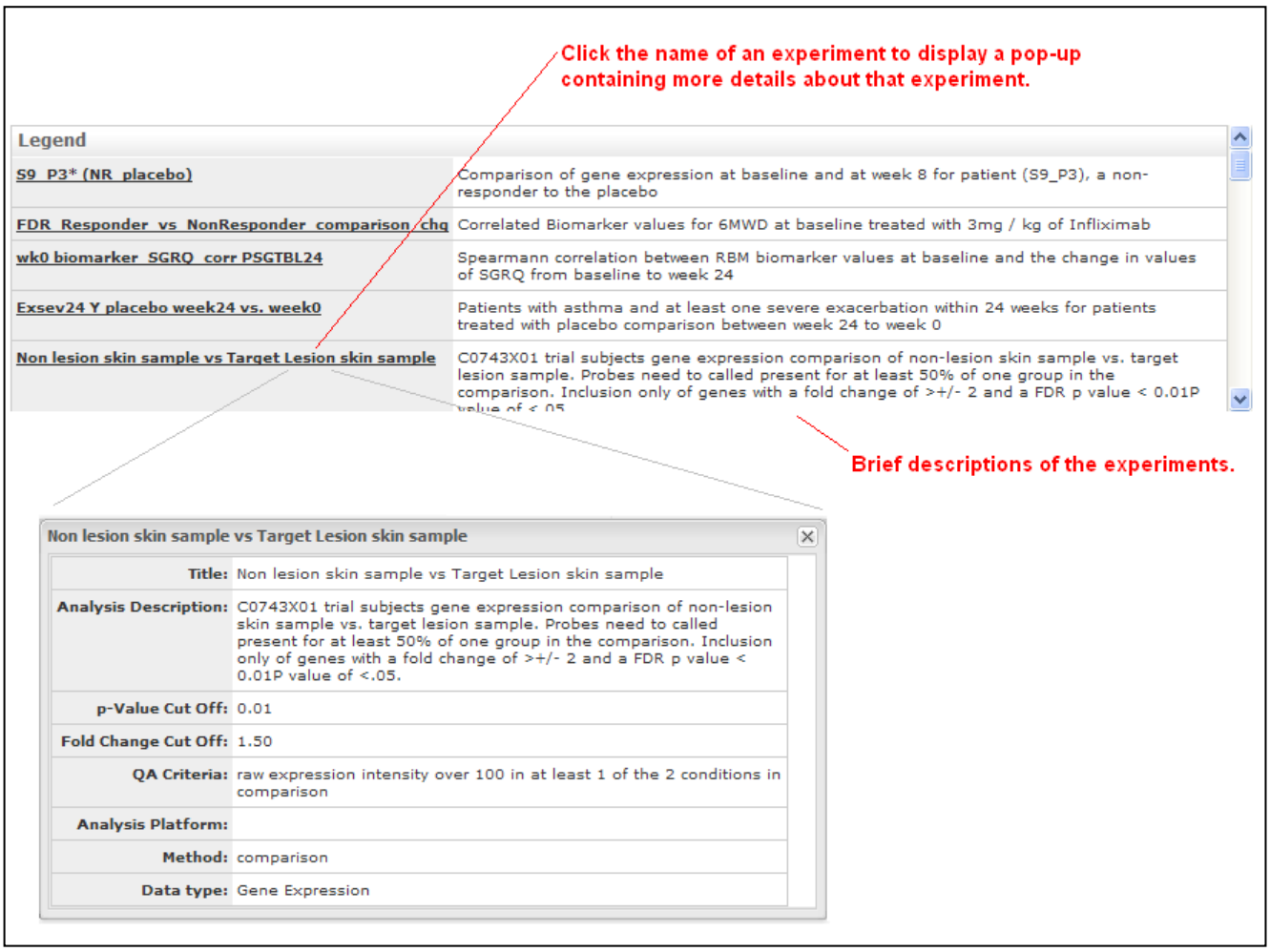
- In the **Filter By** field, do either of the following:
  - Filter the results by a particular gene or pathway.
  - Show the 50 genes that were factors in more of the experiments in the search results than all other genes.
- To re-display the heat map matrixes (if you are currently in Grid view), click **Heatmap View**.
- To see numeric representations of the degree of protein production, click **Grid View**.

  Also, to re-sort the data in the grid according to the data in a particular column, click the column header. The data in the column will be sorted in either ascending or descending order.



- To see a brief description of a particular clinical trial experiment, click **Legend**.

  Also, to display a pop-up box (called a details box) with more detail about an experiment, click the experiment name.

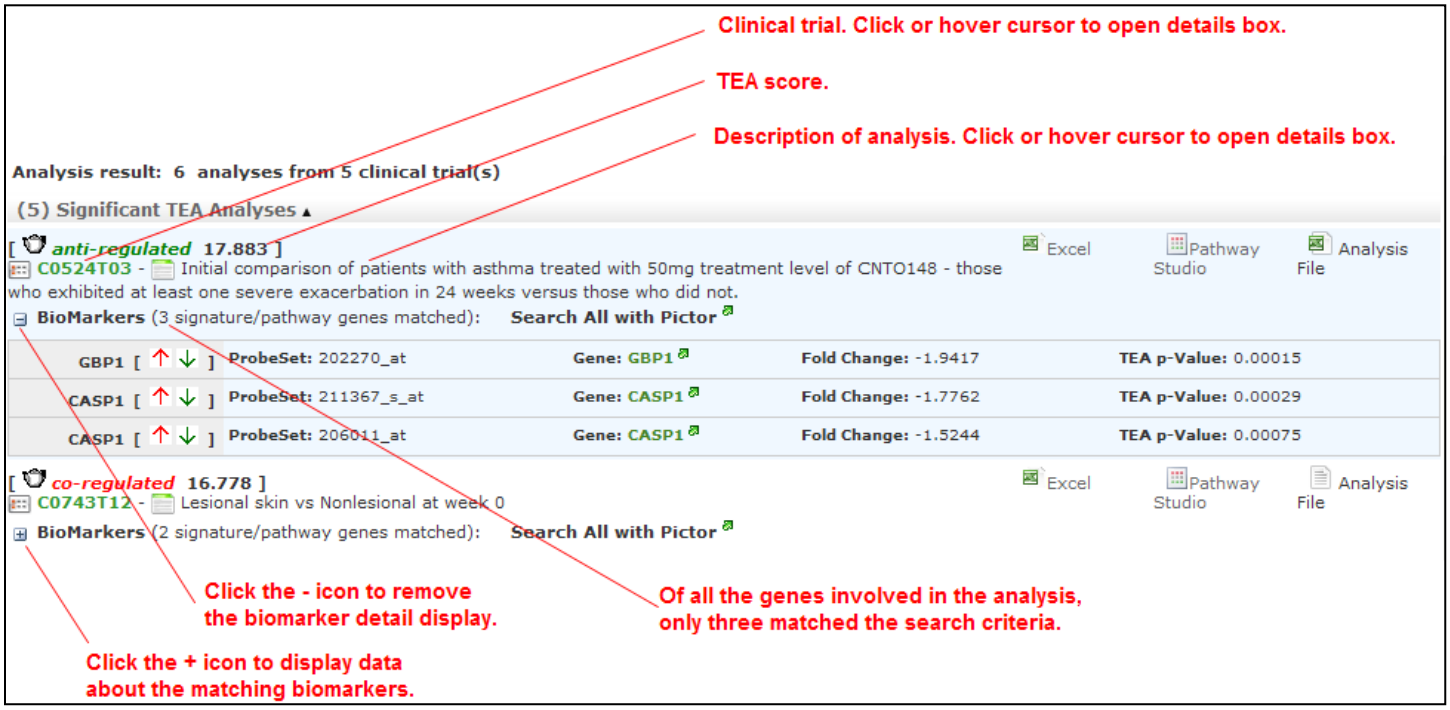


- To close the legend information, click **Legend**.
- To write the data to a Microsoft Excel spreadsheet (XLS), click **Export**.

Analysis View

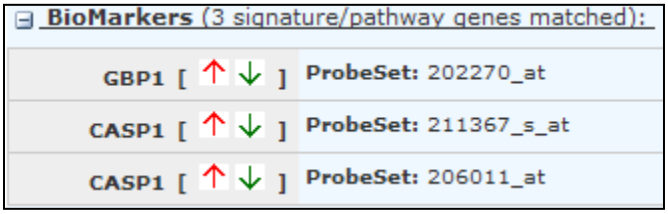
Click the **Analysis View** button to view the statistically significant analyses associated with the clinical trials.

For information on the rules that determine whether an analysis is ranked as statistically significant, see TEA Analyses on page 46.



When you click the **+** icon (⊞) to pull down the list of biomarkers, you see two arrows next to each biomarker name. The arrows have the following meanings:

- The leftmost arrow indicates whether the gene in the signature or list is up-regulated (up arrow) or down-regulated (down-arrow).
- The rightmost arrow indicates whether the gene in the comparison set is up regulated (up arrow) or down-regulated (down arrow).



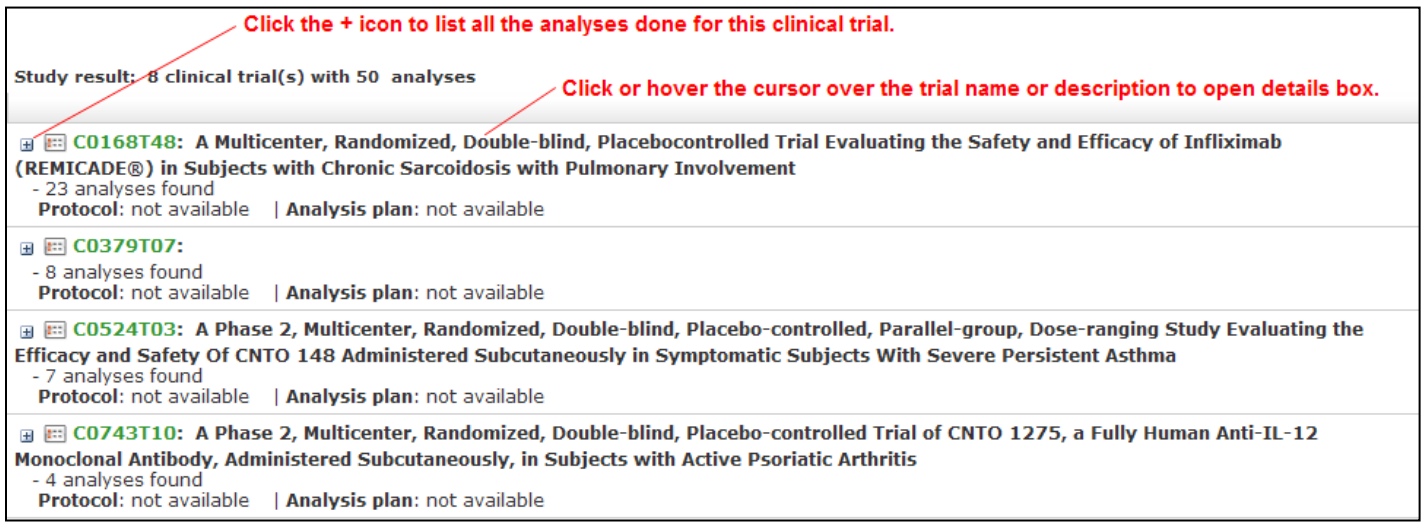
Note: The leftmost arrow has meaning only for searches involving gene signatures or lists. The arrow is not shown for other searches.

Each analysis also includes the following download options:

- **Excel** – Download detailed analysis data (such as probe set, fold change ratio, p-value) to a Microsoft Excel spreadsheet.
- **Pathway Studio** – Download analysis data to a file that can be imported into Ariadne Genomics Pathway Studio.
- **Analysis File** – Download a summary of the analysis to a Microsoft Excel spreadsheet.

Study View

Click the **Study View** button to view the clinical trials that are returned and, when you click the **+** icon (⊞) next to a clinical trial name, *all* the analyses for the clinical trial – that is, those analyses that are considered statistically significant and those that are not.



To drill down from the list of clinical trials:

1. Click the **+** icon (⊞) to the left of the clinical trial name to pull down a list of all the analyses done for the clinical trial.

   The analysis list is similar to the list of the statistically significant analyses in the Analysis View. However:

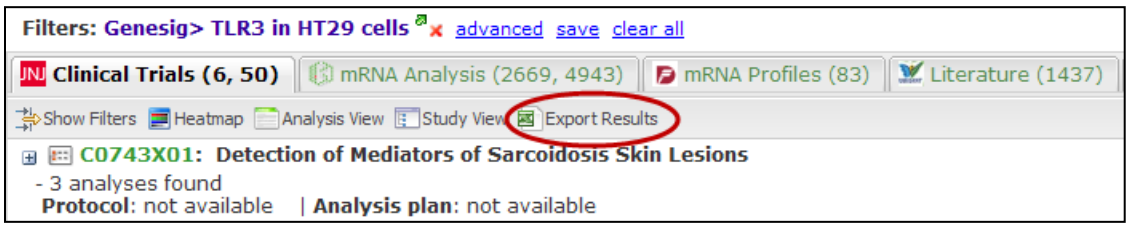
   - Because Study View includes analyses ranked as not statistically significant, TEA scores and the designations co-regulated and anti-regulated are not specified for the analyses in Study View, as they are in Analysis View.
   - Resources such as trial protocols, analysis plans, and other reports may be available for clinical trials.

2. Click the **+** icon (⊞) to the left of the **BioMarker** label to pull down a list of applicable biomarkers for an analysis. Note that:

   - A signature's fold change metric (indicated by a biomarker's leftmost arrow in Analysis View) is not relevant to analyses that are ranked as not statistically significant. Thus, the leftmost arrows are omitted for biomarkers in Study View.
   - The same export options for biomarkers are available in Study View as in Analysis View.

Export Results in Analysis View or Study View

While in either Analysis View or Study View, click the **Export Results** button to export the results data in the view to a Microsoft Excel spreadsheet:



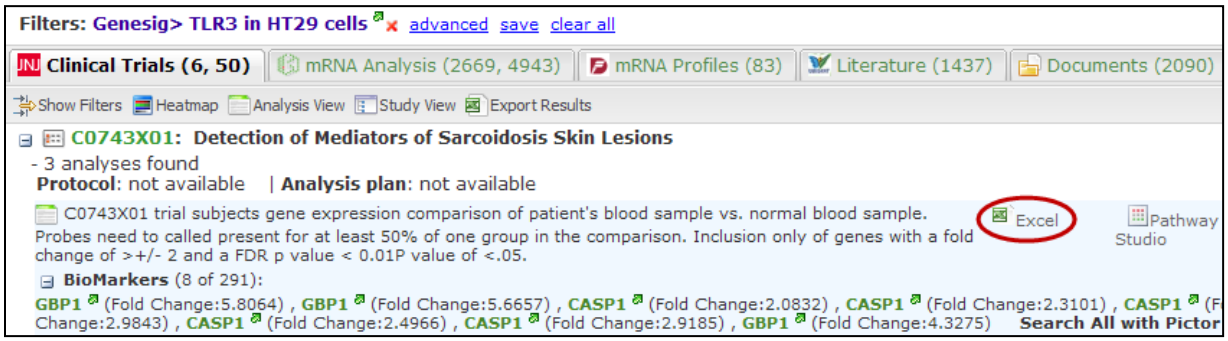
The Export function writes the following information to the spreadsheet:

- Descriptions of each clinical trial returned from the search. This is the same information that appears in a clinical trial details box.
- Information about the analyses associated with each clinical trial returned from the search. Information includes:
  - Analysis information displayed in the search results – for example, analysis description, TEA score, the list of matching biomarkers, and the probe set, fold change value, and TEA p-value associated with each biomarker.
  - Additional information about an analysis, such as descriptions of the biomarkers, biomarker type (such as gene expression), associated diseases, and compounds involved in the clinical trial.

All clinical trial descriptions are written to one worksheet in the file, and all analysis data is written to a second worksheet in the file.

Export Information about a Particular Analysis

To export details about all the biomarkers in a particular analysis, click the **Excel** button to the right of the analysis name – for example:



Note that the number of genes shown in parentheses after the **BioMarkers** label (291 in the above example), which specifies the number of genes included in the analysis, may be less than the number of rows written to the spreadsheet. The Export function writes one row of data for each *probe set*, not each gene, and the same gene may be associated with multiple probe sets.

Additional Resources

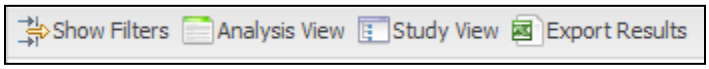
A Clinical Trials search result contains links to the following resources:

Resource Link	Description
Trial name. Example: C0524T03	View information about the trial, including title, owner, and description. The display may contain links to additional information, such as the trial protocol or analysis plan.
Gene. Example: SOCS5	Search the following sites for information about the gene: Entrez Gene | Entrez Global | Pictor | Hydra | GeneCards | TargetCV | Google Scholar
Search All with Pictor	Search Pictor for all the matching biomarkers in a particular analysis.
Excel	Export data (such as gene, probe set, and fold-change ratio) for the matching biomarkers in a particular analysis to Microsoft Excel.
Pathway Studio	Export the analysis data for the matching biomarkers to Ariadne Pathway Studio.
Analysis File	View notes and summary data for the biomarker analysis.
Protocol	Study View only – View the clinical trial's protocol.
Report	Study View only – View the clinical trial's technical report.
Analysis Plan	Study View only – View the clinical trial's biomarker analysis plan.

mRNA Analysis Tab

This result category contains gene expression data derived largely from external experiments and from some internal experiments.

Click the **mRNA Analysis** tab to display the results in this category. The buttons in the following figure appear at the top of the results list. You may see fewer buttons, depending on the results of your particular search:



These buttons give you access to the following views and operations:

- **Show Filters** – Define additional filters to further refine the search results.
- **Analysis View** – View the analyses of the experiments that are ranked as statistically significant analyses.
- **Study View** – View the details of the experiments and, optionally, *all* the analyses for each experiment - that is, those analyses that are considered statistically significant and those that are not.
- **Export Results** – Export descriptions of each experiment, and also all the analysis data from each of the experiments, to a Microsoft Excel file. All descriptions of experiments are written to one worksheet in the file, and all analysis data is written to a second worksheet in the file.

The following sections describe the results of experiments for the gene signature **TLR3 in HT29 cells**. Click the **mRNA Analysis** tab to see the results.

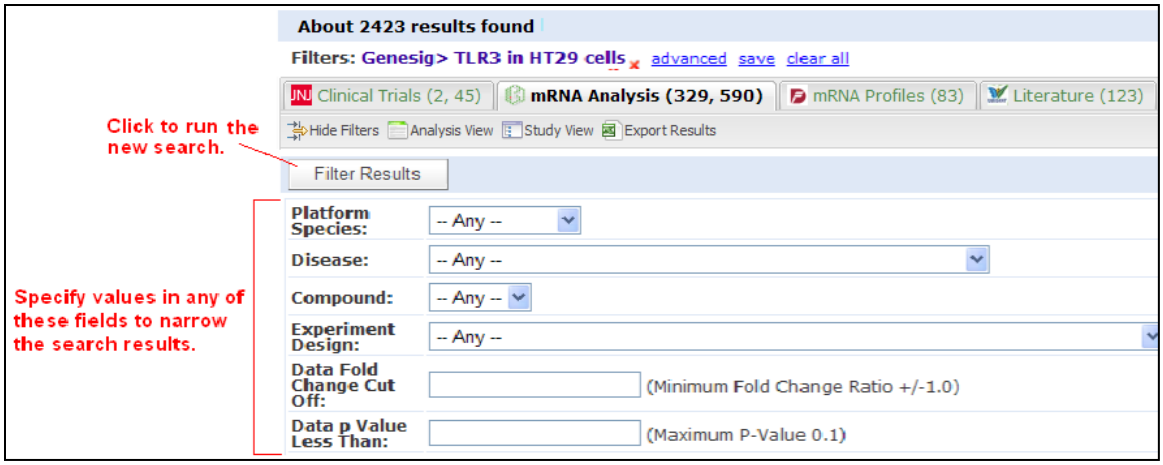
Show Filters

Click the **Show Filters** button to further refine the search results. When you click the button, a window containing filter fields appears (shown below), and the **Show Filters** button is replaced by the **Hide Filters** button.

In the figure below, filter selections are set for the broadest possible search.

To narrow the search:

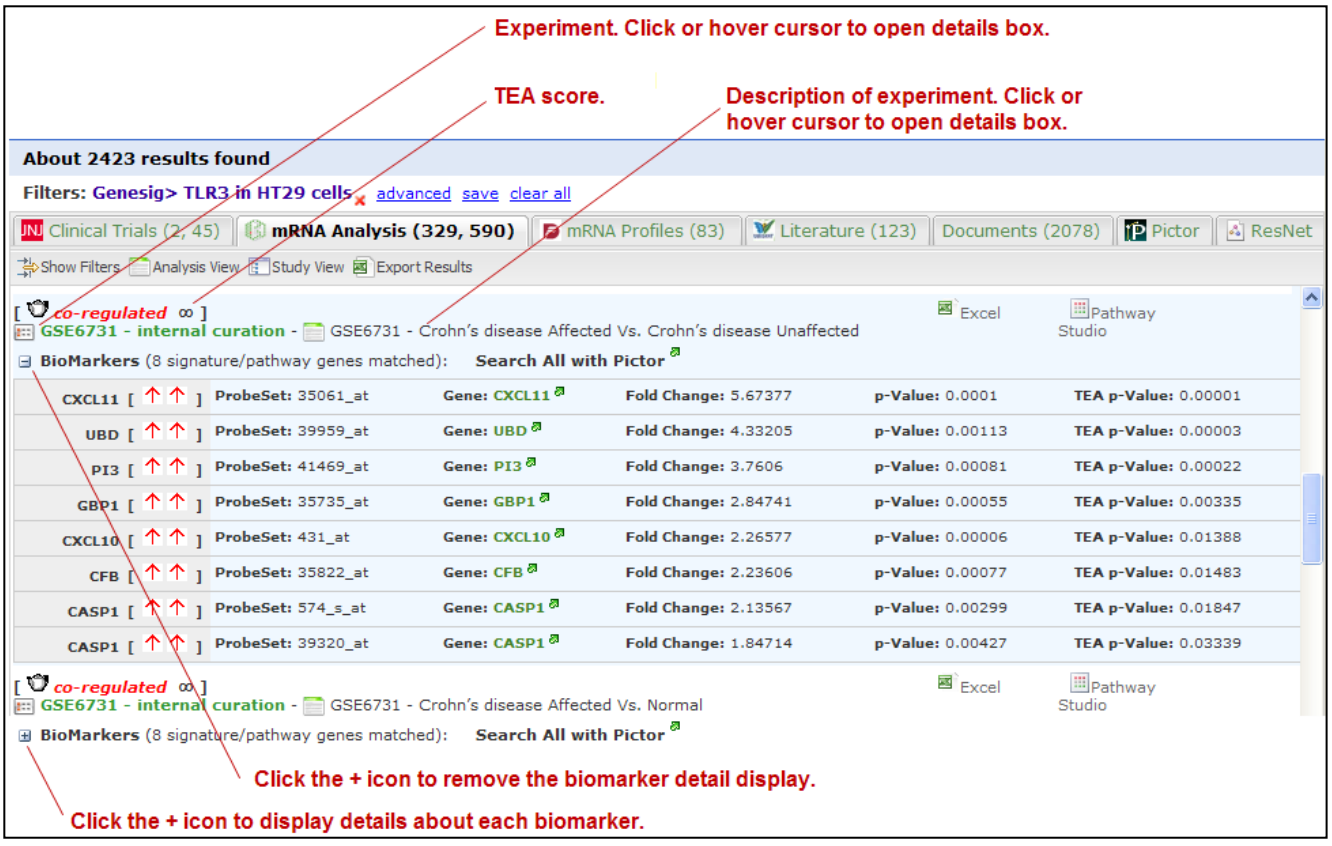
1. Specify one or more filters – for example, specify a particular p-value to search against, and/or select a particular disease from the dropdown list.
2. Click **Filter Results** to start the search.



Analysis View

Click the **Analysis View** button to view the statistically significant analyses associated with mRNA experiments.

For information on the rules that determine how analysis results are ranked, see TEA Analyses on page 46.



When you click the **+** icon ( ) to pull down the list of biomarkers, you see two arrows next to each biomarker name. The arrows have the following meanings:

- The leftmost arrow indicates whether the gene in the signature or list is up-regulated (up arrow) or down-regulated (down-arrow).
- The rightmost arrow indicates whether the gene in the comparison set is up-regulated (up arrow) or down-regulated (down arrow).

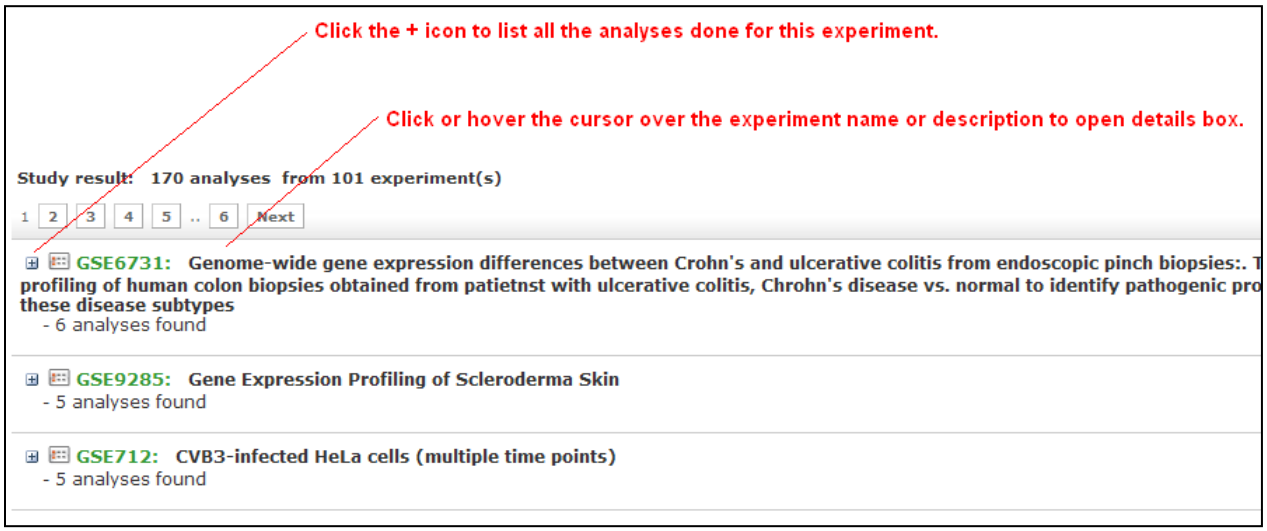
> **Note:** The leftmost arrow has meaning only for searches involving gene signatures or lists. The arrow is not shown for other searches.

Each analysis also includes the following download options:

- **Excel** – Download detailed analysis data (such as probe set, fold change ratio, p-value) to a Microsoft Excel spreadsheet.
- **Pathway Studio** – Download analysis data to a file that can be imported into Ariadne Genomics Pathway Studio.

Study View

Click the **Study View** button to view the mRNA experiments that are returned and, optionally, *all* the analyses for each experiment - that is, those analyses that are considered statistically significant and those that are not.



To drill down from the list of experiments:

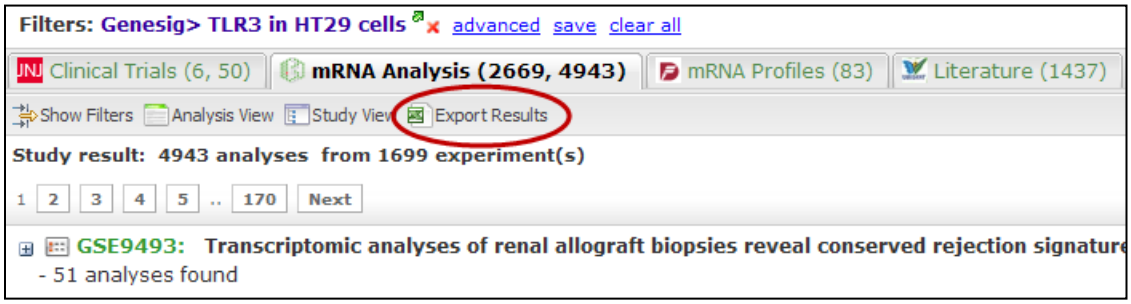
1. Click the **+** icon (⊞) to the left of the experiment name to pull down a list of all the analyses done for the experiment.

   The analysis list is similar to the list of the statistically significant analyses in the Analysis View. However, because Study View includes analyses ranked as not statistically significant, TEA scores and the designations co-regulated and anti-regulated are not specified for the analyses in Study View.

2. Click the **+** icon (⊞) to the left of the **BioMarker** label to pull down a list of applicable biomarkers for an analysis. Note that the same export options for biomarkers are available in Study View as in Analysis View.

Export Results in Analysis View or Study View

While in either Analysis View or Study View, click the **Export Results** button to export the results data in the view to a Microsoft Excel spreadsheet:



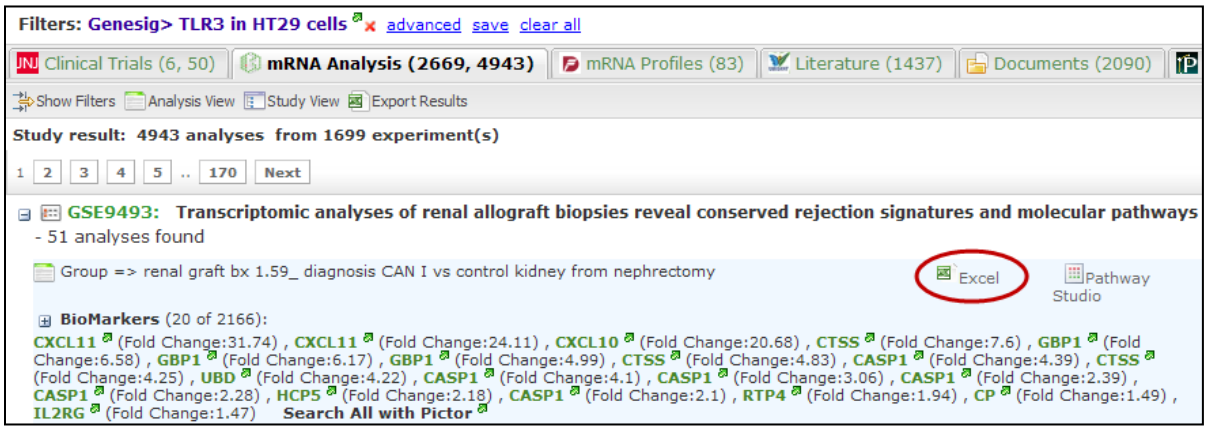
The Export function writes the following information to the spreadsheet:

- Descriptions of each experiment returned from the search. This is the same information that appears in a details box for an experiment. In addition, associated compounds and diseases are exported to the Excel file.
- Information about the analyses associated with each experiment returned from the search. Information includes:
  - Analysis information displayed in the search results – for example, analysis description, TEA score, the list of matching biomarkers, and the probe set, fold change value, p-value, and TEA p-value associated with each biomarker.
  - Additional information about an analysis, such as QA criteria, analysis platform, descriptions of the biomarkers, biomarker type (such as gene expression), associated diseases, and compounds involved in the experiment.

All descriptions of experiments are written to one worksheet in the file, and all analysis data is written to a second worksheet in the file.

Export Information about a Particular Analysis

To export details about all the biomarkers in a particular analysis, click the **Excel** button to the right of the analysis name – for example:



Note that the number of genes shown in parentheses after the **BioMarkers** label (2166 in the above example), which specifies the number of genes included in the analysis, may be less than the number of rows written to the spreadsheet. The Export function writes one row of data for each *probe set*, not each gene, and the same gene may be associated with multiple probe sets.

Mouse Gene Homology in Search Results

Searches can now return experiment results involving mouse genes. If experiment data is collected on a human gene and the corresponding mouse gene, a search against a human gene may potentially return results containing both human and mouse gene expression experiments.

For example information on both can be found by clicking the **Export Results** button in the search results. The **Organism** column in the Excel worksheet indicates whether a particular measurement was made on a human gene or a mouse gene.

The following figure shows part of an Excel worksheet containing the results of a search against the MET gene:

L	M	N	O	P	Q	R	S	T
Diseases	Bio Marker	Description	Organism	Probe Set	Fold Change	RValue	p-Value	TEA p-Value
	MET	met proto-oncogen	HOMO SAPIENS	203510_at	3.48		0.0008	0.0153
Neoplasms by Histologic Type; Leukemia; Neoplasms; Leukemia, Myeloid								
	Met	met proto-oncogen	MUS MUSCULUS	100309_at	3.72		0.0046	0.01828
	MET	met proto-oncogen	HOMO SAPIENS	1609_g_at	-7.19		0.0036	0.01846
	Met	met proto-oncogen	MUS MUSCULUS	1422990_at	7.14		0.00004	0.02186
	Met	met proto-oncogen	MUS MUSCULUS	1422990_at	-2.83		0	0.02559
Acute Disease; Digestive System Diseases; Bacterial Infections and Mycoses; Infection; Sprains and Strains; Lung Diseases; Pathologic Processes; Bacterial Infections;								
	MET	met proto-oncogen	HOMO SAPIENS	213816_s_at	-2.7		0.00008	0.02849
	MET	met proto-oncogen	HOMO SAPIENS	203510_at	-19.29		0.0049	0.02935
Intestinal Neoplasms; Colorectal Neoplasms; Digestive System Diseases; Gastrointestinal Diseases; Neoplasms; Neoplasms by Site; Colonic Diseases; Intestinal Diseas								

Additional Resources

An mRNA Analysis search result contains links to the following resources:

Resource Link	Description
Experiment name Example: GSE3284	View information about the experiment, including title, description, and primary investigator. The display may contain links to additional information, such as NCBI GEO and ArrayExpress data.
QA criteria Example: Compound:Time => 4 h -> bacterial endotoxin vs control	View key parameters of the experiment, such as p-Value and fold-change cutoffs, analysis platform, and methodology.
Gene Example: SOCS5	Search the following sites for information about the gene: Entrez Gene | Entrez Global | Pictor | Hydra | GeneCards | TargetCV | Google Scholar
Search All with Pictor	Search Pictor for all the matching biomarkers in a particular analysis.
Excel	Export data (such as gene, probe set, and fold-change ratio) for the matching biomarkers in a particular analysis to Microsoft Excel.
Pathway Studio	Export the analysis data for the matching biomarkers to Ariadne Pathway Studio.

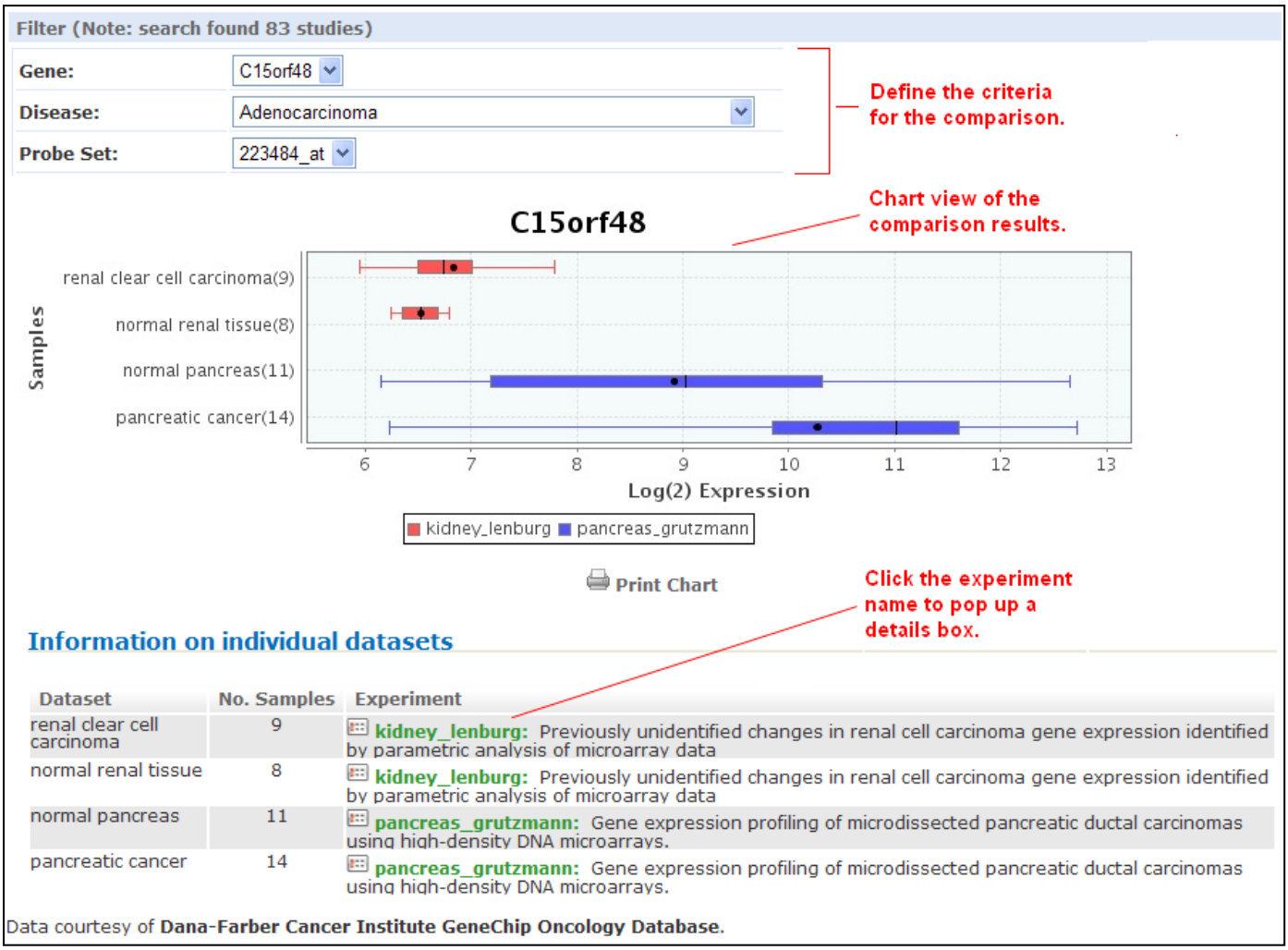
mRNA Profiles Tab

This result category allows mining of the expression levels of a gene or a set of genes in the context of a set of curated public studies involving different cancers, as collated and curated by Dana Farber Cancer Institute.

To see the results in this category, click the **mRNA Profiles** tab.

The number in parentheses on the tab represents the total number of studies that matched the search criteria.

The following figure shows the layout of the mRNA profile result window. The list of genes in the **Gene** dropdown are associated with the gene signature **TLR3 in HT29 cells** (the subject of the search). As you select different genes in the **Gene** dropdown, the diseases and probe sets that have meaning for the selected gene become available in the **Disease** and **Probe Set** dropdowns.



Additional Information about a Study

The mRNA Profile result set includes a list of the study (experiment) names that matched the search criteria for a particular disease and probe set. In the figure above, the studies **kidney_lenburg** and **pancreas_grutzman** contained matches.

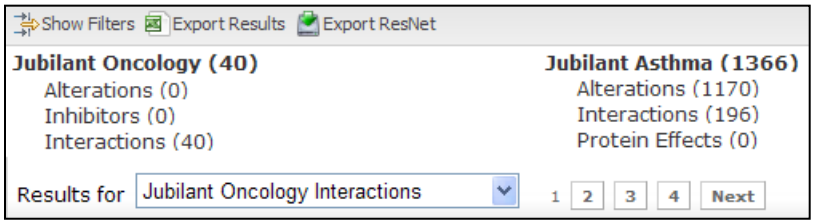
To view information about a study, click the study name. The details box that appears may contain links to more detailed information, such as a PubMed article.

Literature Tab

This result category contains hits from a set of curated articles within a set of selected diseases. Currently, the set includes a selection of genes of particular interest to J&J in different cancers a set of selected genes and their roles in a set of respiratory diseases (asthma, IPF, COPD, etc. The curation services are provided by Jubilant Biosys.

Sources of the articles include PubMed and ClinicalTrials.gov.

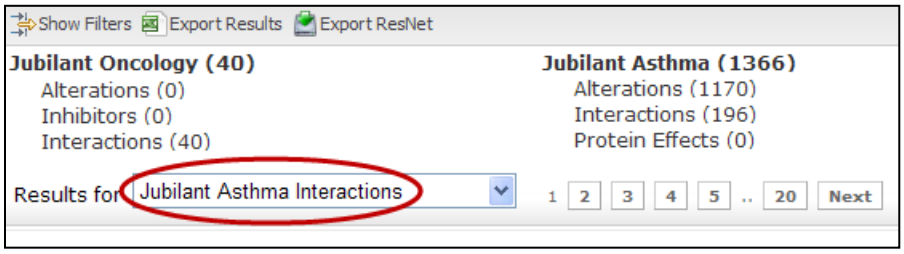
Click the **Literature** tab to display the results in this category. The following figure shows the controls that appear at the top of the results list:



These controls give you access to the following views and operations:

- **Show Filters** button – Define additional filters to further refine the search results.
- **Export Results** button – Write curation data to Microsoft Excel.
- **Export ResNet** button– Write selected facts from Jubilant curated content to Ariadne Genomics' Pathway Studio for network visualization and further in-depth analysis.
- **Results for** dropdown – Specify the type of literature results you want to see in the categories of oncology and asthma. In the figure above, matches for the genes in the gene signature were found in the categories of oncology inhibitors and asthma inhibitors. The **Results for** dropdown lets you select and view results for one of those categories.

The following sections describe the results of a Literature search for the gene signature **TLR3 in HT29 cells**. The results are for the category asthma interactions, as specified in the following figure:



Show Filters

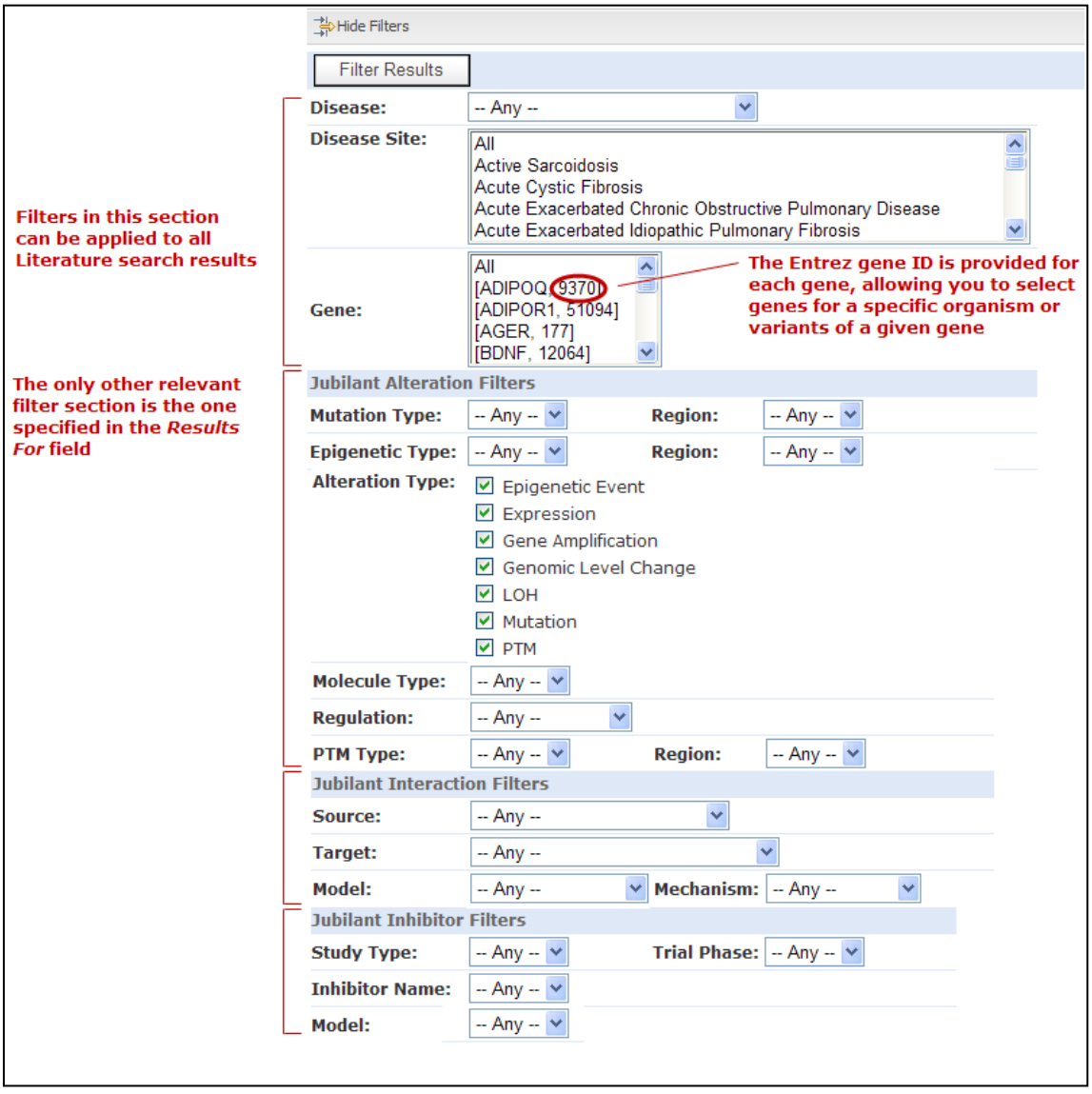
Click the **Show Filters** button to further refine the search results. When you click the button, a window containing filter fields appears (shown below in step 1), and the **Show Filters** button is replaced by the **Hide Filters** button.

In the following figure, filter selections are set for the broadest possible search.

To narrow the search:

1. Specify one or more filters – for example, specify a particular disease and/or one or more particular genes.

   Notice that the filters below the **Disease**, **Disease Site**, and **Gene** fields are divided into sections. These sections correspond to the subcategories Alterations, Inhibitors, and Interactions. The only section that is relevant to a given search is the section that is specified in the **Results for** field (for example, **Jubilant Asthma Interactions** in the figure above).

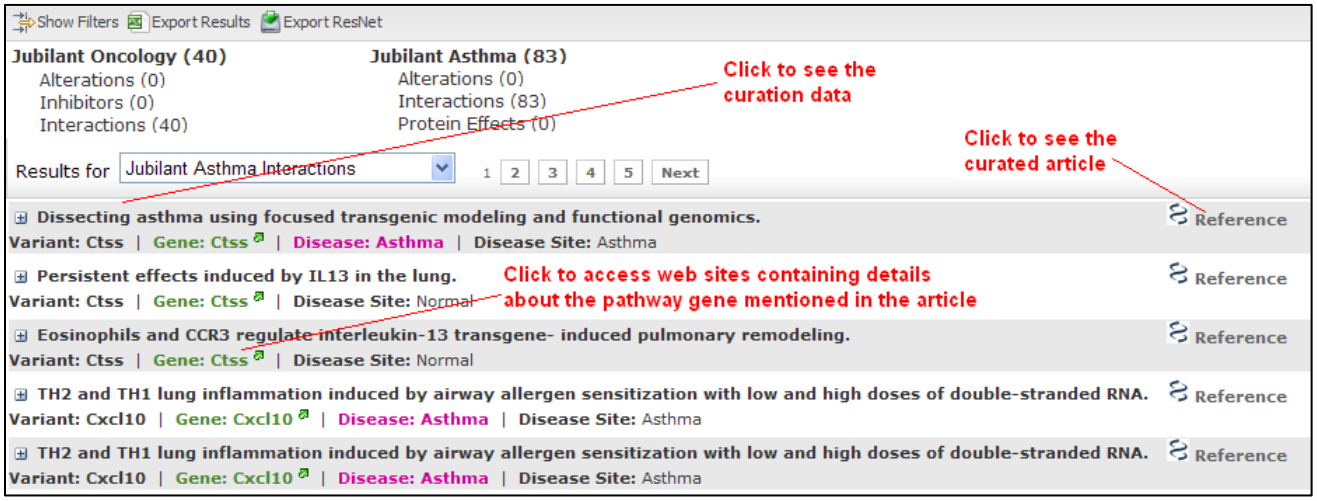


2. When finished setting filters, click **Filter Results** to narrow the search results.

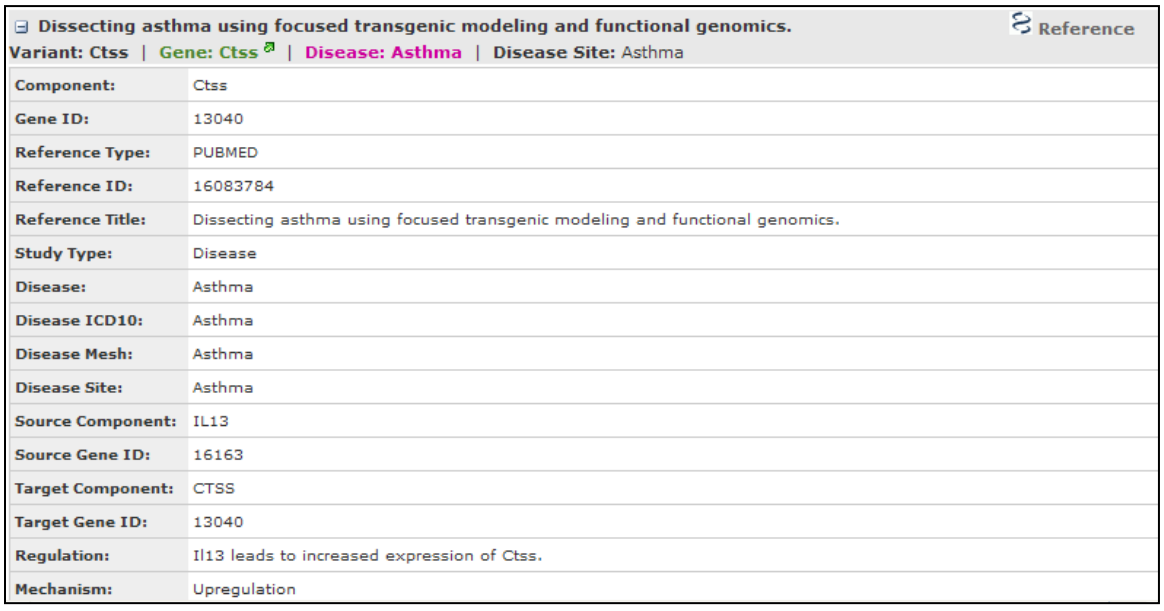
The Results View

The following figure shows a partial results list for a literature search. You can perform the following actions in the literature results view:

- Click **Reference** to see the original curated article.
- Click a gene name to see information about the gene that was referenced in the article.
- Click the title of the article or the **+** icon (⊞) next to the title to see the curation data.



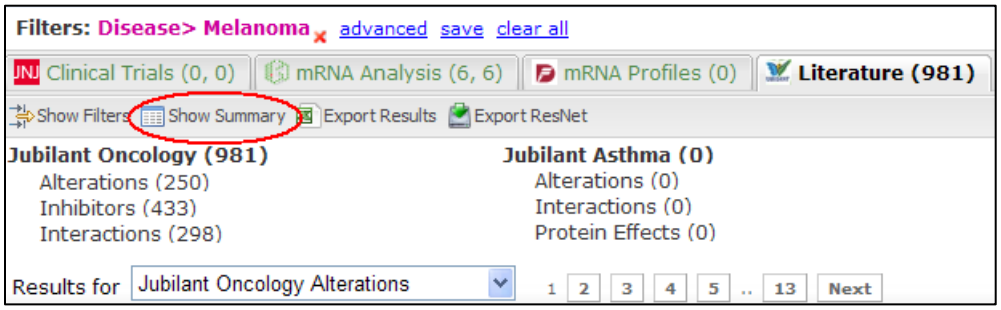
The following figure shows some of the curation data for the article titled **Dissecting asthma using focused transgenic modeling and functional genomics**:



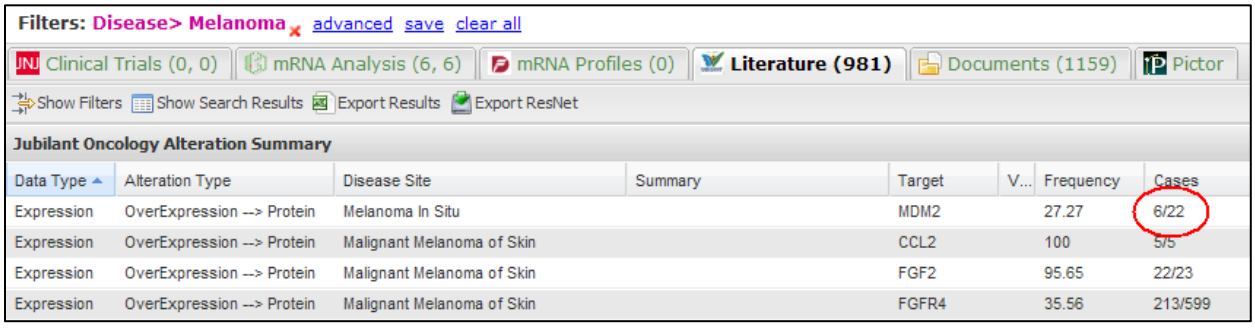
Summary View

When literature results are retrieved for the category **oncology alterations**, you can view the results in summary view. You select summary view by clicking the **Show Summary** button shown in the figure below.

> **Note:** This button appears only for results in the category oncology alterations.

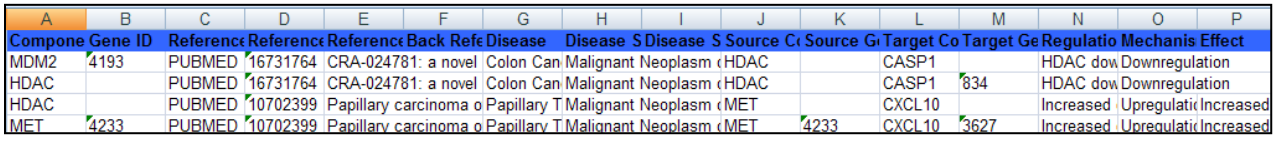


In summary view, each row shows the total cases in the full oncology alterations result that met all the same criteria. For example, in the following figure, 6 of 22 cases (27.27%) involved the data type Expression, alteration type OverExpression -> Protein, disease site Melanoma in Situ, and gene MDM2:

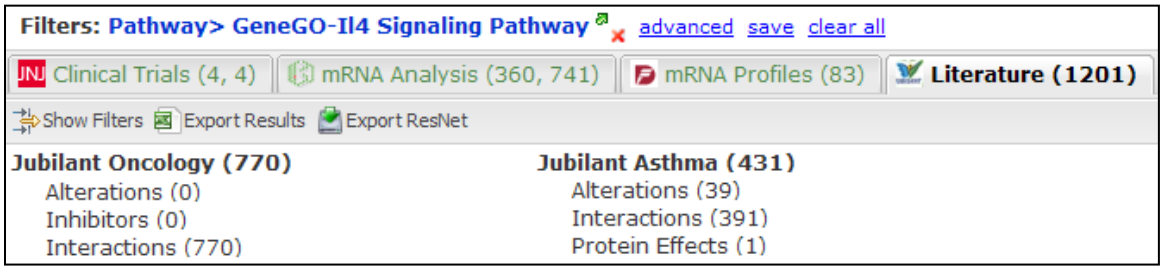


Export Results

Click the **Export Results** button to export the curation data in the results to Microsoft Excel:



Each category of curation data is written to a separate worksheet in the Excel file. For example, given the following search result, the categories Oncology Interactions, Asthma Alterations, Asthma Interactions, and Asthma Protein Effects will be written to separate worksheets in the Excel file:



Export ResNet

Click the **Export ResNet** button to generate an Ariadne Genomics' Pathway Studio visualization of the interactions between the biomarkers returned from a literature search. You can do so either by opening Pathway Studio and importing the literature data now, or by saving the information to a file that can be imported into Pathway Studio later.

When you click **Export ResNet**, tranSMART exports any alterations, inhibitors, and interactions results in the oncology dataset, and any alterations and interactions results in the asthma dataset, to Pathway Studio. Export to Pathway Studio of asthma protein effects is not supported.

For an example of importing curated literature data into Pathway Studio, see the section *Create a Visualization of a Literature Search Result* in the tranSMART *JnJ Training: Advanced* guide.

Additional Resources

A Literature search result contains links to the following resources:

Resource Link	Description
Gene Example: SOCS5	Search the following sites for information about the gene: Entrez Gene | Entrez Global | Pictor | Hydra | GeneCards | TargetCV | Google Scholar
Reference	View the complete article.

Documents Tab

The search results in this category are based on indexing of a set of Johnson & Johnson document repositories:

- Full text indexing of a set of conference abstracts.
- Full text indexing of oncology papers curated by Jubilant.
- Full text indexing of the Biomarker folder. Requires authorization to access the following shared drive:

  `\\eu.jnj.com\rndbedfsroot\All\DrugDiscovery\biomarker`

- Full text indexing of a DIP folder specific to the cMET project. Requires authorization to access the following shared drive:

  `\\na.jnj.com\jnjdfsroot\eu\rndbe\All\DrugDiscovery\_Discovery Projects\`

> **Note:** The Biomarker and DIP repositories require Microsoft Internet Explorer.

Sort Order of Results

Document results are sorted according to the score assigned to each document, from highest score to lowest. The higher the score, the more instances of the search term were found in the document:



> **Note:** Scoring is calculated using the open-source search software Lucene.

Show Filters

The **Show Filters** button appears at the top of the results list. Click this button to further refine the search results.

When you click the button, a window containing filter fields appears (shown below), and the **Show Filters** button is replaced by the **Hide Filters** button.

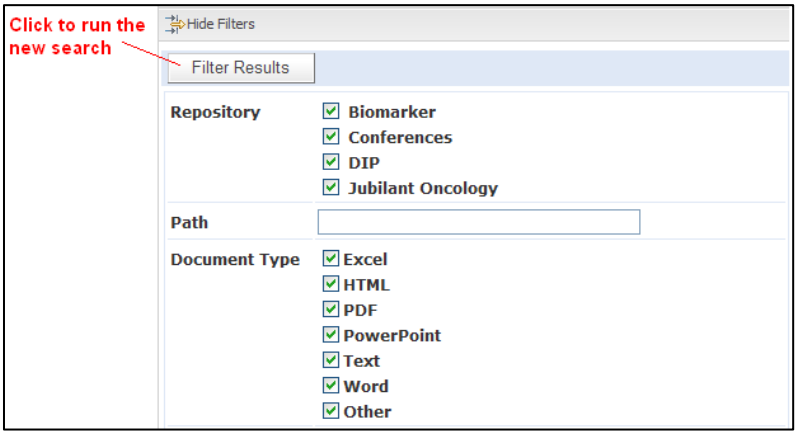
In the figure below, filter selections are set for the broadest possible search.

To narrow the search:

1. Deselect one or more check boxes in the **Repository** and **Document Type** sections.

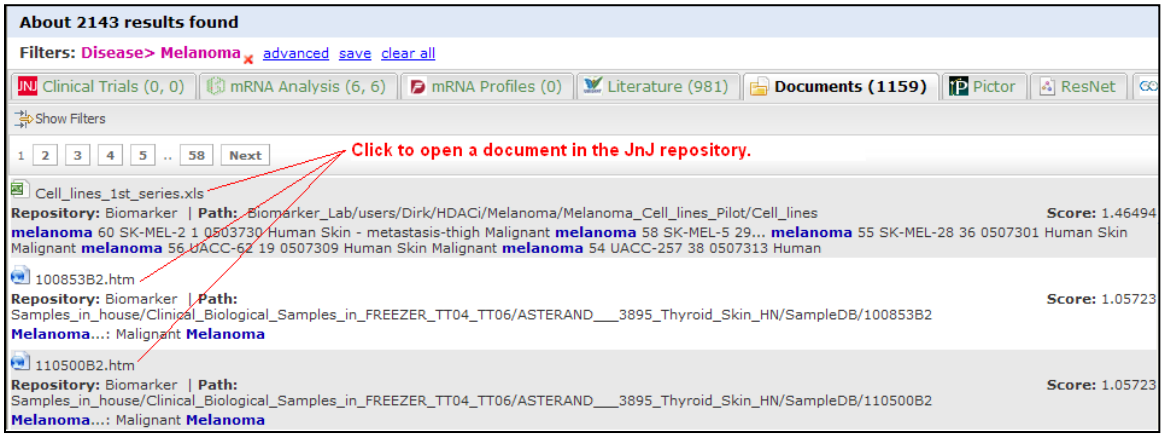
   For example, if you are looking for documents related to conferences, deselect the **Biomarker**, **DIP**, and **Jubilant Oncology** boxes.

2. Click **Filter Results** to start the search.



Open a Document

To open a document listed in the search results, click the document name:



A Document search result contains links to resources in the following repositories:

Repository	Example
Biomarker	N3upreg_benchmark.txt **Repository:** Biomarker | **Path:** Biomarker_Lab/users/Tine/cenix
Conferences	5565_1664.htm **Repository:** Conferences | **Path:** American Association for Cancer Research 2008
DIP	Invitrogen_kinase_13.xls **Repository:** DIP | **Path:** Biology/In vitro and cellular assays
Jubilant Oncology	Cancer cMet in metastasis, motility - review.pdf **Repository:** Jubilant Oncology | **Path:** -

Pictor Tab

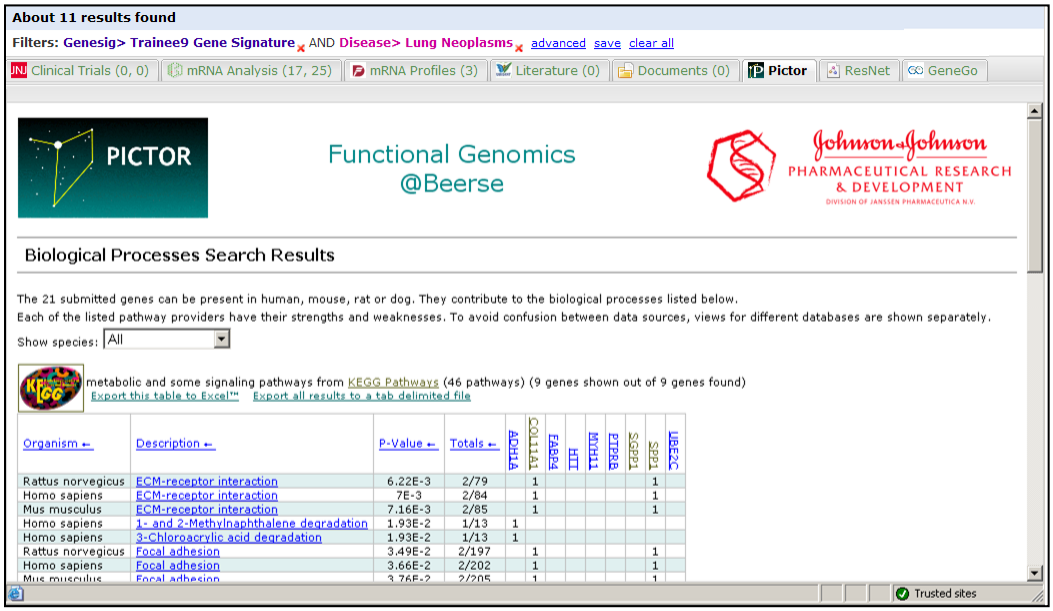
This tab displays the Johnson & Johnson Pictor site in the tranSMART results pane.

> **Note:** If a security warning appears when you click this tab, see Security Warning When Selecting Pictor, ResNet, or GeneGO on page 45.

When you click this tab after a search against a gene, pathway, gene signature or gene list, tranSMART submits the gene symbols to Pictor. Pictor then returns any matches it finds in the pathway databases it searches.

Pictor is essentially a database of biologically meaningful gene lists. When the list does not contain expression data, the submitted genes are compared to all Pictor lists. The Pictor lists which correspond most to the submitted genes are reported in a table, ranked by hypergeometric distribution p-value statistics. When expression data is included in the analysis, Pictor lists are ranked by MeanLogP statistics.

For further information, please see the Pictor documentation.



ResNet Tab

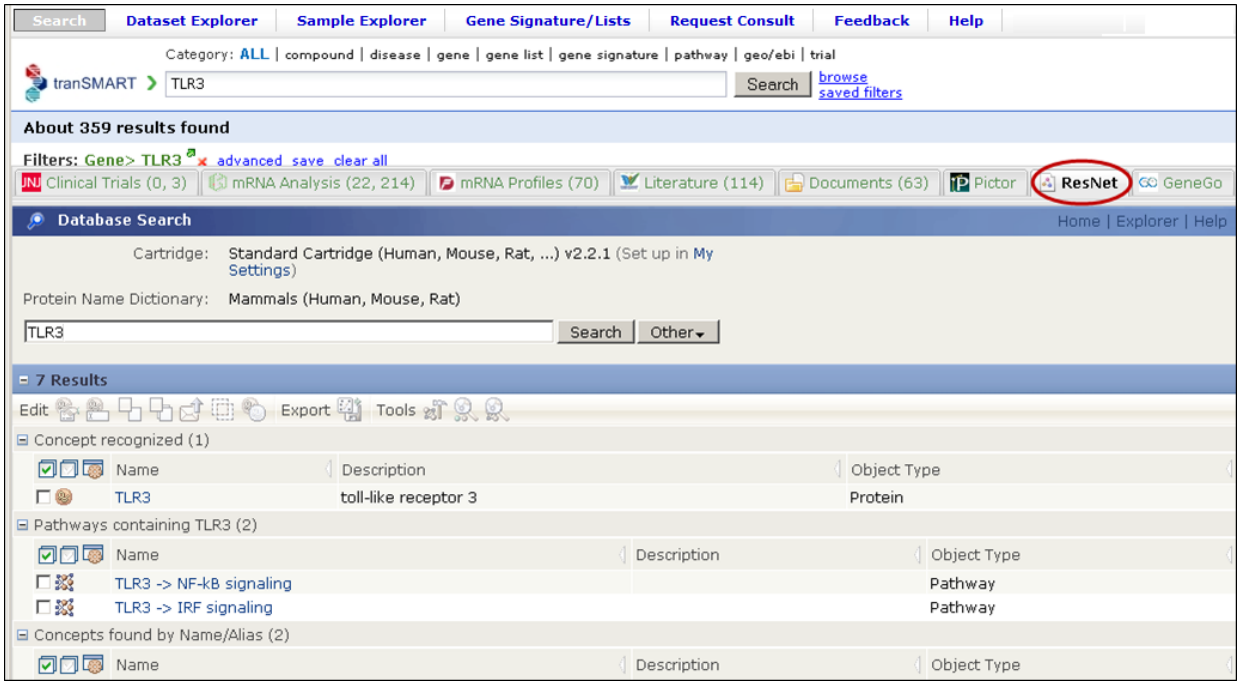
When you click the **ResNet** tab after a search against one or more genes, pathways, and/or diseases, tranSMART displays controls for performing an Ariadne Genomics Medscan search. These controls are displayed in the tranSMART results pane.

> **Note:** If a security warning appears when you click this tab, see Security Warning When Selecting Pictor, ResNet, or GeneGO on page 45.

To begin the Medscan search, click **Search**. Optionally, you can modify the search parameters before you click **Search**.

The first time you click the **ResNet** tab in a tranSMART session, you are prompted for your PWS user ID and password.

The following figure shows a tranSMART search against the TLR3 gene and the Medscan search results:



In the Medscan results, you can perform further operations, such as viewing the target of the search (protein) in a pathway context.

GeneGo Tab

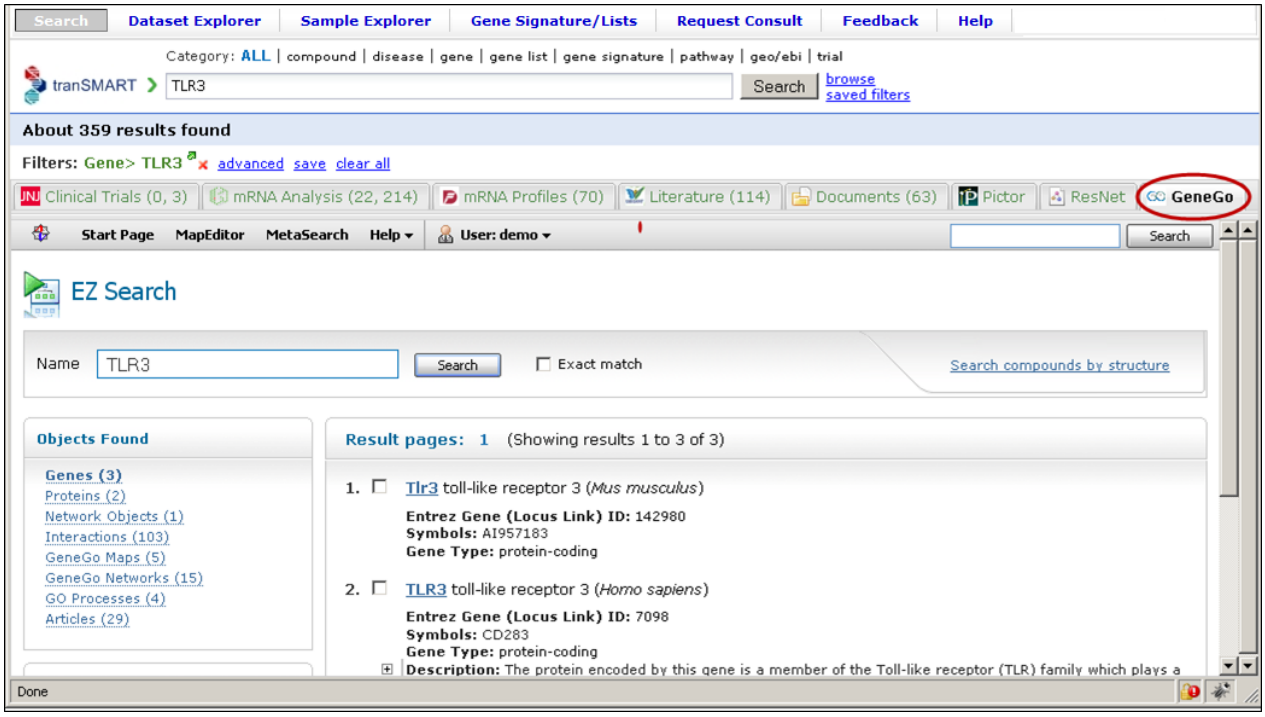
When you click the **GeneGo** tab after a tranSMART search, the search string is passed to GeneGo, and the GeneGo search against the tranSMART search string begins immediately.

> **Note:** If a security warning appears when you click this tab, see Security Warning When Selecting Pictor, ResNet, or GeneGO on page 45.

tranSMART displays the GeneGo user interface in the tranSMART results pane. You can view the initial search results there, and use the GeneGo controls to perform additional GeneGo searches.

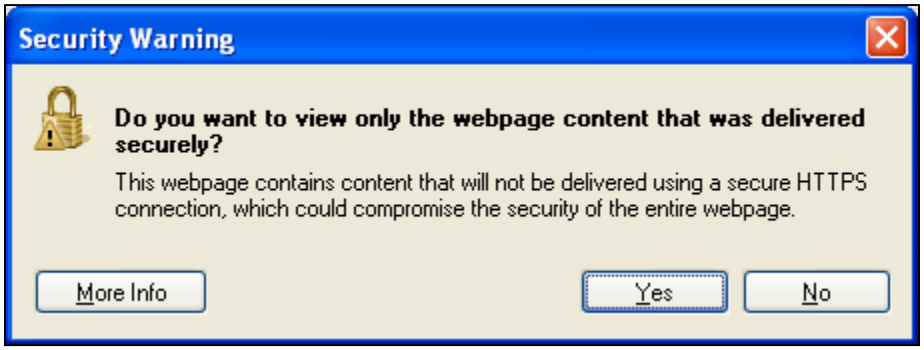
> **Note:** If you are prompted for login credentials after clicking the **GeneGo** tab, hover the mouse cursor over the **GeneGo** tab to see the credentials you need, then supply them in the Authentication Required dialog.

The following figure shows a tranSMART search against the TLR3 gene and a partial view of the GeneGo search results:



Security Warning When Selecting Pictor, ResNet, or GeneGO

If the following dialog appears after you click the **Pictor**, **ResNet**, or **GeneGO** tabs, click **No**:



If you click Yes, the resource you selected will not be displayed.

If your version of Internet Explorer is earlier than version 8, or if you would like more information about this security warning, see the document Configuration Tips.

TEA Analyses

Target Enrichment Analysis (TEA) measures the enrichment of a gene signature, gene list, or pathway in a microarray expression experiment.

> **Note:** For information on how TEA scores are calculated, see Appendix A: How TEA Scores Are Calculated.

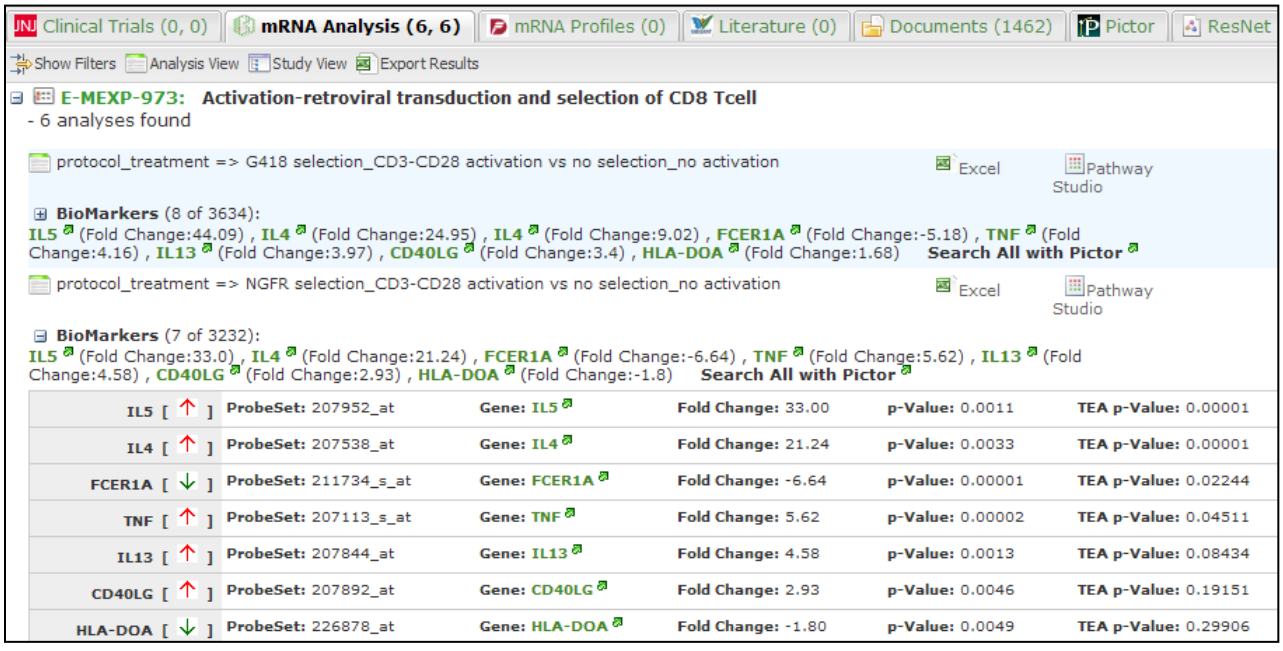
TEA Indicators Applied to Individual Biomarkers

The Study View of Clinical Trial or mRNA Analysis search result lists all experiments that satisfy the search criteria. Further, in Study View, you can list:

- All of an experiment's analyses that satisfy the search criteria
- All of an analysis' biomarkers that satisfy the search criteria

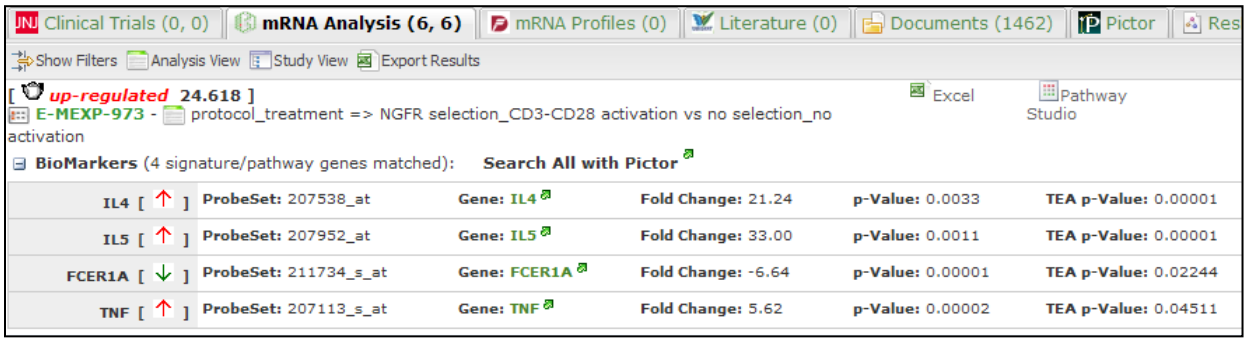
To drill down to the matching analyses in an experiment, click the **+** icon (⊞) next to the experiment name. To drill down to the matching biomarkers in an analysis, click the **+** icon next to the label **BioMarkers** under the analysis name.

The following example shows the experiment **E-MEXP-973** in Study View. The biomarkers for the analysis **protocol_treatment => NGFR selection_CD3-CD28 activation vs no selection_no activation** are displayed:



Notice the rightmost column of biomarker values: **TEA p-Value**. These normalized p-values are intermediate values in the TEA calculation. To be considered a statistically significant analysis, an analysis must have at least one matching biomarker with a TEA p-Value of less than 0.05.

The following figure shows the same experiment and analysis from the figure above, but in Analysis View. Notice that the only biomarkers that are displayed in Analysis View are those with a TEA p-Value below 0.05:



Statistically significant analyses are candidates for display in the Analysis View, after further TEA calculations are performed to determine whether the analysis is a **significant TEA analysis** or an **insignificant TEA analysis**.

TEA Indicators Applied to an Analysis

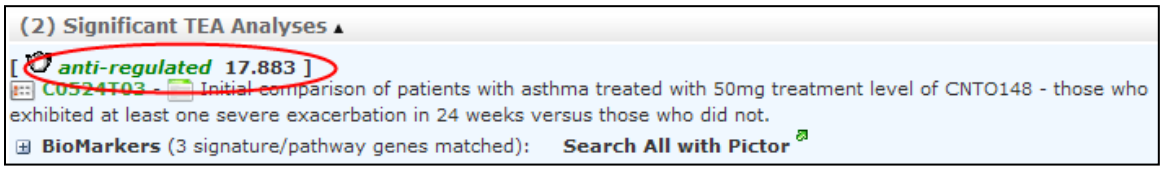
The TEA algorithm assigns an aggregate score to each analysis within an experiment. A TEA score is a binomial distribution of normalized p-values, calculated in the context of the following factors:

- **With gene signatures and gene lists** – The level of co-regulation or anti-regulation of the genes within the gene signature or gene list, as compared with the experiment.
- **With pathways –** The level of up-regulation or down-regulation of the genes within the pathway, as compared with the experiment.

> **Note:** For details on the TEA algorithm, see Appendix A: How TEA Scores Are Calculated.

TEA identifies experiments where the genes in the signature, list, or pathway are *differentially modulated*, indicating that the target is affected by the treatment, disease or other topic examined in the experiment.

In the following figure, the analysis list for a search against a gene signature includes the encircled TEA indicators for the first analysis in the list:



What the TEA Score Means

The TEA score displayed for an analysis of an experiment is not the actual TEA score calculated by the TEA algorithm. TEA scores are typically very small decimal numbers that are not easily human-readable. To aid users in interpreting the relative value of TEA scores, scores are converted to a larger number, as follows:

```
Displayed_TEA_Score  = -log(Actual_TEA_Score)
```

The larger the displayed TEA score, the more statistically significant is the analysis.

Typically, displayed TEA scores for statistically significant analyses of experiments range from 3 to 30 or 40.

Analyses of experiments are grouped into the categories **Significant TEA Analyses** and **Insignificant TEA Analyses**, as follows:

- Significant TEA analyses have a displayed TEA score of >= 2.9957.
- Insignificant TEA analyses have a displayed TEA score of < 2.9957.

What Co-/Anti-Regulation and Up-/Down-Regulation Mean

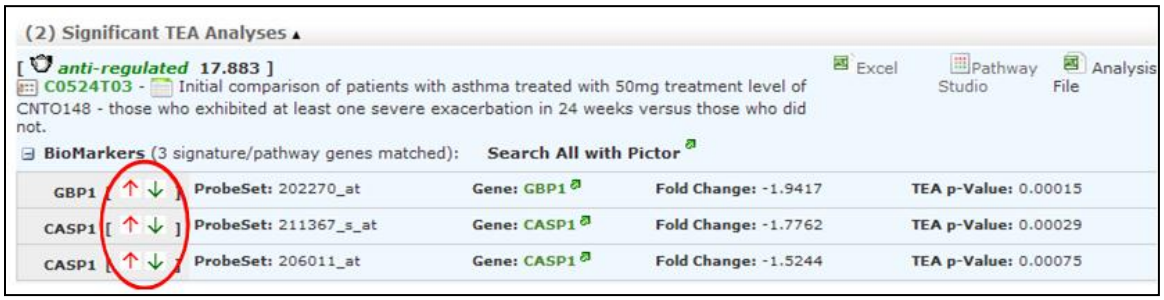
An analysis of a statistically significant experiment returned from a search against a gene signature or list is designated as *co-regulated* or *anti-regulated*. An analysis of a statistically significant experiment returned from a search against a pathway is designated as *up-regulated* or *down-regulated*.

The following table describes what these terms imply in the context of an analysis of a statistically significant experiment:

	Gene Signature/List	**Pathway**
Co-Regulated	Genes that are up-regulated in the signature or list are predominantly up-regulated in the experiment. Genes that are down-regulated in the signature or list are predominantly down-regulated in the experiment.	n/a
Anti-Regulated	Genes that are up-regulated in the signature or list are predominantly down-regulated in the experiment. Genes that are down-regulated in the signature or list are predominantly up-regulated in the experiment.	n/a
Up-Regulated	n/a	Genes in the experiment are predominantly up-regulated.
Down-Regulated	n/a	Genes in the experiment are predominantly down-regulated.

TEA Indicators Applied to an Individual Gene

In an analysis list, TEA indicators for a gene appear as arrows, as shown in the figure below. The leftmost arrow represents the gene expression in the gene signature or list. The rightmost arrow represents the gene expression in the experiment:



Note: The leftmost arrow appears only for gene signatures and gene lists.

The direction of the arrows indicates the following:

- **Up-arrow** – An upward-pointing arrow alongside a gene indicates that the gene is up-regulated in the gene signature/list (leftmost arrow) or in the experiment (rightmost arrow).

  If both arrows point in the same direction, the gene is co-regulated in the signature/list and the experiment. If the arrows point in opposite directions, the gene is anti-regulated.

- **Down-arrow** – A downward-pointing arrow alongside a gene indicates that the gene is down-regulated in the gene signature/list (leftmost arrow) or in the experiment (rightmost arrow).

  If both arrows point in the same direction, the gene is co-regulated in the signature/list and the experiment. If the arrows point in opposite directions, the gene is anti-regulated.

The relationships between TEA indicators for genes and TEA indicators for an experiment are as follows:

- **Co-regulated genes** – Up- or down-regulated genes in the signature/list are similarly up- or down-regulated in the experiment.
- **Anti-regulated genes** – Up- or down-regulated genes in the signature/list are conversely down- or up-regulated in the experiment.

Chapter 3
Dataset Explorer

Dataset Explorer lets you compare data generated for test subjects in two different study groups, based on criteria and points of comparison that you specify. Dataset Explorer is useful to help you test a hypothesis that involves the criteria and points of comparison you select for the comparison.

> **Note:** If you encounter a Java security warning when you attempt to generate a heat map or other Dataset Explorer visualization, see the document Configuration Tips for information.

Overview of the UI

The figure below shows the Dataset Explorer interface. It is divided into two panes:

Left pane

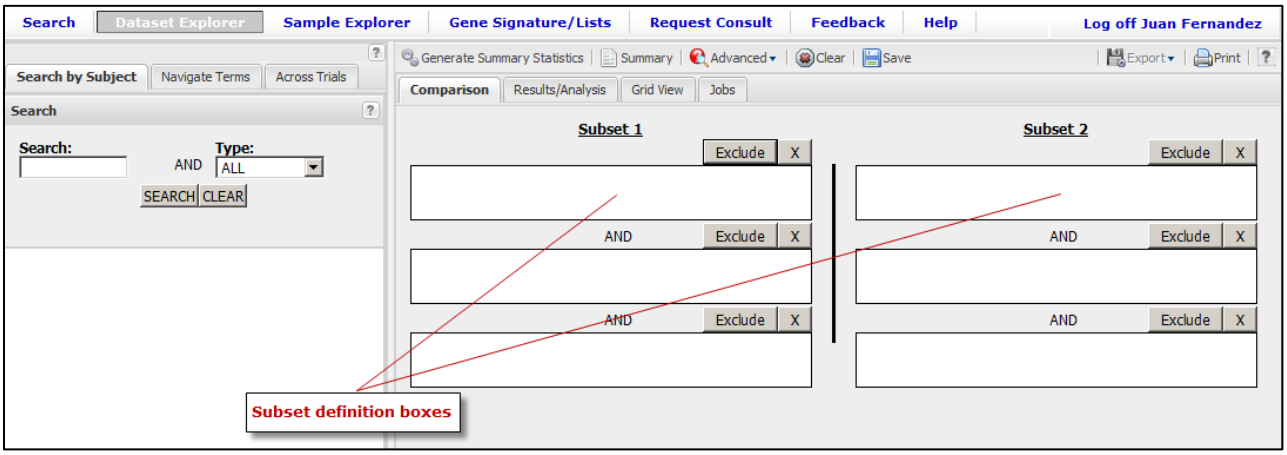
- Lets you select the study of interest.
- Provides a Microsoft Windows Explorer-like navigation tree where you select the criteria for membership in the study groups and the points of comparison between the study groups.

Right pane

- Lets you define the criteria that test subjects must satisfy to become members of one of the two groups being compared. Each of these groups is called a **subset** because it typically contains only some of the subjects in the actual study group involved in the study.

  You define the criteria for the subsets in the subset definition boxes shown below. Subjects who do not satisfy the criteria you define are excluded from the subsets.

- Provides summary data about the subjects being compared, and several different views of the comparison data.



The following table describes the buttons and tabs in the right pane of Dataset Explorer:

Button or Tab	Description
Generate Summary Statistics button	Displays tables and charts that describe demographic information about the subjects in the subsets, and also analyses of criteria included in the subset definitions. The tables and charts are displayed in the Results/Analysis view.
Summary button	Displays a summary of the query criteria you specified. Dataset Explorer uses these criteria to select the subjects for the subsets.
Advanced button	Lets you view subset data in the following ways: ▪ As a heat map of mRNA, RBM, or proteomic data ▪ As a principal component analysis (PCA) of mRNA, RBM, or proteomic data ▪ As a visualization of survival analysis data ▪ As a haploview of SNP data ▪ As a visualization of SNP array data
Clear button	Clears all the criteria in the subset definition boxes.
Save button	Saves the criteria definition. This allows you to re-generate the comparison at a later time without having to reconstruct the criteria that select the subjects for the subsets. For more information, see Saving Comparison Definitions on page 63.

Button or Tab	Description
Export button	Export summary statistics data or expression data to Microsoft Excel after a heat map is generated.
Print button	Print the tables and charts in the Results/Analysis view.
Comparison tab	Removes the currently displayed view (that is, the Results/Analysis view, Grid view, or Haploview) and re-displays the subset definition boxes. This allows you to further refine the subjects for the comparison.
Results/Analysis tab	Displays tables and charts containing comparison and analysis data.
Grid View tab	Displays the comparison and analysis data in grid format.
Jobs tab	Displays previously run analyses. For more information, see Asynchronous Operations on page 130.

Using Dataset Explorer

Four basic tasks are involved in using Dataset Explorer:

- Select the study (clinical trial or experiment) to use in the comparison.
- Specify the criteria for membership in the two study groups.
- Generate summary statistics for the two study groups.
- Specify the points of comparison to apply to the study groups.

Note: You may see the notations **NA** and **Unknown** in the study data. **NA** indicates not applicable, and **Unknown** indicates not available.

Public and Private Studies

Dataset Explorer studies can be either public or private. Public studies are in the **Public Studies** folder of the Dataset Explorer navigation tree. All other studies are private.

You can perform all the operations described in this chapter on public studies. No special privileges are required.

To perform operations described in this chapter on a private study, a tranSMART User Administrator must assign you access rights to the study. Access rights are based on the following access levels:

Access Level	Privileges
VIEW	Define the criteria for the study groups to be compared, generate summary statistics for the study groups, and specify points of comparison for the study groups.
EXPORT	All privileges of the VIEW access level, plus the ability to export comparison data or expression data to a Microsoft Excel spreadsheet.
OWN	All VIEW and EXPORT privileges. This access level can only be assigned to the owner of the study.

If you do not have access rights to the study you want (that is, if the study name is grayed out, as shown in section Navigate Terms Tab on page 58), contact a tranSMART User Administrator. The administrator will contact the study owner to find out if you should be granted VIEW access, EXPORT access, or no access.

> **Note:** Even if you have no access rights to a private study, you can read a description of the study. For information, see Viewing a Study on page 129.

Selecting the Study

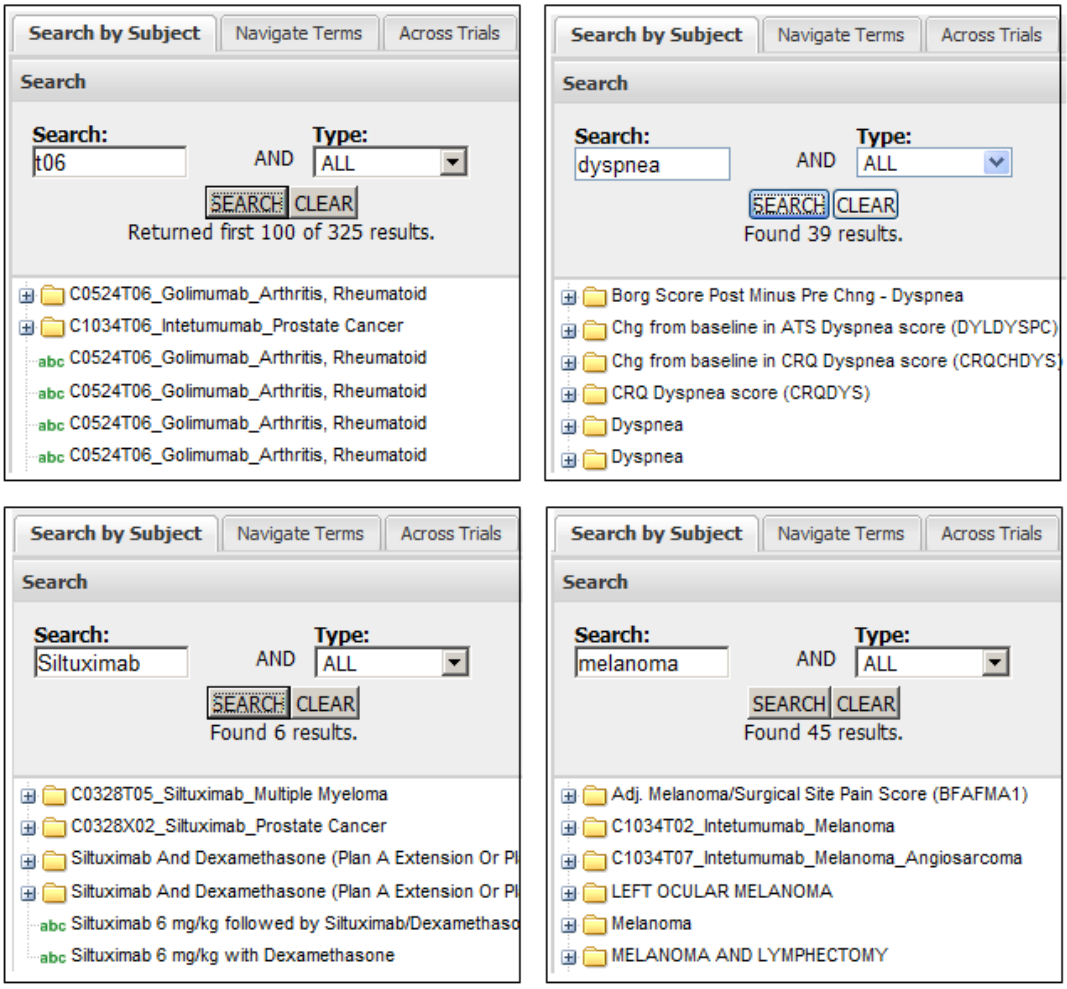
You select the study in the left pane of Dataset Explorer. You have several ways to select the study, based on the tab you choose – Search by Subject or Navigate Terms.

Search by Subject Tab

Use this tab to search for studies using one or a combination of the following fields:

- **Search** field. Lets you specify part or all of a term from a study that is stored in the tranSMART database. Search terms may include part or all of a study name, the text in a branch of the Dataset Explorer navigation tree, or some attribute of a study, such as a compound, a disease, or an area of clinical interest.

  Examples:



If you want to base your search on a study name, note the following naming conventions for studies in Dataset Explorer:

Study Type	Naming Convention
Clinical Trial	Johnson & Johnson trial in the following format: *TrialID_Compound_Condition* Example: `C0328T05_Siltuximab_Multiple Myeloma` In some cases, *TrialID* might include the compound ID.

Experimental Medicine Study	Format may include Johnson and Johnson trial ID or a descriptive name of the study.
	Examples:
	`BRC Antidepressant Study`
	`C0743X01_Sarcoidosis`
	`SebastianJohnston_Asthma_Study`

Internal Studies	Format may include compound name or the study sponsor.
	Examples:
	`Veridex_BreastCancer_2003`
	`A431CellLineSCC_CNTO2559`

Public Studies	Name segments in the following typical format: *StudyFirstAuthor_Condition_GEOid* Example: `Ambs_ProstateCancer_GSE6956`

- **Type** field. Lets you specify a study based on one of the criteria listed in the table below. When you specify a type, a **Terms** dropdown appears, allowing you to further specify the kind of study you're interested in:

Study Type	Study Attributes
ALL	All types of studies.
AREA	Study categories such as: ▪ Cardiovascular ▪ Immunology; Neuroscience ▪ Oncology ▪ Psychiatry
COMPOUND	A compound tested in the study – for example, CNT095, Golimumab, or Infliximab.
DISEASE	A disease of interest in the study – for example, asthma, COPD, depression.

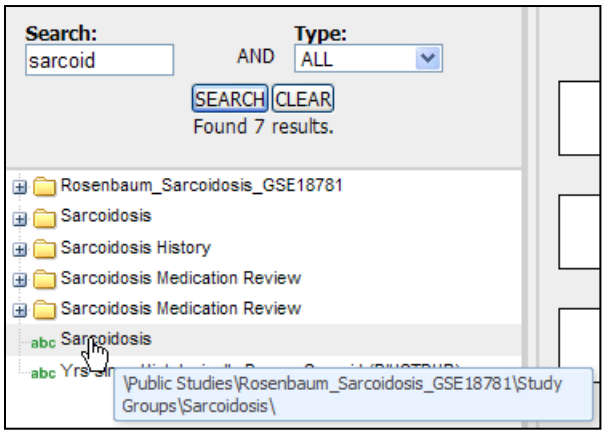
Study Type	Study Attributes
WORKFLOW	A study that involved a particular kind of biomarker, such as gene expression, RBM, SNP.

After you specify the search criteria, click **Search** to run the search.

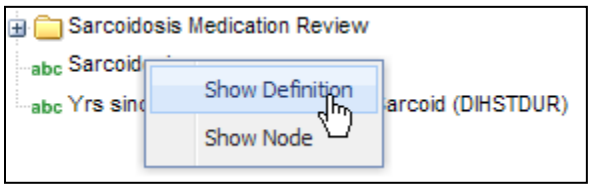
Click **Clear** to remove any existing search criteria and begin a new search.

Selecting and Opening a Study in a Search Result

A search result may include multiple entries. Further, an entry may not indicate the study it is from. To see the name of the study that an entry represents, hover the mouse pointer over the entry – for example:



If you want more information about the study represented by an entry, right-click the entry, then click **Show Definition** to open the details box for the study:



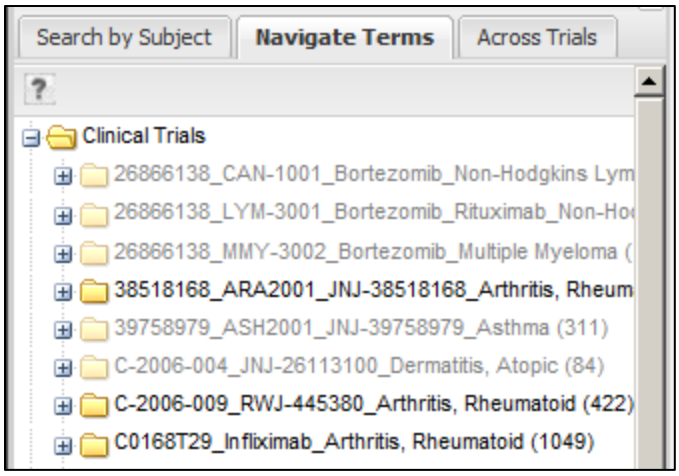
To open a study from an entry in a search result, right-click the entry, then click **Show Node**. The study appears in the Dataset Explorer navigation tree, where you can open any of the branches (nodes) in the study.

> **Note:** You may need to scroll down slightly in the navigation tree to see the study.

Navigate Terms Tab

Use this tab to browse through all the clinical trials and experiments in the navigation tree to select and open the study you want.

Studies that are grayed out are private studies that you are not authorized to access.



To display the details box for a study, right-click the study name and click **Show Definition**. You can display the details box for a study whether or not the study is grayed out.

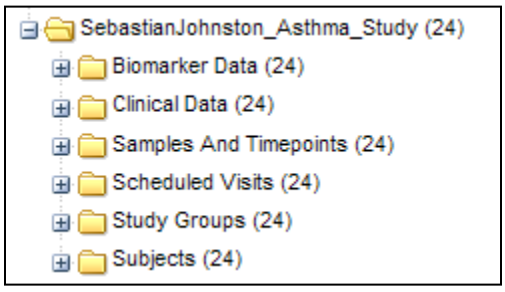
> **Note:** The Release Notes contains a list of all available studies, with the most recently added studies indicated in red text.

Branches and Leaves of the Navigation Tree

The Dataset Explorer navigation tree looks and works much like the Microsoft Windows Explorer. Windows Explorer is a hierarchy of folders, sub-folders, and files. Dataset Explorer is a hierarchy of folders and sub-folders (the branches) and values (the leaves) that reflect aspects of the trial, such as research metrics, compounds used, and patient demographics.

In Dataset Explorer, all levels of the tree, both branches and leaves, are referred to as nodes.

The following figure shows typical top-level nodes of a study. Some studies may not require all of these nodes, and others may require additional nodes (such as Published Conclusions):



The following table describes these top-level nodes of a study:

Node	Description
Biomarker Data	Measurements of biomarkers such as RBM antigens, gene expressions, antibodies and antigens in ELISA tests, and SNPs.
Clinical Data	Primary and secondary endpoints, and other measurements from the study.
Samples and Timepoints	Tested samples (such as tissue or blood) and time periods when the samples were taken.
Scheduled Visits	Periodic stages of the trial during which patients are seen.
Study Groups	Compounds involved in the study, dosages, and regularity with which the compounds were administered. **Note:** With clinical trials, this node is typically named Treatment Groups.
Subjects	Patient information, such as demographics and medical history.

Populating the Study Groups

You populate the study groups by defining criteria that members of each group must satisfy. For example, members of study groups might be required to satisfy a weight or age requirement. Dataset Explorer lets you build a set of criteria for each study group that can be as simple or as complex as you need.

The study groups you define are called *subsets*, because typically, after your criteria are applied, the members of a resulting study group are a subset of the actual study group that participated in the study.

Selecting Criteria for the Study Groups

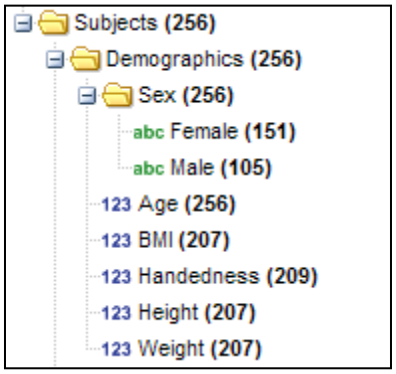
You define the study groups by selecting criteria (called concepts) from the navigation tree and dragging them into the subset definition boxes.

Visual Aids to Help You Select the Criteria

Each concept node in the navigation tree displays the following information about the concept:

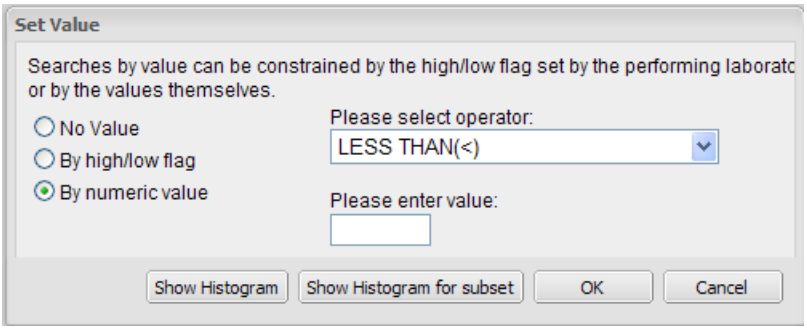
- The numbers in parentheses at each node of the tree indicate the number of subjects to whom that node applies. For example, in the figure below, there are a total of 256 subjects in the study, 151 females and 105 males. Further, height and weight measurements were taken for only 207 of the subjects.

- Some nodes have the icon **abc** before them, and others have the icon **123**.
  - **abc** refers to a concept that is non-numeric – for example, gender.
  - **123** refers to a concept that is numeric – for example, age, height, weight.



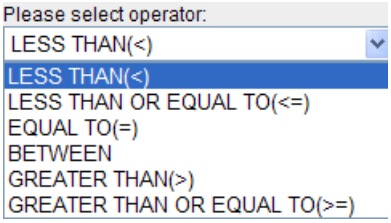
Specifying a Numeric Value

When you drag a non-numeric concept into a subset definition box, the concept immediately becomes a part of the subset's definition. But when you drag a numeric concept into a subset definition box, the Set Value dialog appears:

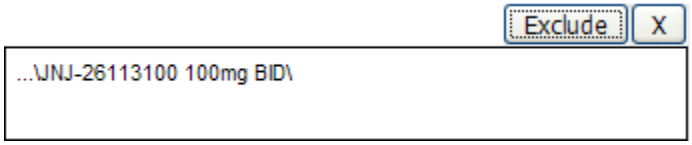


Use the Set Value dialog to specify how you want to constrain the numeric values to use in the subset definition. To do so, first select one of the following choices:

Selection	Description
No Value	Values are not constrained. All the numeric data associated with the concept are factored into the subset definition. If you select **No Value**, no other information is required. Click **OK** to add the concept with all its associated numeric data to the subset.
By high/low flag	If the testing laboratory has grouped the numeric values into High/Low/Normal ranges, select the range to factor into the subset definition. When you select **By high/low flag**, the **Please select range** field appears. Select the range you want and click **OK**.

Selection	Description
By numeric value	Values are constrained by an exact value or a range of values. After you select **By numeric value**: ■ Select one of the following numeric operators in the **Please select operator dropdown**:  ■ **In Please enter value**, type the numeric value that the operator applies to. For example, to constrain the ages of subjects to 50 years or younger, select `LESS THAN OR EQUAL TO(<=)` in the dropdown, then type `50` in the **Please enter value** field. ■ **Click OK.** See the next section for information on viewing the numeric values associated with the concept and that you can select from.

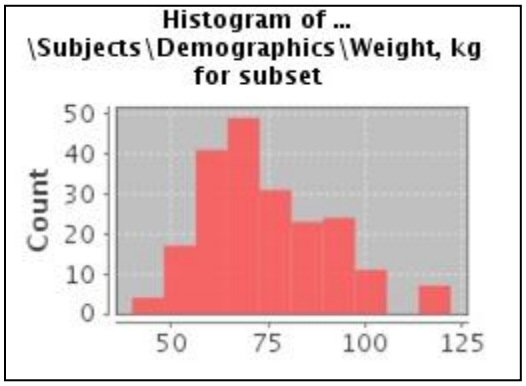
Note: When finished defining the numeric constraint on the Set Value dialog, be sure to click **OK** and not press the **Enter** key. Pressing **Enter** will activate the subset button that has focus – the **Exclude** button in the example below:



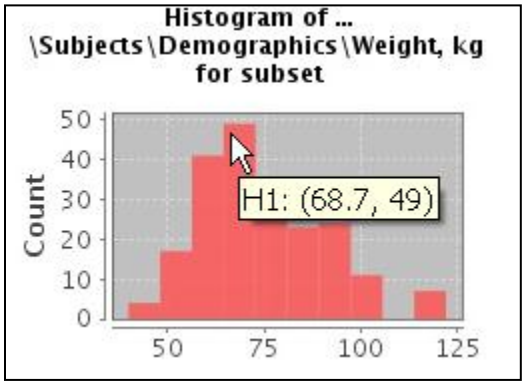
Viewing the Numeric Values Associated with a Concept

Note the buttons **Show Histogram** and **Show Histogram for subset** in the Set Value dialog. The histograms show how the numeric values associated with the concept that you placed in the subset box are distributed among the subjects across both subsets, or in the particular subset you are currently defining, respectively.

A histogram may be helpful in determining the number to set as the constraining factor for a concept. For example, suppose you drag a Weight concept into a subset box, then click **Show Histogram for subset**. In the following histogram of the weights of test subjects, the weights range from about 25 kg to just under 125 kg. The largest bin represents just under 50 subjects. You may want to use these weight parameters to help you determine the value to set for the weight concept.



You can get more specific information about the number of subjects represented by a particular bin and the average of the values in the bin by hovering the mouse cursor over the bin you are interested in. For example, in the following figure, the largest bin represents 49 subjects with an average weight of 68.7 kg:



Saving Comparison Definitions

You may save your search criteria in order to re-generate the comparison at a later time without having to redefine the subsets.

To save search criteria:

1. Run tranSMART, then click the **Dataset Explorer** tab.
2. Select the study of interest.
3. Define the cohorts whose data points will be represented.
4. Click **Save**.



5. Click **Email this comparison**.



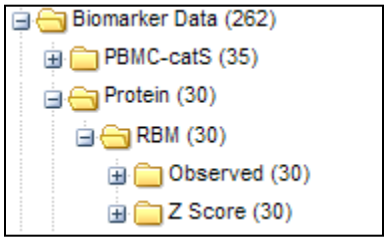
   Your email application will open with a link to the saved comparison.

6. Send the email to yourself so that you can retrieve the comparison later. Optionally, send it to colleagues who might be interested in the comparison.

   When you or someone else clicks the link in the email, Dataset Explorer opens with the subset boxes pre-defined.

Observed Score and Z Score

When you select a concept based on RBM data, you have a choice of viewing the collected data as observed values (the actual values that were recorded during testing) or z-score values (the z-score representations of the observed values):



When you generate a heat map with RBM values, the values are represented as z-scores.

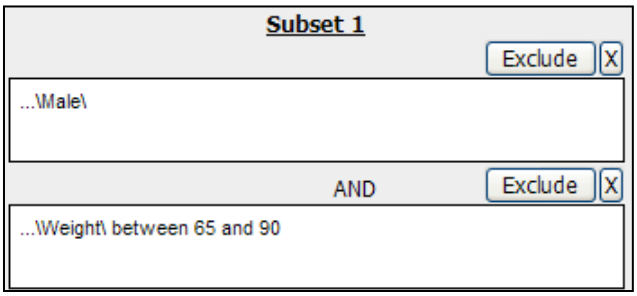
Joining Multiple Criteria for a Subset Definition

Multiple criteria for a subset definition are joined by one of the following logical operators: AND, OR, or AND NOT.

The rules for joining multiple criteria are as follows:

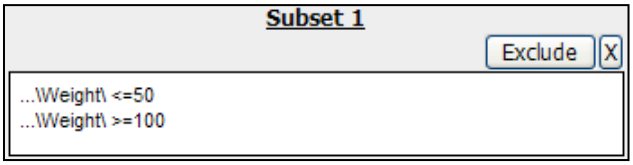
- Criteria in separate subset definition boxes are joined by an AND operator.

  For example, the following definition boxes select only male subjects, AND males whose weights are between 65 kg and 90 kg:



- Criteria within the same subset definition box are joined by an OR operator.

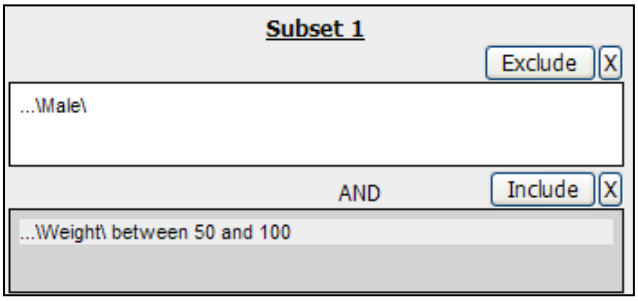
  For example, to use the extreme ends of the weight scale for your weight criterion, you might add the following to a definition box:



  This criterion selects subjects whose weight is either 50 kg or less, OR 100 kg or greater.

- To join a definition box with an AND NOT operator, click the **Exclude** button above the definition box.

  The figure below selects only male subjects, but not those who weigh between 50 kg and 100 kg:

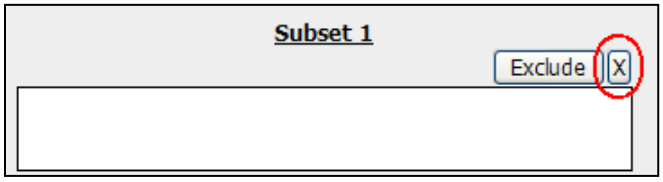


  Note that when you click the **Exclude** button, the button label changes to **Include**, allowing you to join the criteria in the box with an AND operator later if you choose.

Modifying or Deleting Criteria

To delete or modify a criterion in a subset definition box, right-click the criterion and select either **Delete** or **Set Value**.

To remove the entire contents of a subset definition box from the subset definition, click the **X** icon (X) above the box:



Generating Summary Statistics

When you finish defining criteria for the groups to compare – the subsets – click the **Generate Summary Statistics** button.

tranSMART displays tables and charts of information that describe the subsets. The information is displayed in the Results/Analysis view in the following sections:

- A summary of the criteria used to define subsets to compare. Example:

Query Summary for Subset 1	Query Summary for Subset 2
(\\Experimental Medicine Study\Experimental Medicine Study\BRC Depression Study\Subjects\Demographics\Sex\Male\ ) **AND** (\\Experimental Medicine Study\Experimental Medicine Study\BRC Depression Study\Subjects\Demographics\Weight\ BETWEEN 65 and 90)	(\\Experimental Medicine Study\Experimental Medicine Study\BRC Depression Study\Subjects\Demographics\Sex\Male\ ) **AND** (\\Experimental Medicine Study\Experimental Medicine Study\BRC Depression Study\Subjects\Demographics\Weight\ BETWEEN 80 and 120)

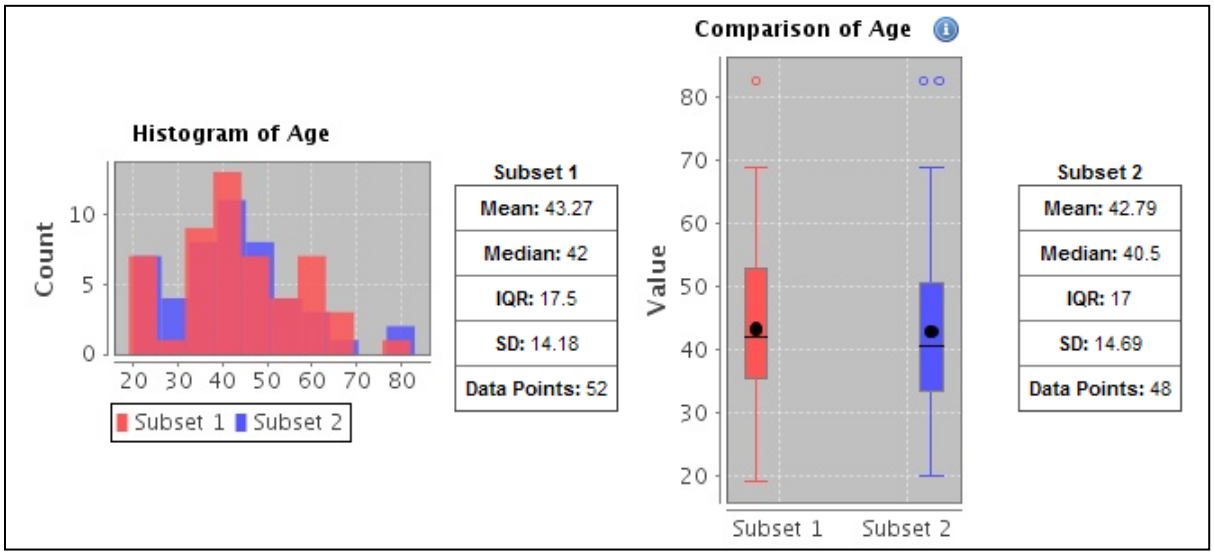
- A table showing the number of subjects in each subset who match the subset criteria. Example:

Subject Totals

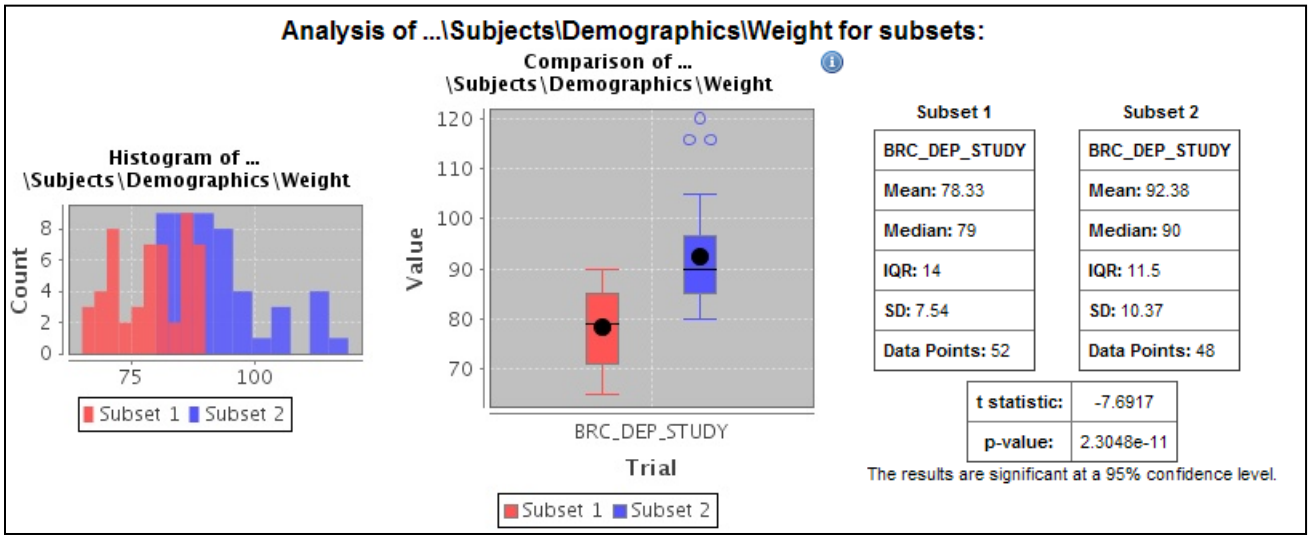
Subset 1	Both	Subset 2
52	25	48

In this example, 52 subjects matched the criteria for Subset 1, and 48 matched the criteria for Subset 2. Further, 25 subjects matched the criteria for both subsets (and thus, were included in both).

- Tables and charts that show how the subjects who match the criteria fit into age, sex, and race demographics. Example (showing the age portion only):



- Analyses of the concepts you added to the subsets from the navigation tree. Example (showing the weight concept):



Significance Tests

The above figure includes the results of significance testing that Dataset Explorer performs:

t statistic:	-7.6917
p-value:	2.3048e-11

The results are significant at a 95% confidence level.

Significance testing is designed to indicate whether the reliability of the statistics is 95% or greater, based on p-value.

Dataset Explorer calculates the significance result using either t-test or chi-squared statistics to determine the p-value:

- For continuous variables (for example, subject weight or age), a t-test compares the observed values in the two subsets.

  tranSMART uses the following Java method to calculate the t-test statistic:

  http://commons.apache.org/math/apidocs/org/apache/commons/math/stat/inference/TTest.html#tTest(double[],%20double[])

- For categorical values (for example, diagnoses), a chi-squared test compares the counts in the two subsets.

  tranSMART uses the following Java method to calculate the chi-squared statistic:

  http://commons.apache.org/math/apidocs/org/apache/commons/math/stat/inference/ChiSquareTest.html#chiSquareTest(long[][])

If there is not enough data to calculate a test, Dataset Explorer displays a message indicating the lack of data. Also, significance test results are not displayed in the following circumstances:

- If two identical subsets are defined. In this case, the significance test results are not meaningful.
- If all subjects in the first subset have one set of values for the categorical value, and all subjects in the second subset have other categorical values. For example, suppose you set Subset 1 to contain only males and Subset 2 to contain only females. Also, suppose that Subset 1 has 15 subjects and Subset 2 has 20. If you then try to show statistics by gender, a table like the following would result:

	Subset 1	**Subset 2**
Female	0	20
Male	15	0

  In this case, the chi-squared function doesn't return meaningful results.

Defining Points of Comparison

Once you establish the subsets of subjects that you want to compare, you can apply one or more points of comparison to the subsets.

A point of comparison is a concept in the navigation tree.

To apply a point of comparison to the subsets:

1. You must already have defined the subsets and have generated summary statistics for the subsets, as described in the previous section.
2. Drag the concept that you want to introduce as the point of comparison from the navigation tree, and drop it anywhere in the Results/Analysis view.

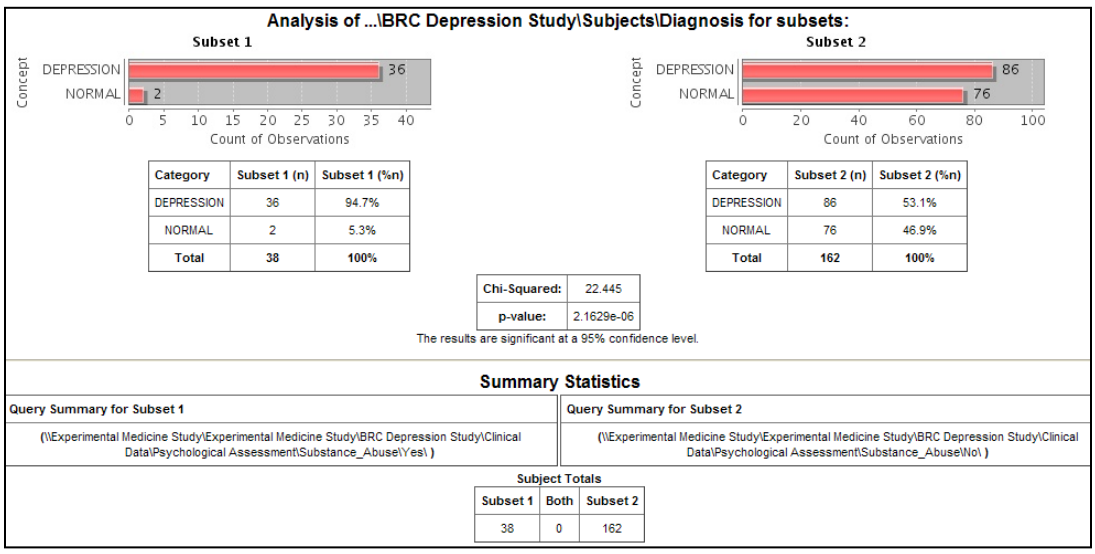
As soon as you drop the point of comparison into the Results/Analysis view, tranSMART begins to compare the subsets based on that point of comparison. When finished, tranSMART displays a side-by-side summary of how the subjects in each subset match or respond to the point of comparison.

Results of a Comparison

In a comparison of subjects in a BRC depression study, suppose Subset 1 contains subjects with a substance abuse problem, and Subset 2 contains subjects with no substance abuse assessment.

After the subsets are defined and summary statistics are generated, a diagnosis of Depression is dropped into the Results/Analysis view as a point of comparison. tranSMART displays a side-by-side comparison of the subjects in each subset, indicating that almost all the subjects with a substance abuse problem have been diagnosed with depression, while that diagnosis for those with no substance abuse problem is more evenly split.

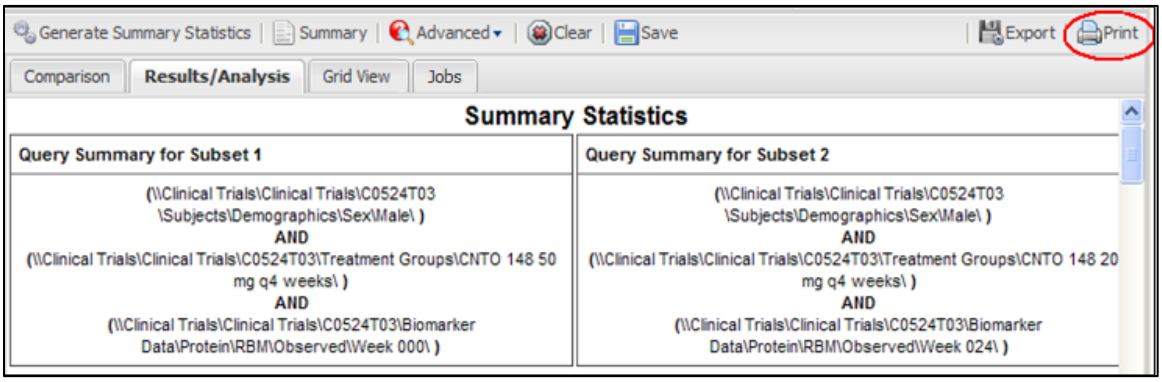
The comparison is placed at the top of the Results/Analysis view, above the demographic definitions plus any other earlier comparisons:



Note: To keep the size of the preceding figure within production limits, the demographics (age, sex, and race) portions of the figure have been excluded.

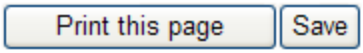
Printing or Saving the Contents of the Results/Analysis View

1. With the Results/Analysis view displayed, click **Print**.



   The entire contents of the Results/Analysis view appears in a separate browser window.

2. Click one of the following buttons at the top of the browser window:



Copying Individual Charts in the Results/Analysis View

If you are interested in a particular chart in the Results/Analysis View, you can copy the chart to a file, as follows:

1. With the Results/Analysis view displayed, click **Print**.

   The entire contents of the Results/Analysis view appears in a separate browser window.

2. Right-click the chart to copy.
3. In the Internet Explorer popup menu, click **Save Picture As**.
4. In the Save Picture dialog, specify the name, location, and the file type for the chart.
5. Click **Save**.

Generating Heat Maps

A heat map generated in Dataset Explorer represents different kinds of data than a heat map generated from a clinical trial search result:

- In Dataset Explorer, a heat map is a matrix of data points for a particular set of biomarkers, such as genes or RBM antigens, at a particular point in time and/or for a particular tissue sample in the study, as measured for each subject in the study.

  Up-regulation is expressed in shades of red. Down-regulation is expressed in shades of blue.

- In a clinical trial search result, a heat map is a matrix of data points for a particular set of genes, as measured in one or more analyses of studies.

  Up-regulation is expressed in shades of red. Down-regulation is expressed in shades of green.

In a Dataset Explorer heat map, the biomarkers appear in the y axis, and the subjects appear in the x axis.

> **Note:** A heat map can display data points for up to 1000 samples.

Types of Heat Maps

You can generate the following types of heat maps:

- Standard heat map – A visualization of biomarker data points (gene expression, protein expression, or RBM), with no indication of patterns, groupings, or differentiation among the data points.

  To generate, select **Heatmap** from the **Advanced** menu.

- Class discovery (hierarchical clustering) heat map – A visualization of patterns of related data points in gene expression or RBM data.

  To generate, select **Hierarchical Clustering** from the **Advanced** menu.

- Class discovery (k-means clustering) heat map – A visualization of groupings of the most closely related data points, based on the number of groupings you specify.

  To generate, select **K-Means Clustering** from the **Advanced** menu.

- Differential Analysis/Marker Selection heat map – A visualization of differentially expressed genes in distinct phenotypes.

  To generate, select **Comparative Marker Selection** from the **Advanced** menu.

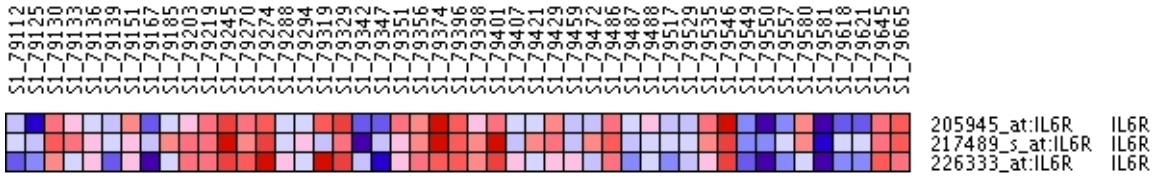
Dataset Explorer uses the Broad Institute's GenePattern genomic analysis platform to generate heat map visualizations.

You may run multiple visualizations in the background while continuing to use Dataset Explorer. For more information on running visualizations in the background, see Asynchronous Operations on page 130.

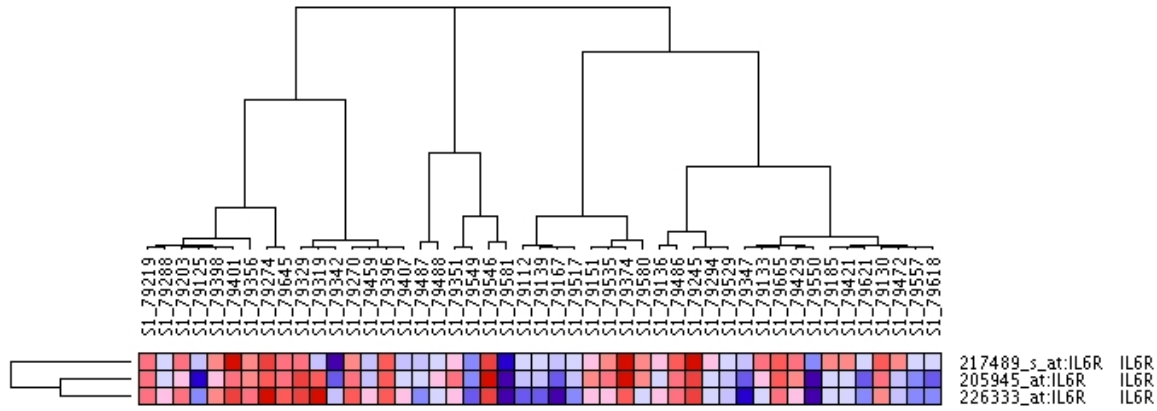
Examples

The following figures illustrate the different types of heat maps you can generate in Dataset Explorer. All were generated from the public study Shaughnessy Multiple Myeloma (GSE2658). The first three examples are visualizations of the gene IL6R in the proliferation group of the study. The fourth example is a Comparative Marker Selection heat map that compares the proliferation group with the MAF/MAFB gene overexpression group.

Standard Heat Map

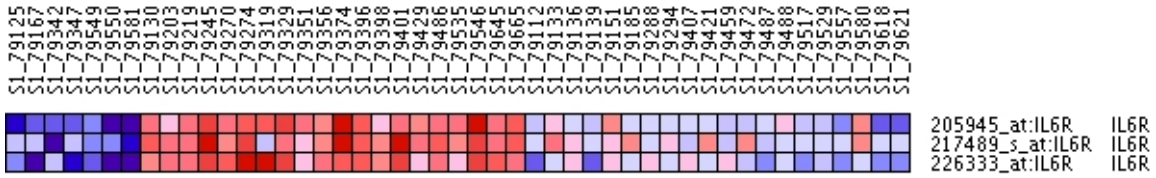


Class Discovery Heat Map – Hierarchical Clustering



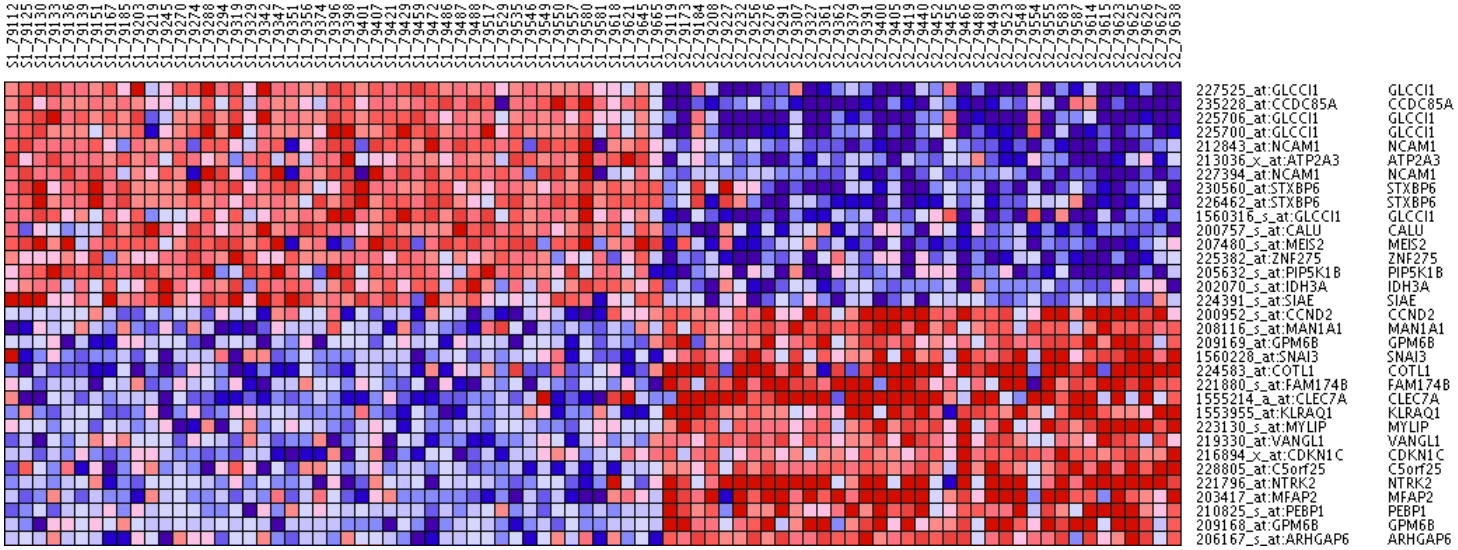
Class Discovery Heat Map – K-Means Clustering

Three clusters:



Comparative Marker Selection Heat Map

Partial view:



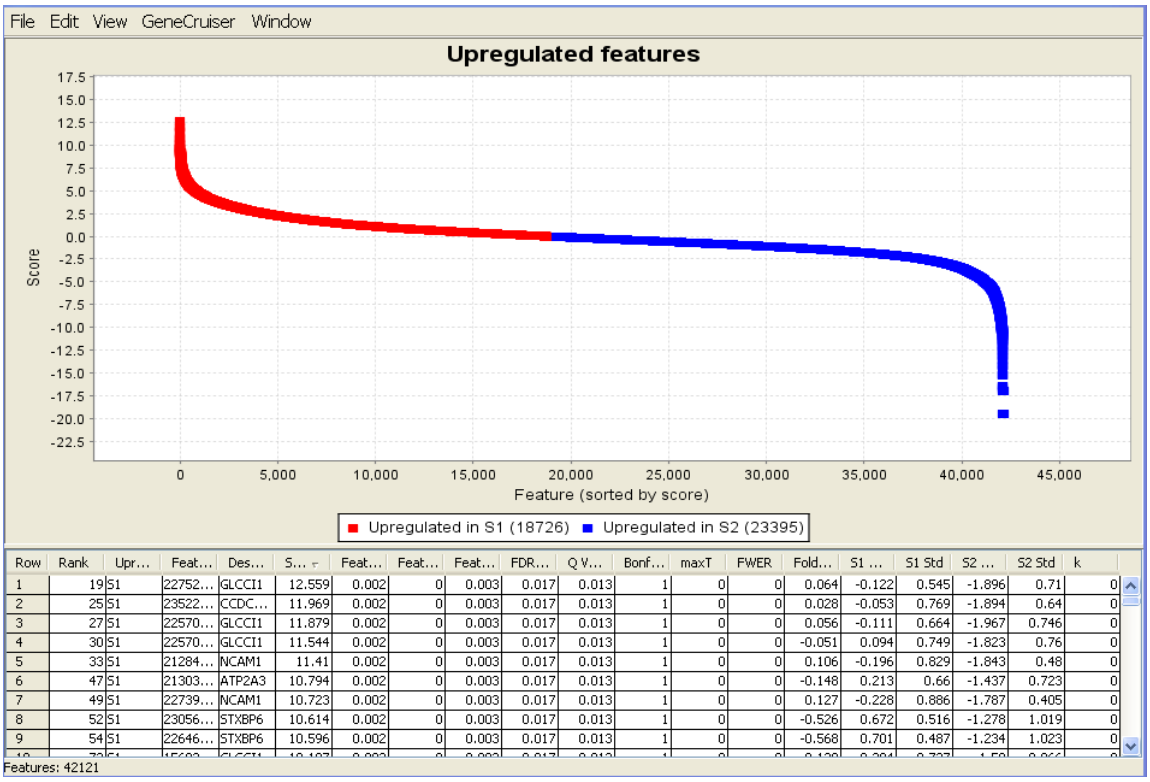
The ComparativeMarkerSelectionViewer

This heat map requires that selection criteria be defined in both subsets.

When you generate a Comparative Marker Selection heat map, the heat map is displayed along with the Broad Institute's GenePattern ComparativeMarkerSelectionViewer. The viewer provides a number of tools and visualizations, including a heat map, for analyzing Comparative Marker Selection results.

> **Note:** Due to the large amount of data being processed, the ComparativeMarkerSelectionViewer may take several minutes to appear.

When the ComparativeMarkerSelectionViewer appears, the viewer's Upregulated Features graph is displayed by default, and a grid containing the Comparative Marker Selection statistical results appears below it:



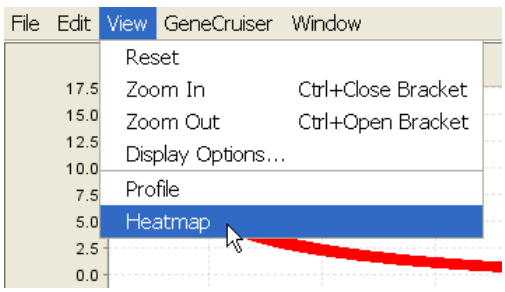
Row	Rank	Upr...	Feat...	Des...	S...	Feat...	Feat...	Feat...	FDR...	Q V...	Bonf...	maxT	FWER	Fold...	S1 ...	S1 Std	S2 ...	S2 Std	k
1	19	S1	22752...	GLCCI1	12.559	0.002	0	0.003	0.017	0.013	1	0	0	0.064	-0.122	0.545	-1.896	0.71	0
2	25	S1	23522...	CCDC...	11.969	0.002	0	0.003	0.017	0.013	1	0	0	0.028	-0.053	0.769	-1.894	0.64	0
3	27	S1	22570...	GLCCI1	11.879	0.002	0	0.003	0.017	0.013	1	0	0	0.056	-0.111	0.664	-1.967	0.746	0
4	30	S1	22570...	GLCCI1	11.544	0.002	0	0.003	0.017	0.013	1	0	0	-0.051	0.094	0.749	-1.823	0.76	0
5	33	S1	21284...	NCAM1	11.41	0.002	0	0.003	0.017	0.013	1	0	0	0.106	-0.196	0.829	-1.843	0.48	0
6	47	S1	21303...	ATP2A3	10.794	0.002	0	0.003	0.017	0.013	1	0	0	-0.148	0.213	0.66	-1.437	0.723	0
7	49	S1	22739...	NCAM1	10.723	0.002	0	0.003	0.017	0.013	1	0	0	0.127	-0.228	0.886	-1.787	0.405	0
8	52	S1	23056...	STXBP6	10.614	0.002	0	0.003	0.017	0.013	1	0	0	-0.526	0.672	0.516	-1.278	1.019	0
9	54	S1	22646...	STXBP6	10.596	0.002	0	0.003	0.017	0.013	1	0	0	-0.568	0.701	0.487	-1.234	1.023	0

> **Note:** The viewer's export function (**File > Saved Derived Dataset)** does not export all values – for example, p-values are not exported. To export the values you want, copy rows/columns from the result grid and paste them into a spreadsheet.

For information on the ComparativeMarkerSelectionViewer's data grid and the tools available with the viewer, see the Broad Institute's *ComparativeMarkerSelectionViewer* document.

To display the heat map from the ComparativeMarkerSelectionViewer:

- Click **View > Heatmap**:



Interactive Heat Maps

Dataset Explorer heat maps generated with Internet Explorer version 8 are interactive.

With an interactive heat map, you can select particular samples and cohorts of interest, then right-click to select among a variety of charts in which to view the data: Profile, Centroid Plot, Histogram, Nearest Neighbors, and Scatter Plot. You can also save the selected data points to a file.

In **Figure 1** below, sample 205945_at has been selected. The user right-clicks the sample name to display the pop-up menu, then clicks the Profile view.

Figure 2 displays the result – a Profile of data points for all cohorts for the sample 205945_at.

> **Note:** If no cohorts are selected, all will be included in the view.

Figure 1

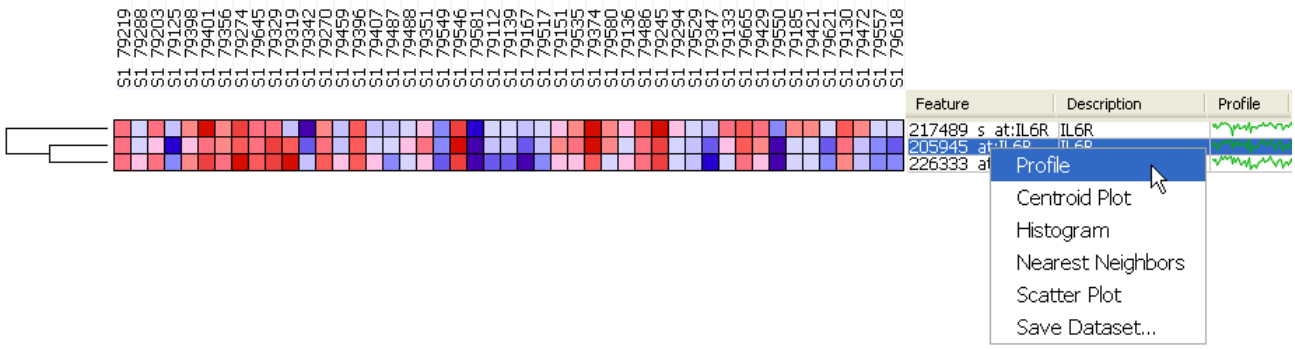
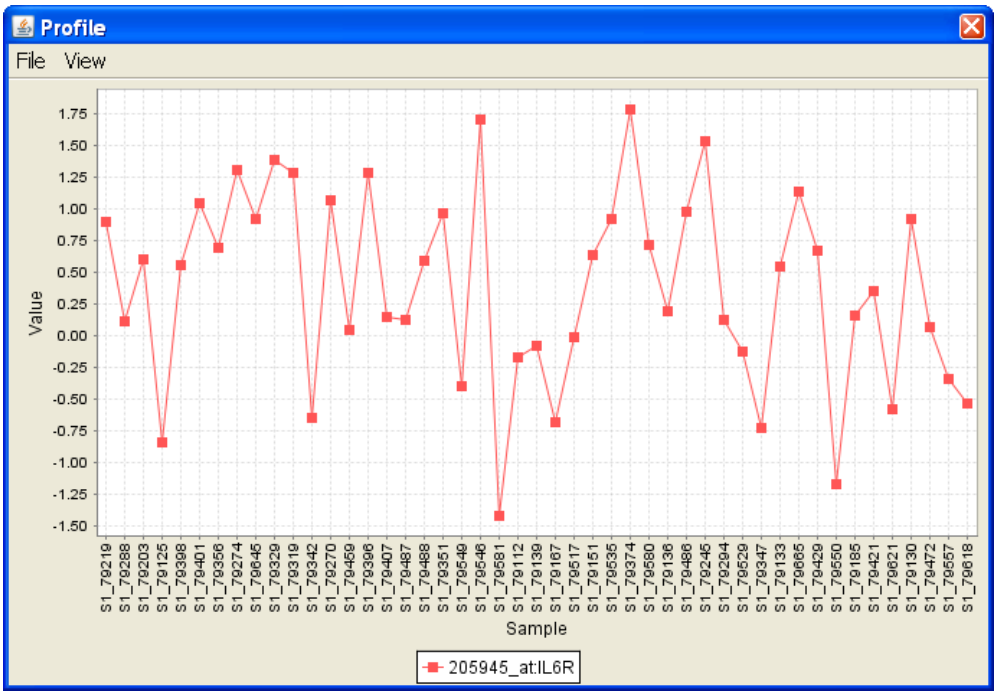


Figure 2



For information on using interactive heat maps, see these Broad Institute documents:

Heat Map Type	Broad Institute Documents
Standard and k-means	HeatMapViewer
Hierarchical clustering	HierarchicalClusteringViewer
Comparative Marker Selection	ComparativeMarkerSelectionViewer

Requirements for Generating Heat Maps

The following table shows the required and optional parameters you specify for each type of heat map. Except where noted in the table, these parameters can be provided in the subset definition boxes, in the Compare Subsets-Pathway Selection dialog (which appears after you select the type of heat map to generate), or in any combination of these locations.

Note: If you do not specify a value for a field, including an optional field, all possible values for that field will be factored into the heat map.

Heat Map Type	Platform	Parameters You Specify
Standard and hierarchical clustering	RBM	▪ Platform (RBM). ▪ Timepoint (for example, Baseline or Week 004). ▪ Optionally, a particular antigen measurement by which to filter the dataset (subset boxes only). ▪ Optionally, a gene or pathway by which to filter the dataset (Compare Subsets-Pathway Selection dialog only).
	mRNA	▪ Platform (MRNA). ▪ GPL Platform. The specific GEO platform – for example, Affymetrix GeneChip Human HGFocus Target Array. ▪ Sample. ▪ Tissue type (optional). ▪ Timepoint (for example, Baseline or Week 004 – not applicable in all cases). ▪ A gene or pathway by which to filter the dataset (Compare Subsets-Pathway Selection dialog only). Optional for standard heat maps, required for hierarchical clustering.
	Proteomics	▪ Platform. ▪ Timepoint (for example, Baseline or Week 004). ▪ Optionally, a gene or pathway by which to filter the dataset (Compare Subsets-Pathway Selection dialog only).

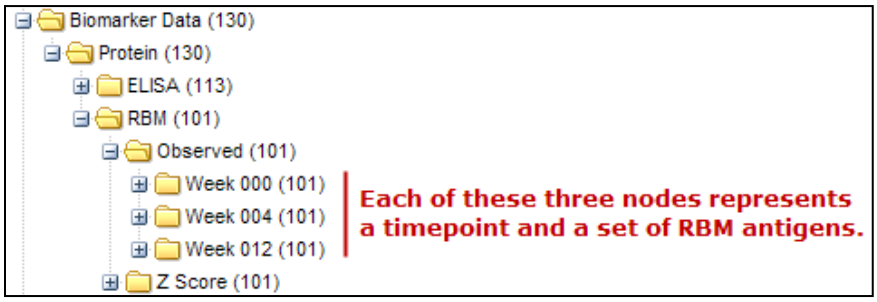
Heat Map Type	Platform	Parameters You Specify
K-means	RBM	▪ Platform (RBM). ▪ Timepoint (for example, Baseline or Week 004). ▪ The number of clusters to display (Compare Subsets-Pathway Selection dialog only). ▪ Optionally, a particular antigen measurement by which to filter the dataset (subset boxes only). ▪ Optionally, a gene or pathway by which to filter the dataset (Compare Subsets-Pathway Selection dialog only).
	mRNA	▪ Platform (MRNA). ▪ GPL Platform. The specific GEO platform – for example, Affymetrix GeneChip Human HGFocus Target Array. ▪ Sample. ▪ Tissue type (optional). ▪ Timepoint (for example, Baseline or Week 004 – not applicable in all cases). ▪ The number of clusters to display (Compare Subsets-Pathway Selection dialog only). ▪ Optionally, a gene or pathway by which to filter the dataset (Compare Subsets-Pathway Selection dialog only).
	Proteomics	▪ Platform. ▪ Timepoint (for example, Baseline or Week 004). ▪ The number of clusters to display (Compare Subsets-Pathway Selection dialog only). ▪ Optionally, a gene or pathway by which to filter the dataset (Compare Subsets-Pathway Selection dialog only).
Comparative Marker Selection (requires two subsets to be defined)	RBM	▪ Platform (RBM). ▪ Timepoint (for example, Baseline or Week 004). ▪ Optionally, a particular antigen measurement by which to filter the dataset (subset boxes only).
	mRNA	▪ Platform (MRNA). ▪ GPL Platform. The specific GEO platform – for example, Affymetrix GeneChip Human HGFocus Target Array. ▪ Sample. ▪ Tissue type (optional). ▪ Timepoint (for example, Baseline or Week 004 – not applicable in all cases).
	Proteomics	▪ Platform ▪ Timepoint (for example, Baseline or Week 004).

Providing Heat Map Parameters in the Subset Boxes

You can drag the following parameters from the Biomarker Data node of the navigation tree into the subset boxes:

- Platform
- mRNA samples/tissues
- Timepoints
- RBM antigens

Typically, the timepoint and the biomarker platform are represented by the same node of the navigation tree – for example:

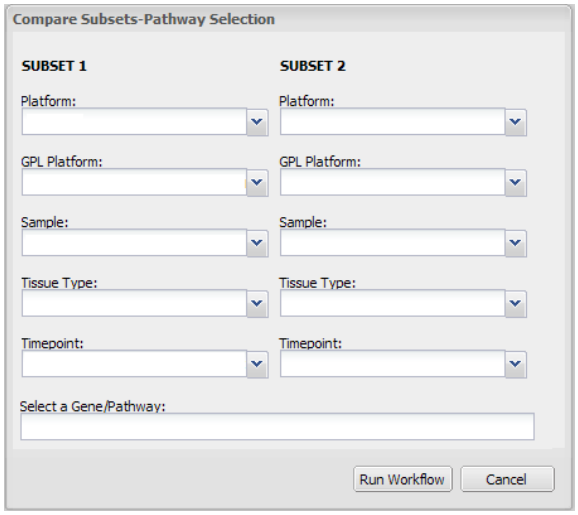


In some cases, the same biomarker platform/timepoint nodes appear in both the **Biomarker Data** and the **Samples and Timepoint** branches of the navigation tree. In those cases, the platform/timepoint node from either branch yields the same heat map results.

Providing Heat Map Parameters in the Compare Subsets-Pathway Selection Dialog

This dialog allows you to vary the input data for heat maps without having to modify the criteria in the subset definition boxes that define the cohorts.

The fields in this dialog will differ, depending on the type of heat map you select. The dialog below appears when you select a standard heat map for an mRNA platform:



Keep in mind these general rules about using the dialog:

- If you have added platform, sample, and/or timepoint criteria to the subset definition(s), Dataset Explorer will attempt to use your selections as default values in the **associated fields of the dialog**.
- You can select one or more values in the **GPL Platform**, **Sample, Tissue Type**, and **Timepoint** fields.
  - To select multiple values in a **GPL Platform**, **Sample, Tissue Type**, or **Timepoint** field, click each value that you want to include. You don't need to hold down the Ctrl key or any other key to select multiple values.
  - To deselect a sample or timepoint value, click it without pressing any other keys.
  - If you deselect all values in one of these fields, all possible values for that field will be factored into the heat map.
- You can make only one selection in a **Platform** field. If you have two subsets defined, the platform must be the same in the **Platform** field for each subset.
- You may run multiple visualizations in the background while continuing to use Dataset Explorer. For more information on running visualizations in the background, see Asynchronous Operations on page 130.

Note: Before clicking **Run Workflow** to generate the heat map, be sure that all fields are defined as you expect. Sometimes a dropdown box will hide one or more fields below it. To "roll-up" the dropdown box, click in an open (non-field) area of the dialog.

Instructions for Generating Heat Maps

The following sections describe how to generate heat maps in Dataset Explorer.

Heat Maps Based on RBM Data

To generate a heat map based on RBM data:

1. Define one or both subsets, as described earlier in this chapter.

   Note: If you intend to generate a Comparative Marker Selection heat map, you must define both subsets.

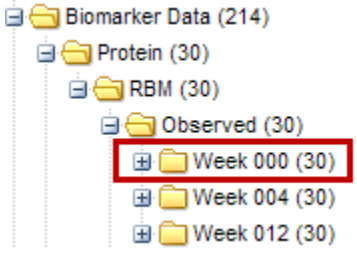
2. In the navigation tree, select one or more timepoints for the RBM platform and drag them into the subset definition boxes.

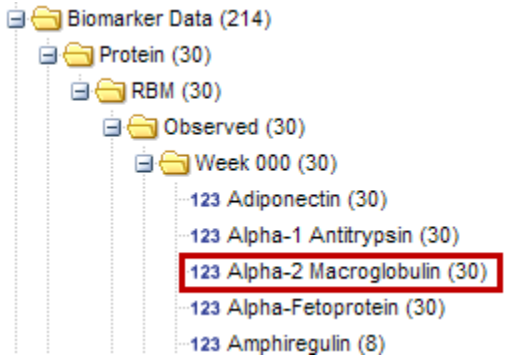
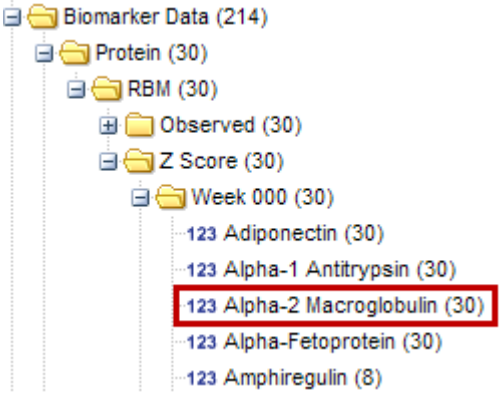
   Note: Alternatively, you can omit some or all of the RBM platform timepoints from the subset definitions, and instead define them in a dialog after you select the type of heat map to generate.

   To select the RBM platform timepoints from the navigation tree and add them to your subset definitions:

   a. Open the **RBM** node under the **Biomarker Data > Protein** branch of the navigation tree for the study of interest.

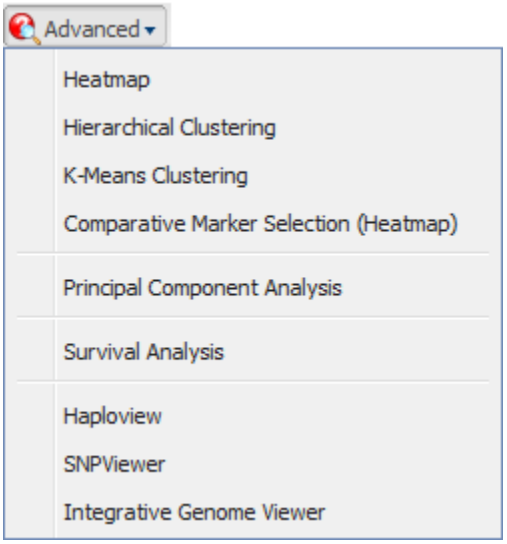
   b. Take **one** of the following actions:

Action	Example	Result
Open **Observed**, then drag Week *nnn* (for example, Week 000 or Week 012) into an empty box of Subset 1.		Bases the heat map on the antigen values observed in the selected timepoint, for all subjects in Subset 1. Optionally, you could have opened **Z Score** with the same result.

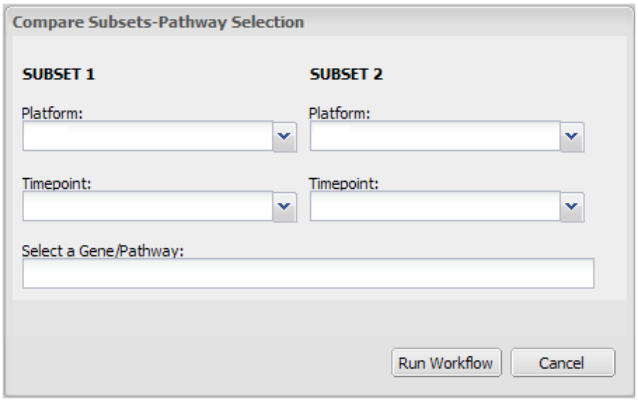
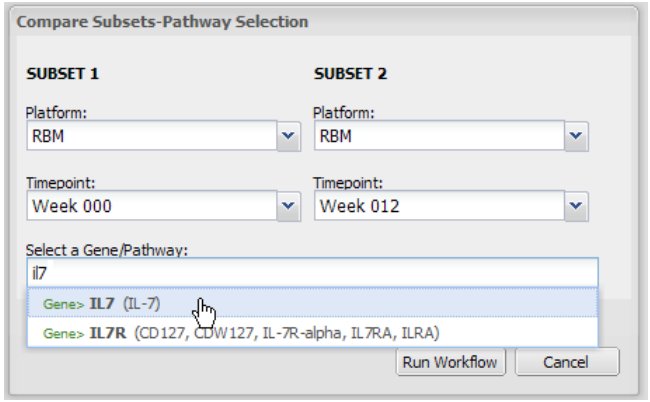
Action	Example	Result
Open Week *nnn* under **Observed**, then drag an antigen of interest into Subset 1. Optionally, repeat this action for other antigens.		The Set Value dialog appears, prompting you to specify an antigen value. The heat map will contain only those subjects in Subset 1 who meet the criteria for the specified antigen(s), as observed during the selected timepoint.
Open Week *nnn* under **Z Score**, then drag an antigen of interest into Subset 1. Optionally, repeat this action for other antigens.		Bases the heat map on the antigen values observed in the selected timepoint, for all subjects in Subset 1. Optionally, you could have opened **Observed** with the same result. **Note:** RBM values are represented in a heat map as z-score values.

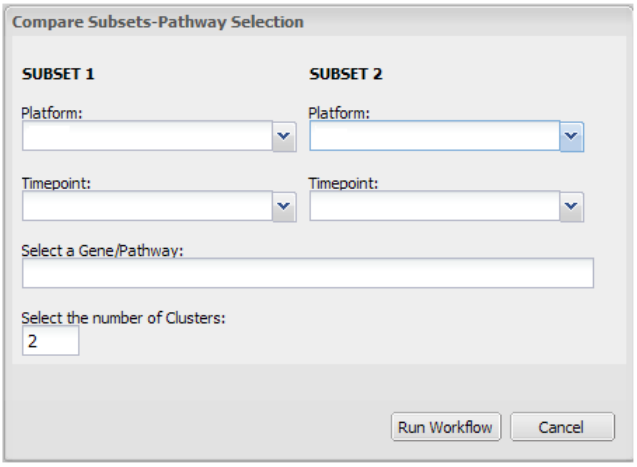
3. Optionally, repeat Step 2.b to include additional RBM data in the heat map for Subset 2.

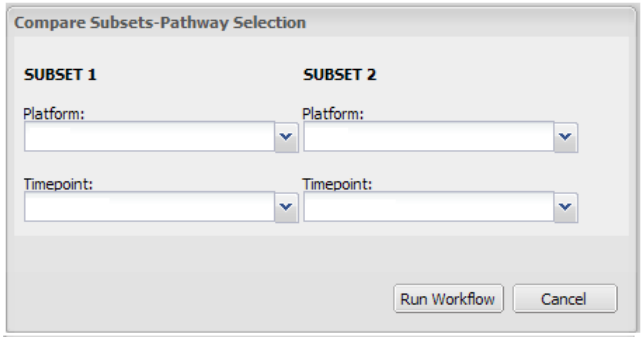
4. Click the Dataset Explorer **Advanced** button, then select the type of heat map you want to generate:



One of the following responses occurs, depending on the type of heat map you selected:

Menu Choice	Response
Heatmap (standard) or Hierarchical Clustering	The Compare Subsets-Pathway Selection dialog appears:  **Actions you take:** ▪ Select the **RBM** platform and the timepoint(s) to use in the heat map, or accept the defaults, if any. Multiple timepoints can be selected within a subset. (Click to select, click again to deselect.) If you leave **Timepoint** blank, all possible timepoints will be factored into the heat map. ▪ To include all antigens in the heat map, leave **Select a Gene/Pathway** blank. ▪ To filter the antigens by a gene or pathway, type part or all of the gene or pathway name in the **Select a Gene/Pathway** field, then select the full name from the dropdown:  ▪ Click **Run Workflow**. After a few seconds, the heat map appears.

Menu Choice	Response
K-Means Clustering	The Compare Subsets-Pathway Selection dialog appears:  **Actions you take:** ▪ Select the **RBM** platform and the timepoint(s) to use in the heat map, or accept the defaults, if any. Multiple timepoints can be selected within a field. (Click to select, click again to deselect.) If you leave **Timepoint** blank, all possible timepoints will be factored into the heat map. ▪ To include all antigens in the heat map, leave **Select a Gene/Pathway** blank. ▪ To filter the antigens by a gene or pathway, type part or all of the gene or pathway name in the **Select a Gene/Pathway** field, then select the full name from the dropdown. ▪ Use the **Select the number of Clusters** field to specify the number of clusters to view in the heat map. ▪ Click **Run Workflow**. After a few seconds, the heat map appears.

Menu Choice	Response
Comparative Marker Selection	The Compare Subsets-Pathway Selection dialog appears:  **Actions you take:** ▪ Select the **RBM** platform and the and timepoint(s) to use in the heat map, or accept the defaults, if any. Multiple timepoints can be selected within a subset. (Click to select, click again to deselect.) If you leave **Timepoint** blank, all possible timepoints will be factored into the heat map. ▪ Click **Run Workflow**. A heat map and a Comparative Marker Selection viewer are displayed. **Note:** Due to the large amount of data being processed, it may take several minutes for the visualizations to appear.

Heat Maps Based on Gene Expression Data

You generate a heat map based on gene expression data from a source such as Affymetrix, Illumina, or Agilent.

To generate a heat map based on gene expression data:

1. Define one or both subsets, as described earlier in this chapter.

   Note: If you intend to generate a Comparative Marker Selection heat map, you must define both subsets.

2. In the navigation tree, select the sample(s) and/or timepoint(s) for the gene expression platform and drag them into the subset definition boxes.

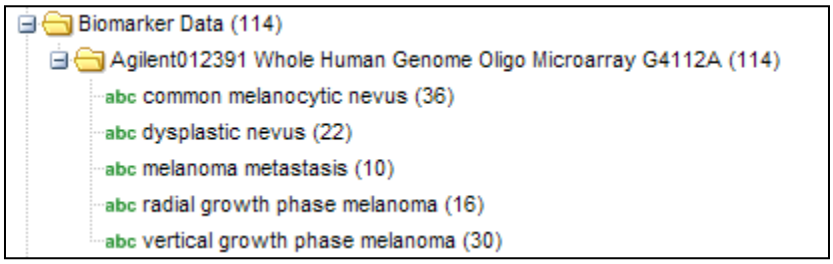
   Note: Alternatively, you can omit some or all of the gene expression samples/ timepoints from the subset definitions, and instead define them in a dialog after you select the type of heat map to generate.

To select gene expression samples and/or timepoints from the navigation tree and add them to your subset definitions:

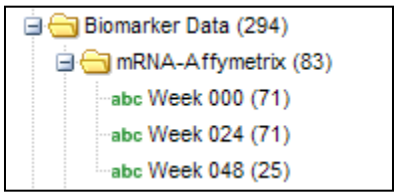
a. Open the appropriate platform node (such as **Affymetrix..., Illumina...,** or **Agilent...**) under the **Biomarker Data** branch of the navigation tree for the study of interest.

b. Drag a node containing a sample type or timepoint into Subset 1 – for example:

Nodes containing sample types:

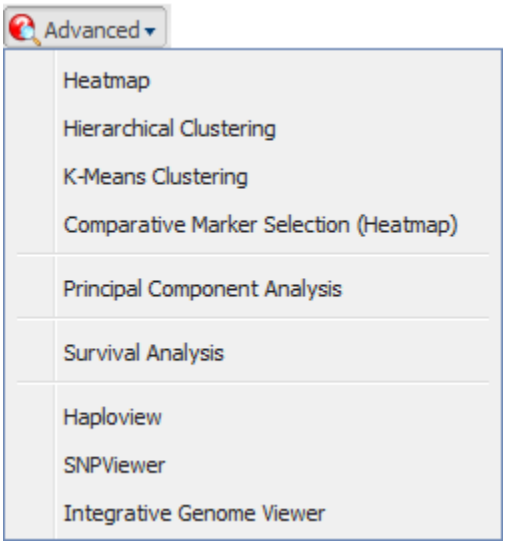


Nodes containing timepoints:

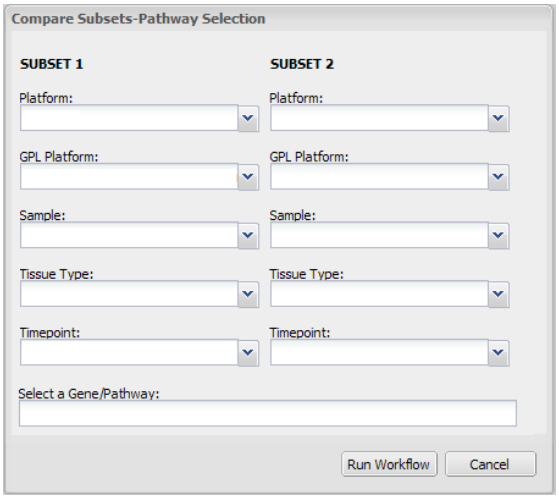
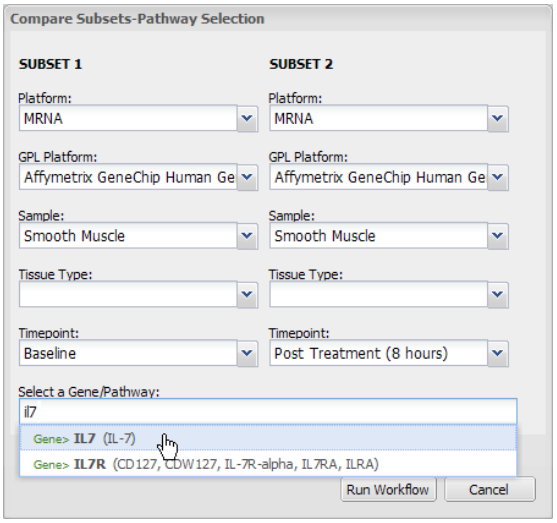


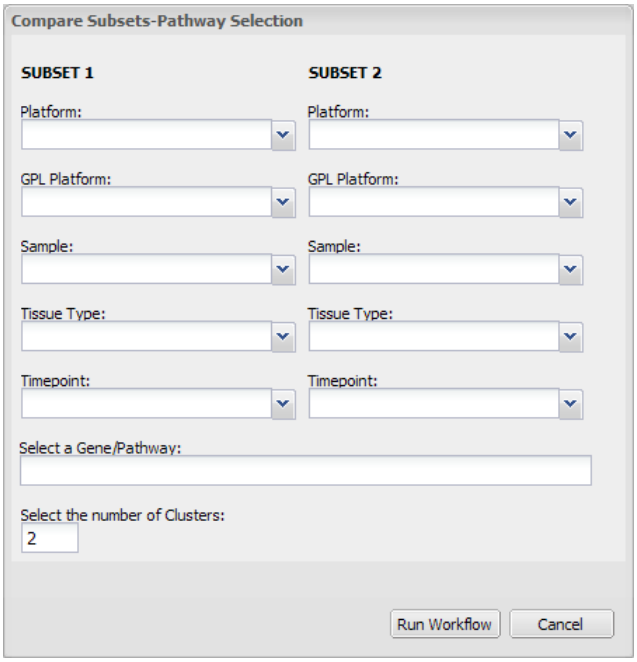
3. Optionally, repeat the previous step to include additional sample types/timepoints in the heat map for Subset 2.

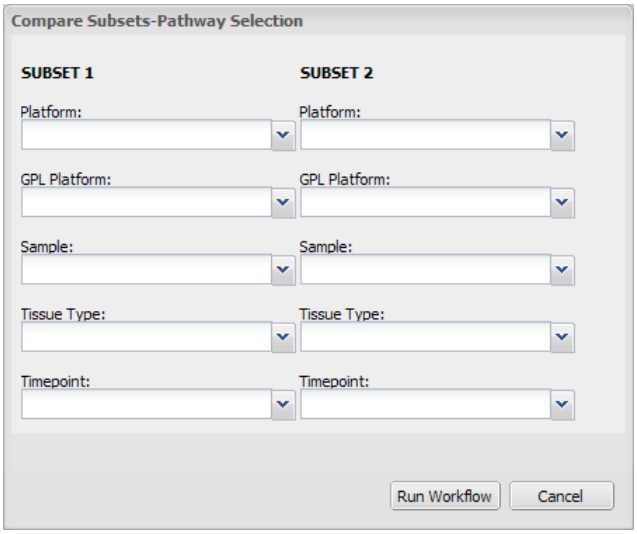
4. Click the Dataset Explorer **Advanced** button, then select the type of heat map you want to generate:



One of the following responses occurs, depending on the type of heat map you selected:

Menu Choice	Response
Heatmap (standard) or Hierarchical Clustering	The Compare Subsets-Pathway Selection dialog appears:  **Actions you take:** ▪ Select the **MRNA** platform and other parameters for the heat map, or accept the defaults, if any. Multiple GPL platforms, samples, tissue types, and timepoints can be selected within a field. (Click to select, click again to deselect.) If you leave **one of the above fields** blank, all possible values for the field will be factored into the heat map. ▪ To filter the dataset by a gene or pathway, type part or all of a gene or pathway name in the **Select a Gene/Pathway** field, then select the full name from the dropdown:  ▪ Click **OK**. After a few seconds, the heat map appears.

Menu Choice	Response
K-Means Clustering	The Compare Subsets-Pathway Selection dialog appears:  **Actions you take:** ▪ Select the **MRNA** platform and other parameters for the heat map, or accept the defaults, if any. Multiple GPL platforms, samples, tissue types, and timepoints can be selected within a field. (Click to select, click again to deselect.) If you leave **one of the above fields** blank, all possible values for the field will be factored into the heat map. ▪ To filter the dataset by a gene or pathway, type part or all of a gene or pathway name in the **Select a Gene/Pathway** field, then select the full name from the dropdown. ▪ Use the **Select the number of Clusters** field to specify the number of clusters to view in the heat map. ▪ Click **Run Workflow**. After a few seconds, the heat map appears.

Menu Choice	Response
Comparative Marker Selection	The Compare Subsets-Pathway Selection dialog appears:  **Actions you take:** ▪ Select the **MRNA** platform and other parameters for the heat map, or accept the defaults, if any. Multiple GPL platforms, samples, tissue types, and timepoints can be selected within a field. (Click to select, click again to deselect.) If you leave **one of the above fields** blank, all possible values for the field will be factored into the heat map. ▪ Click **Run Workflow**. A heat map and a Comparative Marker Selection viewer are displayed. **Note:** Due to the large amount of data being processed, it may take several minutes for the visualizations to appear.

Heat Maps Based on Proteomics Data

To generate a heat map based on proteomics data:

1. Define one or both subsets, as described earlier in this chapter.

   Note: If you intend to generate a Comparative Marker Selection heat map, you must define both subsets.

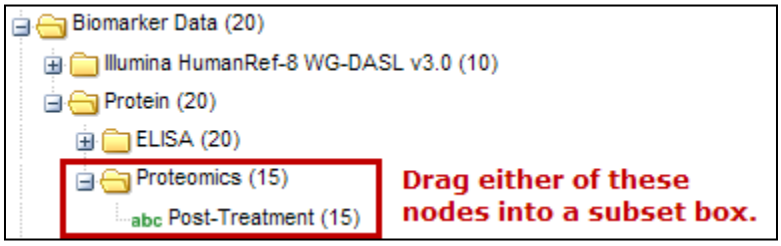
2. In the navigation tree, select the timepoint(s) for the proteomics platform and drag them into the subset definition boxes.

   Note: Alternatively, you can omit some or all of the timepoints from the subset definitions, and instead define them in a dialog after you select the type of heat map to generate.

   To select the timepoints for the proteomics platform from the navigation tree and add them to your subset definitions:

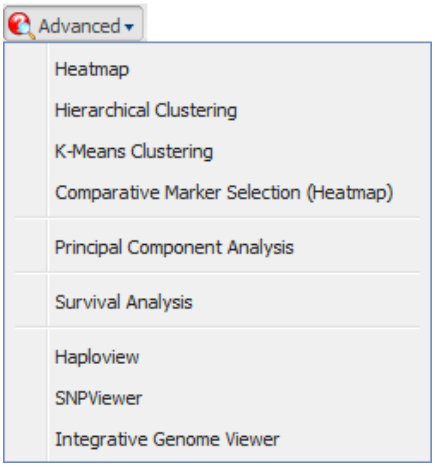
   a. Open the **Protein** node under the **Biomarker Data** branch of the navigation tree for the study of interest.

   b. Drag the proteomics node into Subset 1 – for example:

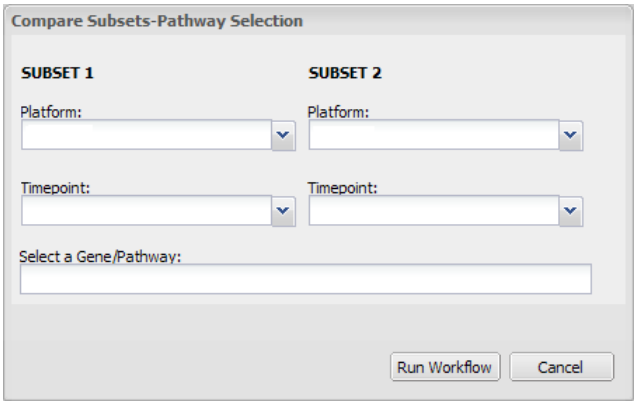
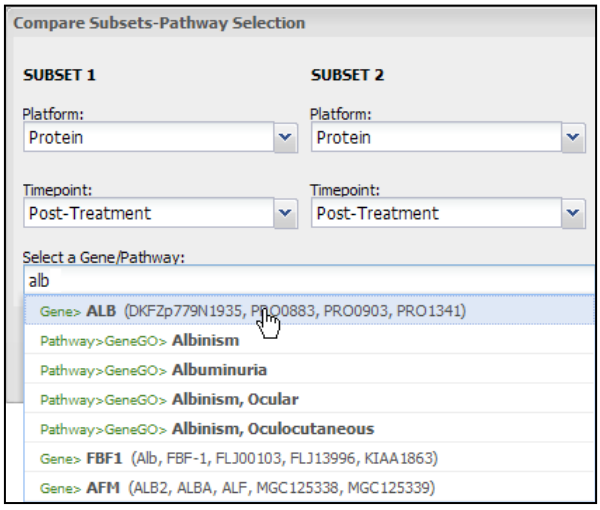


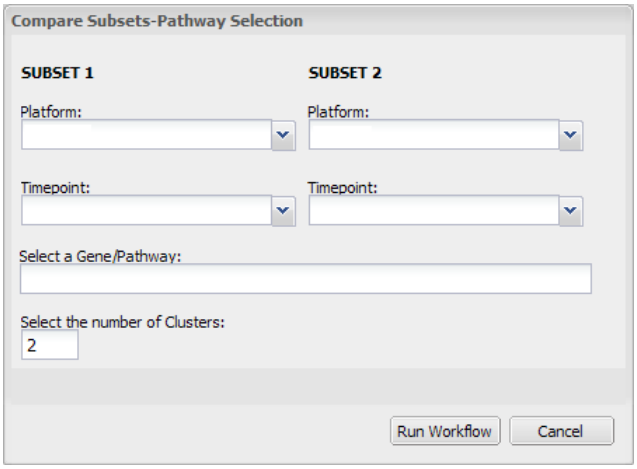
   c. Optionally, repeat the previous step to include additional proteomics data in the heat map for Subset 2.

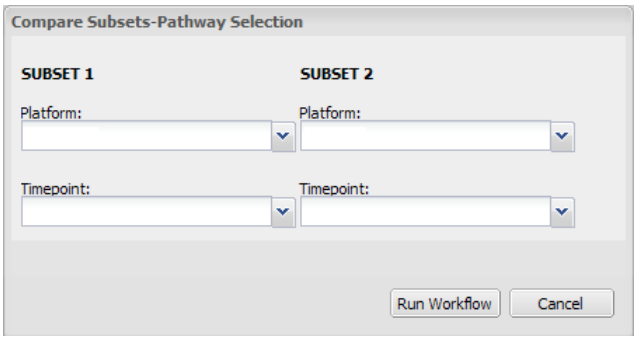
3. Click the Dataset Explorer **Advanced** button, then select the type of heat map you want to generate:



One of the following responses occurs, depending on the type of heat map you selected:

Menu Choice	Response
Heatmap (standard) or Hierarchical Clustering	The Compare Subsets-Pathway Selection dialog appears:  **Actions you take:** ▪ Select the **Protein** platform and the timepoint(s) to use in the heat map, or accept the defaults, if any. Multiple timepoints can be selected within a subset. (Click to select, click again to deselect.) If you leave **Timepoint** blank, all possible timepoints will be factored into the heat map. ▪ To filter the dataset by a gene or pathway, type part or all of the gene or pathway name in the **Select a Gene/Pathway** field, then select the full name from the dropdown:  ▪ Click **Run Workflow**. After a few seconds, the heat map appears.

Menu Choice	Response
K-Means Clustering	The Compare Subsets-Pathway Selection dialog appears:  **Actions you take:** ▪ Select the **Protein** platform and the timepoint(s) to use in the heat map, or accept the defaults, if any. Multiple timepoints can be selected within a subset. (Click to select, click again to deselect.) If you leave **Timepoint** blank, all possible timepoints will be factored into the heat map. ▪ To filter the dataset by a gene or pathway, type part or all of the gene or pathway name in the **Select a Gene/Pathway** field, then select the full name from the dropdown. ▪ Use the **Select the number of Clusters** field to specify the number of clusters to view in the heat map. ▪ Click **Run Workflow**. After a few seconds, the heat map appears.

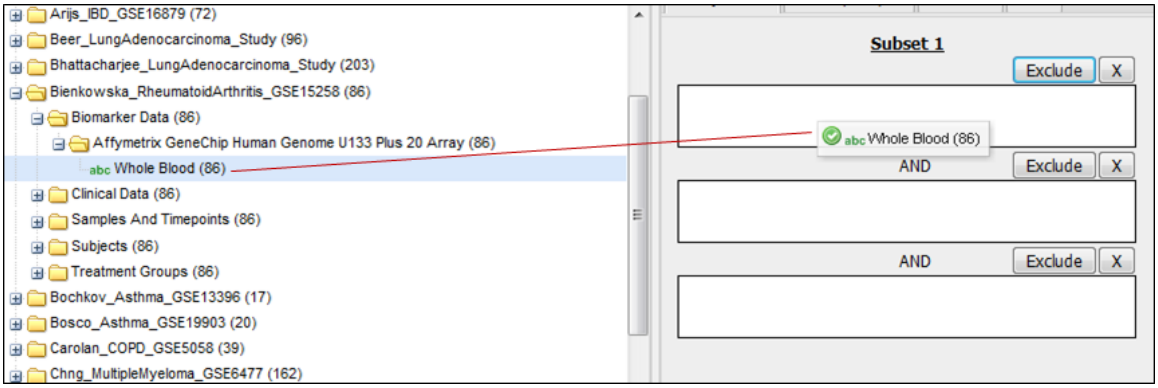
Menu Choice	Response
Comparative Marker Selection	The Compare Subsets-Pathway Selection dialog appears:  **Actions you take:** ▪ Select the **Protein** platform and the timepoint(s) to use in the heat map, or accept the defaults, if any. Multiple timepoints can be selected within a subset. (Click to select, click again to deselect.) If you leave **Timepoint** blank, all possible timepoints will be factored into the heat map. ▪ Click **Run Workflow**. A heat map and a Comparative Marker Selection viewer are displayed. **Note:** Due to the large amount of data being processed, it may take several minutes for the visualizations to appear.

Example

In this example you are interested in analyzing the results of a study on rheumatoid arthritis. You want to see a visualization of gene expression data for the gene REL in two cohorts: those who responded to anti-TnF therapy and those who did not. This example uses Public Study **Bienkowska_RheumatoidArthritis_GSE15258**.

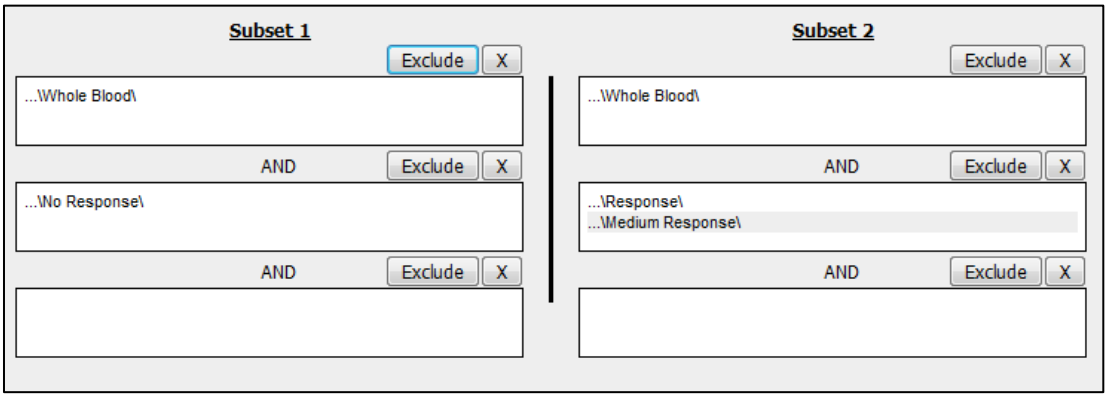
1. Run tranSMART, then click the **Dataset Explorer** tab.
2. Open the study of interest.

3. Drag the array data into an empty box in both subsets:



4. Open the node **Clinical Data**, then open **Response to Anti Tnf Therapy**.

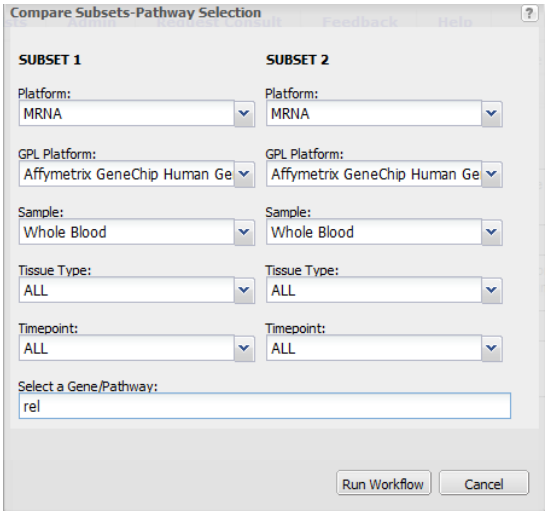
5. Complete the cohort definitions by dragging the response criteria into the subset boxes as shown below:



6. Click **Advanced > Heatmap**.

   The Compare Subsets-Pathway Selection dialog appears.

7. Type **rel** in the **Select a Gene/Pathway** field:



8. Click **Gene>REL** in the dropdown list.

9. Click **Run Workflow**.
10. When prompted, click **OK** to display the heat map.

    A portion of the heat map is shown below:

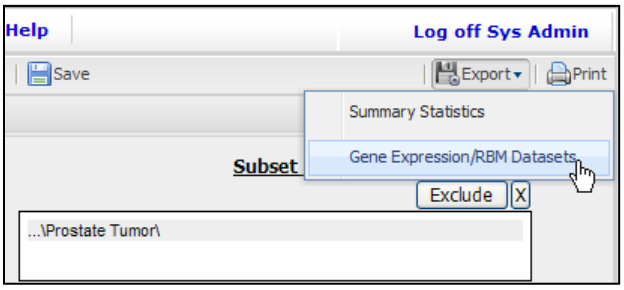


Export Heat Map Data Points

After you generate a heat map containing expression data or RBM data, you can export the data points to a Microsoft Excel spreadsheet.

To export expression or RBM data points:

1. Generate a heat map, as described earlier in this chapter.
2. Click the **Export** button, then click **Gene Expression/RBM Datasets**.



3. Open the spreadsheet for viewing, or save it to a file.

Generating a Principal Component Analysis

In a principal component analysis (PCA) of an mRNA, RBM, or proteomic dataset, the total number of variables in the dataset is reduced to a smaller number of variables – the principle components of the dataset. Principal component variables are calculated from correlated variables in the total dataset.

Dataset Explorer uses the Broad Institute's GenePattern genomic analysis platform to generate PCA analyses. For information about the Broad Institute's PCAViewer, see the document **PCAViewer** on the following site:

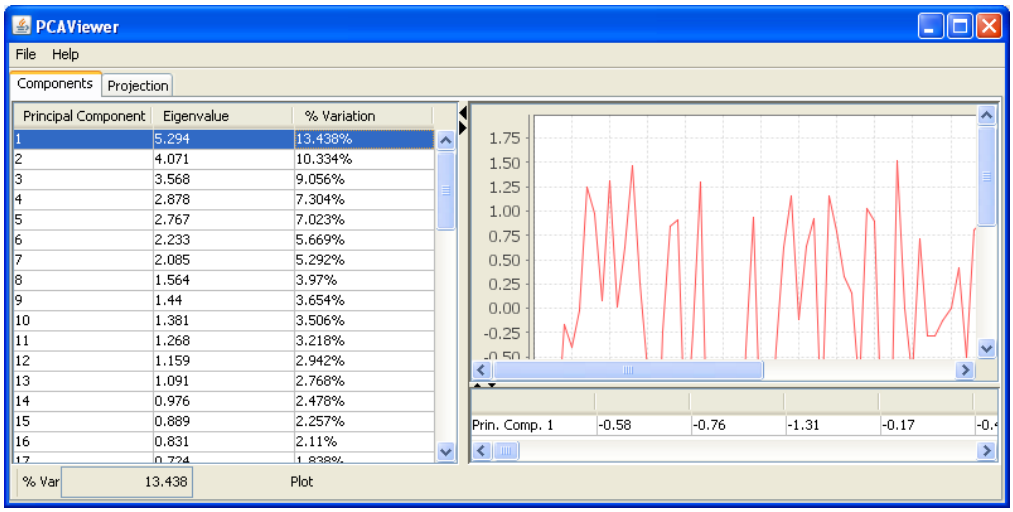
http://www.broadinstitute.org/cgi-bin/cancer/software/genepattern/modules/gp_modules.cgi

You may run multiple visualizations in the background while continuing to use Dataset Explorer. For more information on running visualizations in the background, see Asynchronous Operations on page 130.

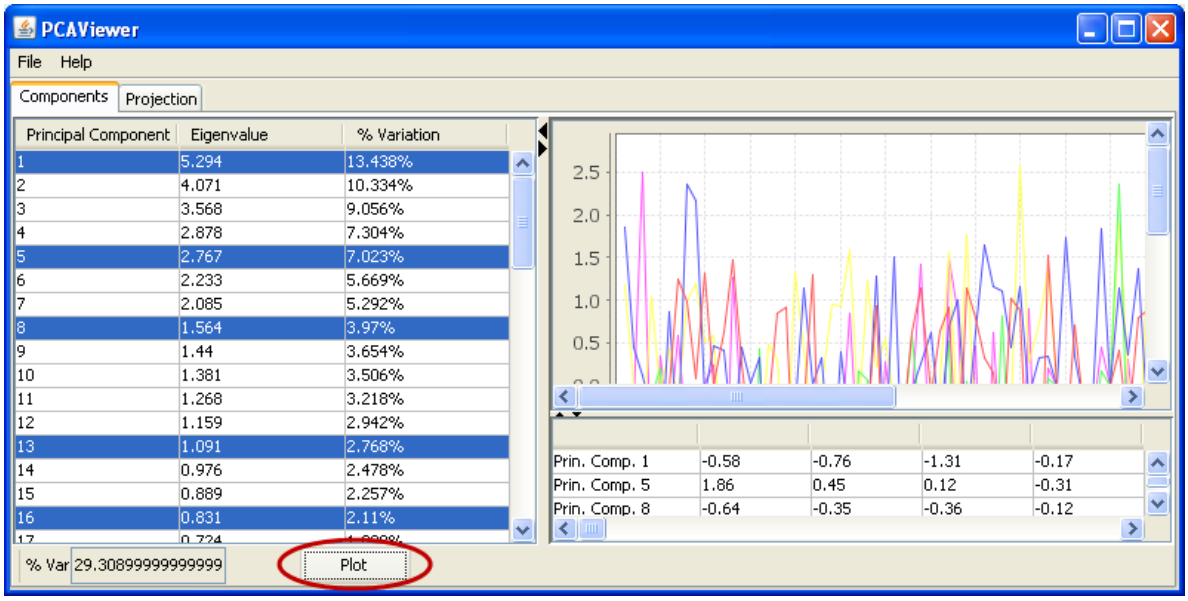
To generate a PCA visualization:

1. Run tranSMART, then click the **Dataset Explorer** tab.
2. Select the study of interest.
3. Define the cohorts whose data points will be represented in the PCA visualization.
4. Click **Advanced > Principal Component Analysis**.
5. In the Compare Subsets-Pathway Selection dialog, specify the platform and other factors, or accept the defaults.
6. Click **Run Workflow**.

   The PCAViewer appears with the contents of the **Components** tab displayed. In this tab, the principal component data appears in the left pane, and a plot of the selected component(s) data, along with a table of each data point in the selected component(s), appears in the right pane.



7. To view plots and data points for multiple principal components:
   a. Hold down the **Ctrl** key.
   b. In the left pane, click each principal component to view.
   c. Click the **Plot** button:



Multi-Dimensional Projections

You can project each individual cohort onto two or three principal components, creating a two-dimensional or three-dimensional projection. You may generate more than one analysis at a time. For more information, see Asynchronous Operations on page 130.

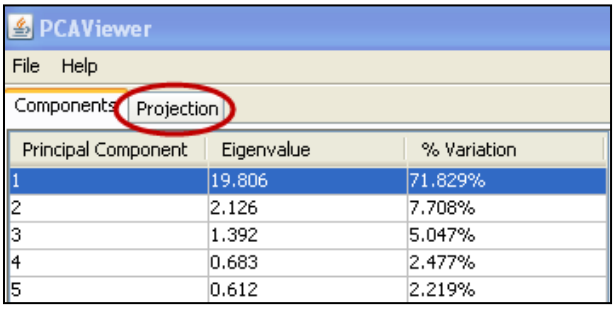
> **Note:** To display a three-dimensional projection, you must have Java 3D installed.

To view a multi-dimensional projection of principal components:

1. Run tranSMART, then click the **Dataset Explorer** tab.
2. Select the study of interest.
3. Define the cohorts whose data points will be represented in the PCA visualization.
4. Click **Advanced > Principal Component Analysis**.
5. In the Compare Subsets-Pathway Selection dialog, specify the platform other variables, or accept the defaults.
6. Click **Run Workflow**.

   The PCAViewer appears.

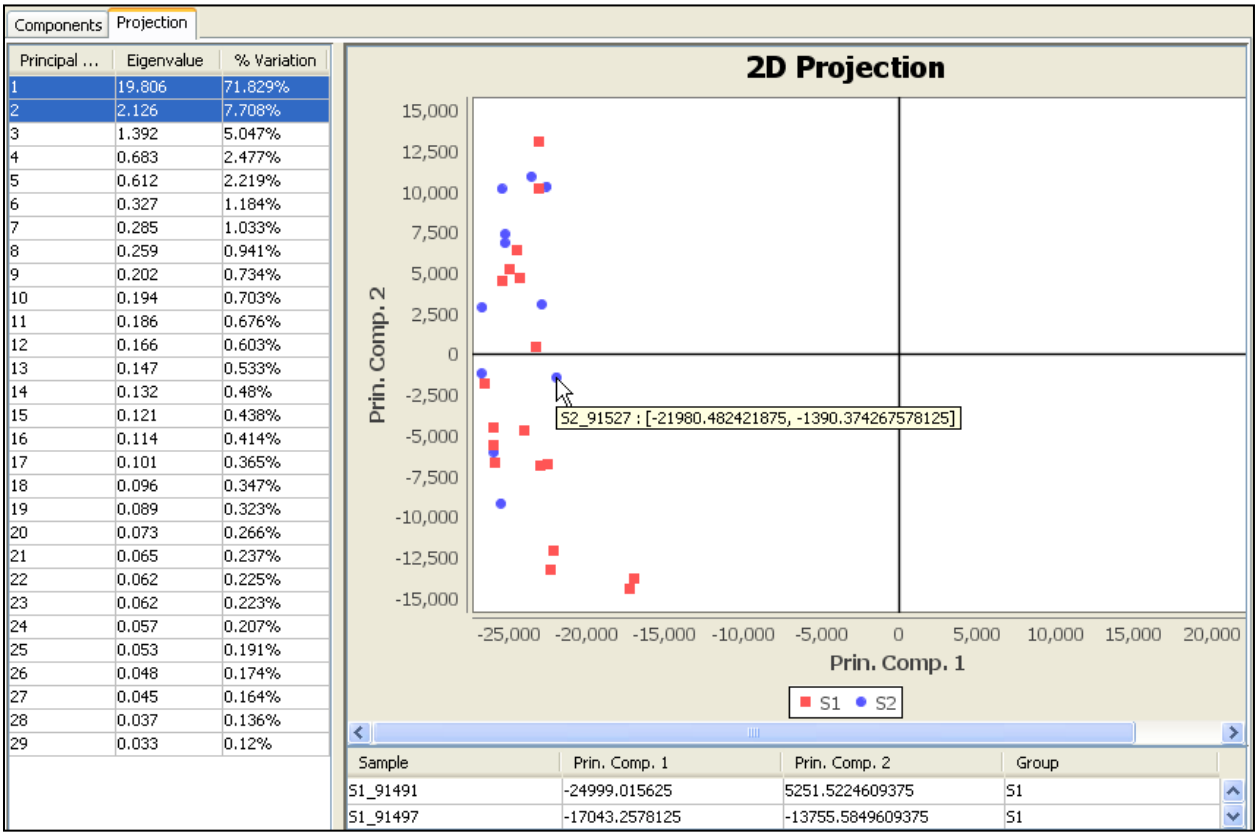
7. Click the **Projection** tab:



8. Select the two or three principal components to use in the visualization.

9. Click the **Plot** button.

   The following figure shows a two-dimensional projection for the first and second principal components. The two different colors in the scatter plot represent the different subsets.



   Note that hovering the mouse pointer over a marker in the scatter plot displays the data that the marker represents.

Generating a Survival Analysis

Dataset Explorer allows you to generate a time-to-event analysis – that is, a relationship between one or more predictive variables and the survival times of individuals in two groups of cohorts.

You may run multiple visualizations in the background while continuing to use Dataset Explorer. For more information on running visualizations in the background, see Asynchronous Operations on page 130.

To generate a survival analysis, you must introduce the following datasets into each group:

- The observed survival times of the individuals in each group.
- At least one variable (such as different medications or different genetic attributes) that distinguishes the two groups.
- The specific event (death) being tracked for the individuals in the study, and optionally, any censoring factors that occurred before the event took place.

  A censoring factor might be the withdrawal of an individual from the study, or the conclusion of the study before the event occurred for a given individual.

  Note: In Dataset Explorer survival analysis datasets, the event is always death. Subjects who are censored are not included in the death count.

Dataset Explorer's survival analysis functionality is based on the Cox regression model.

To generate a survival analysis:

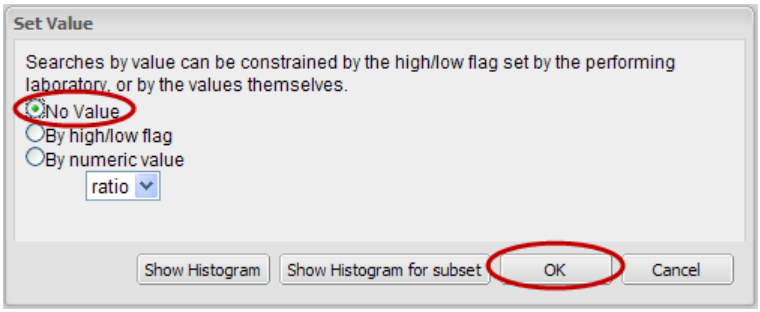
1. Run tranSMART, then click the **Dataset Explorer** tab.
2. Select the study of interest.
3. Define the cohorts for the survival analysis, being sure to include the required datasets indicated above.
4. Click **Advanced > Survival Analysis**.

Example

1. In Dataset Explorer, open public study **Wang_TransBIG_BreastCancer_GSE7390**.
2. Open the node **Clinical Data**.
3. Open the following nested nodes in the following order:
   a. Clinical Data
   b. Overall Survival in Days
4. Drag the time-to-event dataset **Overall Survival in days (OS)** into Subset 1.

   Don't drag the folder into the subset box – just the dataset inside the folder.

5. In the Set Value dialog, select **No Value**, then click **OK**:



   Specifying a value limits the values in the dataset. By not specifying a value, the entire time-to-event dataset is used.

6. Repeat step 4 and step 5 for Subset 2.

7. In the same **Clinical Data** node, drag the **Censored (OSCENS)** folder into empty boxes in Subset 1 and Subset 2.

   The contents of the **Censored (OSCENS)** folder are **No** and **Yes**. This concept introduces the **Event** and **Censored** datasets into the analysis.

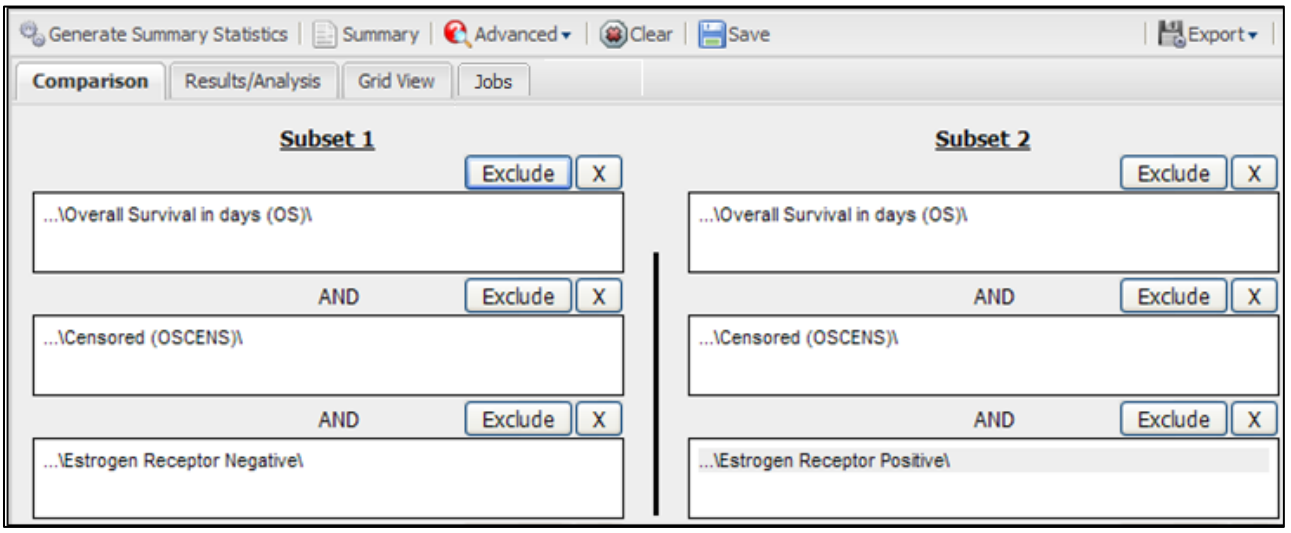
8. Open the following nested nodes in the following order:

   a. Subjects
   b. Medical History
   c. Estrogen Receptor Status

9. Drag **Estrogen Receptor Negative** into an empty box in Subset 1.

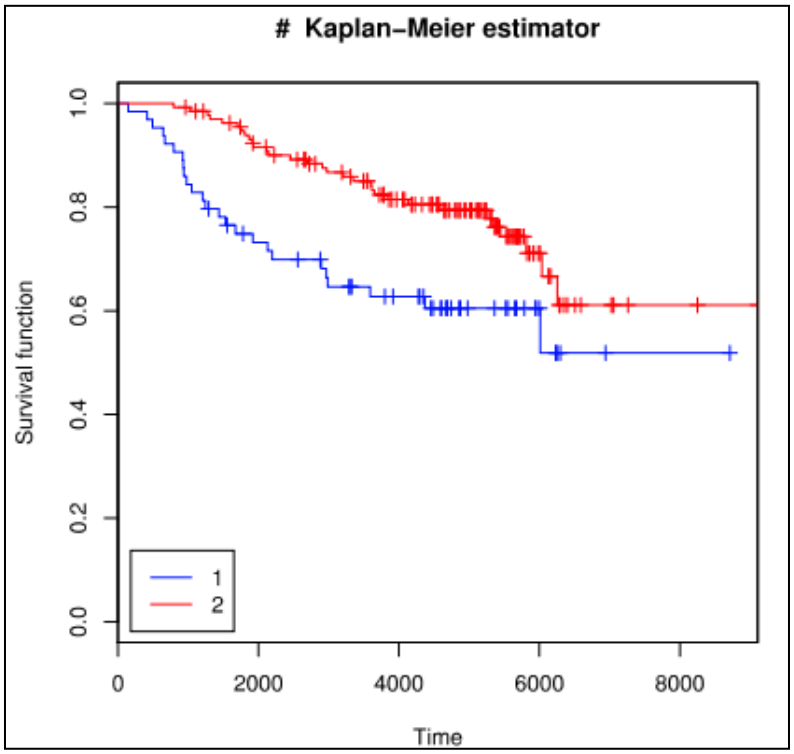
10. Drag **Estrogen Receptor Positive** into an empty box in Subset 2.

    The subset boxes are now defined as follows:



11. Click the **Advanced** tab, then click **Survival Analysis**.

    The analysis includes the following Kaplan-Meier curves:



    In the figure, the x-axis represents survival time in days, and the y-axis represents the percentage of subjects who were still alive at a given point in time during the study.

    Note the hashmarks in the plot lines. These represent censored data – for example, subjects who dropped out of the study before the event (death) occurred.

Hazard Ratio and Relative Risk

The two groups being compared in a survival analysis are sometimes thought of as the treatment group and the control group. The hazard ratio and the relative risk ratio calculated in a survival analysis are ratios of the treatment group over the control group.

In a Dataset Explorer survival analysis, these ratios are based on Subset 2 results over Subset 1 results.

For example, the following table shows the hazard ratio and relative risk calculated from the survival analysis example in the previous section:

Number of Subjects	198
Hazard Ratio (95% CI)	0.476 (0.281 - 0.806)
Relative Risk (p Value)	-0.743 (0.0058)

Based on these ratios, the subjects in Subset 2 (with positive estrogen receptors) have better survivability statistics than the subjects in Subset 1 (with negative estrogen receptors).

Generating a Haploview

A haploview allows you to analyze the differences in allele frequency in two or more loci from one sample to the next. Haploviews are generated for SNP data.

You may run multiple visualizations in the background while continuing to use Dataset Explorer. For more information on running visualizations in the background, see Asynchronous Operations on page 130.

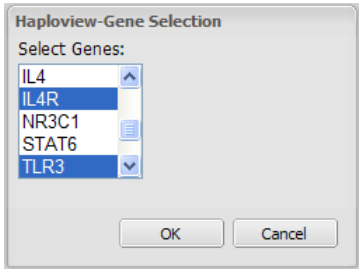
To generate a haploview:

1. Run tranSMART, then click the **Dataset Explorer** tab.
2. Select the study of interest.
3. Define the cohorts whose data points will be represented in the haploview.
4. Do one of the following:

   To base the haploview on one or more genes:

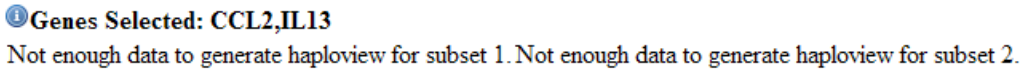
   a. Click **Advanced > Haploview**.

   b. Select one or more genes on which to base the haploview – for example:



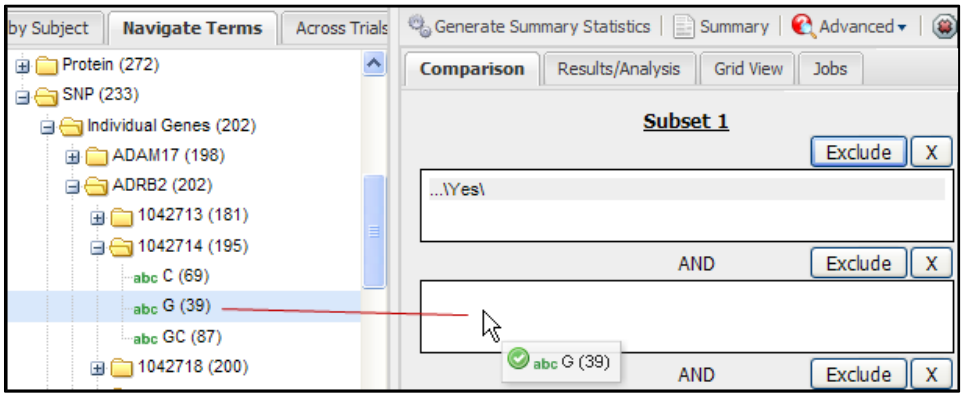
   c. Click **OK**.

Note: If you see a message like the following, a haploview cannot be generated for the gene(s) you selected:



To base the haploview on SNPs or nucleotides:

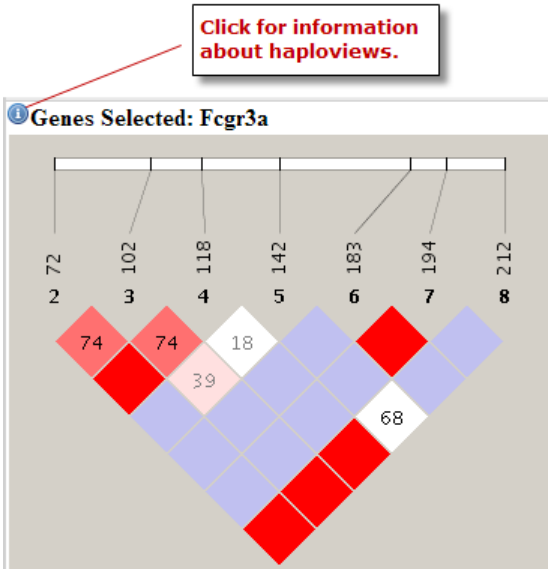
a. Select a SNP ID or nucleotide and drag it into a subset box:



b. Click **Advanced > Haploview**.

Note: Not all trials support the selection of individual SNPs and nucleotides.

For a more detailed description of what a haploview represents, click the Information icon ( ) in the haploview:



Running the SNP Viewer

The SNP Viewer allows you to analyze individual base variations in DNA sequences for normal and tumor tissue samples in a SNP array. The viewer supports both copy number analysis and loss of heterozygosity (LOH) analysis.

SNP array data, such as Affymetrix Genome-Wide Human SNP Array, is required for the SNP Viewer.

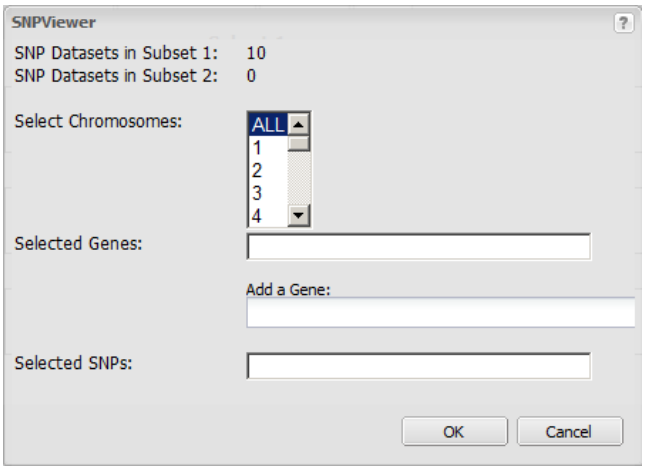
Dataset Explorer uses the Broad Institute's GenePattern genomic analysis platform to generate SNP visualizations. For information about the Broad Institute's SNPViewer, see the document **SNPViewer** on the following site:

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

To run the SNP Viewer:

1. Run tranSMART, then click the **Dataset Explorer** tab.
2. Select the study of interest.
3. Drag the SNP array data into an empty box in one or both subsets.
4. Define the cohorts whose data points will be represented in the SNP visualization.
5. Click **Advanced > SNPViewer**.

   The SNPViewer dialog box appears:



6. Select the data you would like to see in the visualization. Use one of the following filtering methods:

Filtering Method	Description
By Chromosome	Select one or more chromosomes to include in the visualization, or select **ALL** to include all chromosomes. To select multiple chromosomes, click a number, then hold down **Ctrl** and click another. **Note:** If gene or SNP rs IDs are entered, the selection of chromosomes is ignored. **Note:** For performance reasons, only select chromosomes that are of particular interest to you.
By Gene	To select a gene, type all or part of the gene into the **Add a Gene** field. ▪ If you see the gene you want, click to select. ▪ If you do not see the gene you want, type more characters. Separate any additional genes with commas. **Note:** Some genes that you select as filters in tranSMART may not be found in the GenePattern SNPViewer.
By SNP rs ID	To add a SNP rs ID, type the full ID into the **Selected SNPs** field. Separate any additional IDs with commas.
By a combination of genes or SNP rs IDs	Follow the instructions above to add a combination of genes or SNP rs IDs.

7. Click **OK**.

   If any security dialog boxes appear, acknowledge them by clicking **Continue** or **OK**.

   The visualization may require several minutes to display a large dataset, such as when multiple chromosomes are selected.

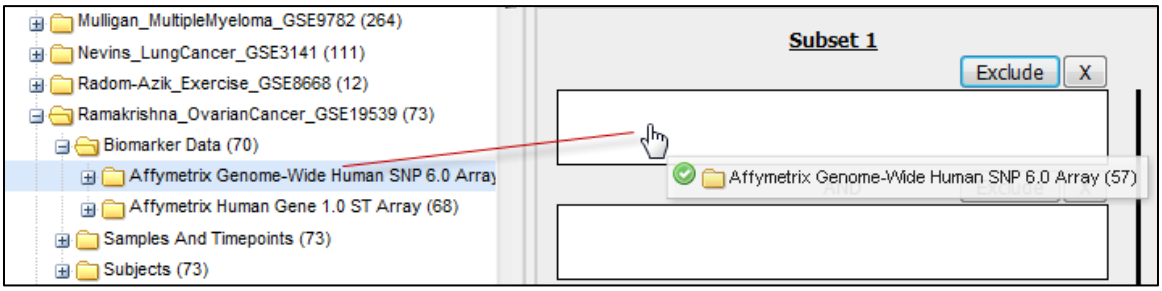
Examples

All examples use biomarker and cohort data from **Public Study GSE19539: Ramakrishna_OvarianCancer**.

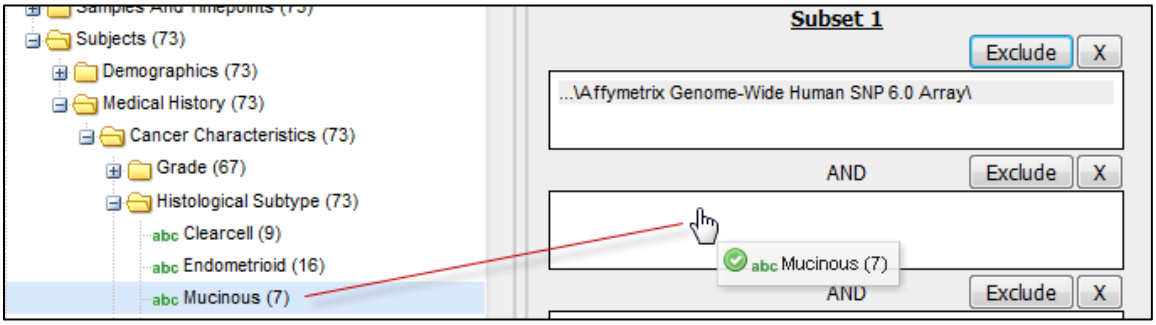
Filter Data by Chromosome

In this example you are interested in viewing SNPs at chromosomes **2** and **4** in mucinous ovarian tumors.

1. Drag the SNP array into an empty subset definition box:

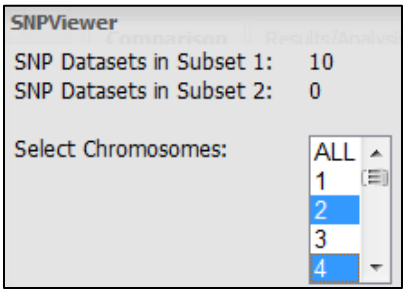


2. Limit the cohort to samples of mucinous tumors only:



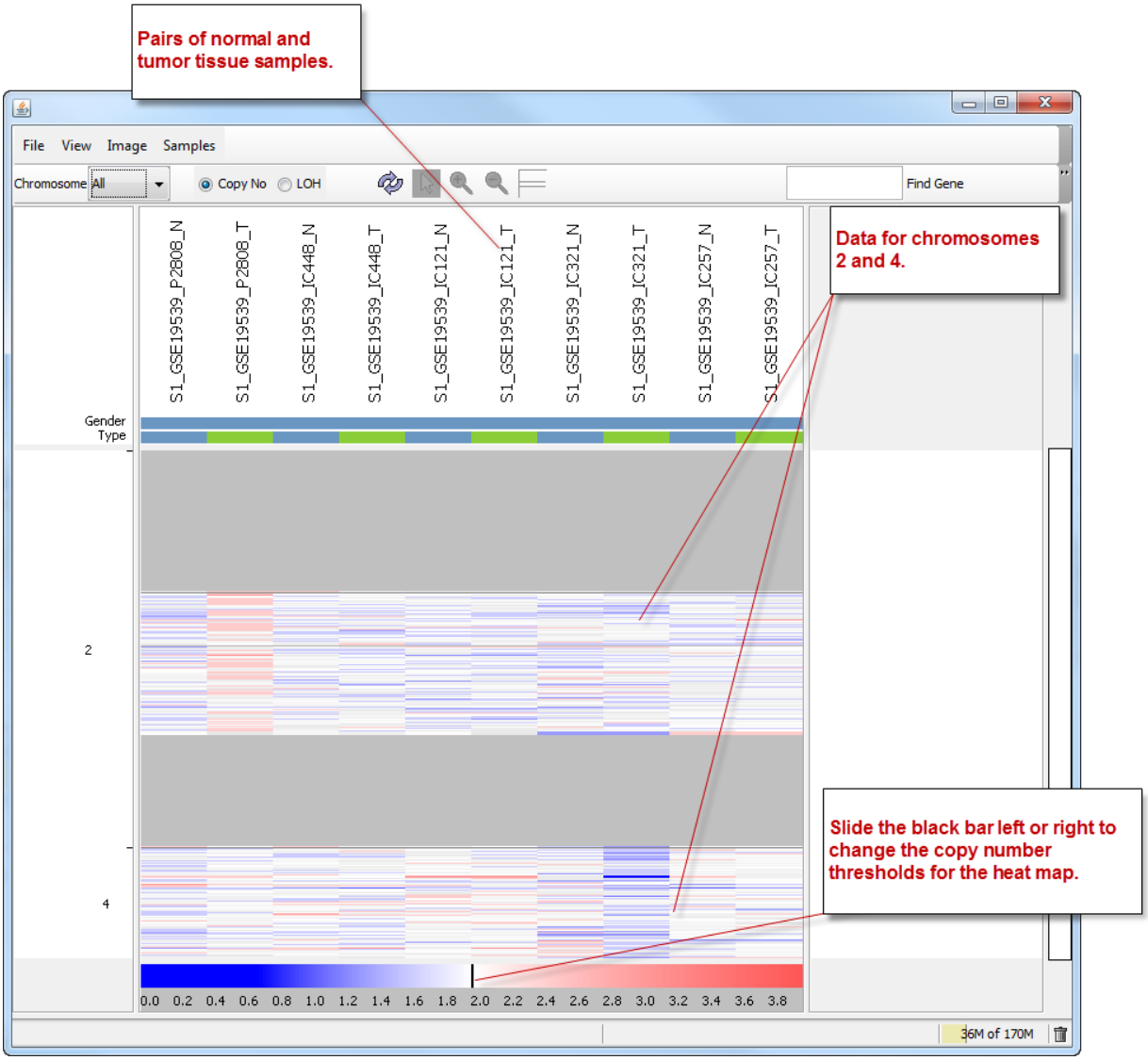
3. Click **Advanced > SNPViewer**.

4. In the SNPViewer dialog box, select chromosomes 2 and 4 (click **2**, then hold down the **Ctrl** key and click **4**):



5. Click **OK**.

   A Workflow Status dialog box appears showing the processing stages. When processing is complete, the following visualization appears:

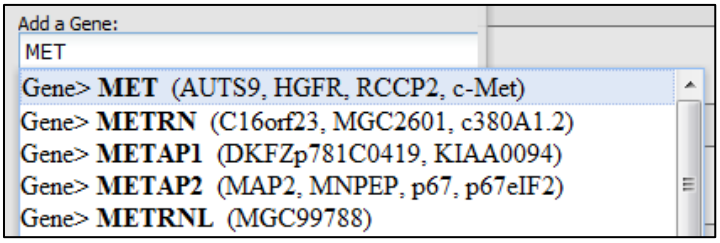


Filter Data by Gene

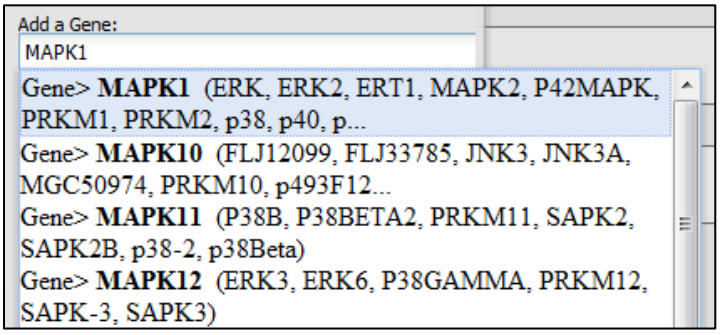
In this example you are interested in viewing genes **MET** and **MAPK1** in mucinous ovarian tumors.

1. Follow Step 1 through Step 3 in the example Filter Data by Chromosome on page 104.

2. In the SNPViewer dialog box, type **MET** into the **Add a Gene** field, then select **MET** from the dropdown box.



3. Repeat Step 2 to add **MAPK1**.



4. Click **OK**.

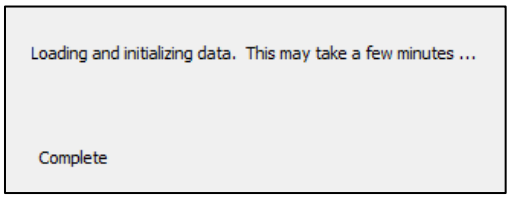
   The following responses occur:

   - A Workflow Status dialog box appears showing the processing stages. When processing is complete, a pop-up window appears that lists the selected genes and their associated SNP rs IDs. The genes you selected are shown in red. The figure below shows a portion of the pop-up:

Selected Genes and SNPs

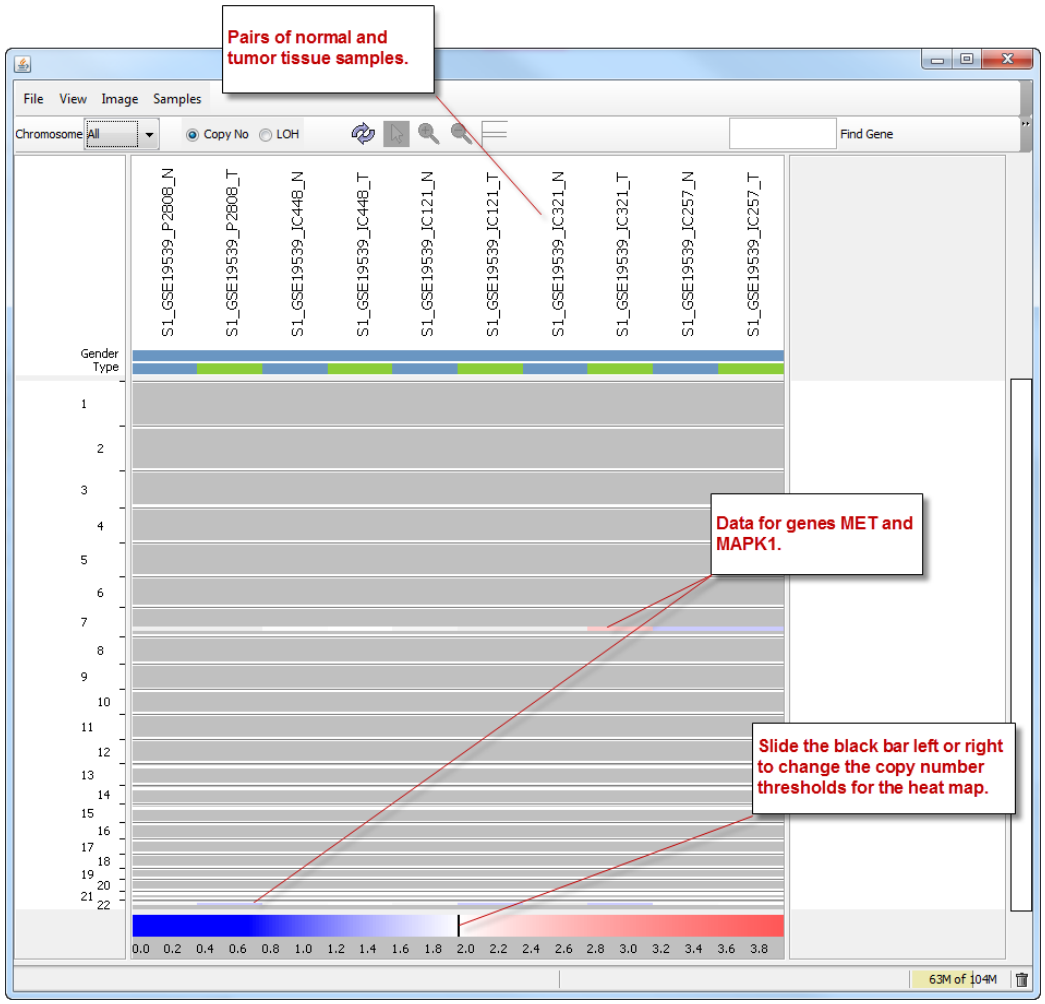
Gene	SNP	Chrom	Position
MET	rs6959106	7	116204058
	rs11974088	7	116205306
	rs7802438	7	116206200
	rs1860588	7	116206486

- The following dialog box appears, advising that the SNPViewer is being prepared for display:



5. When the dialog box displays **Complete** (as shown above), close the dialog box.

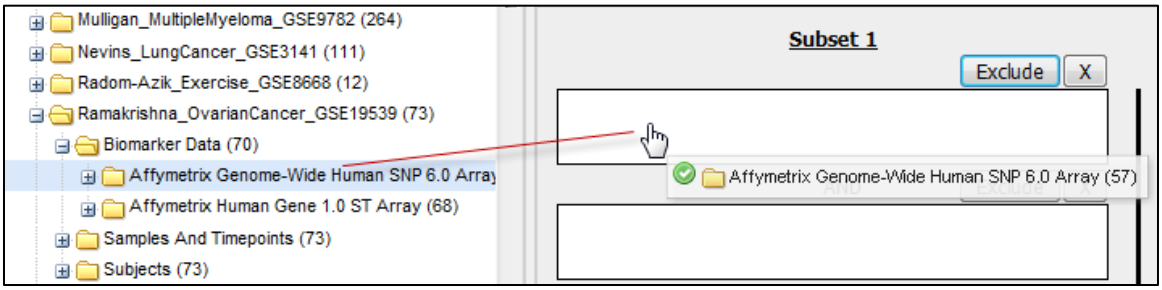
   The visualization is now ready for use. It looks as follows:



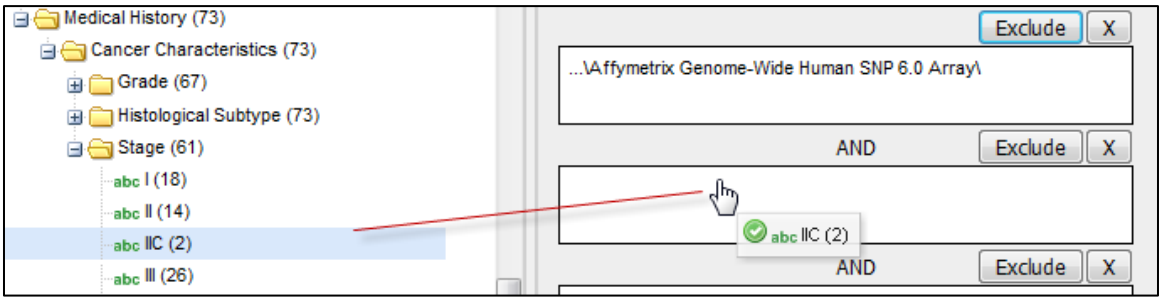
Filter Data by SNP rs ID

In this example you are interested in viewing SNP rs IDs **rs10808181** and **rs28167** in stage IIC ovarian tumors.

1. Drag the SNP array into an empty subset definition box:



2. Limit the cohort to samples of stage IIC ovarian tumors:



3. Click **Advanced > SNPViewer**.

4. Type the full rs IDs (**rs10808181** and **rs28167**) into the **Selected SNPs** field, and separate the IDs with commas.

5. Click **OK**.

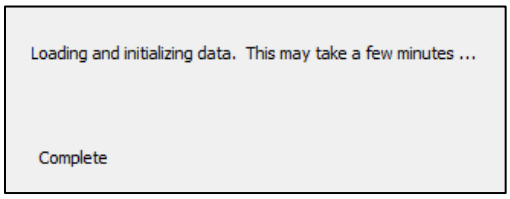
   The following responses occur:

   - A Workflow Status dialog box appears showing the processing stages. When processing is complete, a pop-up window appears that lists the specified SNP rs IDs and their associated genes. The SNP IDs you specified are shown in red:

Selected Genes and SNPs

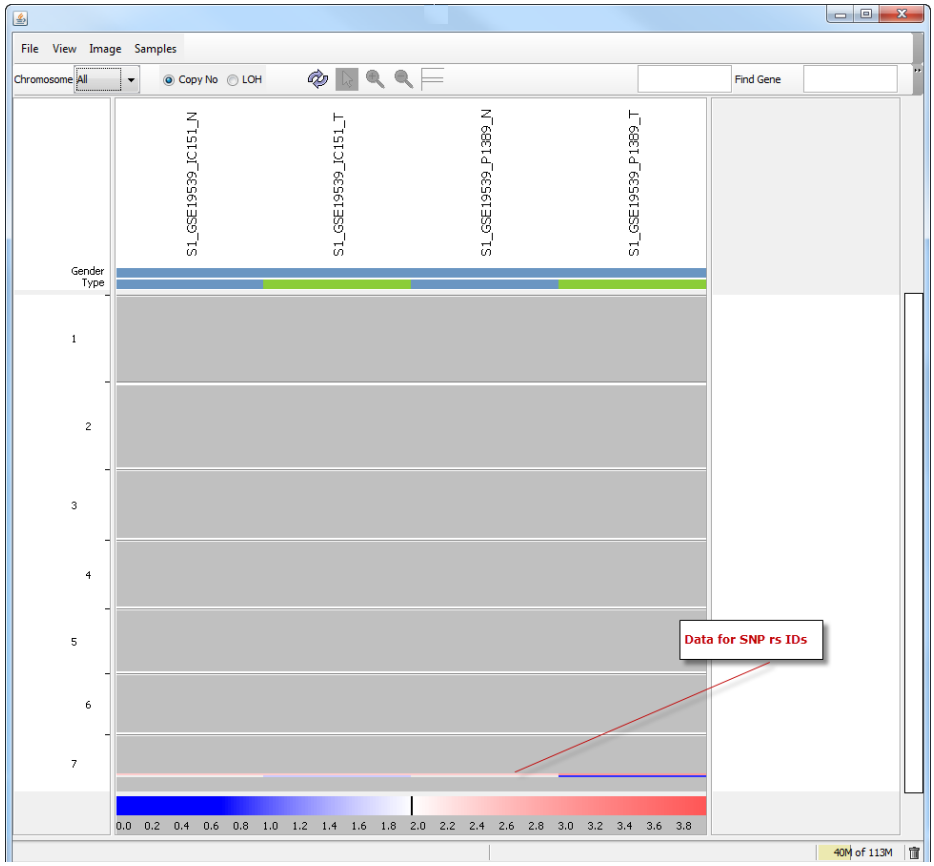
Gene	SNP	Chrom	Position
CAV1	rs10808181	7	116241322
CAPZA2	rs28167	7	116466083

- The following dialog box appears, advising that the SNPViewer is being prepared for display:



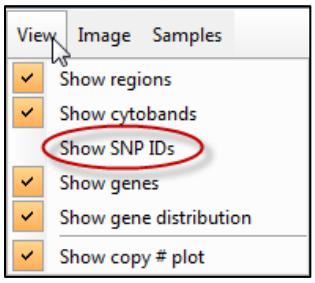
6. When the dialog box displays **Complete** (as shown above), close the dialog box.

   The visualization is now ready for use. It looks as follows:



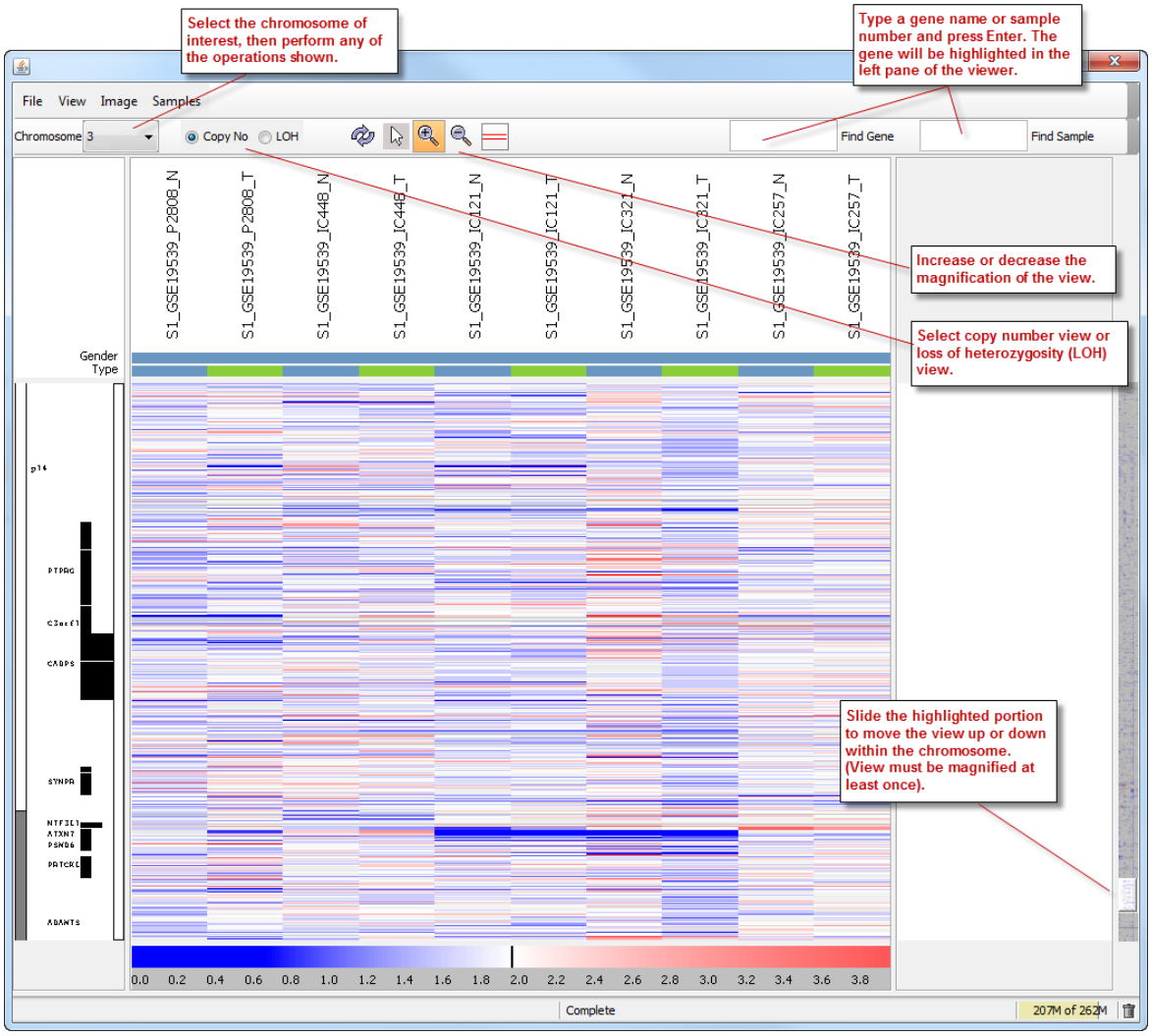
Viewing SNP Data

When viewing SNP data, it is important to note a bug in GenePattern that will cause the visualizer to freeze. The bug occurs when you choose the option **View > Show SNP IDs** within the GenePattern SNP Viewer:

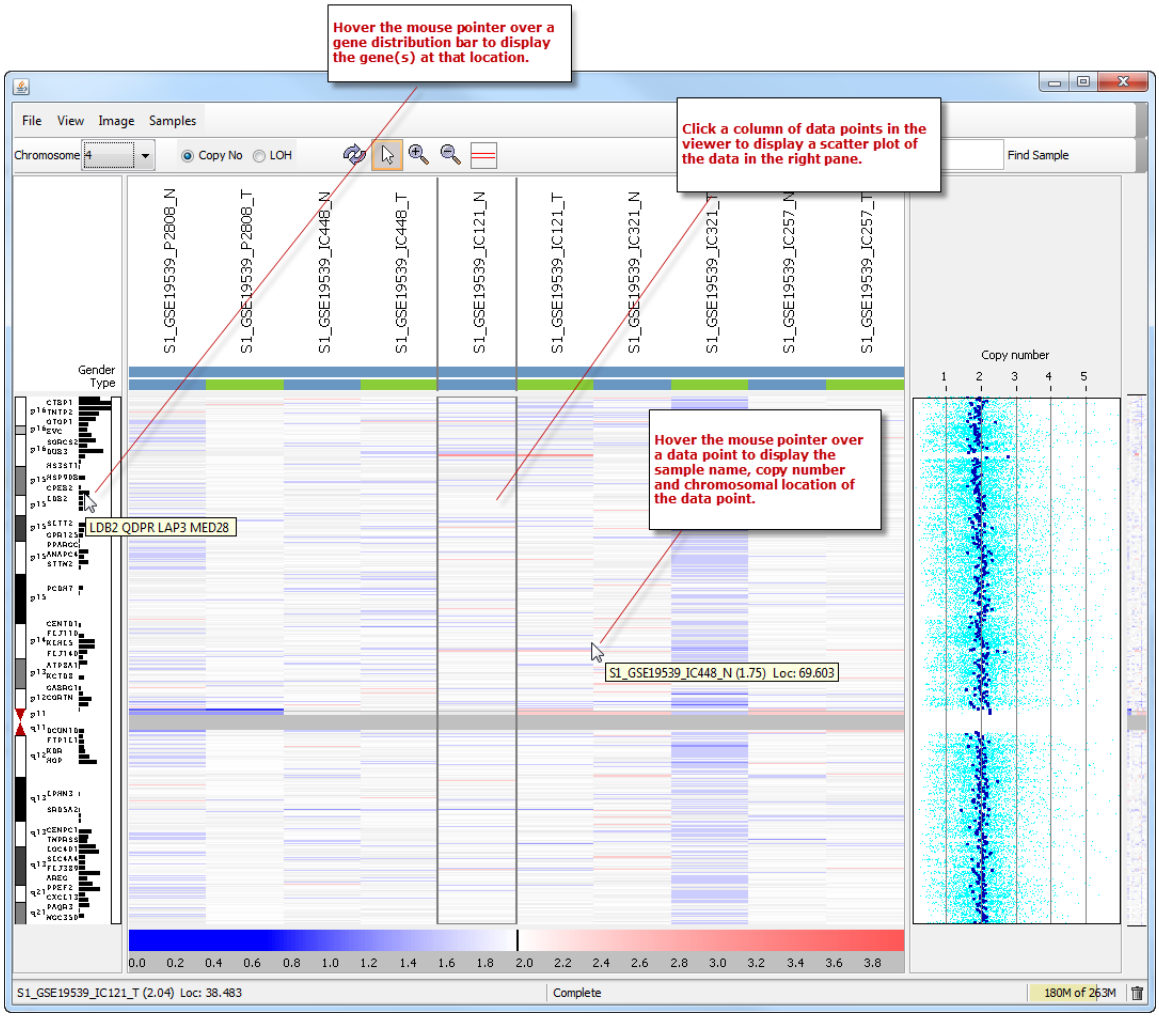


The operations you can perform in the viewer include the following:

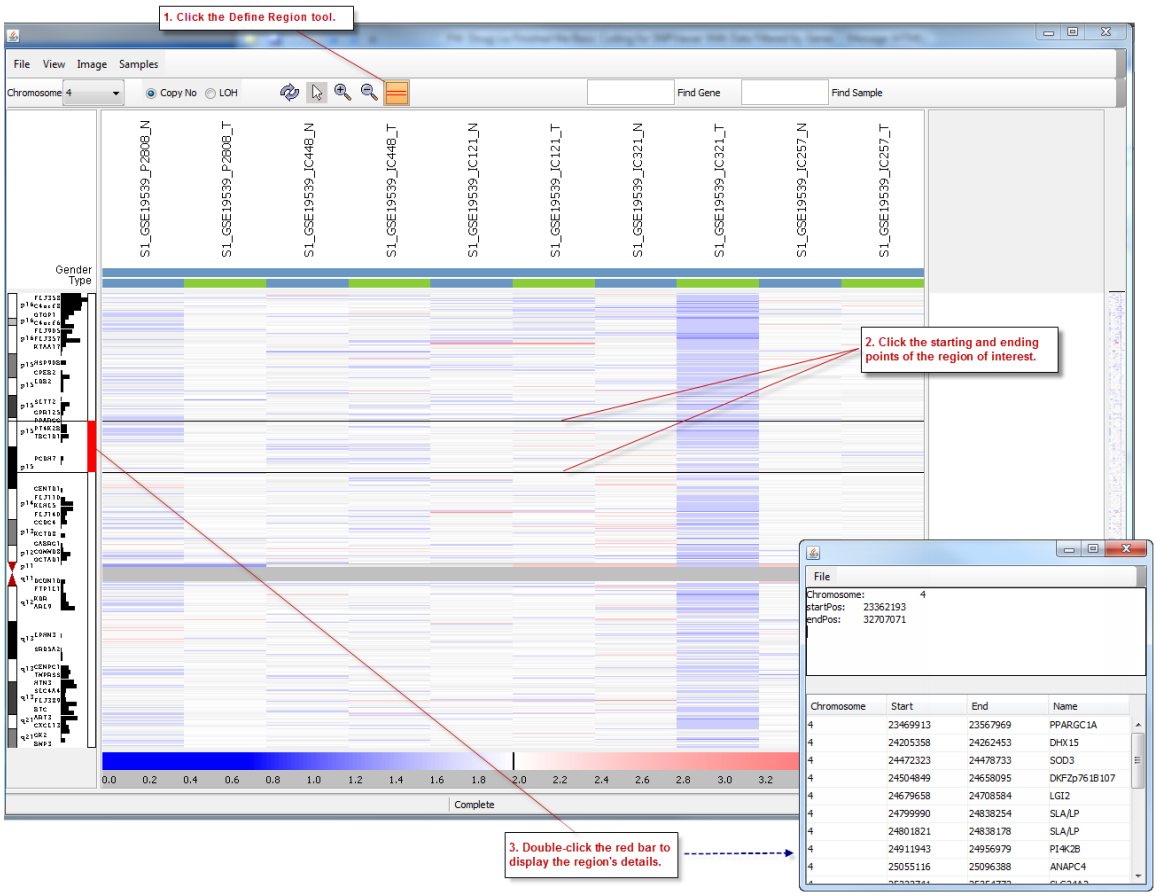
Select a Data Point and Maneuver Within It



View Genes and Data Points Within a Chromosome



Define a Region of Interest and Display Details About the Region



Running the Integrated Genomics Viewer

The Integrated Genomics Viewer is a high-performance visualization tool designed for interactive exploration of large, integrated datasets.

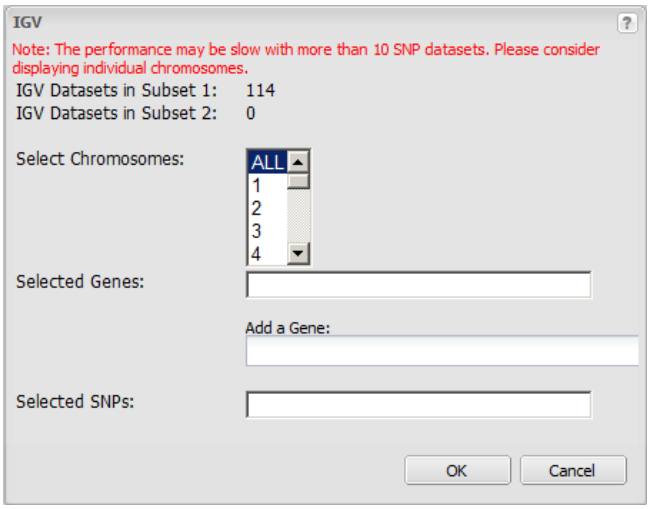
SNP array data, such as Affymetrix Genome-Wide Human SNP Array, is required for the SNPViewer.

Dataset Explorer uses the Broad Institute's IGV genomic analysis platform to generate SNP visualizations. For more information about the Broad Institute's Integrative Genomics Viewer, go to the following site: http://www.broadinstitute.org/igv/.

To run IGV:

1. Run tranSMART, then click the **Dataset Explorer** tab.
2. Select the study of interest.
3. Drag the SNP array data into an empty box in one or both subsets.
4. Define the cohorts whose data points will be represented in the genome viewer.
5. Click **Advanced**, then click **Integrative Genome Viewer**.

   The IGV dialog box appears:



6. Select the data you would like to see in the visualization. Use one of the following filtering methods:

Filtering Method	Description
By Chromosome	Select one or more chromosomes to include in the visualization, or select **ALL** to include all chromosomes. To select multiple chromosomes, click a number, then hold down **Ctrl** and click another. **Note:** If gene or SNP rs IDs are entered, the selection of chromosomes is ignored. **Note:** For performance reasons, only select chromosomes that are of particular interest to you.
By Gene	To select a gene, type all or part of the gene into the **Add a Gene** field. ▪ If you see the gene you want, click to select. ▪ If you do not see the gene you want, type more characters. Separate any additional genes with commas.
By SNP rs ID	To add a SNP rs ID, type the full ID into the **Selected SNPs** field. Separate any additional IDs with commas.
By a combination of genes or SNP rs IDs	Follow the instructions above to add a combination of genes or SNP rs IDs.

7. Click **OK**.

   If any security dialog boxes appear, acknowledge them by clicking **Continue** or **OK**.

   The visualization may require several minutes to display a large dataset, such as when multiple chromosomes are selected.

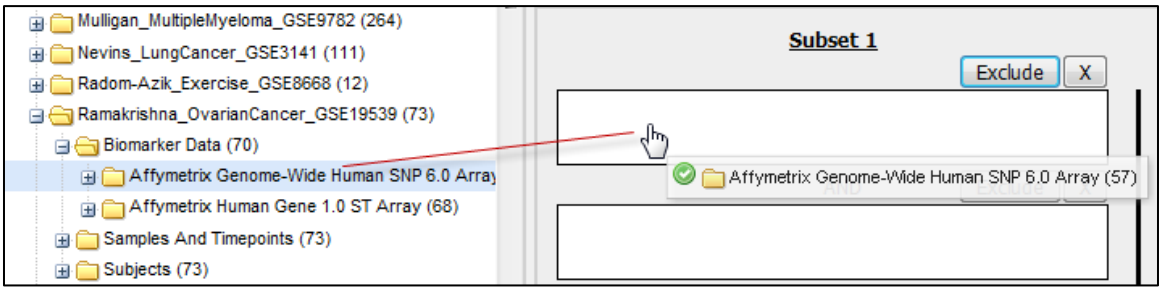
Examples

All examples use biomarker and cohort data from **Public Study GSE19539: Ramakrishna_OvarianCancer**.

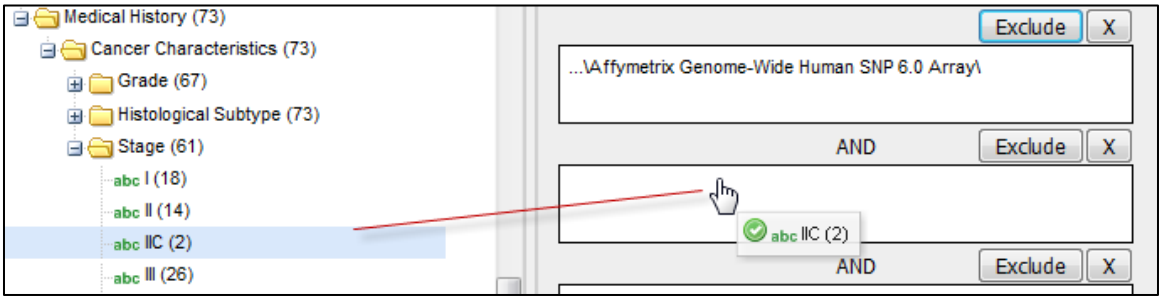
Filter Data by Chromosome

In this example you are interested in viewing chromosomes **1** and **3** in stage IIC ovarian tumors.

1. Drag the SNP array into an empty subset definition box:

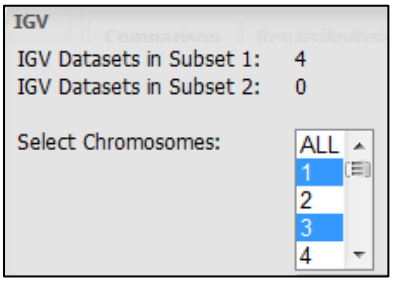


2. Limit the cohort to samples of stage IIC tumors only:



3. Click **Advanced > Integrative Genome Viewer**

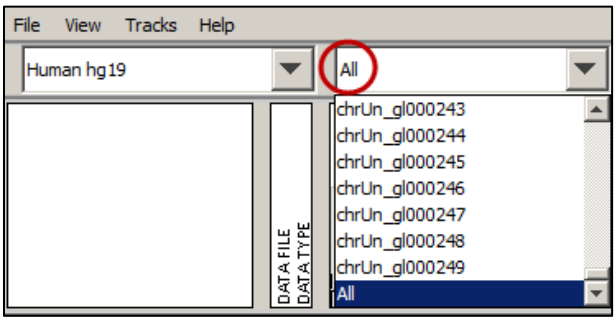
4. In the IGV dialog box, select chromosomes 1 and 3 (click **1**, then hold down the **Ctrl** key and click **3**):



5. Click **OK**.

   A Workflow Status dialog box appears showing the processing stages. When processing is complete, the visualization appears.

6. Select **All** from the chromosome-selection dropdown:



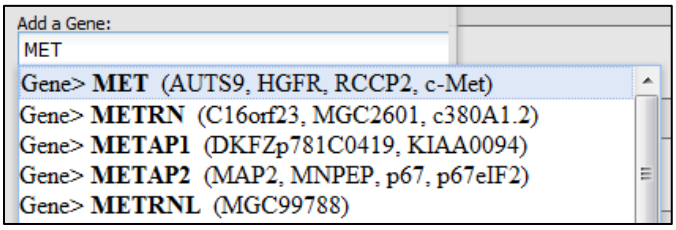
   The visualization appears as follows:



Filter Data by Gene

In this example you are interested in viewing genes **MET** and **MAPK1** in stage IIC ovarian tumors.

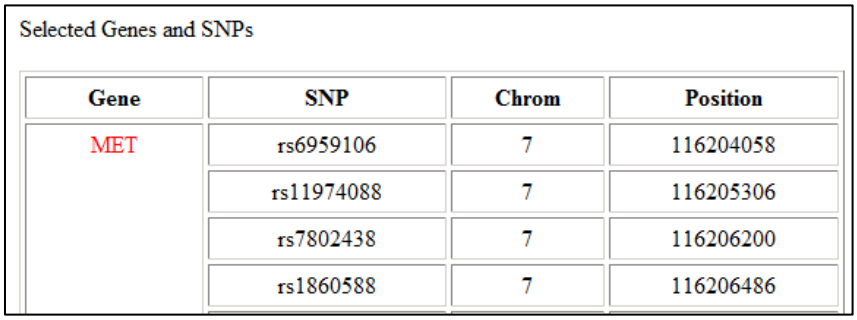
1. Follow Step 1 through Step 3 in the example Filter Data by Chromosome on page 115.
2. In the IGV dialog box, type **MET** into the **Add a Gene** field, then select **MET** from the dropdown box.



3. Repeat Step 2 to add **MAPK1**.
4. Click **OK**.

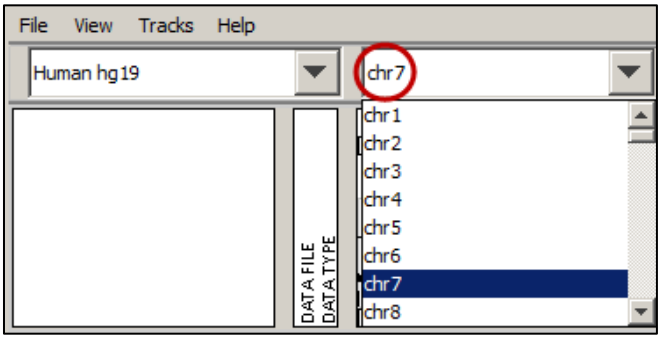
   The following responses occur:

   - A Workflow Status dialog box appears showing the processing stages. When processing is complete, a pop-up window appears that lists the selected genes and their associated SNP rs IDs. The genes you selected are shown in red. The figure below shows a portion of the pop-up:

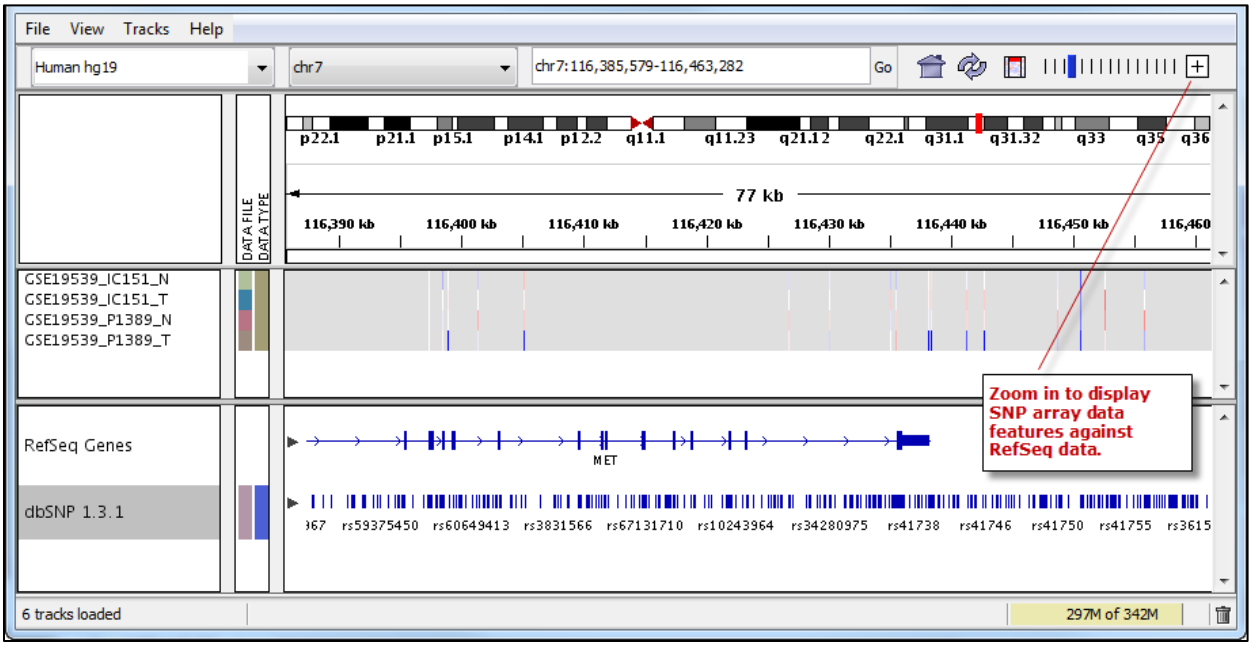


   - Further processing occurs to prepare the IGV for display.

5. When the visualization appears, select **chr 7** from the chromosome-selection dropdown:



The visualization appears as follows:



Filter Data by SNP rs ID

In this example you are interested in viewing SNP rs IDs **rs10808181** and **rs28167** in stage IIC ovarian tumors.

1. Follow Step 1 through Step 3 in the example Filter Data by Chromosome on page 115.
2. Type the full rs IDs (**rs10808181** and **rs28167**) into the **Selected SNPs** field, and separate them with commas.
3. Click **OK**.

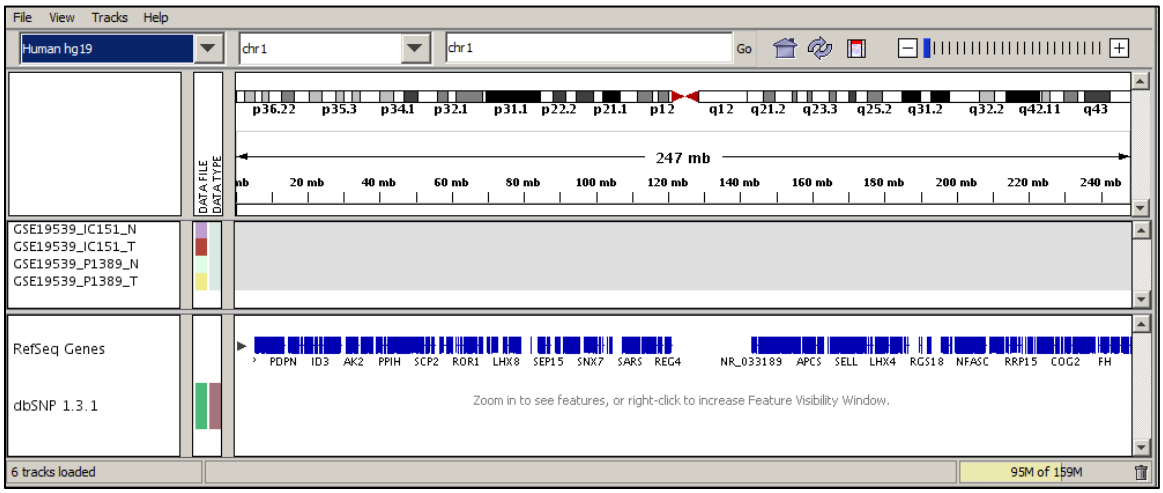
   The following responses occur:

   - A Workflow Status dialog box appears showing the processing stages. When processing is complete, a pop-up window appears that lists the specified SNP IDs and their associated genes. The SNPs you specified are shown in red:

Selected Genes and SNPs

Gene	SNP	Chrom	Position
CAV1	rs10808181	7	116241322
CAPZA2	rs28167	7	116466083

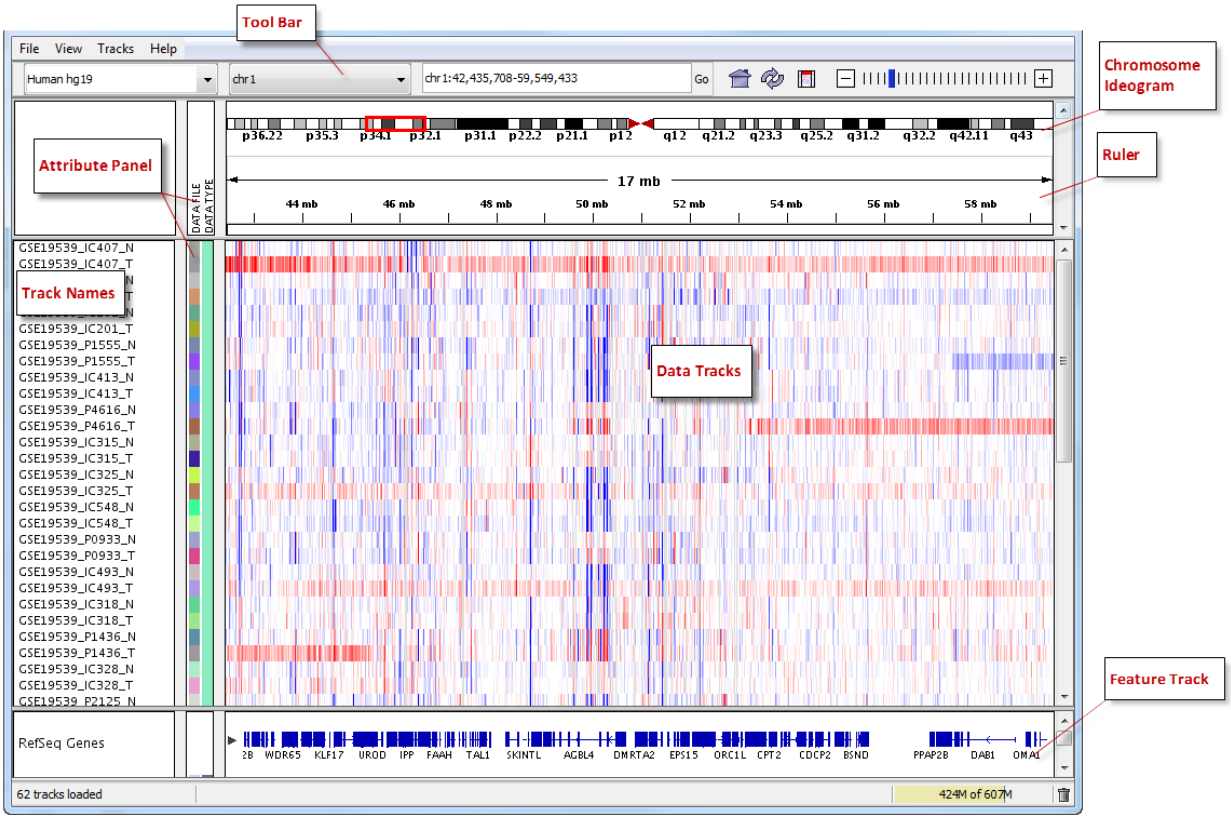
- Further processing occurs to prepare the IGV for display. When the visualization appears, it looks as follows:

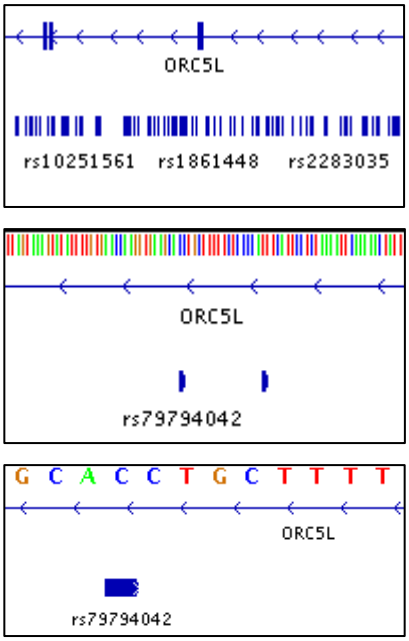


Viewing IGV Data

The operations you can perform inside the viewer include the following:

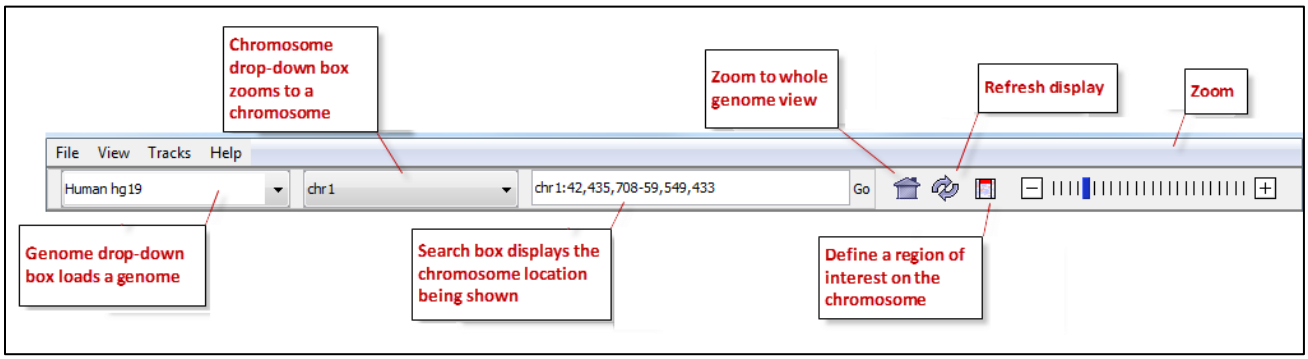
Default Display



Field	Description
Tool Bar	Provides access to commonly used functions. For more information, see the Tool Bar description below.
Chromosome Ideogram	Click anywhere along the chromosome ideogram to display data for that area. The red box on the chromosome ideogram indicates which portion of the chromosome is displayed. When zoomed out to display the full chromosome, the red box disappears from the ideogram.
Ruler	Reflects the visible portion of the chromosome. The tick marks indicate chromosome locations. The span lists the number of bases currently displayed.
Tracks	Tracks display data in horizontal rows. Typically, each track represents one sample or experiment. For each track, IGV displays the track identifier, one or more attributes, and the data.
Feature Track	Features such as genes are displayed here. Drag and drop a track name to display data in the feature track. Depending on the level to which you have zoomed, the display will change: 
Track Identifier	List of track names. Legibility of the names depends on the height of the tracks (the smaller the track, the less legible the identifier is).
Attribute Panel	Attribute names are listed at the top of the attribute panel. Colored blocks represent attribute values. Hover over a colored block to see the attribute value. Click an attribute name to sort tracks based on that attribute value.

Tool Bar

The tool bar provides quick access to locations of particular interest to you. The icons and menu options are described below:



Zoom Functions

Use the tool bar to navigate within IGV. As you use the zoom feature to view a chromosome and then a base pair resolution, the gene tracks show gene names and sequence data. If the sequence data is unavailable, small blocks replace the bases. The zoom slider does not appear when you are viewing the full genome – it reappears when you zoom in to a chromosomal level.

Using the Search Box

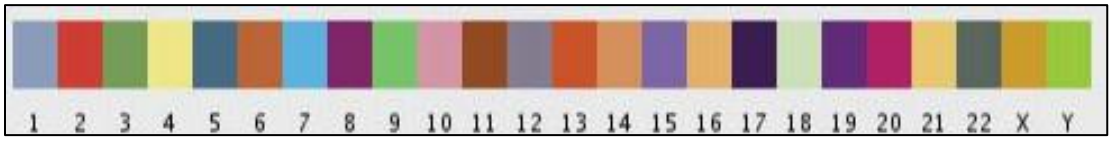
Use the search box to locate:

- A locus (for example, chr5:90,339,00-90,349,000)
- A gene symbol or other feature identifier (for example, DYPD or NM_10000000)
- A track name (for example, secondary_GBM_89)

IGV searches for an exact match to the name entered in the search box. For example, entering **secondary** will not locate the **secondary_GMB_89 track**. If multiple features have the same name, IGV jumps to an arbitrary match.

Chromosome Color Legend

The color legend is used to flag paired end reads with mates on other chromosomes in the attribute panel. The color of the read indicates which chromosome holds its mate. The color legend is shown below:

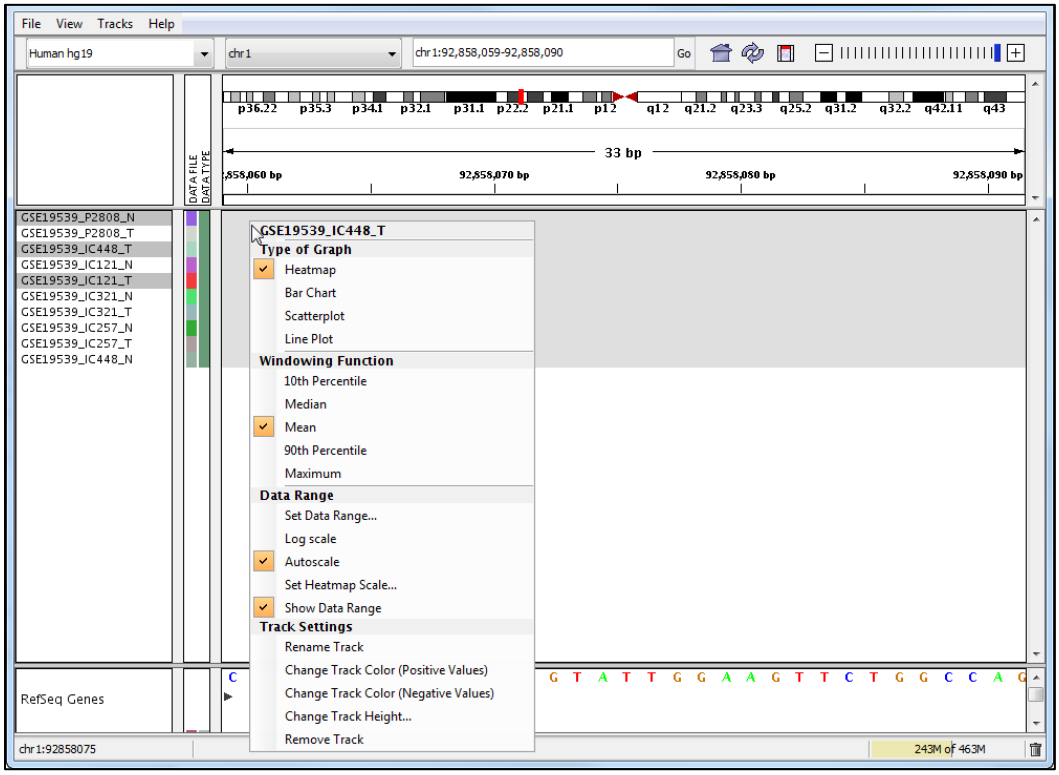


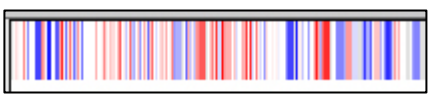
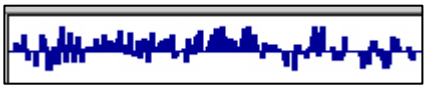
Change the Default Display

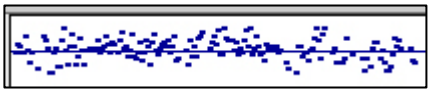
The following section describes how to change the default display to view data that is of interest.

Data Track

Right-click over data tracks to change their display. You may select multiple tracks to edit by using the **Ctrl** key (click a track name, then hold down the **Ctrl** key and click another). Tracks you have selected will be highlighted in grey. The table after the figure below describes the functions you can perform in the display menu.



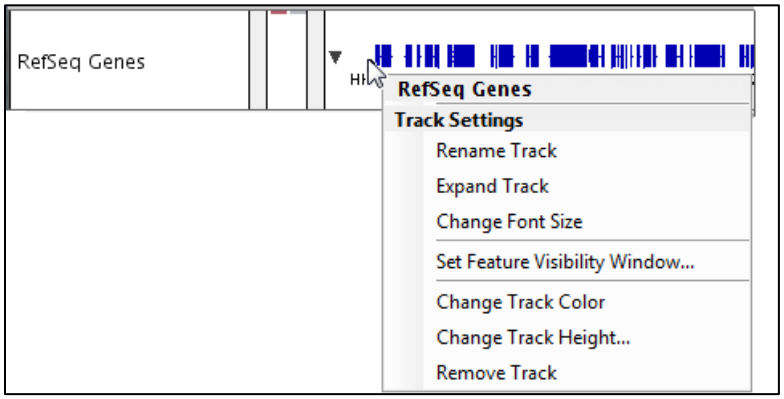
Menu Category	Sub Category	Description
Type of Graph	Heat map	Default Option, displays track data in the form of a heat map: 
	Bar Chart	Displays track data in the form of a bar graph: 

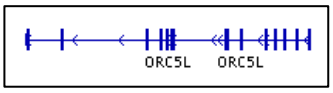
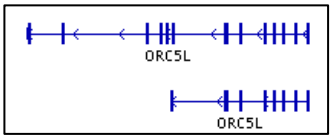
Menu Category	Sub Category	Description
	Scatterplot	Displays track data in the form of a scatterplot: 
	Line Plot	Displays track data in the form of a line plot: 
Windowing Function	10th Percentile	Changes the value represented by each pixel of track data. At all but the lowest zoom levels, each pixel represents a significant amount of data. IGV divides the data to be displays into "windows" of equal length each corresponding to a single pixel, summarizes the values across each window, and then displays the summarized values in the track. Select the function IGV will use to summarize the values. The default window function summarizes values by mean.
	Median	
	Mean	
	90th Percentile	
	Maximum	
Data Range	Set Data Range…	Changes the minimum, baseline, and maximum values of the graph used to display track data.
	Log scale	Plots the chart for that track on a log scale.
	Autoscale	Default option. Toggles the autoscaling function for a given track. With autoscaling enabled, IGV adjusts the plot Y scale to the data range currently in view. Scaling will adjust continually as you navigate through data.
	Set Heat map Scale...	Changes the data range and color of the heat maps used to display track data.
	Show Data Range	Toggles whether the numeric range of values in the view for a given track is displayed; this function works for all charts except heat maps.
Track Settings	Rename Track	Renames a track.
	Change Track Color (Positive Values)	Changes the track color for selected tracks.
	Change Track Color (Negative Values)	Changes the track color for selected tracks.

Menu Category	Sub Category	Description
	Change Track Height…	Changes the track height for selected tracks.
	Remove Track	Removes selected tracks from the display.

Feature Track

Feature tracks identify genomic features. By default all features in a track are drawn on a single line, including features that might overlap, such as alternative isoforms of a transcript. Right-click over the feature track to change its display. The table below describes the functions you can perform in the feature track menu.



Menu Category	Description
Rename Track	Renames a track.
Expand Track/Collapse Track	Displays overlapping features, such as different transcripts of a gene on one line or multiple lines: **Collapsed State (default):**  **Expanded State:** 
Change Font Size	Changes the font size of the feature labels.

Menu Category	Description
Set Feature Visibility Window	Specifies the threshold, in kilobases, for IGV to display features in the window. For example, if you set this at 50kb, IGV will only display features after you have zoomed in to display 50 kb or less in the IGV window.
Change Track Color	Changes the track color for selected tracks.
Change Track Height	Changes the track height for selected tracks.
Remove Track	Removes selected tracks from the display.

Hypothesis Testing Across Multiple Studies

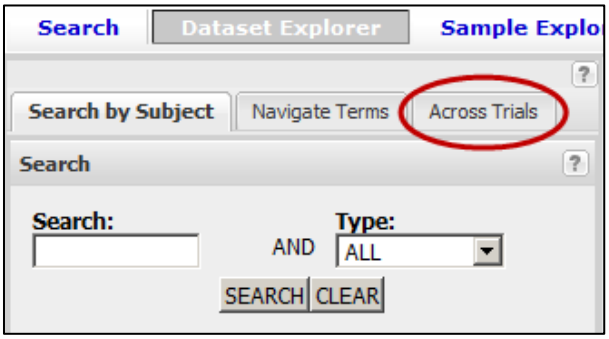
You can perform cohort comparisons across many Johnson & Johnson clinical trials based on CDISC standards. Across-trial comparisons can be performed with RBM data and with certain clinical trial data.

Note: Across-trial comparisons do not support heat maps, principal component analyses, or haploviews.

Across-Trials Comparisons of Clinical Data

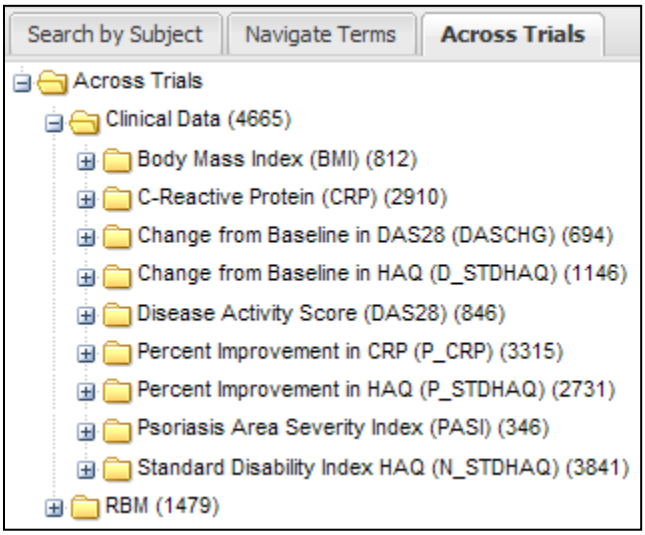
To perform an across-trials comparison of clinical data:

1. Run tranSMART, then click the **Dataset Explorer** tab.
2. In the left pane, click the new **Across Trials** tab:



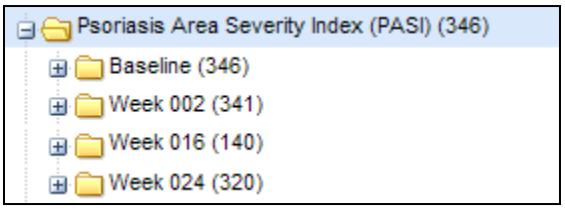
3. Open the **Across Trials** node, then open the **Clinical Data** node.

A list of Clinical Data nodes appears. The nodes represent all the types of clinical data that can be used in across-trials comparisons:



4. Open the node representing the clinical data to use in the comparison.

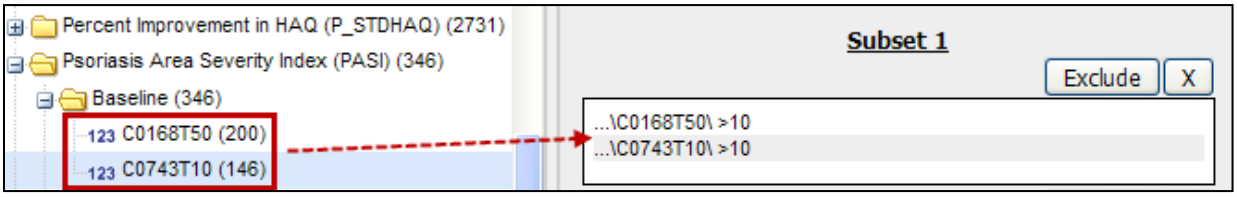
   In the following example, the Psoriasis Area Severity Index (PASI) node is opened:



5. Open the timepoint to use as the starting timepoint for the comparison – for example, Baseline.

6. Drag the studies to use in the comparison into the same subset box in Subset 1.

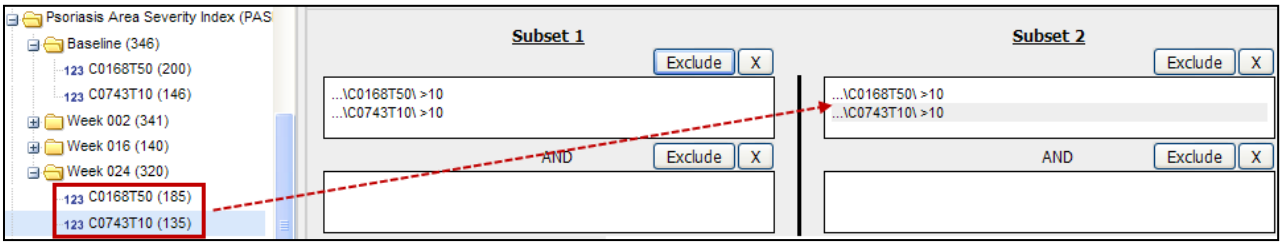
   In the following example, two studies have been dragged into Subset 1 from the Baseline timepoint:



7. Open the timepoint to use as the ending timepoint for the comparison.

8. Drag the studies to use in the comparison into the same subset box in Subset 2.

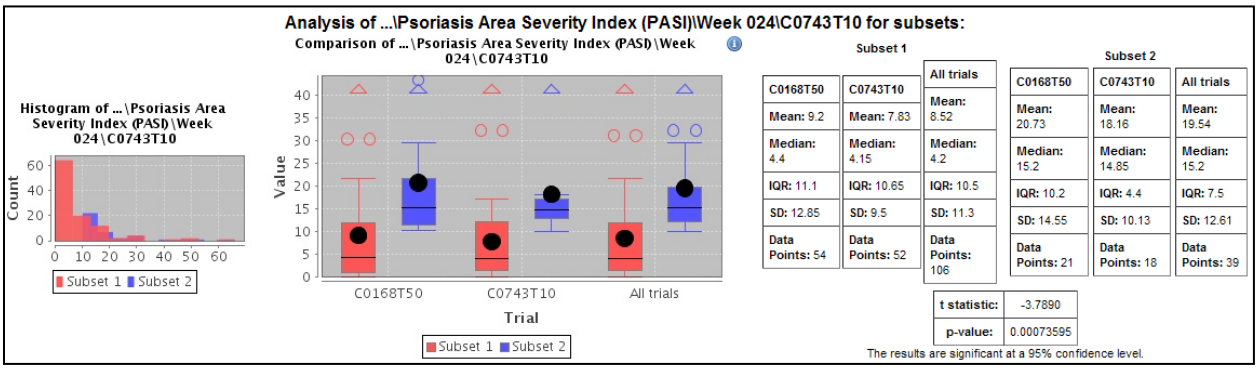
In the following example, the same two studies have been dragged into Subset 2 from the Week 24 timepoint:



9. Click **Generate Summary Statistics**.

   The comparison appears at the bottom of the Results/Analysis view.

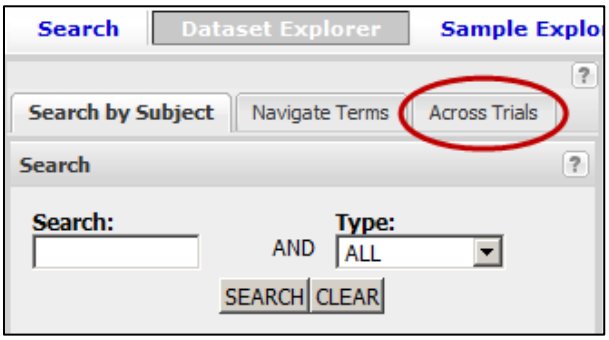
   The following figure shows the Week 24 results for the comparison created in the steps above:



Across-Trials Comparisons of RBM Data

To perform an across-trials comparison of RBM data:

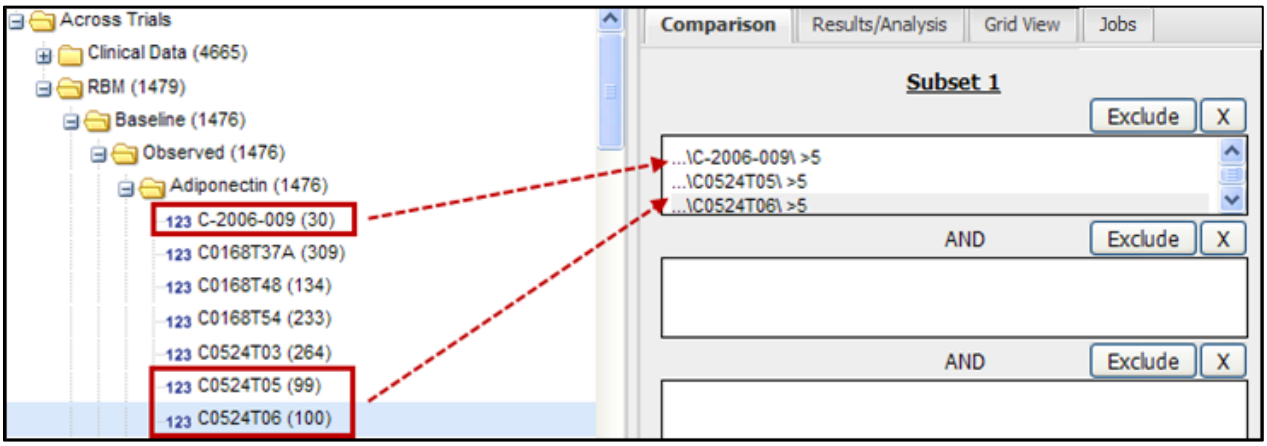
1. Run tranSMART, then click the **Dataset Explorer** tab.
2. In the left pane, click the new **Across Trials** tab:



3. Open the **Across Trials** node, then open the **RBM** node.
4. Open the **Baseline** node, then open **Observed** or **Z Score**.

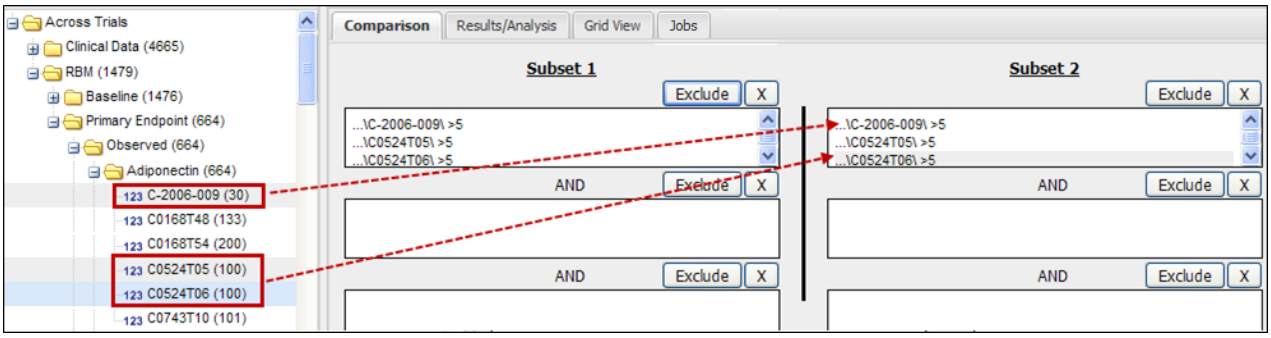
5. Open an antigen node.
6. Drag the studies to use in the comparison into the same subset box in Subset 1.

   In the following example, three studies have been dragged into Subset 1 from the node representing the antigen **Adiponectin**:



7. Open the **Primary Endpoint** node, then open **Observed** or **Z Score**.
8. Open the same antigen node.
9. Drag the studies to use in the comparison into the same subset box in Subset 2.

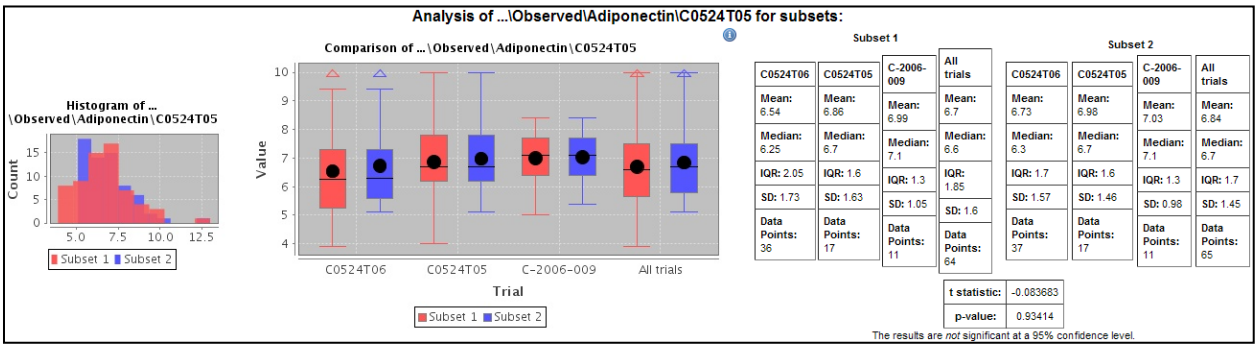
   In the following example, the same three studies have been dragged into Subset 2:



10. Click **Generate Summary Statistics**.

    The comparison appears at the bottom of the Results/Analysis view.

The following figure shows the primary endpoint results for the comparison created in the steps above:

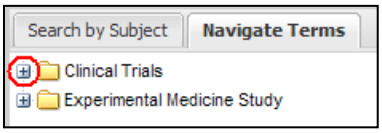


Viewing a Study

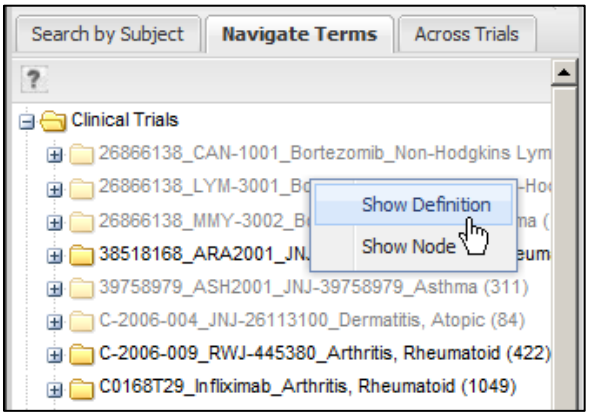
You can view a description of any Dataset Explorer study, whether or not you have access rights to the study.

To view a description of a study:

1. In Dataset Explorer, click the **Navigate Terms** tab.
2. Open the top-level node for the list of studies you are interested in – for example, click the **+** icon (⊞) next to Clinical Trials to open the list of clinical trials:



3. Right-click the particular study you are interested in.
4. Click the **Show Definition** popup:



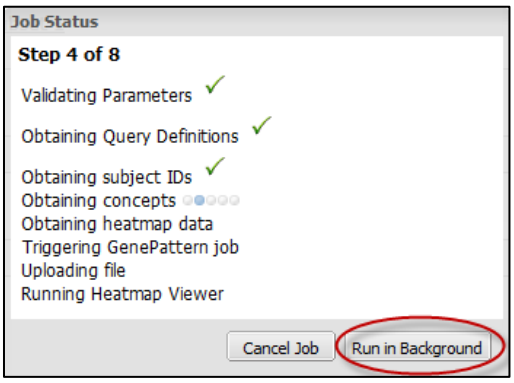
The Show Concept Definition dialog appears, showing the title, description, and other information about the study.

Asynchronous Operations

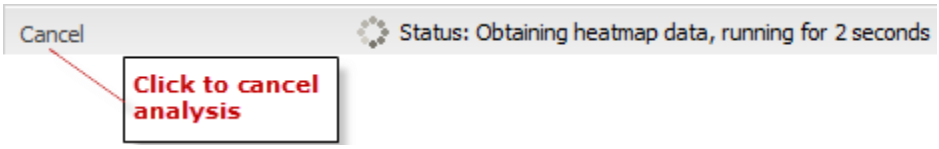
You may run multiple advanced workflow operations in Dataset Explorer asynchronously. Analyses run in the background of the program, allowing you to use other features of the tranSMART application or to perform additional analyses simultaneously within Dataset Explorer.

To run advanced workflow(s) in the background of Dataset Explorer:

1. Select the type of advanced workflow to run from the **Advanced** menu.
2. When the **Job Status** dialog box appears, select **Run in Background**:

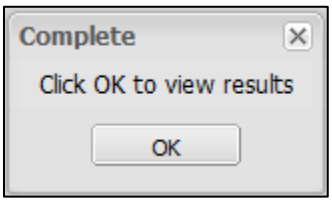


After selecting Run in Background, you will see the status of your job in the lower right corner of your browser:



If you run multiple jobs in the background, the status will cycle through each job in the order that the jobs were started.

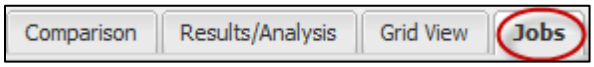
3. Click **OK** to view results:



Note: For smaller analyses you will not see the **Complete** dialog box – the visualizer will appear automatically.

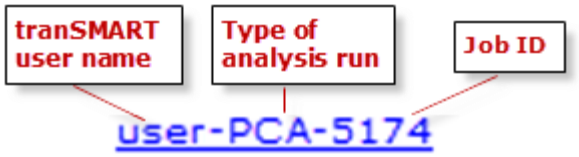
Jobs Tab

The Jobs tab allows you to review analyses you have run previously, and also to see the status of analyses you have chosen to run in the background.

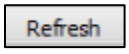


Each advanced workflow that you have run in the past seven days is logged in the Jobs tab in a spreadsheet format.

The columns of information in the Jobs tab are described below:

Column	Description
Name	The name of the analysis run. The format of the name is as follows: 
Status	The status of the analysis. Statuses are explained below: ▪ **Completed** – The job has finished and a visualization is available. ▪ **Started** – The job has been started and is still processing. ▪ **Uploading File** – You have selected to load additional data into your visualization, and the data is still in the process of uploading to tranSMART. ▪ **Error** – The job did not complete due to an error. ▪ **Cancelled** – The job was cancelled and will not complete.
Run Time	The time the analysis took to process.
Started On	The date and time that the analysis was first started.

Note: Click the Refresh button to view any changes that have been made since the Jobs tab initially populated.

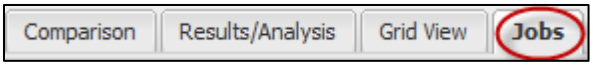


Viewing a Logged Job

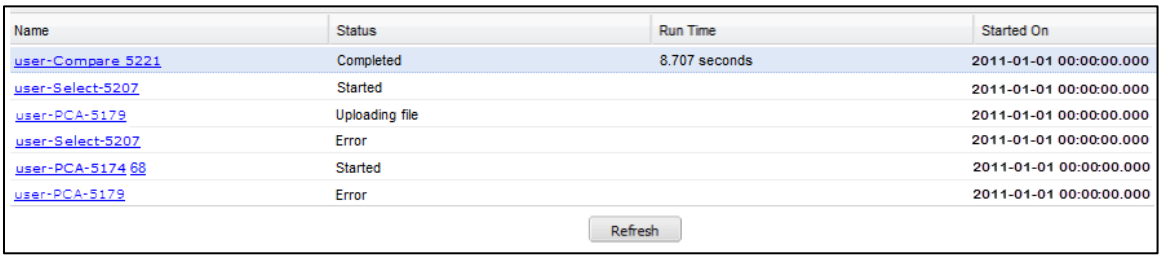
Each advanced analysis that you have run in the previous seven days will be logged in the Jobs tab. You may view the visualization again by selecting it from the list.

To run a logged advanced workflow:

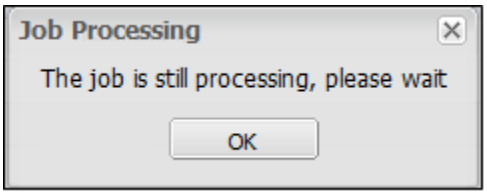
1. Run tranSMART, then click the **Dataset Explorer** tab.
2. In the right pane, click the **Jobs** tab:



3. Click the hyperlink of the analysis you are interested in viewing:



If you click on a job that has not been completed, you will see the following dialog box:



Chapter 4
Sample Explorer

Sample Explorer lets you search for tissue and blood samples of interest so that you can learn more about the samples; for example, you can:

- Look up Biobank IDs for samples so that you can locate them in the Biobank
- Locate the study that produced the samples in the Dataset Explorer
- Project sample data onto a heat map

The Sample Explorer window has two panes:

- **Right pane – Select a primary search filter**

  Lets you begin to search for samples. For information, see Select a Primary Search Filter below.

- **Left pane – Recent Updates**

  Lists up to ten of the most recent sample updates in the database.

  For information about a sample update, including the number and source of updated records, click the item in the list.

Select a Primary Search Filter

This pane of the Sample Explorer window lets you initiate a search for samples by selecting the primary search filter. After you select a search filter, a second Sample Explorer window appears where you can view the search results and refine the search by selecting additional filters.

Search filters are organized in the following categories:

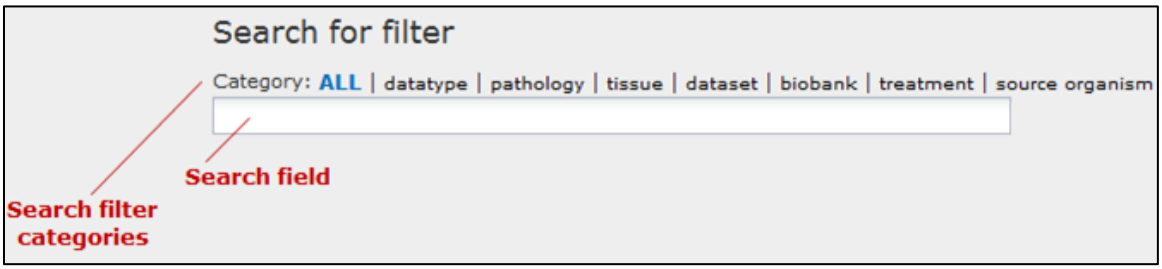
- Data type – Biomarkers such as gene expression, RBM, and SNP
- Pathology – The type of disease associated with the samples
- Tissue – The physical source of the samples, such as liver or colon tissue
- Dataset – The study that generated the samples
- Biobank – Indication of whether the samples are in the Biobank (Yes or No)
- Treatment – The type of treatment, such as a particular compound, being tested in the clinical trial. The treatment may be administered to the test subjects or directly to the samples.
- Source Organism – Indication of whether the samples were collected from a human or other organism.

Note that the number of samples that are associated with a filter appear in parentheses after the filter name.

You can select a select a primary filter by searching or by browsing for the filter.

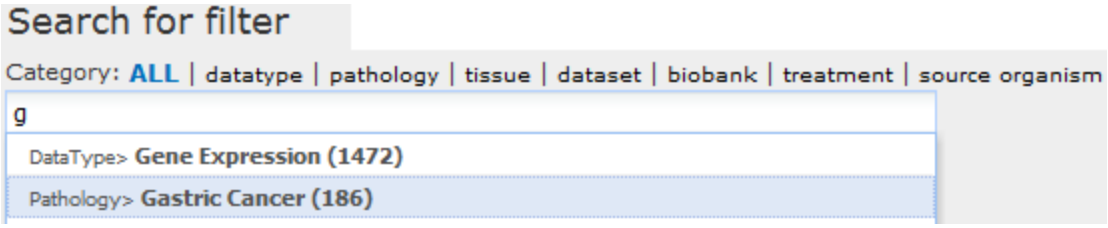
To search for a primary search filter:

1. Click the search filter category to search within, or accept the default of **All** categories:



2. Type part or all of the filter name into the **Search** field.

   The search engine displays a dropdown list containing all the filters within the selected category that begin with the text you typed. For example, if you type the letter **G** in the **Search** field for an all-category search, you might see this:



   Up to 20 filters can be listed. If the filter you want does not appear, type a more complete name in the **Search** field.

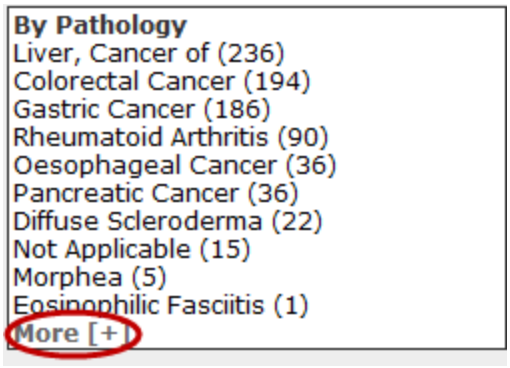
3. When the filter you want appears in the list, click the filter name.

   The search begins immediately, and the results are displayed in a new window (see View and Refine Sample Search Results on page 136).

   Note: You can only initiate a search by clicking a filter name in the dropdown list. You cannot initiate the search by typing the filter name and pressing the **Enter** key.

To browse for a primary search filter:

1. Click a filter name in one of the category browser boxes displayed below the Search filter.
2. If you do not see the filter you want in a particular category, click **More** at the bottom of the box:

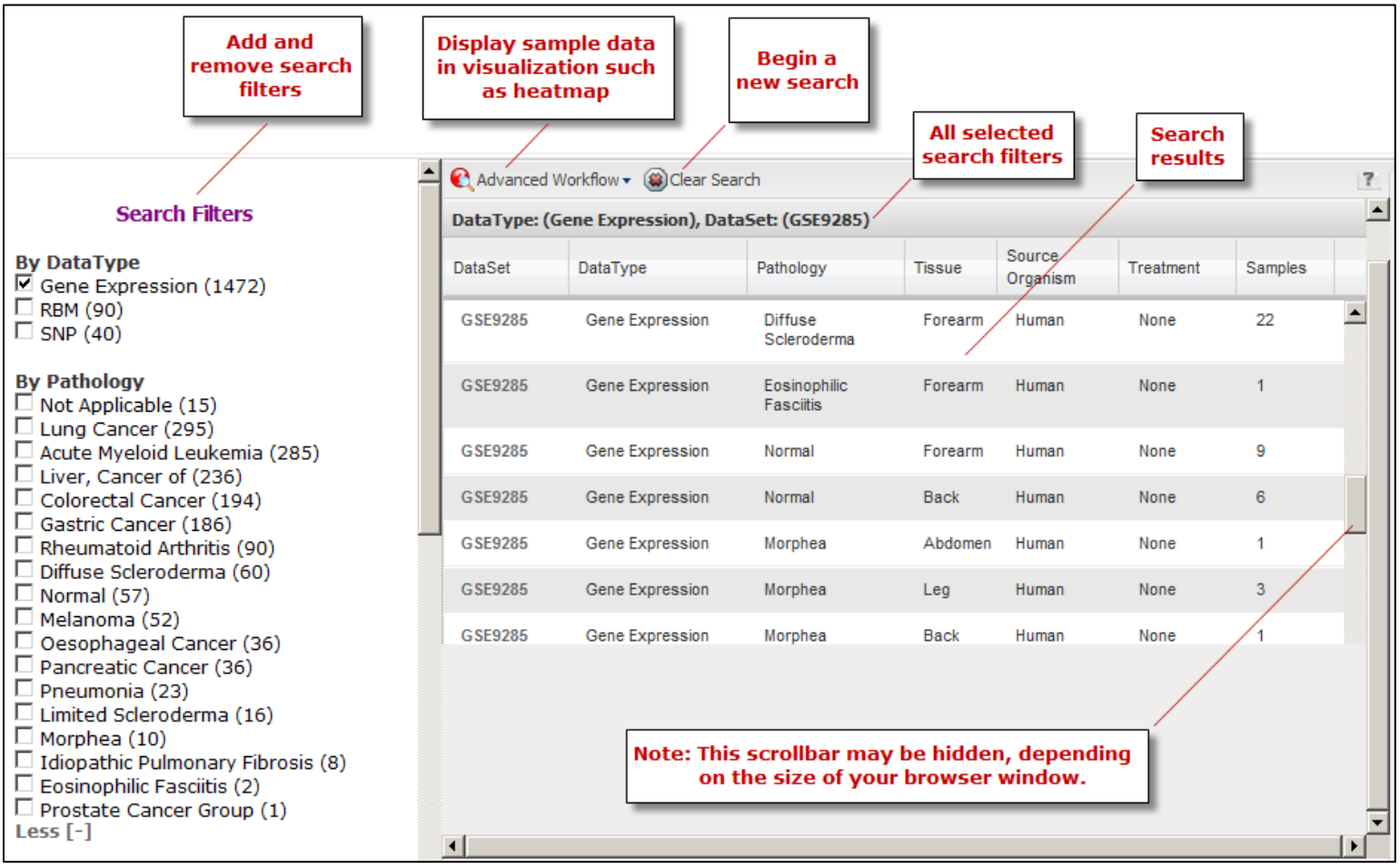


When you click a filter, the search begins immediately, and the results are displayed in a new window (see View and Refine Sample Search Results on page 136).

View and Refine Sample Search Results

After you have selected a primary search filter, a new Sample Explorer window appears, displaying the results of the search. The left pane of the window contains all the search filters, allowing you to narrow the search results.

The following figure illustrates the sections of this Sample Explorer window:



You can perform the following tasks in this Sample Explorer window:

- Select and remove search filters
- Display information to help you find samples in the Biobank
- Locate the study that produced the samples in the Dataset Explorer
- Project sample data onto a heat map
- Re-sort the search results, and add/remove search result columns

Select and Remove Search Filters

You can refine a sample search result by adding and removing search filters, including the primary filter you initially selected. Search filters are listed in the left pane of the Sample Explorer window.

To select or remove a search filter, check or clear the check box next to the filter name.

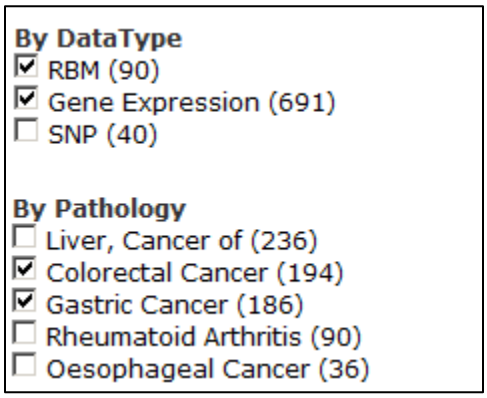
> **Note:** Clicking a filter name rather than the check box next to the name will select that filter and deselect all currently selected filters.

The filters you select are joined together in a search string by the logical operators `AND` and `OR`, as follows:

- Filters within a filter category (such as DataType or Pathology) are joined by `OR`.
- Filters in different filter categories are joined by `AND`.

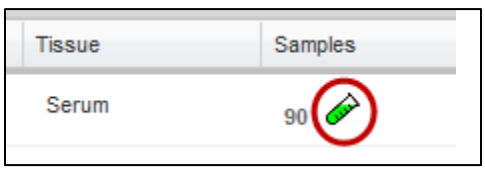
For example, the search string for the filter selections illustrated below is:

```
(RBM OR Gene Expression) AND (Colorectal Cancer OR Gastric Cancer)
```



Find Samples in the Biobank

Many of the samples that you access through the Sample Explorer are in the Biobank. If a dataset contains samples that are in the Biobank, the dataset is flagged with an icon:

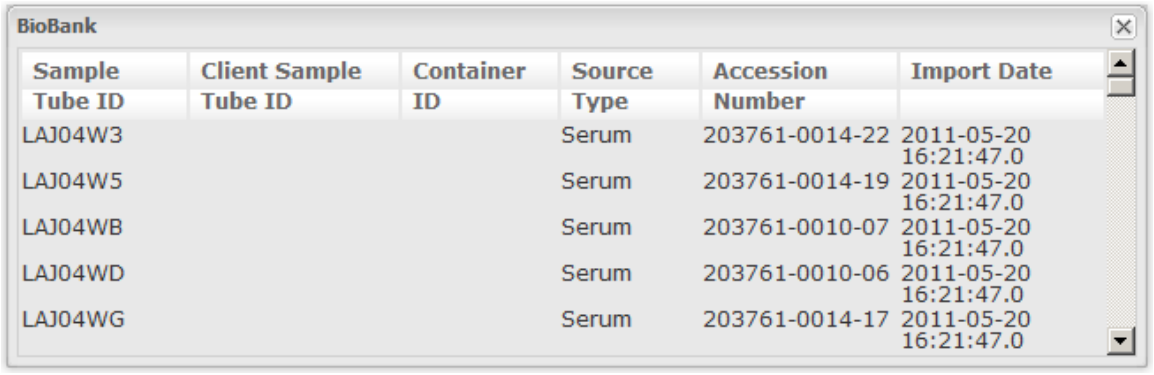


The Sample Explorer lets you display Biobank reference information for samples that are in the Biobank so that you can locate the samples there.

To display Biobank reference information for samples:

1. If the dataset of interest is not included in the result set, refine the search by selecting additional search filters (see Select and Remove Search Filters on page 137).
2. When the dataset of interest appears, check whether it has the Biobank icon displayed in the **Samples** column of the search result.
3. If the icon is displayed, click the number to the left of the icon (the number **90** in the figure above).

   The Biobank dialog box appears, displaying reference information for each of the samples in the dataset:



Note: If you want the Sample Explorer to display only those samples that are in the Biobank, select **Yes** in the **By BioBank** search category.

Locate the Source of the Samples in Dataset Explorer

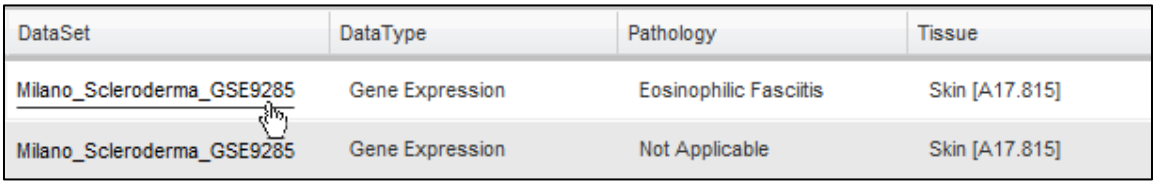
If a dataset of samples was collected for a Dataset Explorer study, you can link back to the study to view information such as the study owner, study description and purpose, demographics of the participants, and other data relevant to the samples.

Note: When you link back to a Dataset Explorer study, and then return to Sample Explorer, the filters you had previously selected in Sample Explorer are cleared.

To link back to the associated Dataset Explorer study:

1. If the dataset of interest is not included in the result set, refine the search by selecting additional search filters (see Select and Remove Search Filters on page 137).

2. When the dataset of interest appears, click the dataset name in the **DataSet** column of the result set:

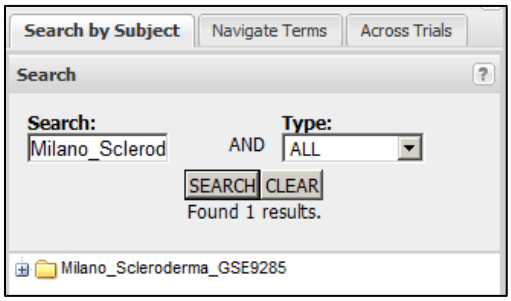


When you click a dataset link, the following actions occur automatically:

a. Dataset Explorer opens.

b. The dataset name you clicked is inserted into the **Search** field of the Dataset Explorer **Search By Subject** tab.

c. The search is immediately executed, and one or more matching studies, or sub-nodes of studies, is listed below the **Search** field:



3. Open and explore the study of interest.

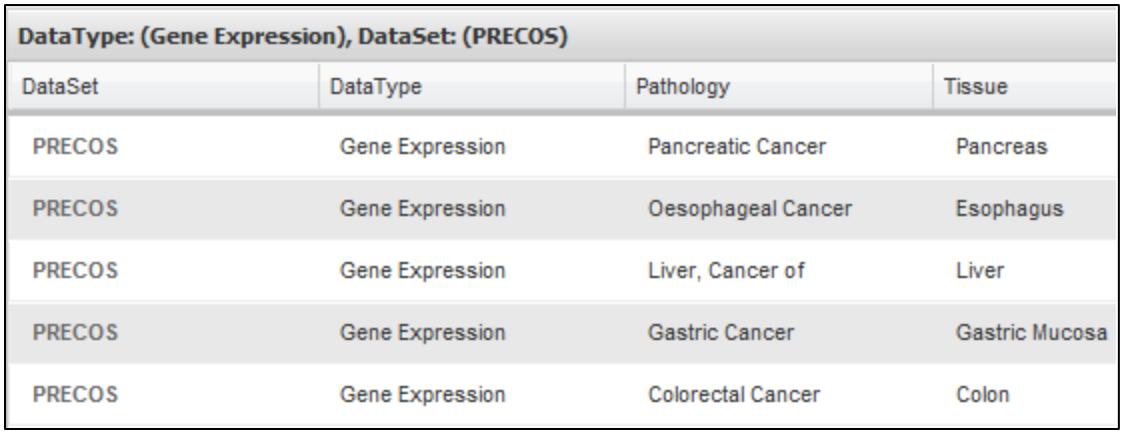
For information, see Branches and Leaves of the Navigation Tree on page 58.

Note: If the study name is grayed out, or an Access Is Restricted warning is displayed when you try to open the study, you have not been granted access to the study. Contact a tranSMART administrator if you want to request access. For more information, see Public and Private Studies on page 54.

Multiple Studies Mapped to a Single Sample Dataset

In some cases, a single Sample Explorer dataset will be associated with multiple Dataset Explorer studies.

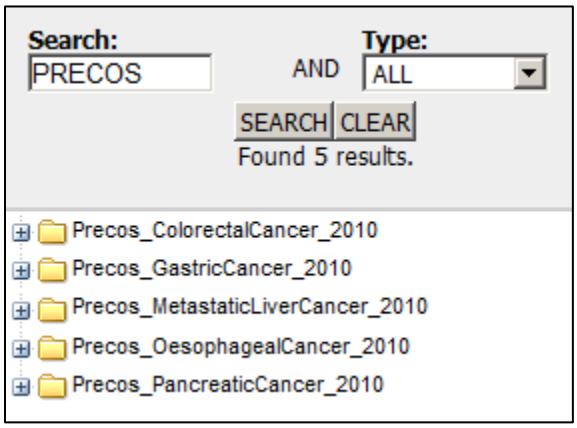
For example, if you click any one of the following **PRECOS** links in the Sample Explorer **DataSet** column…



DataType: (Gene Expression), DataSet: (PRECOS)

DataSet	DataType	Pathology	Tissue
PRECOS	Gene Expression	Pancreatic Cancer	Pancreas
PRECOS	Gene Expression	Oesophageal Cancer	Esophagus
PRECOS	Gene Expression	Liver, Cancer of	Liver
PRECOS	Gene Expression	Gastric Cancer	Gastric Mucosa
PRECOS	Gene Expression	Colorectal Cancer	Colon

…you will see the following Dataset Explorer search result:



In these cases, open the study or studies for which the samples of interest were collected.

Display Sample Data in a Heat Map

Gene expression data, RBM data, and any other data that can be displayed in a Dataset Explorer heat map can also be displayed in a Sample Explorer heat map. You can select one or more datasets for display.

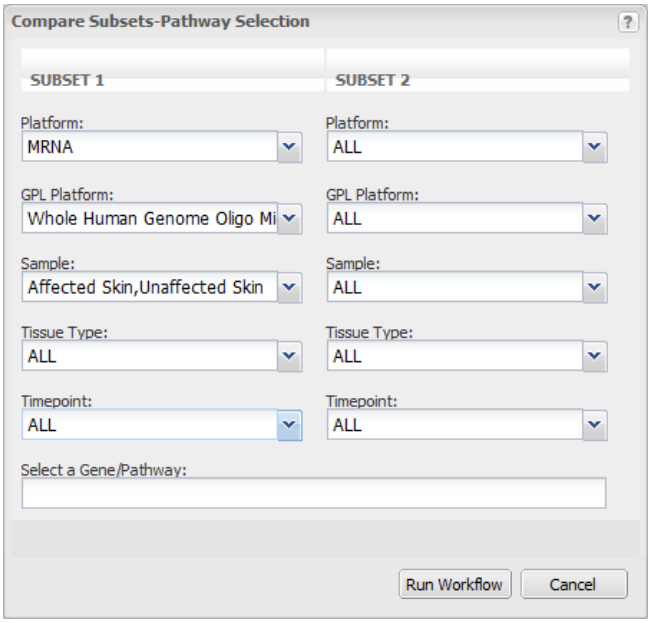
To display sample data in a heat map:

1. If the dataset of interest is not included in the result set, refine the search by selecting additional search filters (see Select and Remove Search Filters on page 137).
2. When a dataset of interest appears, click the dataset's row.

   Note: Do not click the dataset name. Clicking the dataset name will open Dataset Explorer, and you will lose any search filters you selected in Sample Explorer.

3. To display data from multiple datasets, use these techniques:
   - To select non-sequential rows of datasets, press and hold down the **Ctrl** key, then click each row to include.
   - To select sequential rows of datasets, press and hold down the **Ctrl** key to select the top- or bottom-most row, then hold down the **Shift** key and click the row at the opposite end.
4. To deselect a selected row, hold down the **Ctrl** key and click the row again.
5. When all rows to include are selected, click **Advanced Workflow > Heatmap**.

   A dialog box like the following appears:



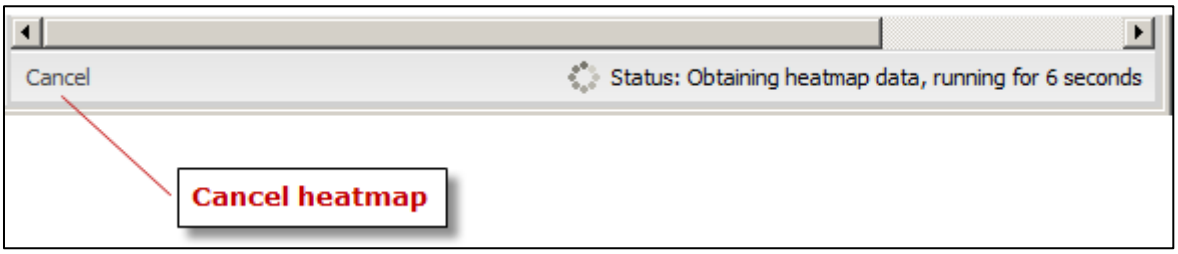
6. Select the parameters for the data to be displayed, or keep the defaults.

   For information about the parameters, see Requirements for Generating Heat Maps on page 75.

   Note: You can only select parameters for Subset 1.

7. Click **Run Workflow**.

   The visualization may require a minute or more to process, depending on the amount of data to be displayed. The processing time is shown at the bottom-right of the window:



8. Click **OK** when the completion dialog box appears.

   Again, it may take some time for the visualization to be downloaded and displayed.

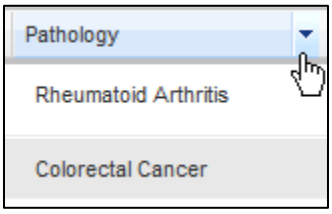
Manage the Sample Search Result List

You can make the following adjustments to the search result list:

Sort by Column

To sort the result list by the contents of a column:

1. Click the right side of the column heading to pull down the menu:

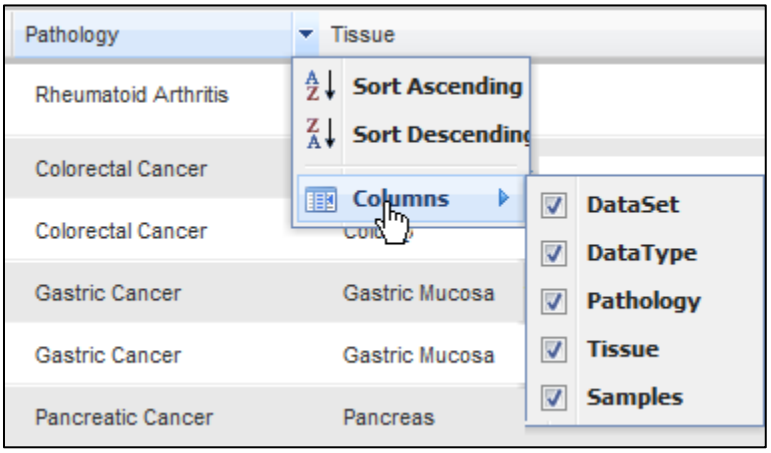


2. Click **Sort Ascending** or **Sort Descending**.

Add and Remove Columns

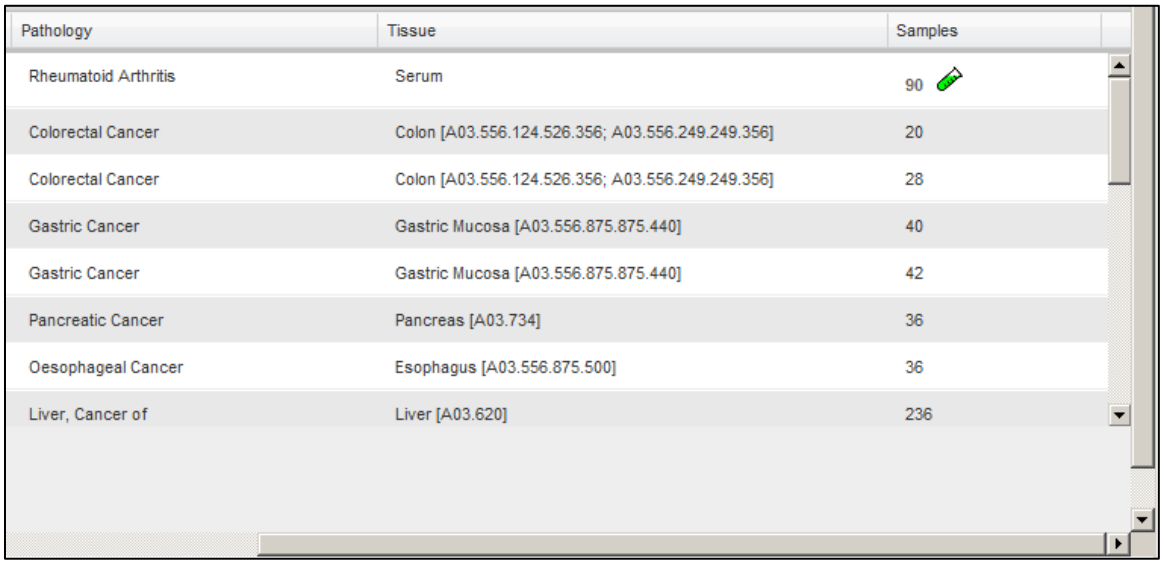
To add and remove columns:

1. Click the right side of the column header to pull down the menu:
2. Hover the mouse pointer over Columns to display the submenu of column headings:



3. Check or clear the check boxes to add or remove a column from the search result.

Note: If there are more rows in the result set than can be displayed at one time, a vertical scroll bar appears at the right of the result set. However, this scroll bar may be hidden from view. To check, move the horizontal scroll bar at the bottom of the window all the way to the right to expose the result set's vertical scroll bar. If the vertical scroll bar is not there, all the rows in the result set currently are displayed.



Pathology	Tissue	Samples
Rheumatoid Arthritis	Serum	90
Colorectal Cancer	Colon [A03.556.124.526.356; A03.556.249.249.356]	20
Colorectal Cancer	Colon [A03.556.124.526.356; A03.556.249.249.356]	28
Gastric Cancer	Gastric Mucosa [A03.556.875.875.440]	40
Gastric Cancer	Gastric Mucosa [A03.556.875.875.440]	42
Pancreatic Cancer	Pancreas [A03.734]	36
Oesophageal Cancer	Esophagus [A03.556.875.500]	36
Liver, Cancer of	Liver [A03.620]	236

Chapter 5

Gene Signatures and Gene Lists

The tranSMART gene signature wizard guides you through the process of creating a gene signature or gene list. You specify whether the gene signature or list is publicly available to other tranSMART users or is reserved for your private use.

Once you create the gene signature or list, it can be used in tranSMART searches to find clinical studies and experiments where the differentially regulated genes overlap with the genes contained in the gene signature or list. This will generate a set of hypotheses about diseases or treatments that may have similar genes disregulated, and that can help you develop a further set of experiments.

Note: This chapter uses the term "gene signature" to refer to both gene signatures and gene lists.

Creating a Gene Signature

There are two basic tasks involved in creating a gene signature:

1. Add the list of genes for the gene signature to a text file.

   Genes can be indicated by gene symbol or by their associated probe set ID.

2. Use the gene signature wizard to define the information on which the gene signature is based, such as species, source of data, and test type, and also to import into the gene signature definition the text file containing the genes.

Step 1. Adding the Genes to a Text File

The gene signature wizard expects to import the genes for the gene signature from a tab-separated text file. The file must contain one, and possibly two, columns of information:

- First column – A list of gene symbols or probe set IDs.
- Optional second column – The fold change ratios associated with the gene symbols or probe set IDs.

  The fold change ratios can be either **actual values** (for example, 12.8 or -12.8) or one of the following **composite values:**

  - **-1**. All down-regulated gene expressions.
  - **1**. All up-regulated gene expressions.

- **0**. No change.

The following table shows the different ways you can specify the genes for your gene signature:

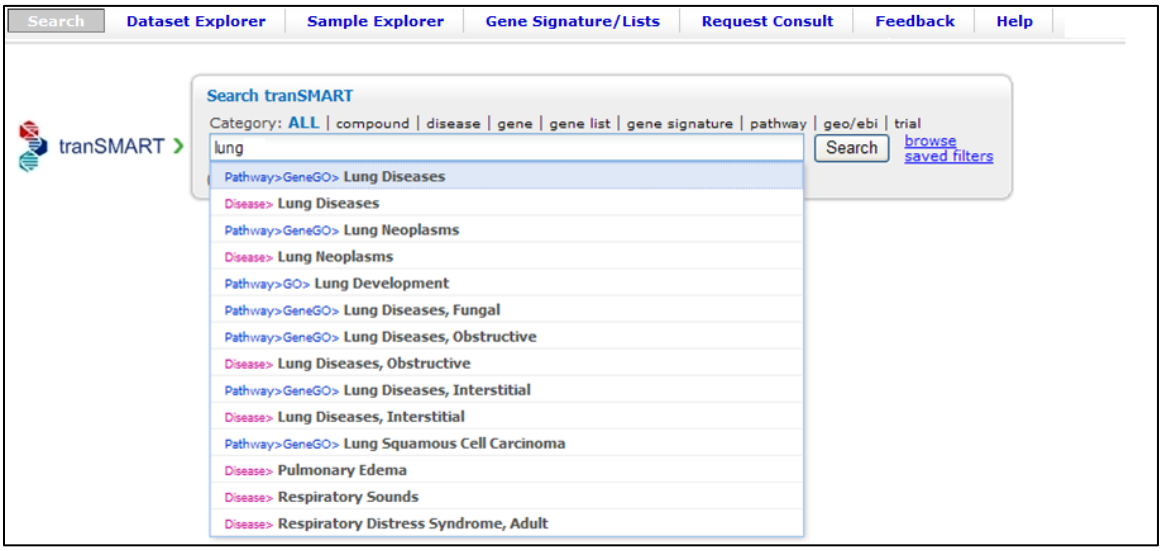
Contents of File	Format	Examples
Gene symbols only	*GeneSymbol*	TCN1 IL1RN KIAA1199 G0S2
Gene symbols, actual fold change	*GeneSymbol*<tab>*ActualFC*	CXCL5 -19.19385797 IL8RB -18.21493625 FPR1 -17.6056338 FCGR3A -15.69858713
Gene symbols, composite fold change	*GeneSymbol*<tab>*CompositeFC*	CXCL5 -1 IL8RB -1 MMP3 0 SOD2 1
Probe set IDs only	*ProbesetID*	224301_x_at 1398191_at Dr.2473.1.A1_at A_24_P93251
Probe set IDs, actual fold change	*ProbesetID*<tab>*ActualFC*	224301_x_at -19.19385797 1398191_at -18.21493625 Dr.2473.1.A1_at -17.6056338 A_24_P93251 -15.69858713
Probe set IDs, composite fold change	*ProbesetID*<tab>*CompositeFC*	224301_x_at -1 1398191_at 0 Dr.2473.1.A1_at 1 A_24_P93251 -1

Using tranSMART to Select Genes

You can use the tranSMART Search tool to help you select the list of genes for your gene signature. For example, suppose you are interested in lung adenocarcinoma, and want to create a gene signature consisting of genes that were strongly up-regulated in an experiment involving lung adenocarcinoma patients.

You can use tranSMART to select the genes for the gene signature as follows:

1. In tranSMART, click the **Search** tab to display the Search window.
2. Type **lung** in the Search field:



Lung adenocarcinoma is not listed in the dropdown list of search filters, but lung neoplasms is listed.

3. Click Disease> Lung Neoplasms.

   In a few seconds, the search result appears.

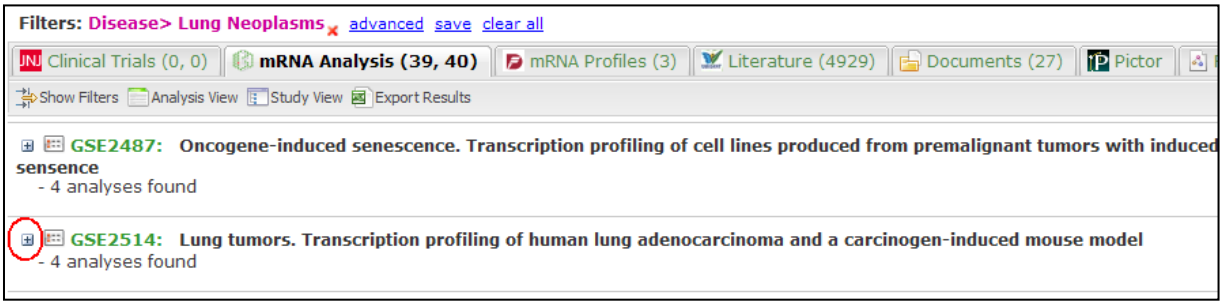
4. Click the **mRNA Analysis** tab, then click the **Study View** button:



tranSMART displays a list of all the experiments related to lung neoplasms.

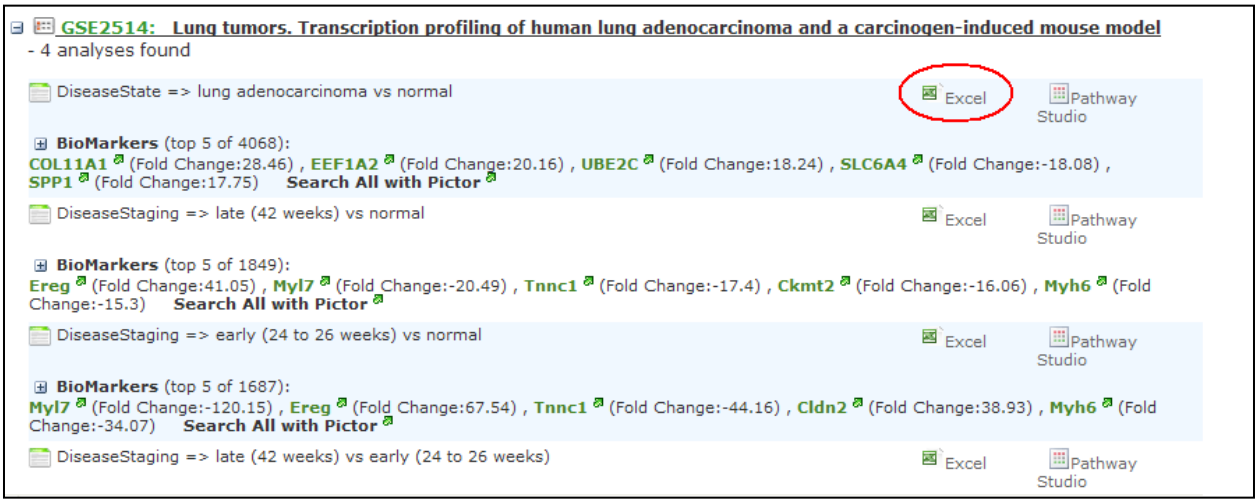
5. Scroll through the list of experiments until you find the one to use as the basis of your gene signature.

6. Click the **+** icon (⊞) to the left of the experiment name:



A list of the analyses based on this experiment appears.

7. Click the **Excel** button for the analysis that you want to use for your gene signature**:**



This action exports the analysis information, including the gene expression data, to a Microsoft Excel file:

8. Click **Open** in the File Download dialog.

Excel starts up and displays the analysis data – for example:

	A	B	C	D	E	F
1	Analysis	Probe Set	Fold Cha	p-Value	adjusted p-value	Gene
2	DiseaseState => lung adenocarci	374_f_at	1.21	0.0227	0.0689	DDT
3	DiseaseState => lung adenocarci	374_f_at	1.21	0.0227	0.0689	DDTL
4	DiseaseState => lung adenocarci	1056_s_at	-1.51	0.0039	0.0166	IL16
5	DiseaseState => lung adenocarci	40599_at	-1.48	0.0093	0.0337	IL16
6	DiseaseState => lung adenocarci	31540_at	2.24	0.0004	0.0026	TNFRSF9
7	DiseaseState => lung adenocarci	40476_s_at	1.42	0	0.00006	FOXK2
8	DiseaseState => lung adenocarci	36189_at	1.85	0	0	ILF2

9. Remove any rows containing genes that you do not want to include in the gene signature.

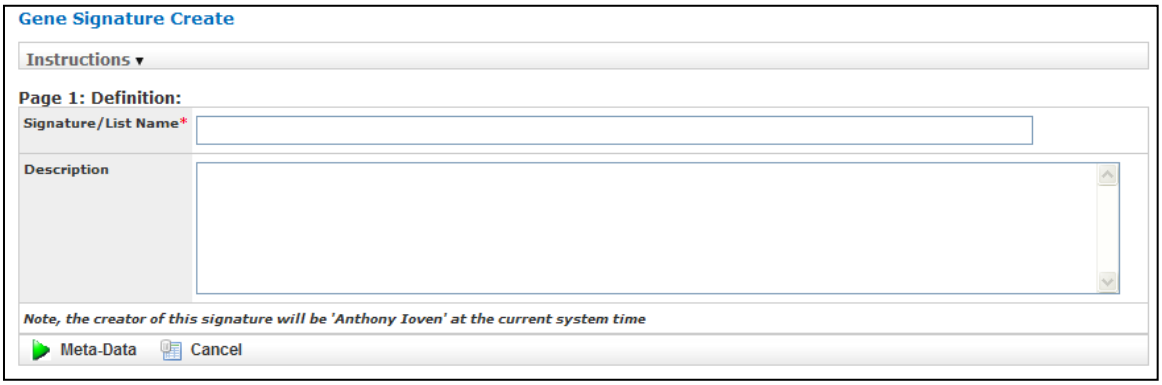
10. Remove the column headings and all data from the spreadsheet except for the gene symbol and fold change ratio, then export the remaining data to a tab-separated text file using the Excel **Save as type** option **Text (Tab delimited) (*.txt)**.

> **Note:** For an example of using Microsoft Excel 2007 to pare down the list of genes and to export the gene symbols and fold change ratios to a properly formatted text file, see the section *Create a Gene Signature* in the tranSMART *JnJ Training: Advanced* guide.

Step 2. Creating the Gene Signature

1. In tranSMART, click the **Gene Signature/Lists** tab.
2. Click the **New Signature** button.

   The first page of the gene signature wizard appears:

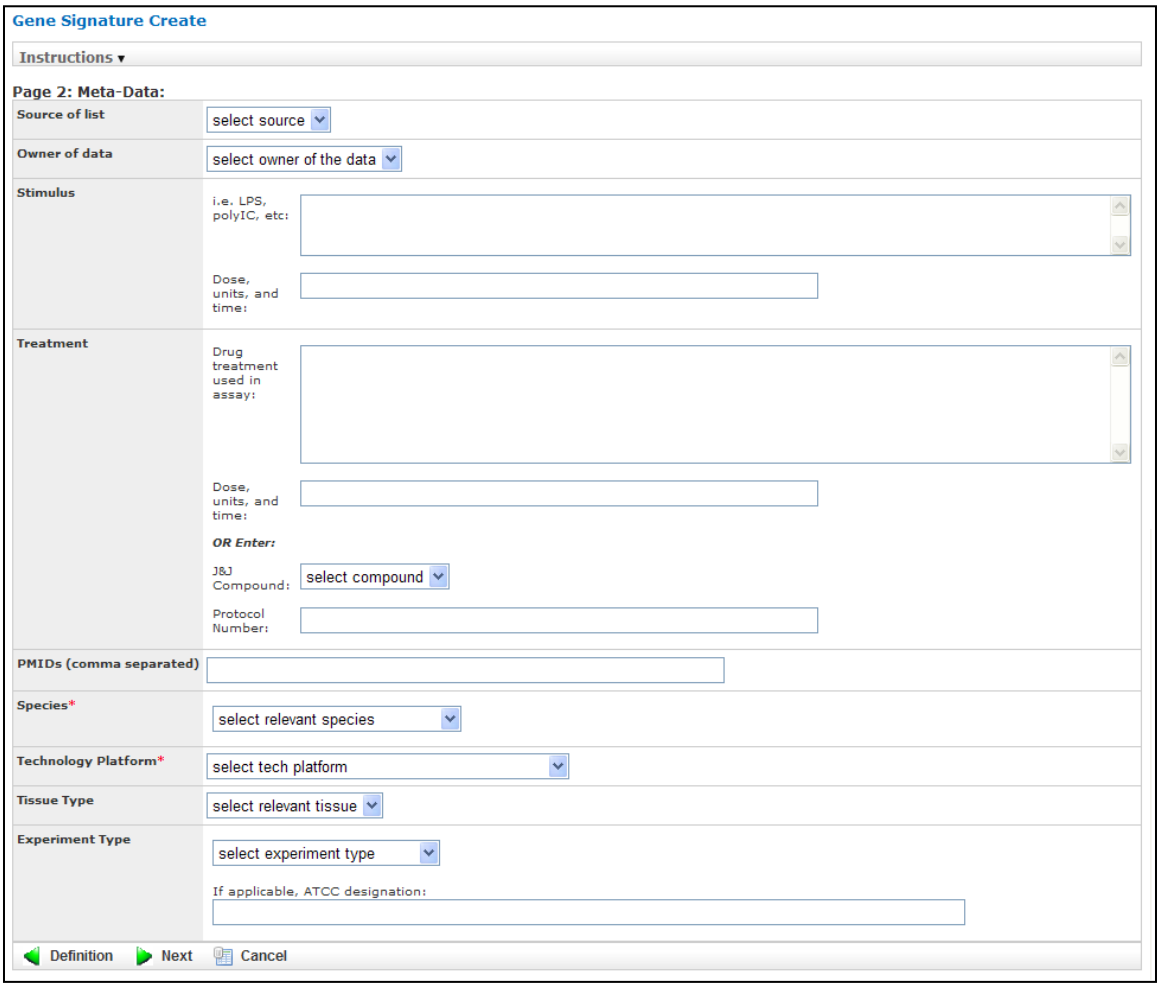


> **Note:** Required fields on gene signature wizard pages are marked with a red asterisk (*).

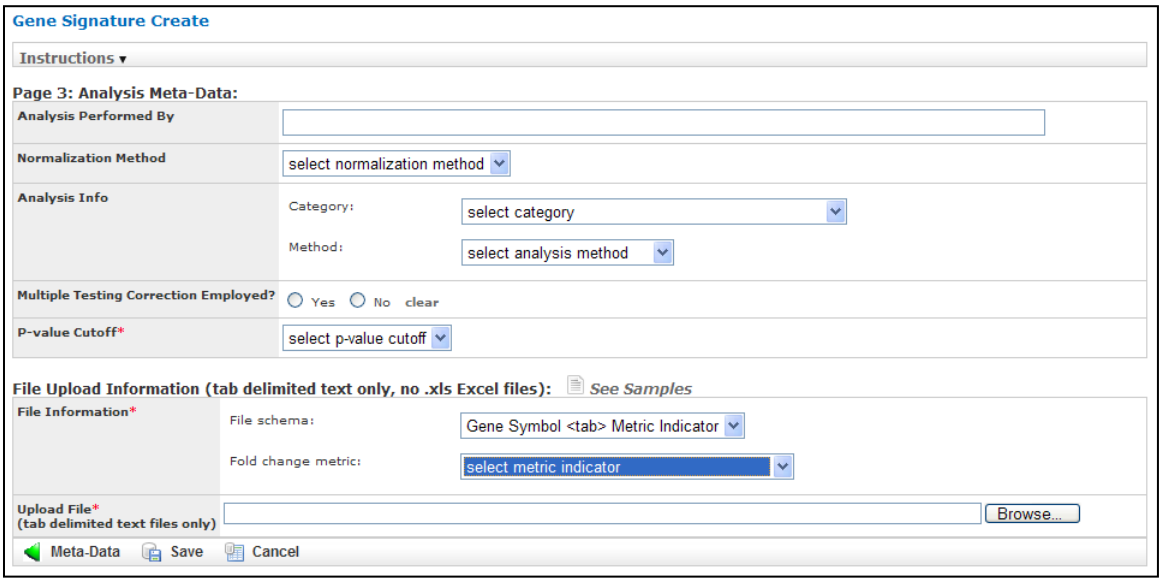
   You can find additional information about the gene signature wizard by clicking **Information** on any wizard page.

3. Specify a name (required) and an optional description for your gene signature, then click Meta-Data to proceed to the next gene wizard page.

The second page of the gene signature wizard appears:



4. Specify values in the required fields **Species** and **Technology Platform**, and also in any other relevant fields, then click Next to proceed to the final gene signature wizard page:



5. Specify values in the required field **P-value Cutoff**.

6. In the section **File Upload Information**, describe the text file you created in the section Step 1. Adding the Genes to a Text File on page 145, using the required fields **File Information** and **Upload File**:

   - In the **File schema** section of **File Information**, select **Gene Symbol <tab> Metric Indicator** or **Probe Set Symbol <tab> Metric Indicator**, depending on the method you chose to specify the genes.

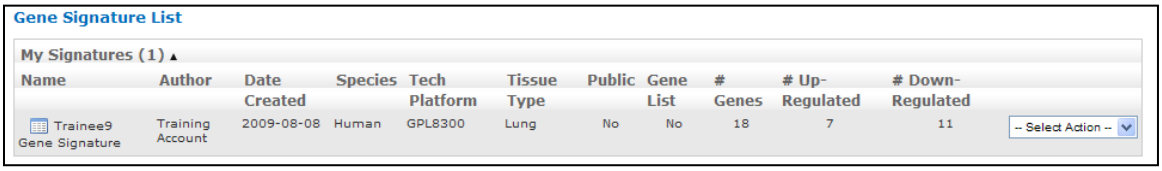
   - In the **Fold change metric** section of **File Information**, select one of the following choices from the dropdown:

Fold Change Metric Indicator	Description
Actual fold change	The text file contains actual fold change values for each gene symbol or probe set ID.
Not used	The text file contains gene symbols or probe set ID only. There are no associated fold change values.
-1 (down), 1 (up), 0 (optional for unchanged)	The fold change values are not actual values. They simply represent whether the gene expression was down-regulated (-1), up-regulated (1), or unchanged (0).

   - In **Upload File**, specify the path and name of the file that contains the genes to import. Use the **Browse** button to select the file from the navigation tree.

7. Specify values in any other relevant fields on this gene wizard page, then click **Save** to save the gene signature.

   The new gene signature appears in the **Gene Signature List** at the top of the Gene Signature/List view:



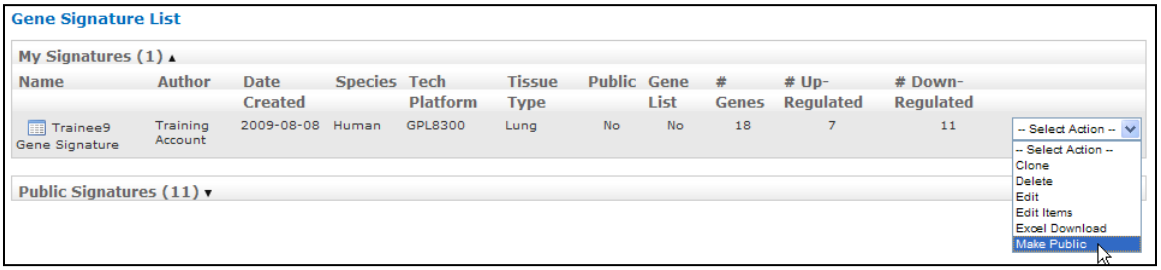
Making a New Gene Signature Public

By default, a newly created gene signature is private.

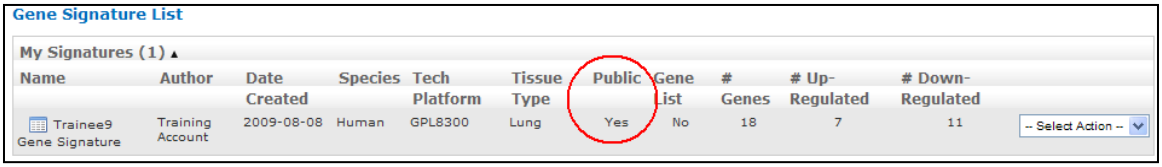
To make a gene signature public:

1. In the **Gene Signature List**, click the **Select Action** dropdown to the right of the gene signature you just created.

2. Click **Make Public** in the dropdown list:



   After you click **Make Public**, the value in the **Public** column for the gene signature changes from **No** to **Yes**:



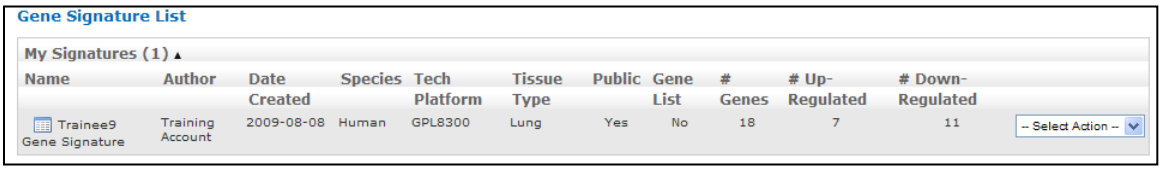
Note: tranSMART users assigned the role `ROLE_ADMIN` have access to both public and private gene signatures.

Performing Actions on Your Gene Signatures

To edit or perform other actions on a gene signature in your gene signature list:

1. In tranSMART, click the **Gene Signature/Lists** tab.

   The **Gene Signature List** appears, containing all the genes you have created:



2. Click the **Select Action** dropdown for the gene signature you are acting on. The dropdown contains all the actions you can perform on the gene signature:

Action	Description
Clone	Create an exact duplicate of the gene signature definition (*except* for the text file containing the gene symbols and fold change values), and display the definition in the gene signature wizard. Cloning a gene signature helps you create a new gene signature with a similar definition to an existing one. However, it is expected you will import a different set of genes into the gene signature.
Delete	Delete the gene signature.
Edit	Open the gene signature in the gene signature wizard for editing. The gene signature wizard displays all the information in the gene signature, including the reference to the text file containing the list of genes and fold change values. If you want to choose a different text file, click the following label:  To save any changes you make during editing, you must click the **Save** button on the third page of the wizard.
Edit Items	Add, delete, or modify one or more genes in the text file containing the gene symbols and fold change values.
Excel Download	Generate the entire contents of the gene signature, including the information in the text file containing the gene symbols and fold change values, to a Microsoft Excel spreadsheet. The gene signature definition and gene symbols/fold change values are written to separate spreadsheets.

Action	Description
Make Public	Make a private gene signature public. **Note:** To make a public gene signature private, edit the gene signature and set the **Public?** field to **No** on the first page of the gene signature wizard: 

Performing Actions on Other Users' Signatures

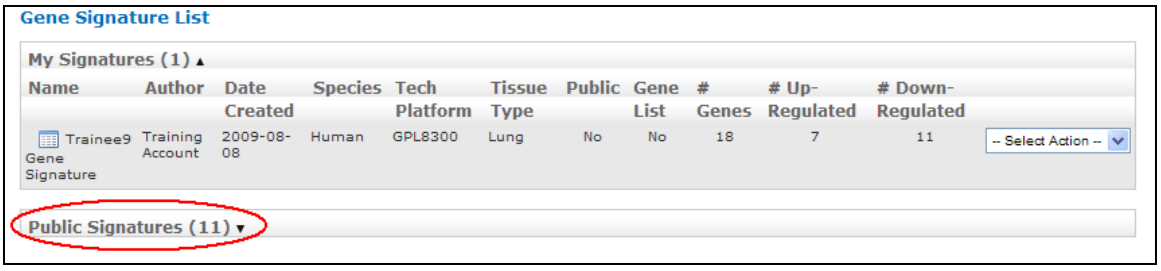
You can perform actions on gene signatures that other tranSMART users have created. The gene signatures you can access and the actions you can perform on them depend on the role assigned to your tranSMART user ID, as follows:

Role	Authorized Actions
`ROLE_ADMIN`	All actions on all gene signatures, both public and private.

`ROLE_SPECTATOR`
`ROLE_STUDY_OWNER`
`ROLE_DATASET_EXPLORER_ADMIN`
	Only **Clone** and **Excel Download**, and only on public gene signatures.

To edit or perform actions on a gene signature other than your own:

1. In tranSMART, click the **Gene Signature/Lists** tab.
2. Click **Public Signatures** to open the list of public gene signatures:



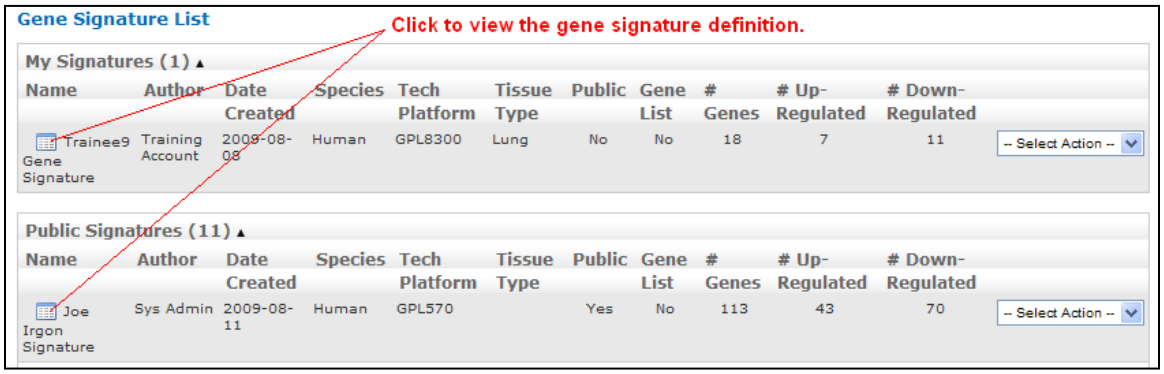
> **Note:** tranSMART users assigned the role `ROLE_ADMIN` will see **Other Signatures** instead of **Public Signatures**.

3. Click the **Select Action** dropdown for the gene signature you want to act on.
4. Select the action you want to perform on the gene signature.

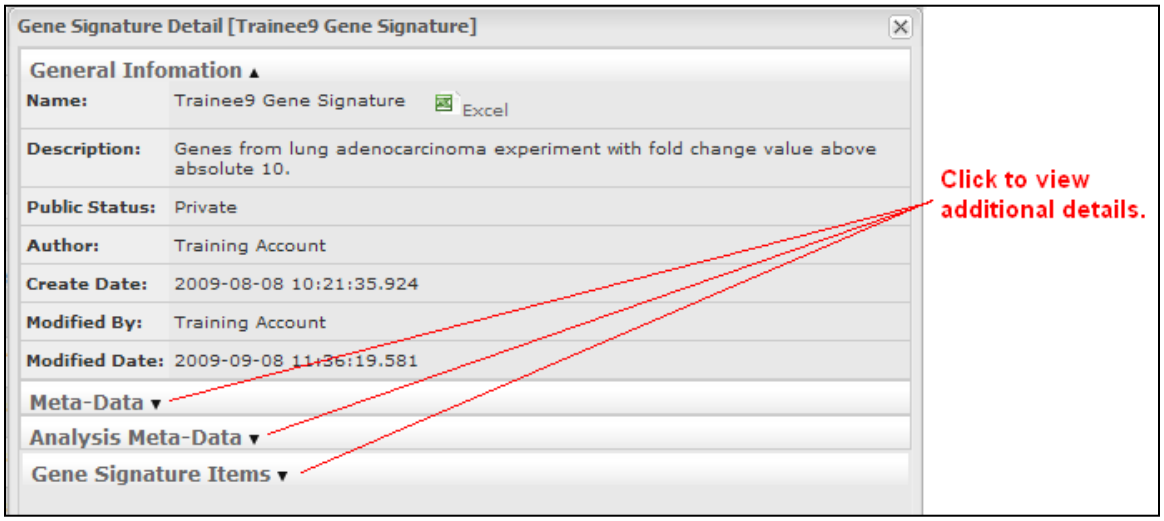
Viewing a Gene Signature Definition

You can view the definition of a gene signature, including its list of genes and fold change values, for any gene signature you are authorized to access.

To view a gene signature definition, click the **Detail** icon ( ) next to the gene signature name:



The Gene Signature Detail dialog appears, containing the gene signature definition:



Chapter 6
Other Tasks

In addition to Search, Dataset Explorer, and Gene Signature/List, the tranSMART toolbar includes the following tool tabs:

- **Request Consult**

  Email a request for information that cannot be found in the data warehouse.

- **Feedback**

  Email questions, problem reports, enhancement requests, or any other feedback about the tranSMART application.

- **Help**

  Display links to the tranSMART online and PDF documentation sets.

- **Admin**

  Perform user administration functions such as creating users, assigning roles to users, and enabling or disabling user accounts.

  This tool tab is visible only to users who are assigned the role `ROLE_ADMIN`.

Appendix A

How TEA Scores Are Calculated

This appendix summarizes the operations tranSMART performs to calculate the overall TEA score for an experiment, and the data inputs that the calculation requires. Pseudocode representations of the operations being performed are included where they may clarify the operation.

Data Inputs to the TEA algorithm

One of the following:

- A gene signature or gene list containing any number of genes, with a binary up- or down-regulation flag based on fold change.
- A pathway containing any number of genes.

And:

- A gene search result list for each signature, list, or pathway gene. Result lists contain experimental comparisons.

Operations

1. Compute the average fold change and standard deviation for all genes in the comparison.
2. Compute a normalized p-value (`NPV`) for each gene in the comparison, based on its fold change (`fc`) value, and the above average (`ave`) and standard deviation (`std`) values. Use a normal distribution function (`CDF`):

```
if (fc > 0)
    NPV = 1.0 - CDF(fc, ave, std)
 else
    NPV = 1.0 - CDF(-fc, ave, std)
if NPV < 1.0e-15, set to 1.0e-15
```

3. For each gene in the gene signature, list, or pathway, search against experimental comparisons and extract those comparisons where the gene's normalized p-value is less than 0.05. This returns a comparison list.

4. Iterate through the comparison list. For each comparison (`C`), add the normalized p-value to one of two arrays of sums (`pv_sum`), as follows:

   - For gene signatures and gene lists, add the gene's normalized p-value to:
     - `pv_sum(C, up)` if the gene's fold change in the signature (`svc`) *and* in the comparison (`cfc`) are co-regulated.
     - `pv_sum(C, down)` if the gene's fold change in the signature *and* in the comparison are anti-regulated.
   - For pathways, add the gene's normalized p-value to:
     - `pv_sum(C, up)` if the gene's comparison fold change (`cfc`) is up-regulated.
     - `pv_sum(C, down)` if the gene's comparison fold change is down-regulated.

   Also, use the logarithm of the normalized p-value to make the final TEA score more human readable:

```
if (gene_signature OR gene_list)
   if ( (sfc > 0 AND cfc > 0) OR (sfc < 0 AND cfc < 0) )
       pv_sum(C, up) += -Math.log(NPV)
       pv_count(C, up)++
     else
       pv_sum(C, down) += -Math.log(NPV)
       pv_count(C, down)++

if (gene_pathway)
   if (cfc > 0)
       pv_sum(C, up) += -Math.log(NPV)
       pv_count(C, up)++
     else
       pv_sum(C, down) += -Math.log(NPV)
       pv_count(C, down)++
```

5. Compute the min-LogP average (`pv_ave`) for each sum:

```
pv_ave(C, up) = Math.exp(-pv_sum(C,up) / pv_count(C,up) )
pv_ave(C, down) = Math.exp(-pv_sum(C,down) / pv_count(C,down) )
```

6. Compute a TEA score (`pv_tea`) for each min-LogP average through a binomial distribution function:

```
pv_tea(C, up) = 1.0 - Binom( N, pv_count(C,up), pv_ave(C,up) )
pv_tea(C, down) = 1.0 - Binom( N, pv_count(C,down), pv_ave(C,down) )
```

Result

- **TEA score**: For gene signatures, lists, and pathways, the final TEA score is the more significant `pv_tea` value (the lower of the two `pv_tea` values).
- A gene signature or list is determined to be co-regulated or anti-regulated as follows:
  - **Co-regulated**: The more significant `pv_tea` value was derived from the sums associated with co-regulated fold change values (`pv_sum(C, up)`).
  - **Anti-regulated**: The more significant `pv_tea` value was derived from the sums associated with anti-regulated fold change values (`pv_sum(C, down)`).
- A pathway is determined to be up-regulated or down-regulated as follows:
  - **Up-regulated**: The more significant `pv_tea` value was derived from the sums associated with up-regulated fold change values (`pv_sum(C, up)`).
  - **Down-regulated**: The more significant `pv_tea` value was derived from the sums associated with down-regulated fold change values (`pv_sum(C, down)`).

Result

Appendix B

Rules for Loading OmicSoft Data

The following rules determine whether OmicSoft data is loaded into tranSMART databases:

Rules for loading data into table BIO_ASSAY_ANALYSIS_DATA

Load the data if one of the following two criteria is satisfied:

- fold_change_ratio is >= 1.0 OR <=-1.0

  AND

  preferred_pvalue is null OR <= 0.1

- fold_change_ratio is null

  AND

  At least one of the following values is not null:

  - r_value

    Pearson product-moment correlation coefficient.
  - rho_value

    Spearman rank correlation coefficient
  - cut_value
  - results_value

Rules for loading data into table BIO_ASSAY_ANALYSIS_DATA_TEA

Load the data if:

```
TEA_NORMALIZED_PVALUE <= 0.05
```

For information about the TEA algorithm, see Appendix A: How TEA Scores Are Calculated.