Pfizer

tranSMART

User's Guide

December 21, 2012
Edition 1

Recombinant
By Deloitte.

Any blank pages in this document are intentionally inserted to allow correct double-sided printing.

Contents

Chapter 1: Getting Started with tranSMART 1

Feature Overview 1

Faceted Search 2

Dataset Explorer 2

Sample Explorer.. 3

Gene Signature Wizard 3

Upload Data 3

Logging In... 4

Tools 4

Opening a Particular Tool at Login 5

Chapter 2: Faceted Search 7

Overview of the Faceted Search UI 7

Defining Search Filters 8

Using the Filter Browser 11

Managing Active Filters 15

Viewing Search Results 16

Analysis View.. 17

Table View 22

Manhattan Plot.. 23

Chapter 3: Dataset Explorer 25

Overview of the Dataset Explorer UI 25

Using Dataset Explorer 27

Public and Private Studies 27

Selecting the Study 28

Populating the Study Groups 31

Generating Summary Statistics 36

Defining Points of Comparison 39

Printing or Saving the Contents of the Results/Analysis View 40

Copying Individual Charts in the Results/Analysis View 40

Viewing a Study 40

Exporting Dataset Explorer Findings 41

Generating Advanced Analyses and Visualizations 43
Generating Heatmaps 44
Generating Advanced Analyses 54
Data Binning.. 66
High Dimensional Data.. 71
Other Features.. 73
The Jobs Tab 75
Viewing a Logged Job 76

Chapter 4: Sample Explorer..77
Select a Primary Search Filter 77
View and Refine Sample Search Results 80
Select and Remove Search Filters.. 81
Locate the Source of the Samples in Dataset Explorer............................ 82
Manage the Sample Search Result List.. 83

Chapter 5: Gene Signatures and Gene Lists...85
Creating a Gene Signature 85
Step 1. Adding the Genes to a Text File 85
Step 2. Creating the Gene Signature 87
Performing Actions on Your Gene Signatures 91
Performing Actions on Other Users' Signatures 92
Viewing a Gene Signature Definition 93

Chapter 6: Analysis Data Upload 95
Uploading Analysis Data 95
File Templates 97
Handling an Upload Failure 98

Appendix A: Glossary 99
Glossary... 99

Chapter 1

Getting Started with tranSMART

The tranSMART application reflects the efforts of various informatics groups 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 biomarker data from gene expression, RBM, and SNP experiments involving GWAS, Metabolic GWAS, and eQTL data types, and including both raw and analyzed data.

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

Another major tranSMART feature, called the Dataset Explorer, allows authorized users to create and study cohorts of patients that have been involved in completed clinical research efforts. Dataset Explorer also provides users with tools to compare an individual (or group) in one study against a person or cohort in another study.

ℹ 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 features:

- Faceted Search
- Dataset Explorer
- Sample Explorer
- Gene Signature Wizard
- Upload Data

Faceted Search

tranSMART provides a Google-like search tool that lets you search across multiple data sources for information related to items of interest, such as diseases, genes, SNPs, and genomic regions.

The scope of a search can include clinical studies, externally conducted experiments, and in vivo/in vitro studies.

Search tool functionality includes:

- Searching within a particular category, such as diseases, genes, or gene signatures, or SNPs, or searching across all categories.
- Building complex search criteria that let you precisely define what to search for.
- Refining search results by p-value and/or search keyword.
- Exporting search results to a comma-separated text file, Manhattan Plot, or QQ Plot.

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 heatmap 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.
- Performing advanced analyses and displaying results in various formats (scatter plot with linear regression, box plot with analysis of variance, etc.)
- Exporting a study or subset of a study to analyze in an external tool.

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).

Once you find samples of interest, you can link back to the Dataset Explorer study for which the samples were collected.

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 deregulated.

Stored gene signatures can also be used in the analyses functionality of Dataset Explorer.

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.

Upload Data

The Upload Data feature lets you upload analysis data for a particular study by filling out a browser-based form and referencing the location of the file that contains the data. Templates are provided for each of the supported data types (GWAS, Metabolic GWAS, eQTL) to ensure that your data is in the proper format for uploading.

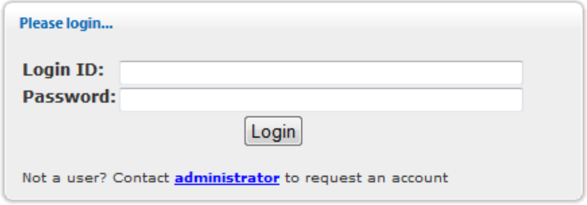
Logging In

To log into tranSMART:

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

   http://amre1al306.pcld.pfizer.com/transmart/search

   The login screen appears:



Please login...

Login ID:

Password:

Not a user? Contact administrator to request an account

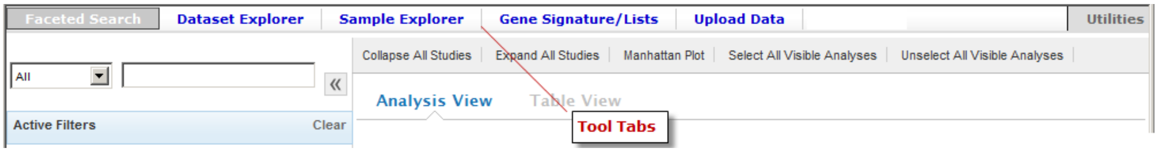
2. Type your tranSMART login credentials, then click **Login**.

Tools

tranSMART provides the following tools:

- **Faceted Search** – Search across studies and analyses for research data related to search filters that you specify.
- **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.
- **Upload Data** – Upload analysis data for a study.
- **Utilities** – Contains the following submenus:
  - **Help** – Opens the tranSMART online Help.
  - **Contact Us –** Email questions, problem reports, enhancement requests, or any other feedback about the tranSMART application.
  - **About** – Displays the version of tranSMART.

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 Faceted 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:

The addresses below are case-sensitive.

- Faceted Search tool – either of the following:

  http://amre1al306.pcld.pfizer.com/transmart/

  http://amre1al306.pcld.pfizer.com/transmart/search

- Dataset Explorer tool

  http://amre1al306.pcld.pfizer.com/transmart/datasetExplorer

- Sample Explorer tool

  http://amre1al306.pcld.pfizer.com/transmart/sampleExplorer

- Gene Signature/Lists tool

  http://amre1al306.pcld.pfizer.com/transmart/geneSignature

- Upload Data tool

  http://amre1al306.pcld.pfizer.com/transmart/uploadData

Chapter 2

Faceted Search

The tranSMART Faceted Search feature provides a fine-grained search capability from a single user interface into Genome Wide Associations studies of interest.

You define a search query by specifying search keywords, by selecting search filters from one or more filter browsers, or by any combination of these methods. tranSMART conducts the search across multiple studies and analyses, and allows you to view the results in a single analysis of a study or in an aggregated result that spans multiple analyses and studies. You can also export results to a file, to a Manhattan Plot, and to a QQ Plot.

Faceted Search supports filters based on one or more of the following kinds of information:

- Keywords that you specify, such as part of a study or analysis name
- Reference SNP (RS) identifiers
- Individual genes and all genes in a gene signature
- Chromosomes and a specific position within a chromosome
- A range of base pairs around a specified gene or chromosomal position
- p-value thresholds
- Diseases and observations
- Data types on which analysis data is based: GWAS (gene expression), Metabolic GWAS, eQTL

Overview of the Faceted Search UI

The figure below shows the Faceted Search interface. It is divided into two panes:

Left pane

Use this pane to define search filters to retrieve the studies, analyses, and analysis data of interest.

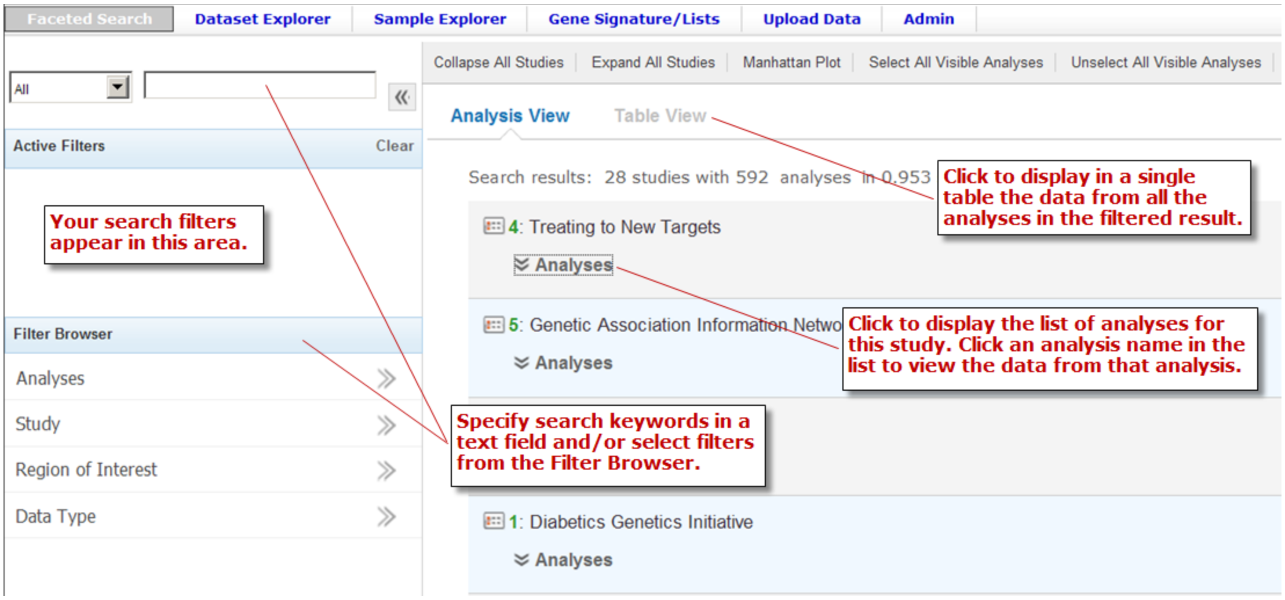
Right pane

Use this pane to:

- View the results of your search in Analysis View, where you view data from individual analyses, or Table View, where you view data from multiple analyses.

- Narrow analysis data results by a p-value and/or keyword that you specify.
- Export analysis data to a text file.
- View analysis data in a QQ Plot or Manhattan Plot, and optionally export the plot to a file.
- View the metadata associated with each study and analysis.

The default Faceted Search page is shown below:



Your search filters appear in this area.

Click to display in a single table the data from all the analyses in the filtered result.

Click to display the list of analyses for this study. Click an analysis name in the list to view the data from that analysis.

Specify search keywords in a text field and/or select filters from the Filter Browser.

Defining Search Filters

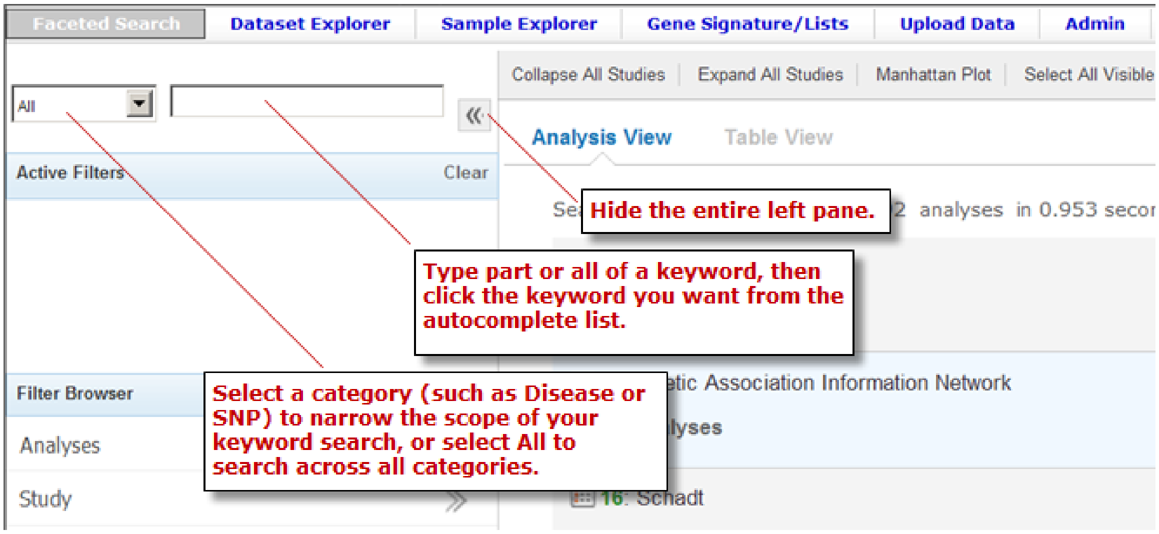
You define search filters to retrieve just the studies, analyses, and analysis data that interest you.

This section describes how to use search keywords to define search filters. You can also select search filters from the Filter Browser (page 11).

The search filters you define are displayed in the Active Filters area (page 15).

Keyword Search

The following figure shows the controls for defining a keyword search:



Hide the entire left pane.

Type part or all of a keyword, then click the keyword you want from the autocomplete list.

Select a category (such as Disease or SNP) to narrow the scope of your keyword search, or select All to search across all categories.

To define a keyword search:

1. Select one of the categories in the category dropdown control, or select **All** to search across all categories.

   Categories are:

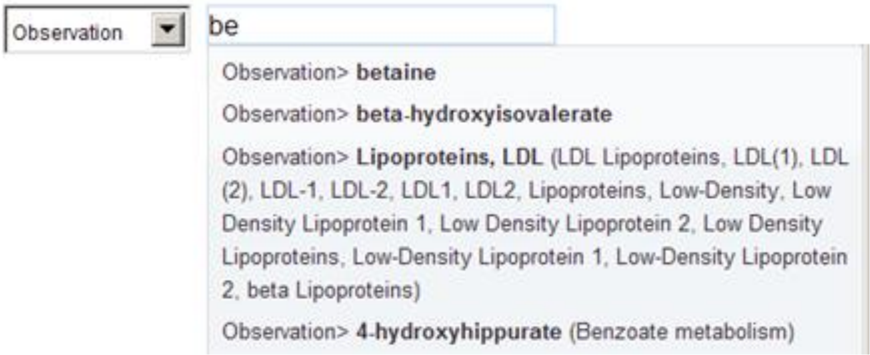
Category	Search Scope
All	All categories.
Observation	Studies and analyses associated with the medical observation you specify.
Disease	Studies and analyses associated with the disease you specify.
Data_type	Studies and analyses associated with GWAS, eQTL, or Metabolic GWAS data.
Snp	Analysis data that includes the reference SNP (RS) number you specify.
Gene	Analysis data that includes the gene you specify.
Genesig	Analysis data that includes any gene in the gene signature you specify. Gene signatures are defined through the tranSMART Gene Signature/Lists tab (see Chapter 5: Gene Signatures and Gene Lists).

Gene searches return all matches of the gene, not just results that are statistically significant.

2. Specify part or all of a search keyword in the text field to the right of the category dropdown.

   When you type at least two characters in the field, tranSMART begins to search within the specified category and lists keywords that begin with those characters. Not case sensitive.

   For example, the following figure shows the keywords displayed when the characters **be** are typed and the category **Observation** is selected.



Observation> **betaine**

Observation> **beta-hydroxyisovalerate**

Observation> **Lipoproteins, LDL** (LDL Lipoproteins, LDL(1), LDL (2), LDL-1, LDL-2, LDL1, LDL2, Lipoproteins, Low-Density, Low Density Lipoprotein 1, Low Density Lipoprotein 2, Low Density Lipoproteins, Low-Density Lipoprotein 1, Low-Density Lipoprotein 2, beta Lipoproteins)

Observation> **4-hydroxyhippurate** (Benzoate metabolism)

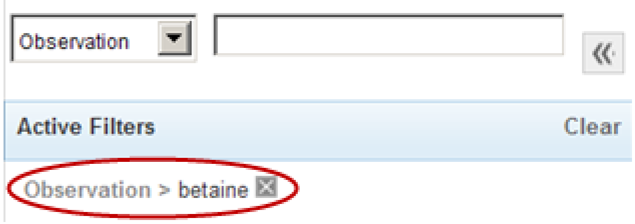
   Up to 15 keywords can be displayed at one time. If you don’t see the one you want, type more characters into the field.

   Note that the search looks for matches based on the characters at the beginning of the keywords (in bold) or the keyword synonyms (in parentheses).

3. Click the keyword you want, but do not press Enter or Return.

   When you click the keyword, the following actions occur:

   - The search begins immediately, and a result is displayed in the right pane.
   - The associated search filter appears in the Active Filters area:



   You can add more filters by repeating the steps above, by selecting filters from the Filter Browser, or by any combination of these actions.

Search filters for SNPs, genes, and gene signatures do not filter out studies and analyses that omit the specified SNP or gene. However, the only records returned for an analysis are those that contain the specified SNP or gene. If an analysis contains no references to the SNP or gene, no records are returned for that analysis.

Using the Filter Browser

The Filter Browser lets you select one or more search filters to include in your search query.

tranSMART adds all of your search filters, including those you define through the keyword search field (see Defining Search Filters on page 8), to the Active Filters area (see Managing Active Filters on page 15).

The following tables shows the types of filters available through the Filter Browser:

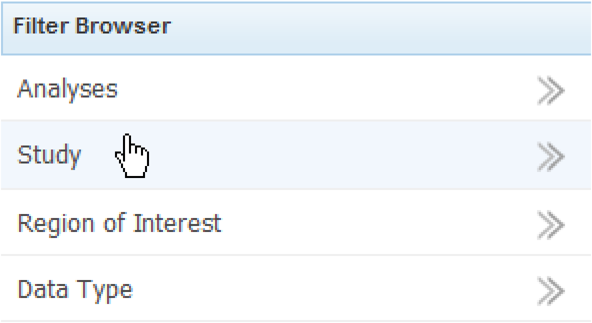
Filter Type	**Description**
Analyses	Lets you select one or more analyses from a list. Only those studies that contain the selected analyses will be listed in the right pane. Each listed study will contain only the selected analyses.
Study	Lets you select one or more studies from a list. All selected studies will be listed in the right pane. Each listed study will include all associated analyses.
Region of Interest	Lets you specify a facet of analysis data to use as a search filter; for example, a gene, a SNP RS identifier, or a chromosome. You can further refine the filter by specifying a range of base pairs above or below (or both) the specified facet of data. Additionally, you can specify a particular Human Genome version to use as the basis of the data. For information on specifying this type of filter, see Region of Interest on page 13.
Data Type	Lets you select studies and analyses associated with one or more data types: GWAS, eQTL, Metabolic GWAS.

ℹ A search is constrained by *all* the filters specified in the Active Filters area. The results indicated in the above table assume that there are no other filters specified in Active Filters.

To select one or more filters from the Filter Browser:

1. Under Filter Browser, click one of the filter types.

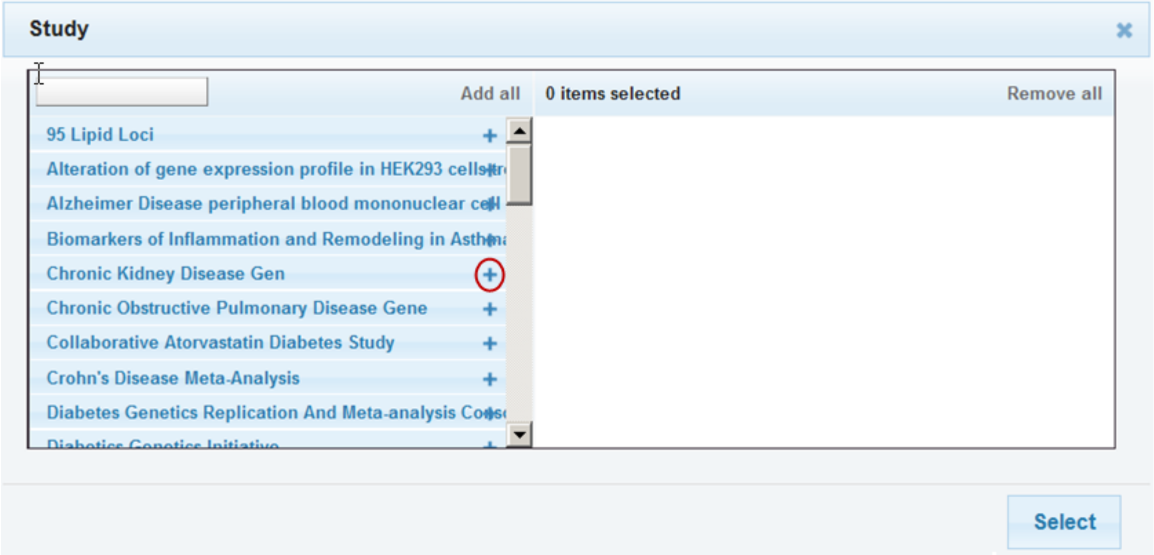
   In the figure below, the Study browser is selected:



   The selected browser opens.

   Browsers for Analyses, Study, and Data Type work the same way. For information on specifying Region of Interest filters, see Region of Interest on page 13.

2. Filters are listed in the left part of the browser. Select a filter by clicking the plus sign (**+**) to the right of the filter name:


Study
Add all
0 items selected
Remove all
95 Lipid Loci
Alteration of gene expression profile in HEK293 cells tr
Alzheimer Disease peripheral blood mononuclear cell
Biomarkers of Inflammation and Remodeling in Asthma
Chronic Kidney Disease Gen
Chronic Obstructive Pulmonary Disease Gene
Collaborative Atorvastatin Diabetes Study
Crohn's Disease Meta-Analysis
Diabetes Genetics Replication And Meta-analysis Cons
Diabetics Genetics Initiative
Select

   The selected filter is added to the right part of the browser.

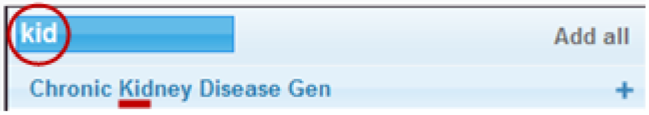
   With filters that have long names, note that:

   - The plus sign (**+**) is partially obscured by the name, but you can still click it.
   - Hovering the mouse pointer over the name displays the full name.

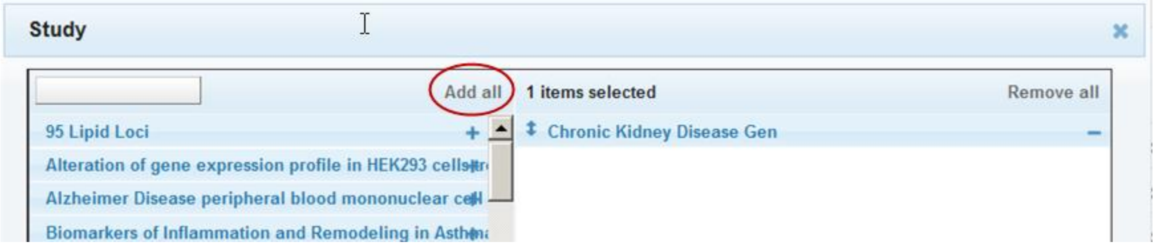
3. Optionally:

□ Add more filters within this browser by repeating the previous step.

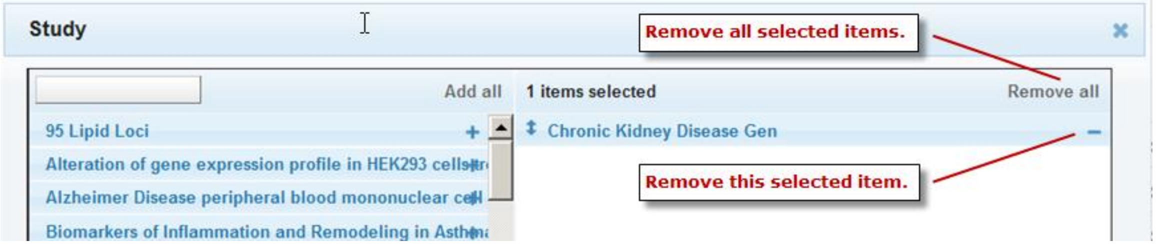
  □ Narrow the list of filters by typing characters into the text box at the top left of the browser. tranSMART lists all those filters whose names include the characters, in a contiguous string, anywhere in the name (not case-sensitive):



  □ Select all the filters in the list by clicking **Add all**:


Study
Add all
1 items selected
Remove all
95 Lipid Loci
Alteration of gene expression profile in HEK293 cells tr
Alzheimer Disease peripheral blood mononuclear cell
Biomarkers of Inflammation and Remodeling in Asthma
Chronic Kidney Disease Gen

  □ Remove a selected filter by clicking the minus sign (**–**) to the right of the selected filter name, or remove all selected filters by clicking **Remove all**:


Study
Remove all selected items.
Add all
1 items selected
Remove all
95 Lipid Loci
Alteration of gene expression profile in HEK293 cells tr
Alzheimer Disease peripheral blood mononuclear cell
Biomarkers of Inflammation and Remodeling in Asthma
Chronic Kidney Disease Gen
Remove this selected item.

4. When finished selecting filters from this browser, click **Select** at the bottom right of the browser. Your selections will be added to the Active Filters area.

5. Optionally, select filters from a different browser by repeating the above steps.

   All of the selected filters will become part of the same search query and be included in the Active Filters area.

Region of Interest

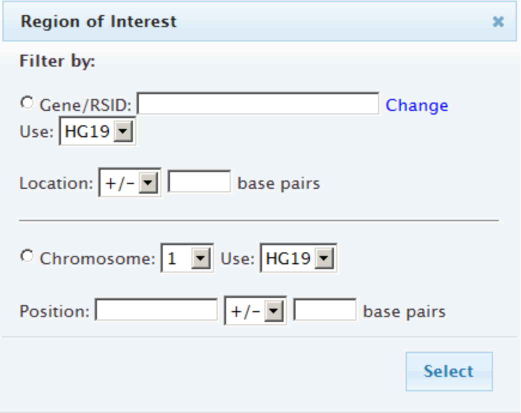
The Region of Interest browser lets you specify a particular area of the human genome as a search filter.

Search filters for regions of interest do not filter out studies and analyses that omit the region of interest. However, the only records returned for an analysis are those that contain the specified region of interest. If an analysis does not reference the region of interest, no records are returned for that analysis.

To specify a region of interest:

1. Under Filter Browser, click **Region of Interest**.

   The Region of Interest browser appears:



Region of Interest

Filter by:

○ Gene/RSID: Change
Use: HG19

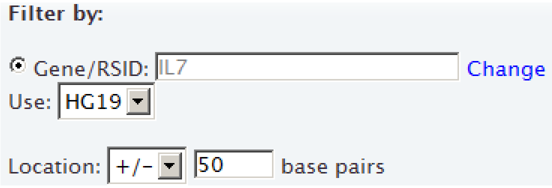
Location: +/- base pairs

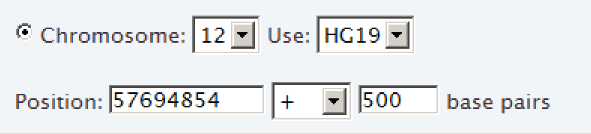
○ Chromosome: 1 Use: HG19

Position: +/- base pairs

Select

2. Define the filter as described in the following table:

Filter by	Description
Gene	1. Select the **Gene/RSID** radio button. 2. Type two or more characters in the field after the **Gene/RSID** label (not case sensitive). When you type at least two characters in the field, tranSMART begins to search for gene names or synonyms that begin with the characters you typed. Up to 15 keywords are displayed. If you don’t see the one you want, type more characters into the field. 3. Click the gene of interest. To select a different gene, click **Change**, then repeat the above step. 4. Optionally, in the **Use** field, select the Human Genome version to use as the basis of this search. The default is the current version. 5. Optionally, in **Location**, specify the number of base pairs above, below, or both above and below the specified gene to include in the region of interest. If you do not specify a location, the region of interest will be the specified gene only. For example, the following selects a region that spans 50 base pairs above and below the gene IL7, based on Human Genome version 19:  Filter by: ● Gene/RSID: IL7 Change Use: HG19 Location: +/- 50 base pairs 6. When finished defining the region of interest, click **Select**. The filter is added to the search query in the Active Filters area.

Filter by	Description
RS Identifier	Define the region of interest based on an RS identifier the same way you would define one for a gene. In step 2, type **rs** followed by at least one numeric character.
Chromosome	1. Select the **Chromosome** radio button. 2. Select the number of the chromosome of interest from the dropdown list. 3. Optionally, in the **Use** field, select the Human Genome version to use as the basis of this search. The default is the current version. 4. Optionally, in the **Position** text box, type the *exact* position number of the base pair of interest. If you do not specify a position, the region of interest will be the entire chromosome. 5. Optionally, in the two fields after the **Position** text box, specify the number of base pairs above, below, or both above and below the specified chromosomal position to include in the region of interest. If you specify a position but not a range of base pairs, the region of interest will be the specified position within the chromosome. For example, the following selects a region of interest that spans the base pair at position 57694854 and the 500 base pairs above it within chromosome 12, based on Human Genome version 19:  ● Chromosome: 12 Use: HG19 Position: 57694854 + 500 base pairs 6. When finished defining the region of interest, click **Select**. The filter is added to the search query in the Active Filters area.

3. Optionally, repeat the above steps to add an additional region of interest to the search query.

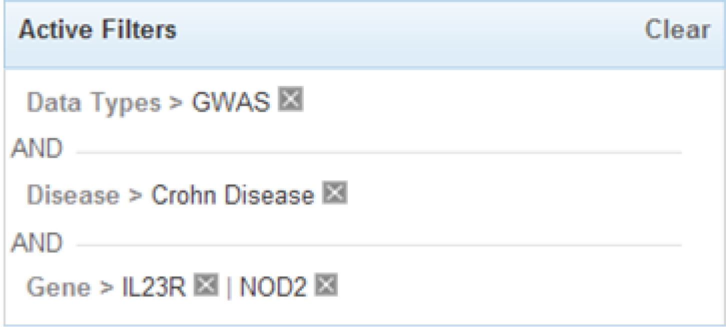
Region of interest searches return all matches within the region of interest, not just results that are statistically significant.

Managing Active Filters

The Active Filters area displays the entire search query that you build using the keyword search feature (see Defining Search Filters on page 8) and filter browser feature (see Using the Filter Browser on page 11).

Each filter that you define is added to the search query. Each time you add a filter to the search query, the result set in the right side of the Faceted Search page is modified to satisfy the entire search query.

The following search query in Active Filters will return all studies and analyses involving both Crohn's disease and GWAS data. Additionally, the only records in the analysis data will be those containing the gene IL23R or NOD2:



Active Filters Clear

Data Types > GWAS ☒

AND

Disease > Crohn Disease ☒

AND

Gene > IL23R ☒ | NOD2 ☒

Note the following about Active Filters:

- Filters of different types (for example, Disease filters and Data Types filters) are joined in the search query by a logical `AND` operator.

  To be included in a result set, items must satisfy all filters joined by `AND`.

- Filters of the same type (for example, the two genes in the figure above) are joined by a logical `OR` operator (represented by a vertical bar character, |).

  Any item joined by an `OR` operator can be included in a result set.

- To remove an individual filter from the search query, click the ☒ icon after the item name.

- To delete the entire search query, click **Clear** in the upper right corner of the Active Filters area.

Viewing Search Results

Search results appear in the right pane of the Faceted Search page.

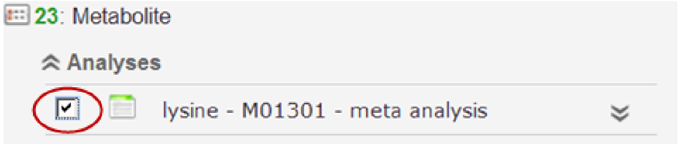
You can view search results in the following forms:

- Analysis View (page 17), including QQ Plots of analysis data.
- Table View (page 22)
- Manhattan Plot (page 23)

You can also export Analysis View and Table View data and visualizations to files.

Tabs on the Faceted Search Page

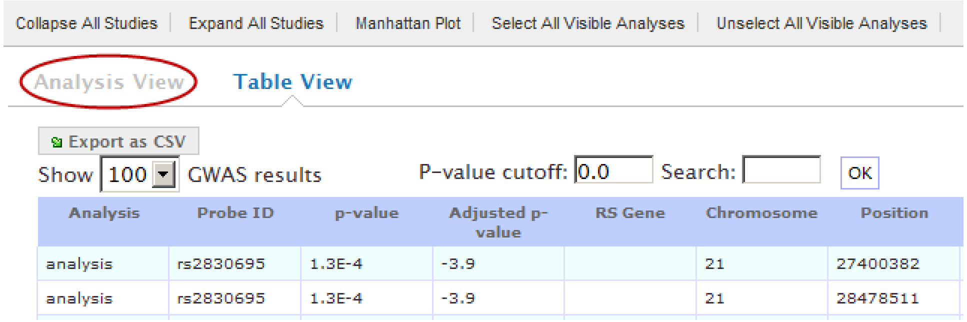
The following tabs are always visible on the Faceted Search page:

Tab	Description
Collapse All Studies	Hides the analysis names that appear under the names of listed studies.
Expand All Studies	Lists the names of each study's analyses under the study's name. Only the names of analyses that satisfy the search query in Active Filters are listed.
Manhattan Plot	Launches the Manhattan Plot application, which will display data from all selected analyses. A selected analysis is one whose check box next to its name is checked:  23: Metabolite Analyses ☑ lysine - M01301 - meta analysis
Select All Visible Analyses	Selects the check boxes for all analyses in all listed studies.
Unselect All Visible Analyses	Unchecks the check boxes for all visible studies. Data from unselected analyses will not appear in Manhattan Plots.

You only need to select analyses when generating a Manhattan Plot. Selecting analyses has no effect on any other functions.

Analysis View

Analysis View is the default view on the Faceted Search page. When Analysis View is not displayed and you want to display it, click the **Analysis View** button:



Show 100 GWAS results P-value cutoff: 0.0 Search:

Analysis	Probe ID	p-value	Adjusted p-value	RS Gene	Chromosome	Position
analysis	rs2830695	1.3E-4	-3.9		21	27400382
analysis	rs2830695	1.3E-4	-3.9		21	28478511

Tasks

You can perform the following tasks in Analysis View.

- Browse the list of studies, view information about a study, and expand the list of the analyses of a study.

  See Browse the Study List on page 18.

- View metadata for a particular analysis.

  See View Metadata for an Analysis on page 19.

- View the data in a particular analysis, filter the data, export the data to a comma-separated text file, and display the data in a QQ Plot.

  See View, Filter, and Export Analysis Data on page 19.

> You can display analysis data for multiple analyses in a Manhattan plot. For information, see Manhattan Plot on page 23.

For information about the studies, analyses, and analysis data that are displayed in Analysis View, as determined by the search query, see Notes about the Contents of Analysis View on page 21.

Browse the Study List

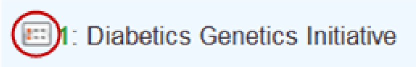
Before a search query is defined in Active Filters, the Faceted Search page is displayed in Analysis View with all studies listed. You can view the entire list of studies using the scroll bar on the page.

As you add search filters to the Active Filters area, the studies that appear in the list, and the list of analyses that can be opened for a study, are determined by the search filters you define.

You can perform the following tasks for a study:

- View metadata for the study.

  To do so, click the icon to the left of the study name:

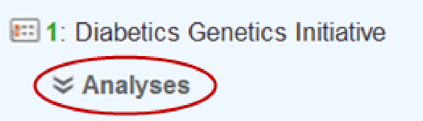
  

  1: Diabetics Genetics Initiative

  Information about the study is displayed, such as the description of the study, the institution that conducted the study, and data availability.

- Pull down a list of the study’s analyses that satisfy the current search query in Active Filters. This is called “expanding” the study.

To do so, click the **Analyses** button under the study name:

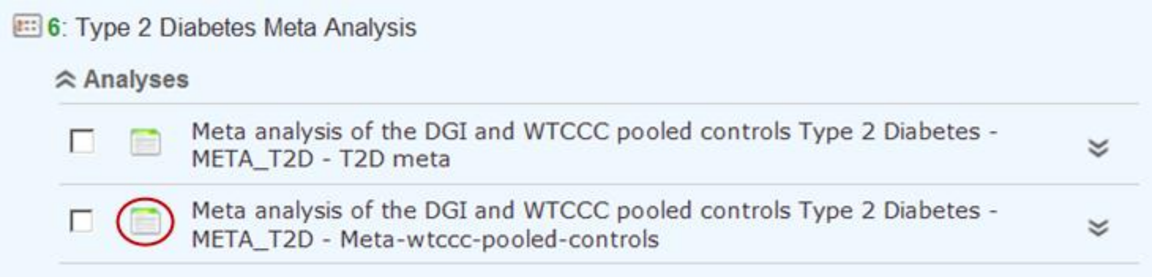


Optionally, pull down the analyses for all listed studies by clicking the **Expand All Studies** tab at the top of the Faceted Browser page.

View Metadata for an Analysis

You can view a variety of information about an analysis, such as description of the analysis, type of data collected, data sample size, tissue type, cell type, and analysis platform.

To do so, click the icon to the left of the analysis name:



□ Meta analysis of the DGI and WTCCC pooled controls Type 2 Diabetes - META_T2D - T2D meta

□ Meta analysis of the DGI and WTCCC pooled controls Type 2 Diabetes - META_T2D - Meta-wtccc-pooled-controls

View, Filter, and Export Analysis Data

This section describes how to:

- Display the data for a particular analysis of a study
- Filter the data according to p-value and/or search keyword.
- Export the data to a comma-separated text file.
- Export the data to a QQ Plot.

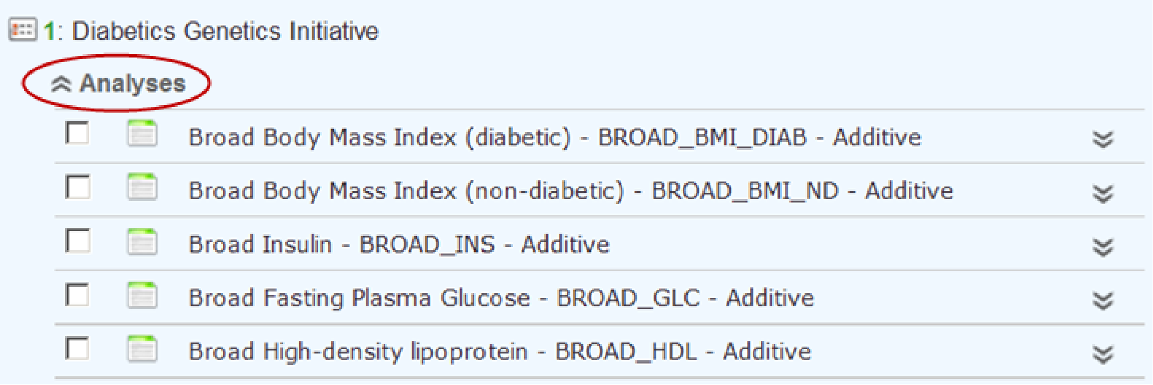
Typically, before you view analysis data, you will define a search query to narrow the lists of studies and analyses that appear in Analysis View.

To upload analysis data for a study, see Chapter 6: Analysis Data Upload.

To view analysis data, and optionally filter and export the data:

1. In Analysis View, navigate to the study that contains the analysis.
2. Click the **Analyses** button under the study name to expand the list of analyses for the study.

A list appears containing the study's analyses that satisfy the search query in Active Filters:



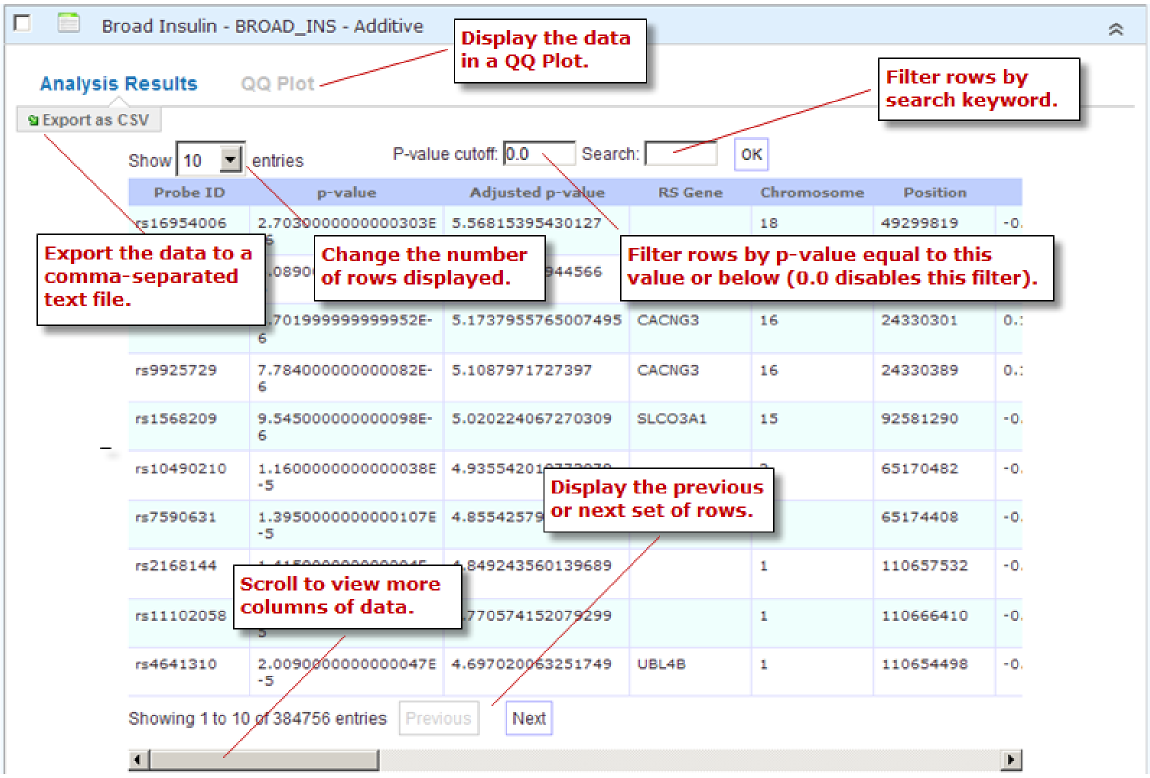
1: Diabetics Genetics Initiative

Analyses

- □ Broad Body Mass Index (diabetic) - BROAD_BMI_DIAB - Additive
- □ Broad Body Mass Index (non-diabetic) - BROAD_BMI_ND - Additive
- □ Broad Insulin - BROAD_INS - Additive
- □ Broad Fasting Plasma Glucose - BROAD_GLC - Additive
- □ Broad High-density lipoprotein - BROAD_HDL - Additive

3. Click the name of the analysis of interest.

   The rows of analysis data appear below the analysis name:



□ Broad Insulin - BROAD_INS - Additive

Show 10 entries   P-value cutoff: 0.0   Search:

Probe ID	p-value	Adjusted p-value	RS Gene	Chromosome	Position	
rs16954006	2.7030000000000303E-6	5.56815395430127		18	49299819	-0.
[illegible]	.0890[illegible]	[illegible]944566	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	.7019999999999952E-6	5.1737955765007495	CACNG3	16	24330301	0.:
rs9925729	7.784000000000082E-6	5.1087971727397	CACNG3	16	24330389	0.:
rs1568209	9.545000000000098E-6	5.020224067270309	SLCO3A1	15	92581290	-0.
rs10490210	1.1600000000000038E-5	4.93554201[illegible]	[illegible]	[illegible]	65170482	-0.
rs7590631	1.3950000000000107E-5	4.85542579[illegible]	[illegible]	[illegible]	65174408	-0.
rs2168144	[illegible]	[illegible].849243560139689		1	110657532	-0.
rs11102058	[illegible]	[illegible].770574152079299		1	110666410	-0.
rs4641310	2.0090000000000047E-5	4.697020063251749	UBL4B	1	110654498	-0.

Showing 1 to 10 of 384756 entries

Display the data in a QQ Plot.

Filter rows by search keyword.

Export the data to a comma-separated text file.

Change the number of rows displayed.

Filter rows by p-value equal to this value or below (0.0 disables this filter).

Display the previous or next set of rows.

Scroll to view more columns of data.

4. Optionally, filter the data results through one or both of the following methods and then click **OK** (do not press Enter or Return):

   - □ Specify a p-value in the **P-value cutoff** field.

     Only those rows whose **p-value** column contains a p-value at or below the specified p-value are returned.

     Setting **P-value-cutoff** to **0.0** disables the p-value filter.

   - □ Specify a search keyword in the **Search** field. All data columns are searchable.

5. Optionally, click **Export as CSV** to export the filtered data to a comma-separated text file.

6. Optionally, click **QQ Plot** to display the filtered data in a QQ Plot.

Notes about the Contents of Analysis View

- ■ A filter that specifies a gene, gene signature, RS identifier, or chromosome will not filter out studies or analyses that do not satisfy the filter. However, the data in an analysis will be limited to records that do satisfy the filter.

  tranSMART displays the following message above analysis data that has been filtered by this type of information:

  ```
  These results have been filtered according to gene/chromosome region
  criteria.
  ```

- ■ The only studies that are listed in Analysis View, and the only analyses that are listed under a study, are those that satisfy the non-genetic and non-chromosomal filters in Active Filters.

  For example, if you select a particular data type from the Data Type browser, only those studies and analyses that are associated with that data type are listed.

- ■ The data displayed for a particular analysis must satisfy the search query in Active Filters.

  For example, if your search query consists of a single filter for gene IL7, the only records that can be displayed for any analysis are those with IL7 in column **RS Gene**.

  If no records contain the IL7 gene, tranSMART displays the following message:

  ```
  No entries to display
  ```

Table View

Table View lets you perform the following tasks:

- View analysis data from multiple analyses in a single table.
- Filter the rows of analysis data by p-value and/or a search keyword.
- Export the analysis data to a comma-separated text file.

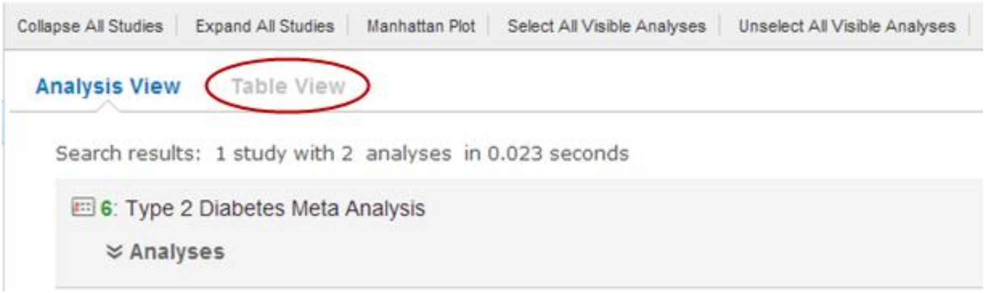
The contents of Table View are determined by the filters in the Active Filters area. Selecting an individual analysis by checking the check box next to the analysis name in Analysis View will not cause the analysis to be included in Table View.

To view analysis data in Table View:

1. Define search filters that will retrieve the records you want to view.

   Be sure to filter your search as narrowly as possible. Not only will this result in a table that contains only the most pertinent data, but it will reduce the time required to retrieve and display the data.

2. Click the **Table View** button:



Search results: 1 study with 2 analyses in 0.023 seconds

6: Type 2 Diabetes Meta Analysis

3. Optionally, filter the data results through one or both of the following methods and then click **OK** (do not press Enter or Return):

   - Specify a p-value in the **P-value cutoff** field.

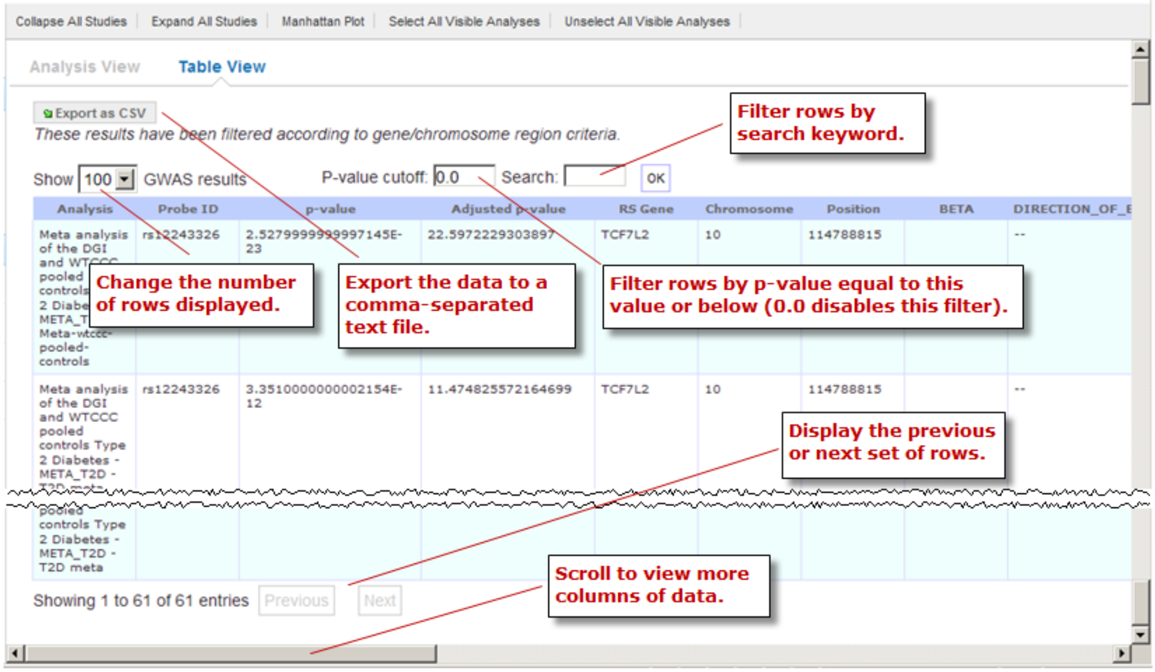
     Only those rows whose **p-value** column contains a p-value at or below the specified p-value are returned.

     Setting **P-value-cutoff** to **0.0** disables the p-value filter.

   - Specify a search keyword in the **Search** field. All data columns are searchable.

     Setting a p-value or search keyword in Analysis View for a particular analysis will not filter the data that appears in Table View. To filter Table View records by these parameters, you must define the filters in Table View itself.

4. Optionally, click **Export as CSV** to export the filtered data to a comma-separated text file.



These results have been filtered according to gene/chromosome region criteria.

Show 100 GWAS results P-value cutoff: 0.0 Search:

Analysis	Probe ID	p-value	Adjusted p-value	RS Gene	Chromosome	Position	BETA	DIRECTION_OF_E
Meta analysis of the DGI and WTCCC pooled controls 2 Diabe META_T Meta-wtccc-pooled-controls	rs12243326	2.5279999999997145E-23	22.5972229303897	TCF7L2	10	114788815		--
Meta analysis of the DGI and WTCCC pooled controls Type 2 Diabetes - META_T2D - T2D meta	rs12243326	3.3510000000002154E-12	11.474825572164699	TCF7L2	10	114788815		--
pooled controls Type 2 Diabetes - META_T2D - T2D meta								

Showing 1 to 61 of 61 entries

Filter rows by search keyword.

Change the number of rows displayed.

Export the data to a comma-separated text file.

Filter rows by p-value equal to this value or below (0.0 disables this filter).

Display the previous or next set of rows.

Scroll to view more columns of data.

Manhattan Plot

You can view data from selected analyses in a Manhattan Plot. Manhattan Plots are displayed by the Pfizer Gene Wide Association Visual Analyzer (GWAVA).

To display analysis data in a Manhattan Plot:

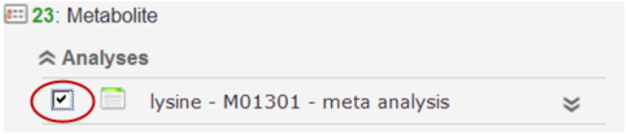
1. Optionally, define search filters using the keyword search and Filter Browser features.

   Doing so will reduce the number of studies and analyses that you will need to browse through in Analysis View when selecting the analyses to include in the Manhattan Plot.

   If you define any gene or gene signature filters, those genes will appear in the GWAVA Gene Model Selection window.

2. In Analysis View, do one of the following:

   - Select the check box next to each study whose data will be included in the Manhattan Plot:

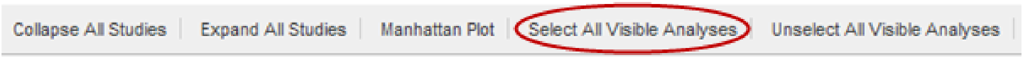
     23: Metabolite

     Analyses

     ☑ lysine - M01301 - meta analysis

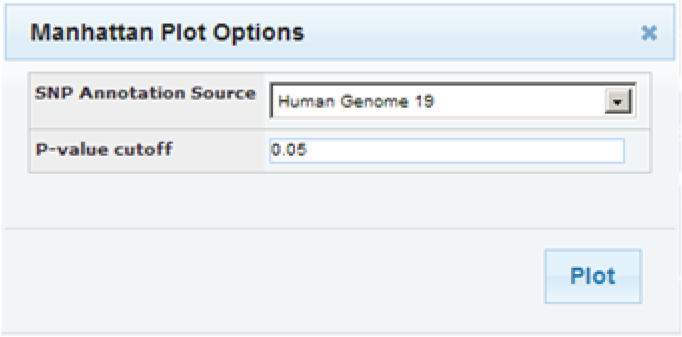
     At least one analysis must be selected.

   - Click the **Select All Visible Analyses** tab to select all analyses for all listed studies.

3. Click the **Manhattan Plot** tab.

4. In the Manhattan Plot Options dialog box, select the human genome version to use as the basis for the selected data, and optionally, specify a p-value cutoff:

   SNP Annotation Source: Human Genome 19

   P-value cutoff: 0.05

   If you specify a p-value cutoff, the only data included in the Manhattan Plot will be from records containing the specified p-value or below.

5. Click **Plot**.

   The GWAVA application opens.

6. Use the GWAVA interface to further define the data to display in the Manhattan Plot, and then click the **Results** tab to display the data. Refer to the GWAVA Help if necessary.

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 that you select.

Overview of the Dataset Explorer UI

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

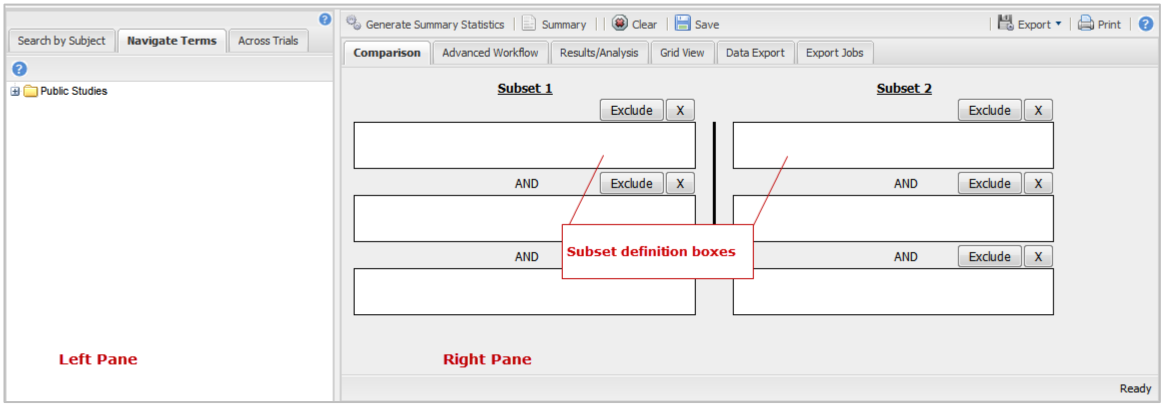
Left pane

- 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.



Subset definition boxes

Left Pane

Right Pane

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.
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 34.
Export button	Exports summary statistics data or expression data to Microsoft Excel.
Print button	Prints the tables and charts in the Results/Analysis view.
Comparison tab	Removes the currently displayed view (that is, the Results/Analysis view, or Grid view) and re-displays the subset definition boxes. This allows you to further refine the subjects for the comparison.
Advanced Workflow tab	Displays advanced analyses and visualizations that you can perform on data.
Results/Analysis tab	Displays tables and charts containing comparison and analysis data generated from the "Summary Statistics" workflow.
Grid View tab	Displays the comparison and analysis data in grid format.
Jobs tab	Displays previously run analyses.
Data Export tab	Allows you to select data to export for further analysis in an external tool.
Export Jobs tab	Displays previously exported jobs.

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. Note some analyses only allow for the specification of one group at this time.
- Generate summary statistics for the two study groups.
- Specify the points of comparison to apply to the study groups.

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 can be found both in the **Public Studies** folder of the Dataset Explorer navigation tree, as well as in the research-specific folders.

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 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), contact a tranSMART Administrator. The administrator will contact the study owner to find out if you should be granted VIEW access, EXPORT access, or no access.

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 40.

Selecting the Study

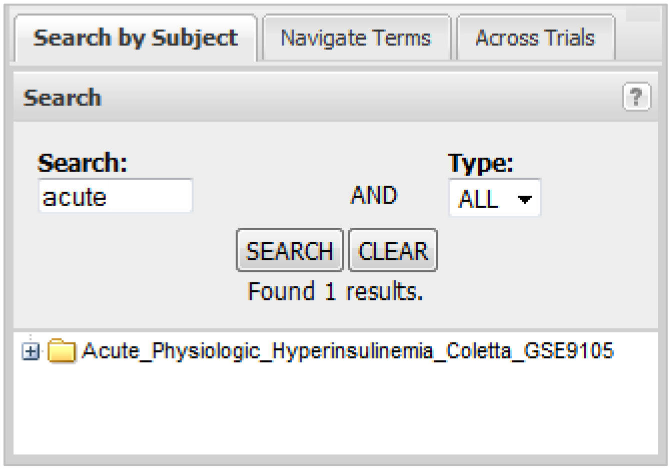
You select the study in the left pane of Dataset Explorer. You have several ways to select the study, based on the tab you choose – Search by Subject or Navigate Terms.

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 disease or an area of clinical interest.

  Example:



Search: acute

AND

Type: ALL

Found 1 results.

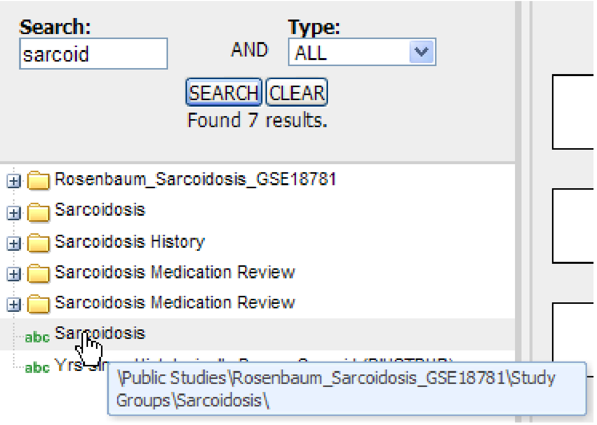
Acute_Physiologic_Hyperinsulinemia_Coletta_GSE9105

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

Study Type	**Naming Convention**
Public Studies	Name segments in the following typical format: *Condition_StudyFirstAuthor_GEOid* Example: `ProstateCancer_Ambs_GSE6956` If you prefer, you can rearrange the order of the segments (for example, author segment first). The name structure is determined by the ETL process that loads the data into the i2b2 database.

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:



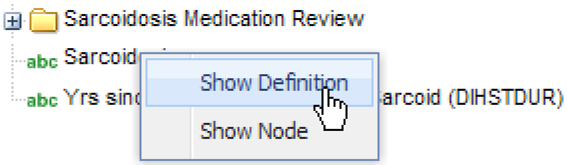
Search: sarcoid AND Type: ALL

Found 7 results.

- Rosenbaum_Sarcoidosis_GSE18781
- Sarcoidosis
- Sarcoidosis History
- Sarcoidosis Medication Review
- Sarcoidosis Medication Review
- abc Sarcoidosis
- abc Yrs si

\Public Studies\Rosenbaum_Sarcoidosis_GSE18781\Study Groups\Sarcoidosis\

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:



- Sarcoidosis Medication Review
- abc Sarcoid
- abc Yrs sinc arcoid (DIHSTDUR)

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.

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 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.

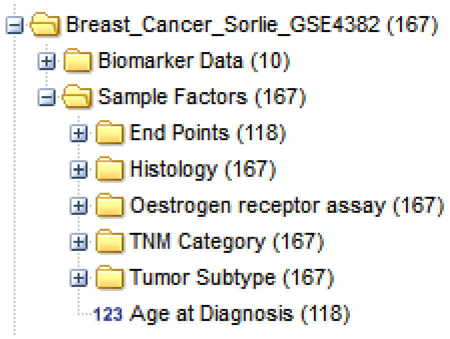
To display the details box for a study, right-click the study name and click **Show Definition**. You can display the details box for a study whether or not the study is grayed out.

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):



Breast_Cancer_Sorlie_GSE4382 (167)
Biomarker Data (10)
Sample Factors (167)
End Points (118)
Histology (167)
Oestrogen receptor assay (167)
TNM Category (167)
Tumor Subtype (167)
123 Age at Diagnosis (118)

The following table describes possible 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.
Design Factors	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.
Sample Factors	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 full 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.

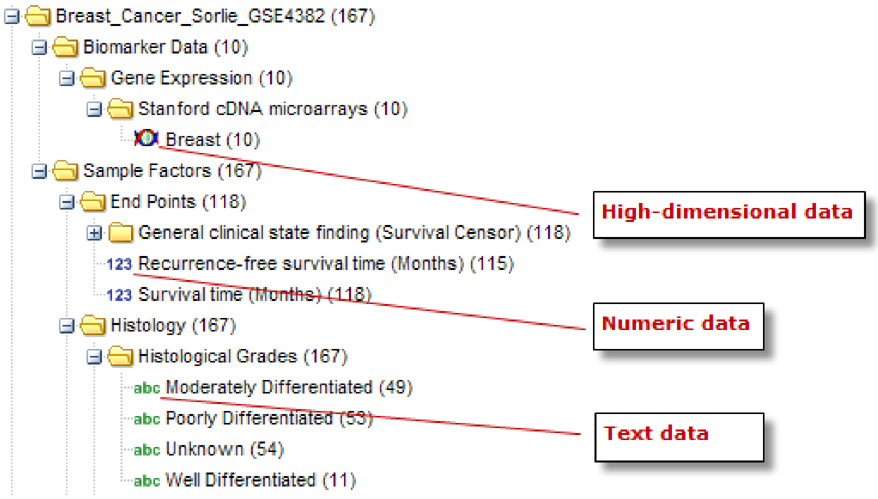
Visual Aids to Help You Select the Criteria

Each concept node in the navigation tree includes the following information:

- The numbers in parentheses at each node of the tree indicate the number of subjects represented by the node. For example, the following indicates that there are 167 subjects in the study.

Breast_Cancer_Sorlie_GSE4382 (167)

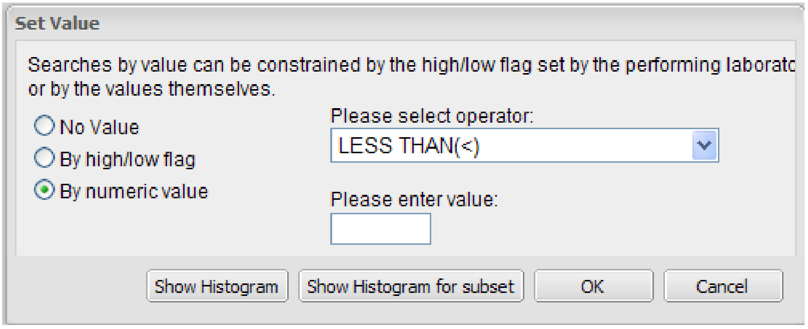
- Data values are represented in one of three ways: by number, by text, or by high dimensional data (SNP, gene expression, etc. stored as *arrays*). Each data value type is flagged by an icon, as shown below:



```
Breast_Cancer_Sorlie_GSE4382 (167)
  Biomarker Data (10)
    Gene Expression (10)
      Stanford cDNA microarrays (10)
        Breast (10)
  Sample Factors (167)
    End Points (118)
      General clinical state finding (Survival Censor) (118)
      123 Recurrence-free survival time (Months) (115)
      123 Survival time (Months) (118)
    Histology (167)
      Histological Grades (167)
        abc Moderately Differentiated (49)
        abc Poorly Differentiated (53)
        abc Unknown (54)
        abc Well Differentiated (11)
```

Specifying a Numeric Value

When you drag a numeric concept into a subset definition box, the Set Value dialog appears:



Set Value

Searches by value can be constrained by the high/low flag set by the performing laborato or by the values themselves.

○ No Value

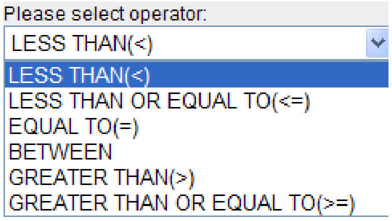
○ By high/low flag

● By numeric value

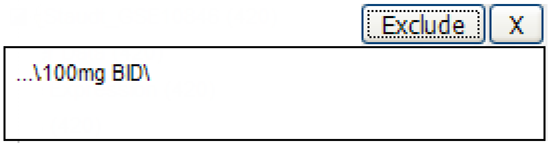
Please select operator: LESS THAN(<)

Please enter value:

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**.
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**:  Please select operator: LESS THAN(<) LESS THAN(<) LESS THAN OR EQUAL TO(<=) EQUAL TO(=) BETWEEN GREATER THAN(>) GREATER THAN OR EQUAL TO(>=) ▪ 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.

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:

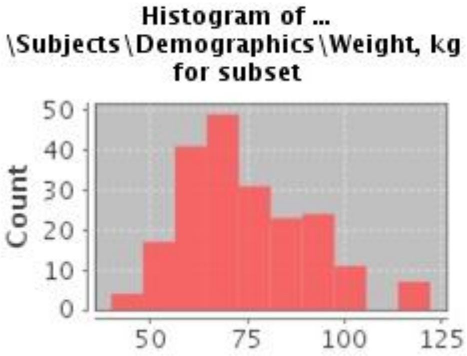


...\100mg BID\

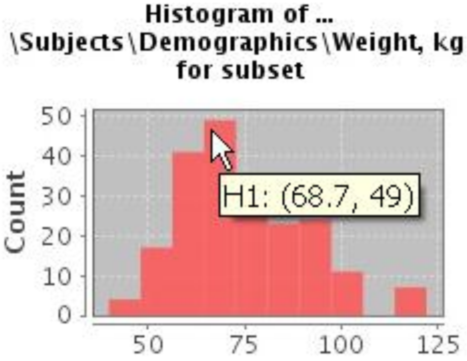
Viewing the Numeric Values Associated with a Concept

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:

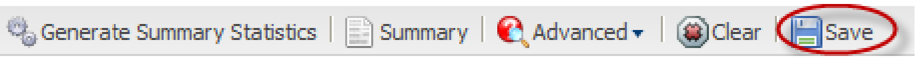


Saving Comparison Definitions

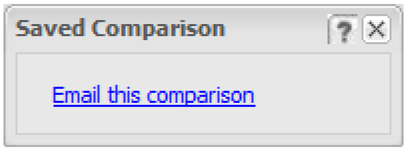
You may save your search criteria in order to regenerate the comparison at a later time without having to redefine the subsets.

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.

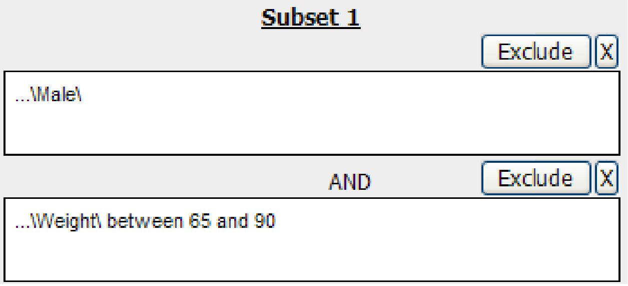
Joining Multiple Criteria for a Subset Definition

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:

  Subset 1

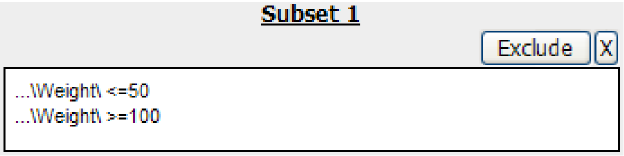
  ...\Male\

  AND

  ...\Weight\ between 65 and 90

- 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:

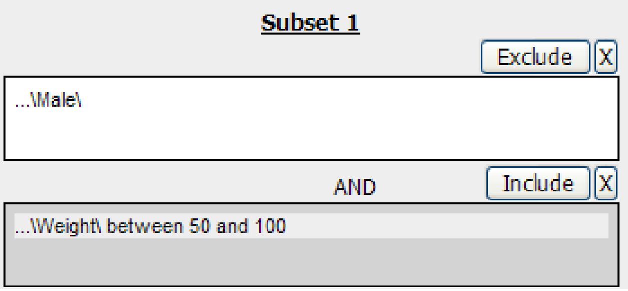
  Subset 1

  ...\Weight\ <=50
  ...\Weight\ >=100

  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:

  Subset 1

  ...\Male\

  AND

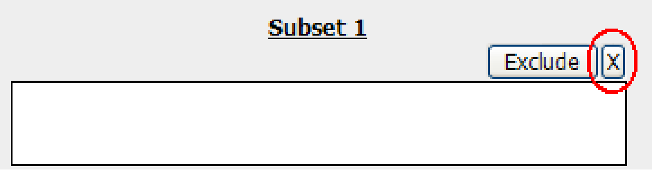
  ...\Weight\ between 50 and 100

  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:

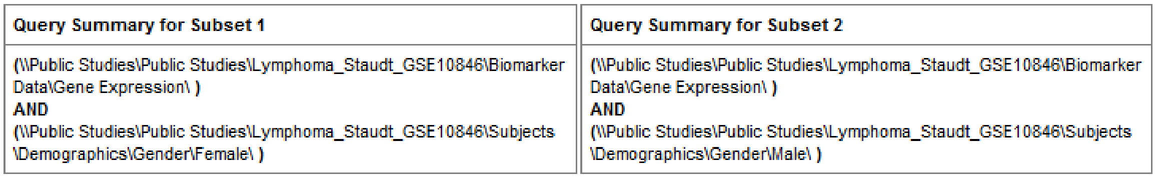


Generating Summary Statistics

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

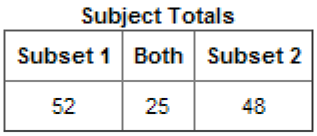
tranSMART displays tables and charts of information that describe the subsets. The information is displayed in the Results/Analysis view in the following sections:

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



Query Summary for Subset 1	Query Summary for Subset 2
(\\Public Studies\Public Studies\Lymphoma_Staudt_GSE10846\Biomarker Data\Gene Expression\ ) AND (\\Public Studies\Public Studies\Lymphoma_Staudt_GSE10846\Subjects \Demographics\Gender\Female\ )	(\\Public Studies\Public Studies\Lymphoma_Staudt_GSE10846\Biomarker Data\Gene Expression\ ) AND (\\Public Studies\Public Studies\Lymphoma_Staudt_GSE10846\Subjects \Demographics\Gender\Male\ )

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

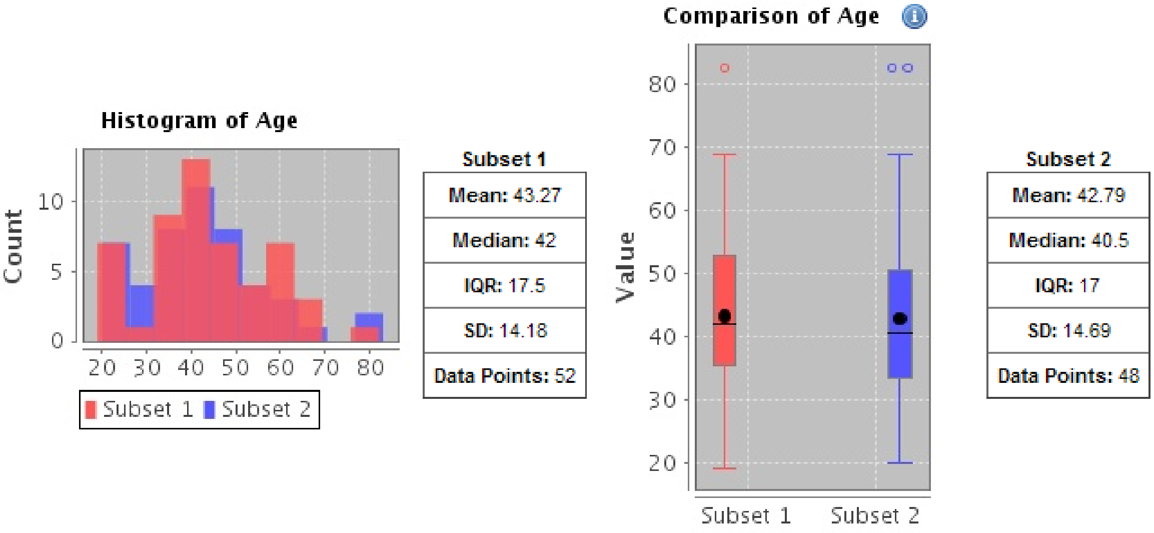


Subject Totals

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).

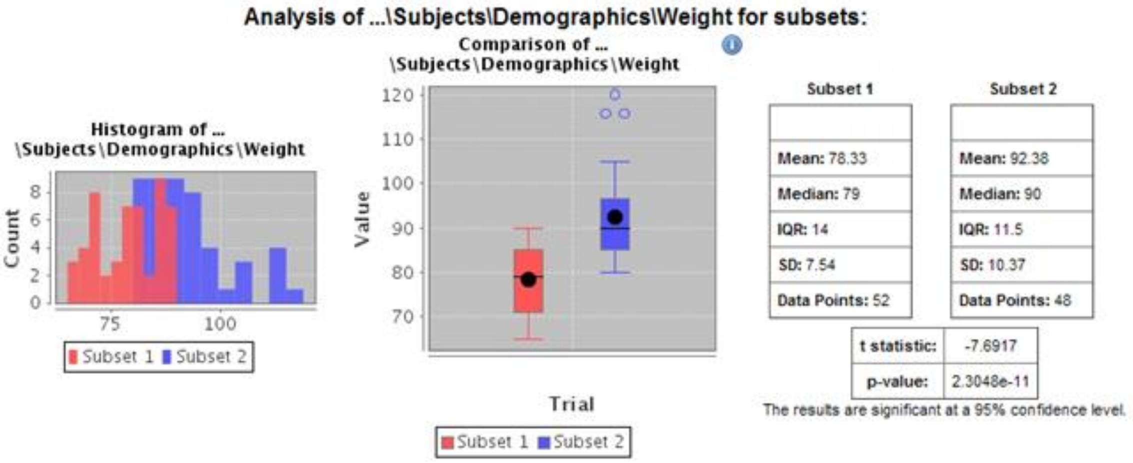
- Tables and charts that show how the subjects who match the criteria fit into age, sex, and race demographics. Example (showing the age portion only):



Subset 1
Mean: 43.27
Median: 42
IQR: 17.5
SD: 14.18
Data Points: 52

Subset 2
Mean: 42.79
Median: 40.5
IQR: 17
SD: 14.69
Data Points: 48

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


Analysis of ...\Subjects\Demographics\Weight for subsets:

Subset 1
Mean: 78.33
Median: 79
IQR: 14
SD: 7.54
Data Points: 52

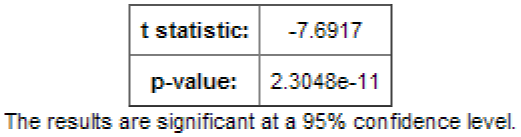
Subset 2
Mean: 92.38
Median: 90
IQR: 11.5
SD: 10.37
Data Points: 48

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

The results are significant at a 95% confidence level.

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 insufficient quantity 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.

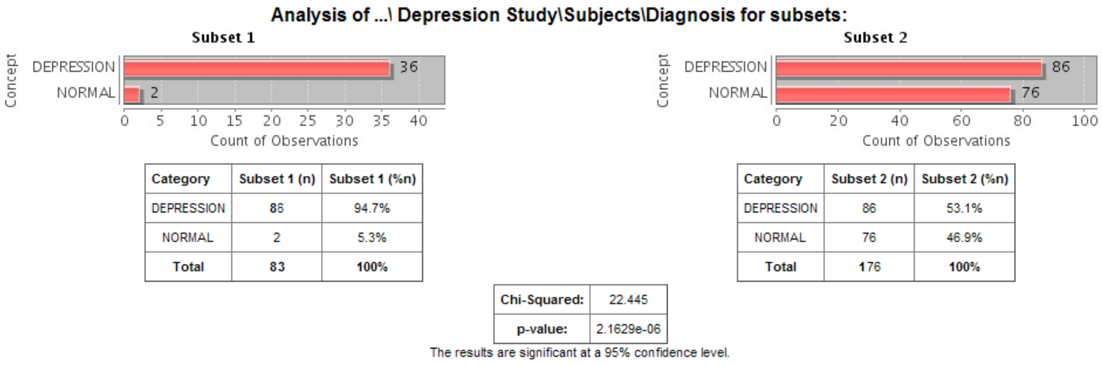
As soon as you drop the point of comparison into the Results/Analysis view, tranSMART begins to compare the subsets based on that point of comparison. When finished, tranSMART displays a side-by-side summary of how the subjects in each subset match or respond to the point of comparison.

Results of a Comparison

In a comparison of subjects in a psychological 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:



Category	Subset 1 (n)	Subset 1 (%n)
DEPRESSION	86	94.7%
NORMAL	2	5.3%
Total	**83**	**100%**

Category	Subset 2 (n)	Subset 2 (%n)
DEPRESSION	86	53.1%
NORMAL	76	46.9%
Total	**176**	**100%**

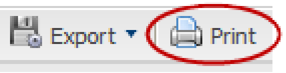
Chi-Squared:	22.445
p-value:	2.1629e-06

The results are significant at a 95% confidence level.

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

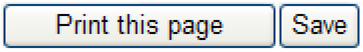
Printing or Saving the Contents of the Results/Analysis View

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

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

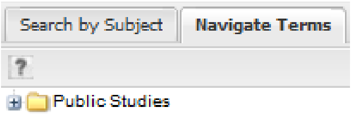
2. Right-click the chart to copy.
3. In the Internet Explorer popup menu, click **Save Picture As**.
4. In the Save Picture dialog, specify the name, location, and the file type for the chart.
5. Click **Save**.

Viewing a Study

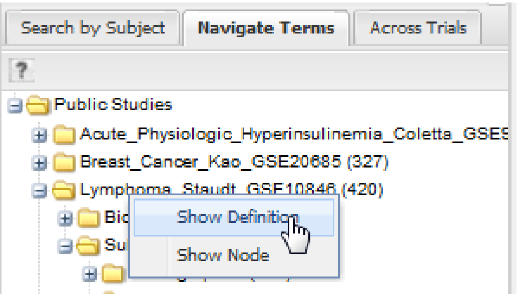
You can view a description of any Dataset Explorer study, whether or not you have access rights to the study.

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 Public Studies to open the list of experiments:

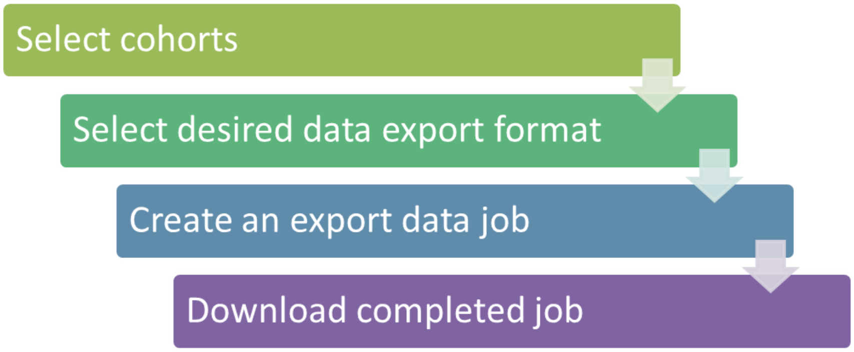
3. Right-click the particular study you are interested in.
4. Click the **Show Definition** popup:


Search by Subject
Navigate Terms
Across Trials
?
Public Studies
Acute_Physiologic_Hyperinsulinemia_Coletta_GSE
Breast_Cancer_Kao_GSE20685 (327)
Lymphoma_Staudt_GSE10846 (420)
Bio
Su
Show Definition
Show Node

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

Exporting Dataset Explorer Findings

The Data Export tab allows you to export your data locally for further analysis in several different formats. Exporting data using this tool involves the following high-level tasks:



Supported file formats include:

- Clinical and low dimensional biomarker data
- Gene expression data
- SNP data
- Gene set enrichment analysis (GSEA)

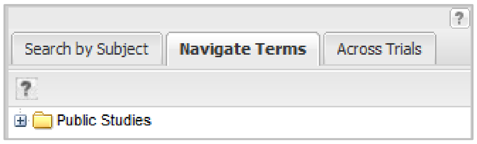
For more information on GSEA data files, visit the following site: http://www.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats

To export Dataset Explorer findings to your local machine:

1. Click the tranSMART **Dataset Explorer** tab to display the Dataset Explorer window.

2. In the left pane of the Dataset Explorer window, click the **Navigate Terms** tab.

   The navigation tree appears, showing the categories of available studies:

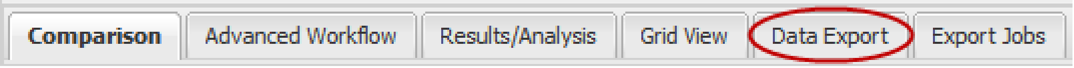
   

3. Select the study of interest.

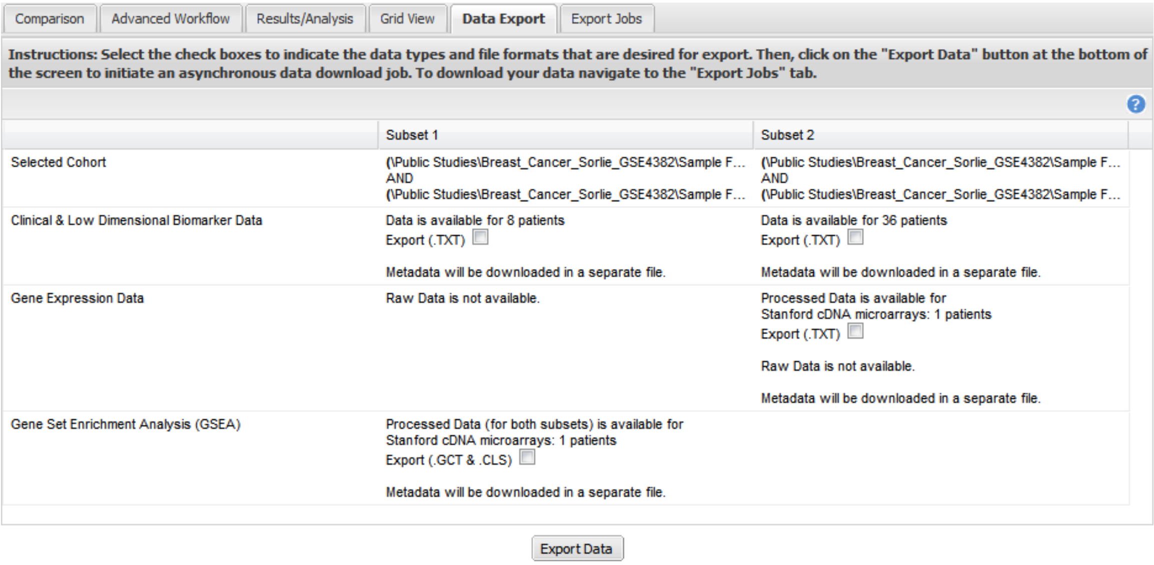
4. Define the cohorts whose data points are of interest.

   Now that the subsets are defined, you are ready to export data from the study that applies to the subsets.

5. Click the **Data Export** tab:

   The Data Export page appears with your selected cohorts:

   Instructions: Select the check boxes to indicate the data types and file formats that are desired for export. Then, click on the "Export Data" button at the bottom of the screen to initiate an asynchronous data download job. To download your data navigate to the "Export Jobs" tab.

	Subset 1	Subset 2
Selected Cohort	(\Public Studies\Breast_Cancer_Sorlie_GSE4382\Sample F... AND (\Public Studies\Breast_Cancer_Sorlie_GSE4382\Sample F...	(\Public Studies\Breast_Cancer_Sorlie_GSE4382\Sample F... AND (\Public Studies\Breast_Cancer_Sorlie_GSE4382\Sample F...
Clinical & Low Dimensional Biomarker Data	Data is available for 8 patients Export (.TXT) □ Metadata will be downloaded in a separate file.	Data is available for 36 patients Export (.TXT) □ Metadata will be downloaded in a separate file.
Gene Expression Data	Raw Data is not available.	Processed Data is available for Stanford cDNA microarrays: 1 patients Export (.TXT) □ Raw Data is not available. Metadata will be downloaded in a separate file.
Gene Set Enrichment Analysis (GSEA)	Processed Data (for both subsets) is available for Stanford cDNA microarrays: 1 patients Export (.GCT & .CLS) □ Metadata will be downloaded in a separate file.	

6. Select the check boxes to indicate the data types and file formats that are desired for export.

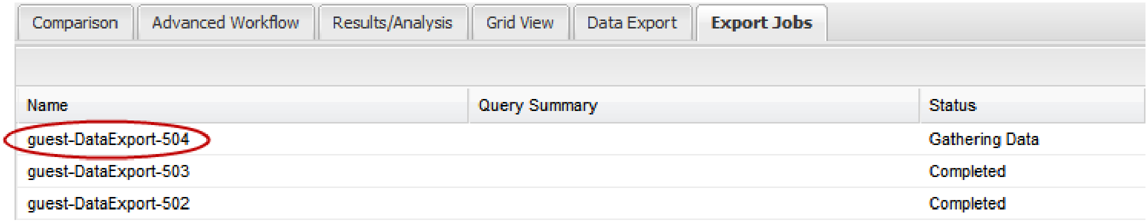
7. Click **Export Data** at the bottom of the tranSMART browser window.

   The command will now start a job. You may choose to have the job run in the background in order to continue with other analyses and cohort selection while the job completes. The job could take several minutes depending on the amount of data selected.

8. Click the **Export Jobs** tab to access completed jobs or to check the status of a pending job.

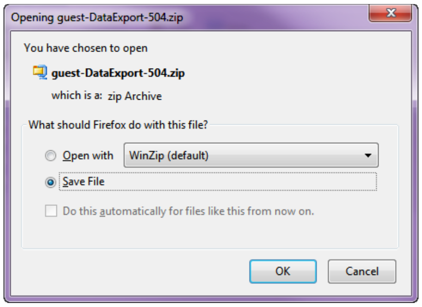
   Jobs follow the naming convention *User - Type of Job Run - Job ID.*

9. Click the name of the job you processed:



Name	Query Summary	Status
guest-DataExport-504		Gathering Data
guest-DataExport-503		Completed
guest-DataExport-502		Completed

   The Open File dialog box appears:


Opening guest-DataExport-504.zip

You have chosen to open

guest-DataExport-504.zip

which is a: zip Archive

What should Firefox do with this file?

○ Open with WinZip (default)

● Save File

□ Do this automatically for files like this from now on.

10. Select **Save File**, then click **OK**.

    Your file will be sent to the **Downloads** folder on your local machine in a .zip file. The .zip file contains separate folders for subsets, clinical data, gene expression data, and other factors you may have specified during cohort selection.

Generating Advanced Analyses and Visualizations

Advanced analyses and visualizations offered with tranSMART allow a user to produce the following within Dataset Explorer:

- Heatmaps
  - Standard Heatmap (page 44)
  - Hierarchical Clustering (page 46)
  - K-Means Clustering (page 49)
  - Marker Selection (page 52)
- Advanced Analyses
  - Box Plot with ANOVA (page 55)
  - Principal Component Analysis (page 57)

- □ Scatter Plot with Linear Regression (page 60)
- □ Survival Analysis (page 62)
- □ Table with Fisher Test Analysis (page 65)

Dataset Explorer uses the R software environment for statistical computing and to generate analyses and visualizations. For more information, visit http://www.r-project.org.

Generating Heatmaps

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

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

A heatmap can display data points for up to 1000 samples.

Dataset Explorer uses the R software environment for statistical computing and to generate analyses and visualizations. For more information, visit http://www.r-project.org.

You can generate the following types of heatmaps:

- Standard Heatmap (below)
- Hierarchical Clustering (page 46)
- K-Means Clustering (page 49)
- Marker Selection (page 52)

Standard Heatmap

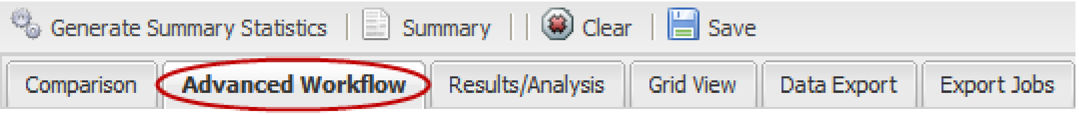
A standard heatmap is a visualization of biomarker data points with no indication of patterns, groupings, or differentiation among the data points.

To generate a standard heatmap:

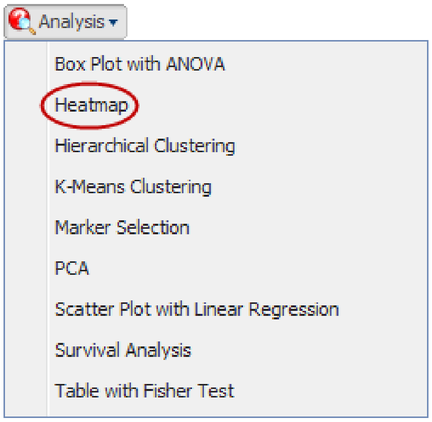
1. Run tranSMART, then click the **Dataset Explorer** tab.
2. Define the cohorts you wish to analyze by dragging one or more concepts from a study into empty subset definition boxes. For more information, see Populating the Study Groups on page 31.

   To compare two subsets, you may drag an additional concept into the Subset 2 comparison box.

3. Click the **Advanced Workflow** tab:

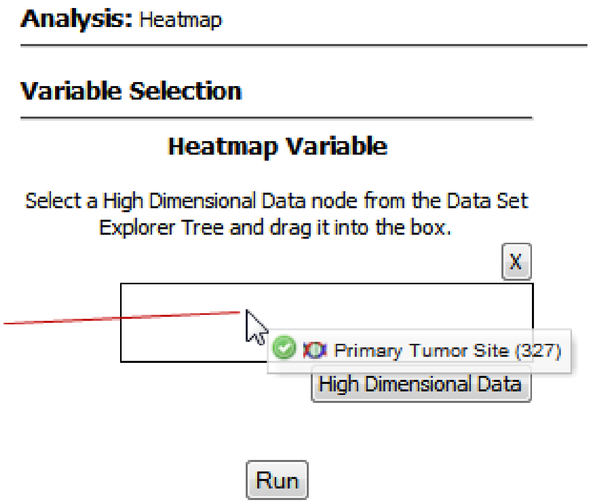


4. Select **Heatmap** from the **Analysis** dropdown menu:



The Variable Selection section appears.

5. Define the heatmap variable by selecting a high dimensional data node from the Dataset Explorer tree and dragging it into the Heatmap Variable definition box:



Select a High Dimensional Data node from the Data Set Explorer Tree and drag it into the box.

High dimensional data nodes are indicated by the icon (⊠) to the left of study data.

6. Click the **High Dimensional Data** button.

The Compare Subsets-Pathway Selection dialog appears.

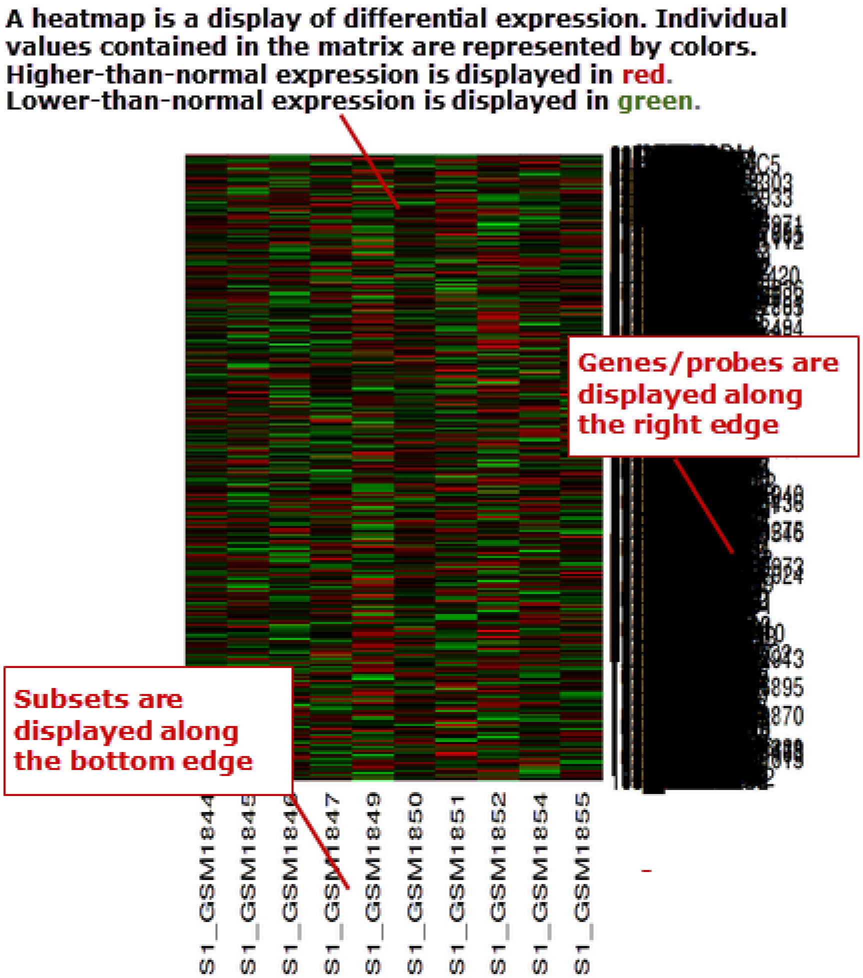
7. Specify the platform and other factors of interest.

For more information, see High Dimensional Data on page 71.

8. Click **Apply Selections**.

9. Click **Run**.

   Your analysis appears below:



A heatmap is a display of differential expression. Individual values contained in the matrix are represented by colors. Higher-than-normal expression is displayed in red. Lower-than-normal expression is displayed in green.

Genes/probes are displayed along the right edge

Subsets are displayed along the bottom edge

Hierarchical Clustering

Hierarchical clustering is a visualization of patterns of related data points in gene expression data.

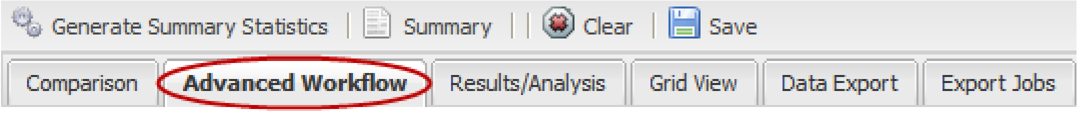
To generate a hierarchical clustering heatmap:

1. Run tranSMART, then click the **Dataset Explorer** tab.

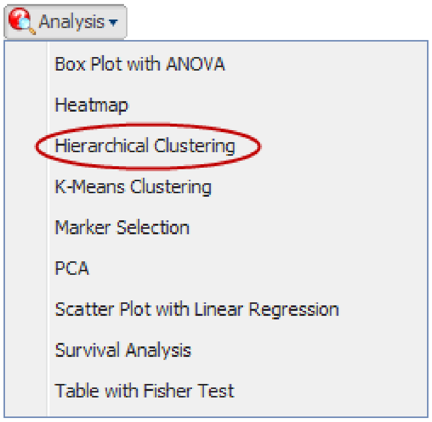
2. Define the cohorts you wish to analyze by dragging one or more concepts from a study into empty subset definition boxes. For more information, see Populating the Study Groups on page 31.

To compare two subsets, you may drag an additional concept into the Subset 2 comparison box.

3. Click the **Advanced Workflow** tab:

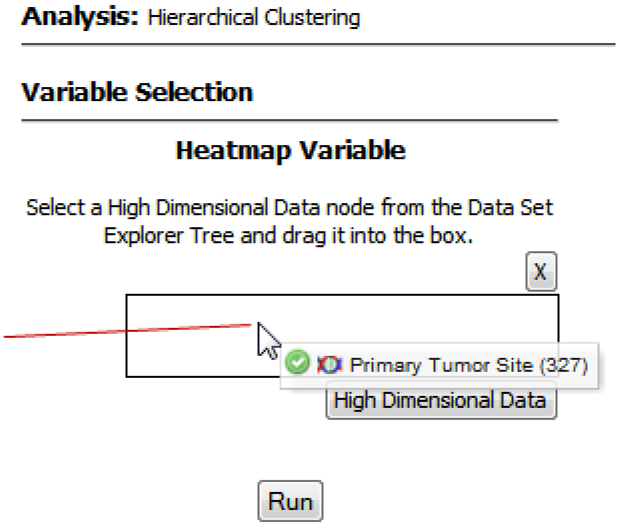


4. Select **Hierarchical Clustering** from the **Analysis** dropdown menu:



The Variable Selection section appears.

5. Define the heatmap variable by selecting a high dimensional data node from the Dataset Explorer tree and dragging it into the Heatmap Variable definition box:



Analysis: Hierarchical Clustering

Variable Selection

Heatmap Variable

Select a High Dimensional Data node from the Data Set Explorer Tree and drag it into the box.

High dimensional data nodes are indicated by the icon ( ) to the left of study data.

6. Click the **High Dimensional Data** button.

The Compare Subsets-Pathway Selection dialog appears.

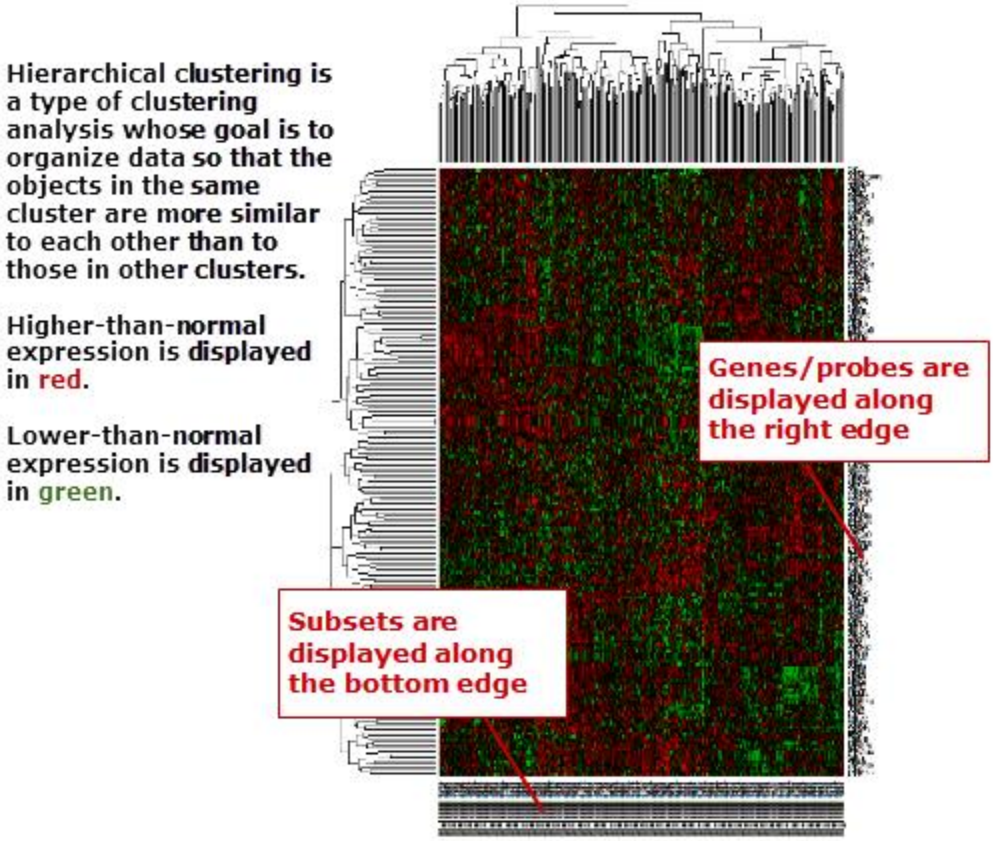
7. Specify the platform and other factors of interest.

   For more information, see High Dimensional Data on page 71.

8. Click **Apply Selections**.

9. Click **Run**.

   Your analysis appears below:



Hierarchical clustering is a type of clustering analysis whose goal is to organize data so that the objects in the same cluster are more similar to each other than to those in other clusters.

Higher-than-normal expression is displayed in red.

Lower-than-normal expression is displayed in green.

Genes/probes are displayed along the right edge

Subsets are displayed along the bottom edge

To read more about Hierarchical Clustering, visit: http://www.ics.uci.edu/~eppstein/280/cluster.html

K-Means Clustering

K-Means clustering is a visualization of groupings of the most closely related data points, based on the number of groupings you specify.

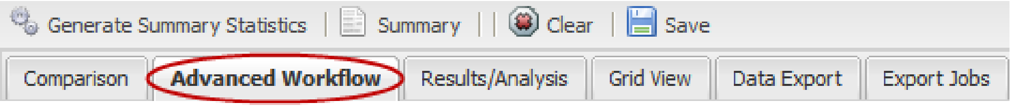
The K-Means analysis clusters columns together – rows are not clustered.

To generate a k-means clustering heatmap:

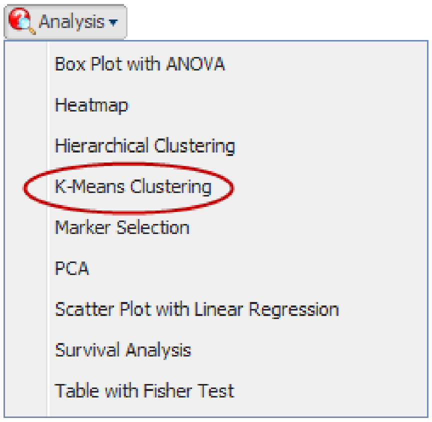
1. Run tranSMART, then click the **Dataset Explorer** tab.
2. Define the cohorts you wish to analyze by dragging one or more concepts from a study into empty subset definition boxes. For more information, see Populating the Study Groups on page 31.

To compare two subsets, you may drag an additional concept into the Subset 2 comparison box.

3. Click the **Advanced Workflow** tab:

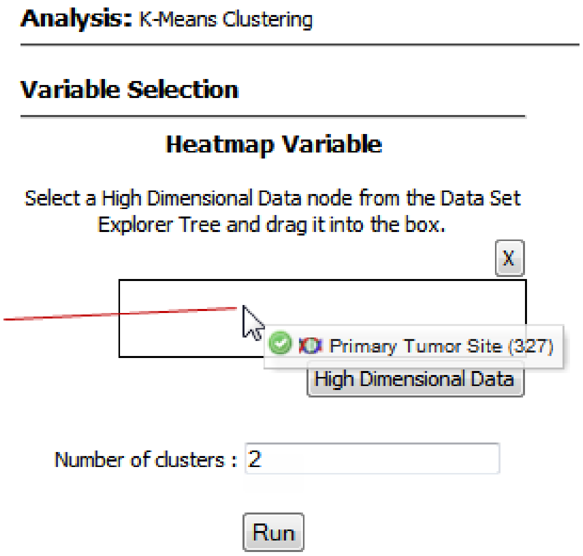


4. Select **K-Means Clustering** from the **Analysis** dropdown menu:



The Variable Selection section appears.

5. Define the heatmap variable by selecting a high dimensional data node from the Dataset Explorer tree and dragging it into the Heatmap Variable definition box:



Analysis: K-Means Clustering

Variable Selection

Heatmap Variable

Select a High Dimensional Data node from the Data Set Explorer Tree and drag it into the box.

Number of clusters : 2

High dimensional data nodes are indicated by the icon (⊗) to the left of study data.

6. Click the **High Dimensional Data** button.

   The Compare Subsets-Pathway Selection dialog appears.

7. Specify the platform and other factors of interest.

   For more information, see High Dimensional Data on page 71.

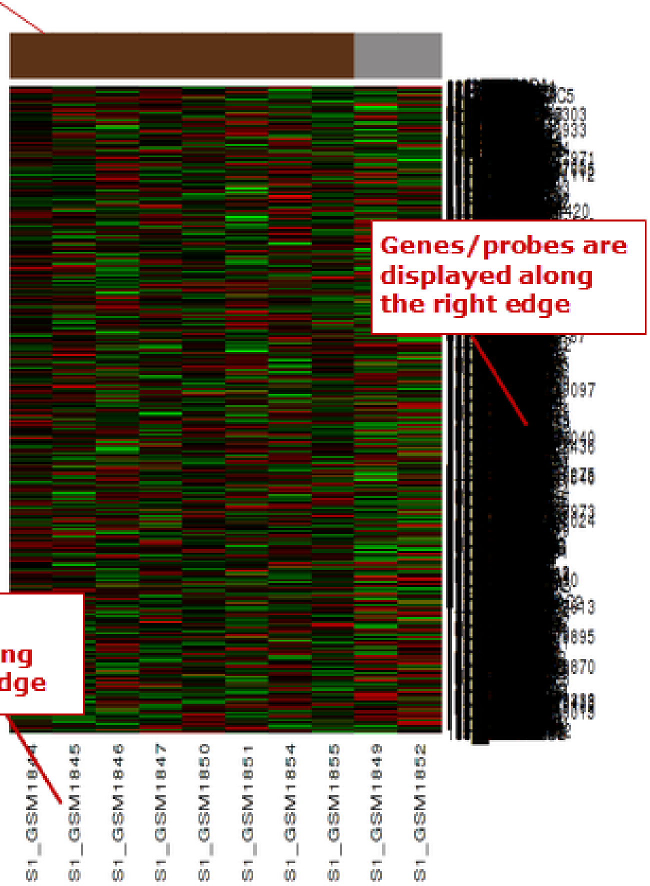
8. Click **Apply Selections**.

9. In the **Number of clusters** field, type a numerical value.

10. Click **Run**.

    Your analysis appears below:

Rows/columns that fall in the same cluster are represented by the distinct colors on the row/column side bars at the top of the heatmap



Genes/probes are displayed along the right edge

Subsets are displayed along the bottom edge

The K-Means clustering heatmap clusters genes and/or samples into a specified number of clusters. The result is *k* clusters, each centered around a randomly-selected data point.

Higher-than-normal expression is displayed in red.

Lower-than-normal expression is displayed in green.

To read more about K-Means Clustering, visit: http://www.ics.uci.edu/~eppstein/280/cluster.html

Marker Selection

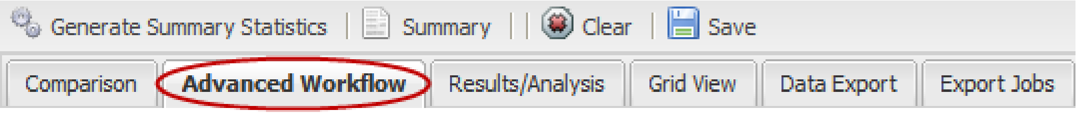
A marker selection heatmap is a visualization of differentially expressed genes in distinct phenotypes. Specifically, the algorithm determines the set of genes which is most differently expressed between the two subsets. This list of differentially expressed genes is subsequently presented in a table, along with a variety of accompanying statistics.

To generate a marker selection heatmap:

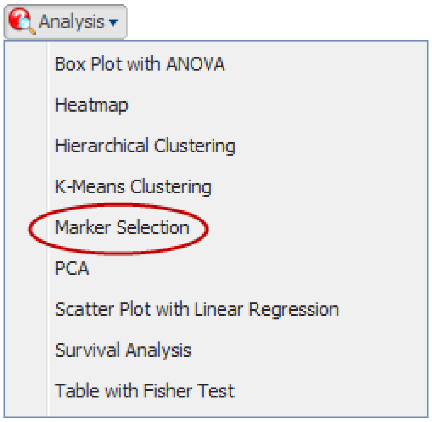
1. Run tranSMART, then click the **Dataset Explorer** tab.
2. Define the cohorts you wish to analyze by dragging one or more concepts from a study into empty subset definition boxes. For more information, see Populating the Study Groups on page 31.

   ⚠ Two subsets must be specified when using a Marker Selection heatmap.

3. Click the **Advanced Workflow** tab:

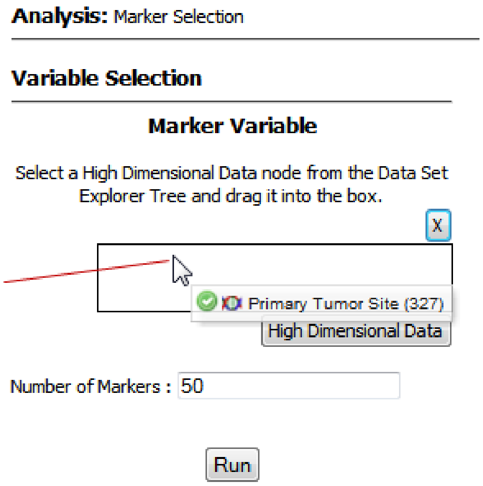


4. Select **Marker Selection** from the **Analysis** dropdown menu:



   The Variable Selection section appears.

5. Define the required variable by selecting a high dimensional data node from the Dataset Explorer tree and dragging it into the Marker Variable definition box:



Analysis: Marker Selection

Variable Selection

Marker Variable

Select a High Dimensional Data node from the Data Set Explorer Tree and drag it into the box.

Number of Markers : 50

High dimensional data nodes are indicated by the icon (■) to the left of study data.

6. Click the **High Dimensional Data** button.

   The Compare Subsets-Pathway Selection dialog appears.

7. Specify the platform and other factors of interest.

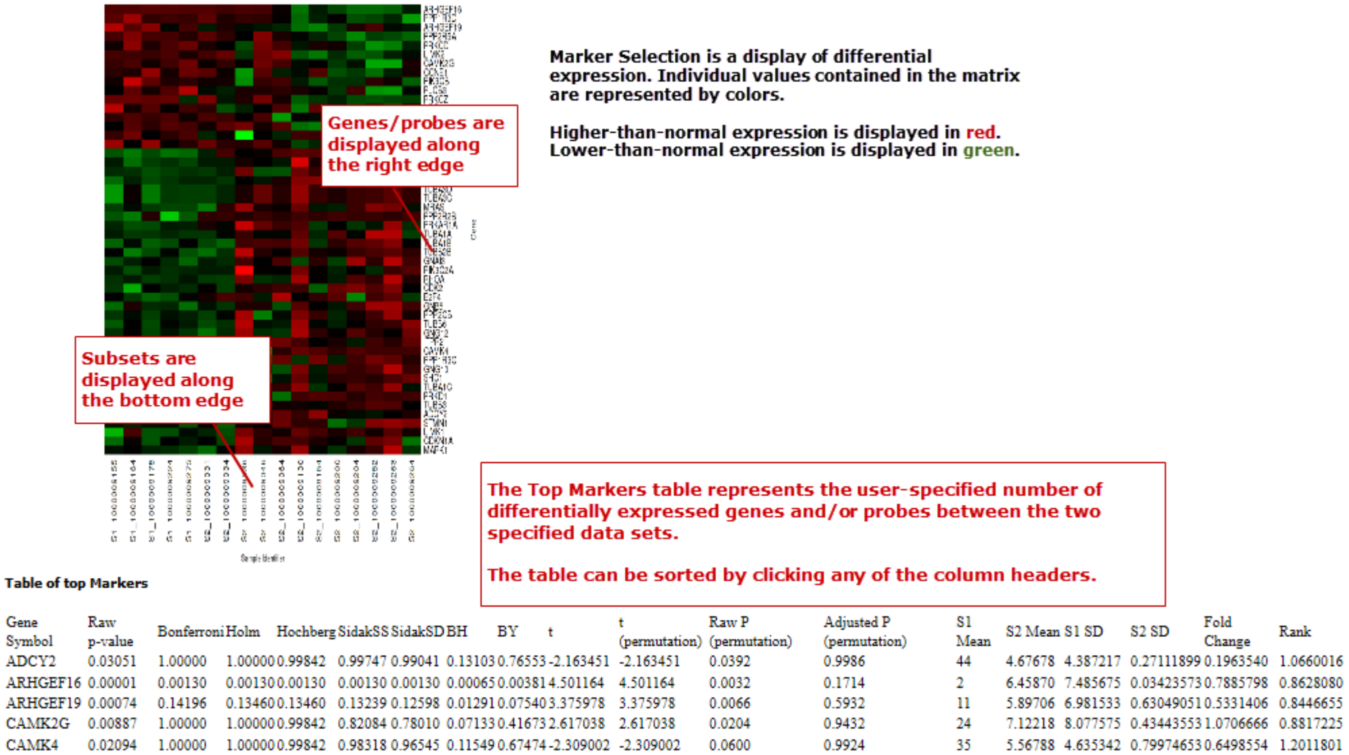
   For more information, see High Dimensional Data on page 71.

8. Click **Apply Selections**.

   In the **Number of Markers** field, type a numerical value. This will determine the number of differentially expressed genes that are returned.

9. Click **Run**.

   Your analysis appears below:



Genes/probes are displayed along the right edge

Subsets are displayed along the bottom edge

Marker Selection is a display of differential expression. Individual values contained in the matrix are represented by colors.

Higher-than-normal expression is displayed in red. Lower-than-normal expression is displayed in green.

The Top Markers table represents the user-specified number of differentially expressed genes and/or probes between the two specified data sets.

The table can be sorted by clicking any of the column headers.

Table of top Markers

Gene Symbol	Raw p-value	Bonferroni	Holm	Hochberg	SidakSS	SidakSD	BH	BY	t	t (permutation)	Raw P (permutation)	Adjusted P (permutation)	S1 Mean	S2 Mean	S1 SD	S2 SD	Fold Change	Rank
ADCY2	0.03051	1.00000	1.00000	0.99842	0.99747	0.99041	0.13103	0.76553	-2.163451	-2.163451	0.0392	0.9986	44	4.67678	4.387217	0.27111899	0.1963540	1.0660016
ARHGEF16	0.00001	0.00130	0.00130	0.00130	0.00130	0.00130	0.00065	0.00381	4.501164	4.501164	0.0032	0.1714	2	6.45870	7.485675	0.03423573	0.7885798	0.8628080
ARHGEF19	0.00074	0.14196	0.13460	0.13460	0.13239	0.12598	0.01291	0.07540	3.375978	3.375978	0.0066	0.5932	11	5.89706	6.981533	0.63049051	0.5331406	0.8446655
CAMK2G	0.00887	1.00000	1.00000	0.99842	0.82084	0.78010	0.07133	0.41673	2.617038	2.617038	0.0204	0.9432	24	7.12218	8.077575	0.43443553	1.0706666	0.8817225
CAMK4	0.02094	1.00000	1.00000	0.99842	0.98318	0.96545	0.11549	0.67474	-2.309002	-2.309002	0.0600	0.9924	35	5.56788	4.635342	0.79974653	0.6498554	1.2011801

For more information on the analyses used in Marker Selection, visit: http://mathworld.wolfram.com/bonferronicorrection.html

Generating Advanced Analyses

Advanced analyses include:

- Box Plot with ANOVA (page 55)
- Principal Component Analysis (page 57)
- Scatter Plot with Linear Regression (page 60)
- Survival Analysis (page 62)
- Table with Fisher Test Analysis (page 65)

Box Plot with ANOVA

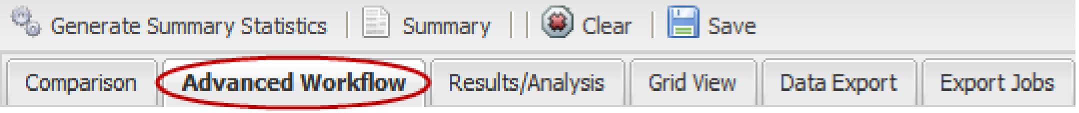
A box plot with ANOVA analysis displays a box and whisker plot with corresponding analysis of variance in the sample(s).

To perform a box plot with ANOVA analysis:

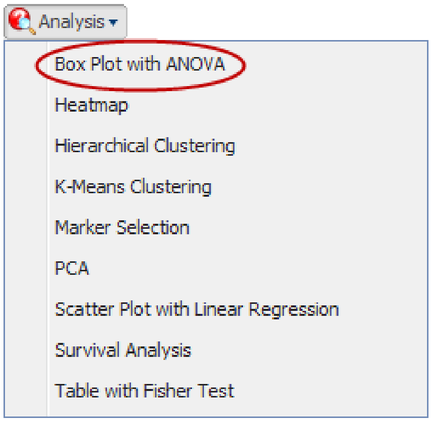
1. Run tranSMART, then click the **Dataset Explorer** tab.
2. Define the cohorts you wish to analyze by dragging one or more concepts from a study into empty subset definition boxes. For more information, see Populating the Study Groups on page 31.

   Only one subset may be specified in this analysis. Information in Subset 2 will be ignored.

3. Click the **Advanced Workflow** tab:



4. Select **Box Plot with ANOVA** from the **Analysis** dropdown menu:



   The Variable Selection section appears. You will need to define what variables in the study are independent, and what variables are dependent. At least one of the variables should be continuous (for example, Age), and one should be a categorical value (for example, Tissue Type).

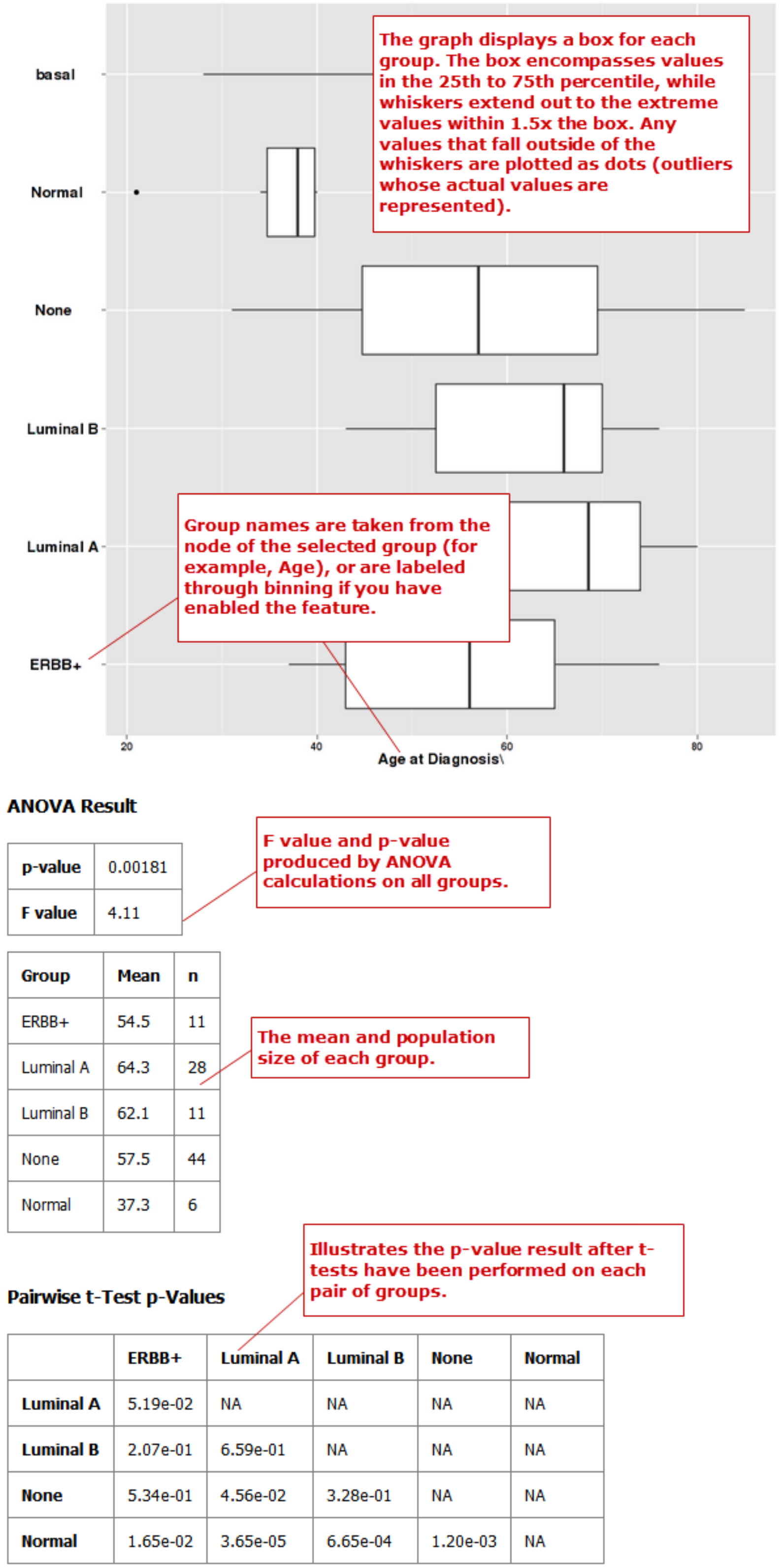
   If the *independent variable* defines the groups, boxes will be plotted horizontally. If the *dependent variable* defines the groups, boxes will be plotted vertically.

5. Define the variables.

   In this example, the data binning feature is not used. For future reference, data binning refers to a pre-processing technique used to reduce minor observation errors. Clusters of data are replaced by a value representative of that cluster (the central value). For information on binning, see Data Binning on page 66.

6. Click **Run**.

   Your analysis appears below:



The graph displays a box for each group. The box encompasses values in the 25th to 75th percentile, while whiskers extend out to the extreme values within 1.5x the box. Any values that fall outside of the whiskers are plotted as dots (outliers whose actual values are represented).

Group names are taken from the node of the selected group (for example, Age), or are labeled through binning if you have enabled the feature.

ANOVA Result

p-value	0.00181
F value	4.11

F value and p-value produced by ANOVA calculations on all groups.

Group	Mean	n
ERBB+	54.5	11
Luminal A	64.3	28
Luminal B	62.1	11
None	57.5	44
Normal	37.3	6

The mean and population size of each group.

Pairwise t-Test p-Values

Illustrates the p-value result after t-tests have been performed on each pair of groups.

	ERBB+	Luminal A	Luminal B	None	Normal
Luminal A	5.19e-02	NA	NA	NA	NA
Luminal B	2.07e-01	6.59e-01	NA	NA	NA
None	5.34e-01	4.56e-02	3.28e-01	NA	NA
Normal	1.65e-02	3.65e-05	6.65e-04	1.20e-03	NA

Principal Component Analysis

In a principal component analysis (PCA), the total number of variables in the dataset is reduced to a smaller number of variables – the principle components of the dataset.

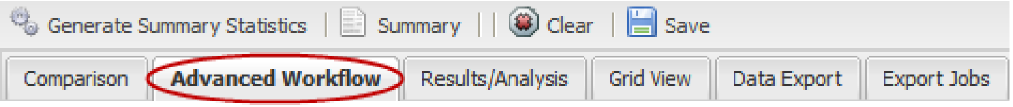
Principal component variables are calculated from correlated variables in the total dataset. In other words, the principal component analysis is a workflow used to identify variance in a dataset. The analysis can be run on an entire microarray chip, or on a pathway.

To perform a principal component analysis:

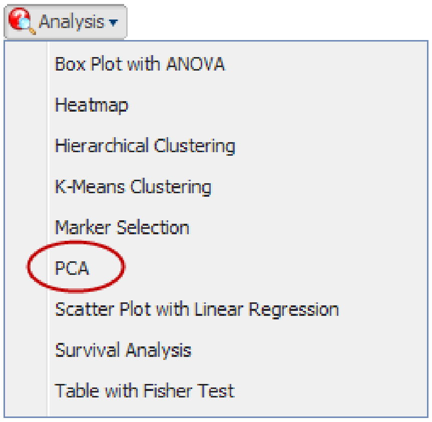
1. Run tranSMART, then click the **Dataset Explorer** tab.
2. Define the cohorts you wish to analyze by dragging one or more concepts from a study into empty subset definition boxes. For more information, see Populating the Study Groups on page 31.

   Only one subset may be specified in this analysis. Information in Subset 2 will be ignored.

3. Click the **Advanced Workflow** tab:

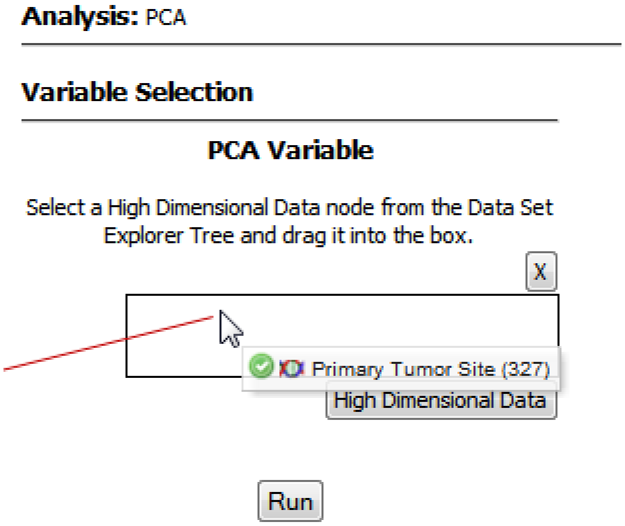


4. Select **PCA** from the **Analysis** dropdown menu:



   The Variable Selection section appears.

5. Define the heatmap variable by selecting a high dimensional data node from the Dataset Explorer tree and dragging it into the PCA Variable definition box:



Analysis: PCA

Variable Selection

PCA Variable

Select a High Dimensional Data node from the Data Set Explorer Tree and drag it into the box.

High dimensional data nodes are indicated by the (●) icon to the left of study data.

6. Click the **High Dimensional Data** button.

   The Compare Subsets-Pathway Selection dialog appears.

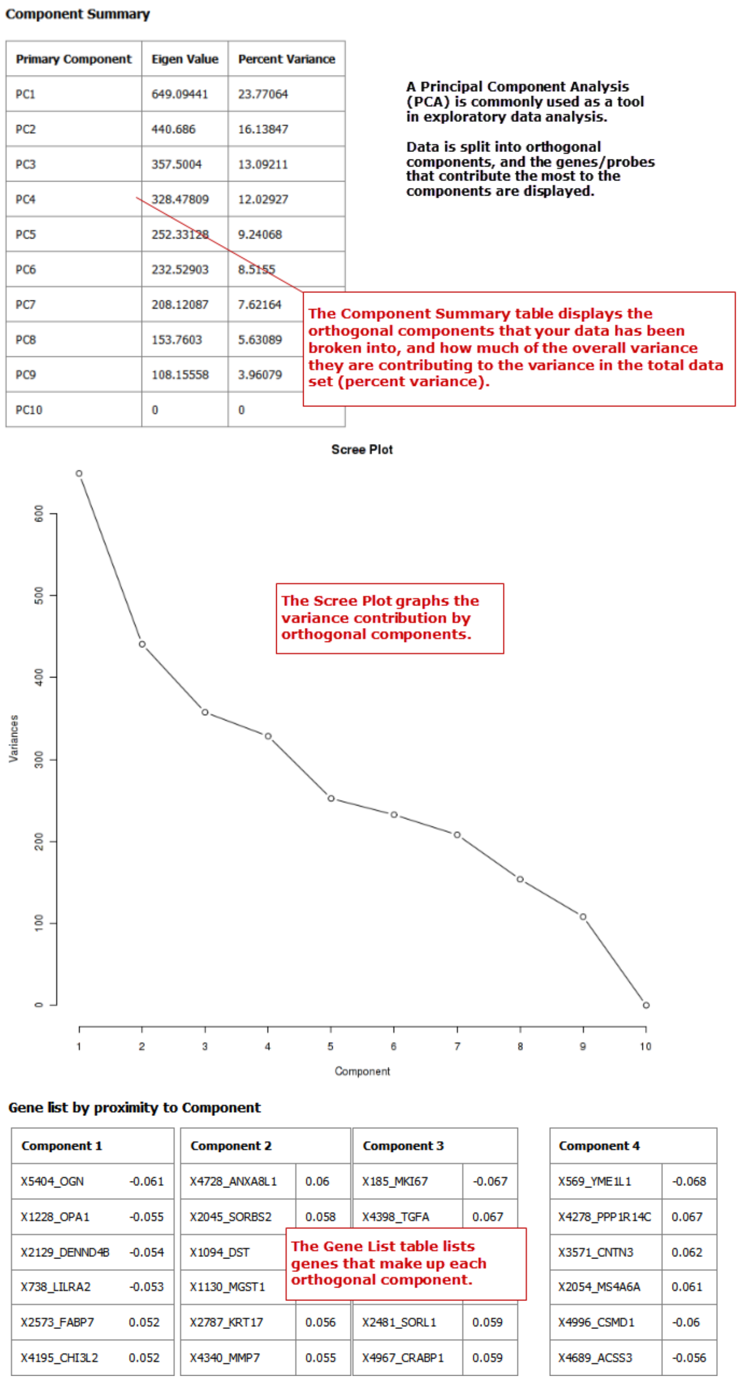
7. Specify the platform and other factors of interest.

   For more information, see High Dimensional Data on page 71.

8. Click **Apply Selections**.

9. Click **Run**.

   Your analysis appears below:



Component Summary

Primary Component	Eigen Value	Percent Variance
PC1	649.09441	23.77064
PC2	440.686	16.13847
PC3	357.5004	13.09211
PC4	328.47809	12.02927
PC5	252.33128	9.24068
PC6	232.52903	8.5155
PC7	208.12087	7.62164
PC8	153.7603	5.63089
PC9	108.15558	3.96079
PC10	0	0

A Principal Component Analysis (PCA) is commonly used as a tool in exploratory data analysis.

Data is split into orthogonal components, and the genes/probes that contribute the most to the components are displayed.

The Component Summary table displays the orthogonal components that your data has been broken into, and how much of the overall variance they are contributing to the variance in the total data set (percent variance).

Scree Plot

The Scree Plot graphs the variance contribution by orthogonal components.

Gene list by proximity to Component

Component 1		Component 2		Component 3		Component 4	
X5404_OGN	-0.061	X4728_ANXA8L1	0.06	X185_MKI67	-0.067	X569_YME1L1	-0.068
X1228_OPA1	-0.055	X2045_SORBS2	0.058	X4398_TGFA	0.067	X4278_PPP1R14C	0.067
X2129_DENND4B	-0.054	X1094_DST				X3571_CNTN3	0.062
X738_LILRA2	-0.053	X1130_MGST1				X2054_MS4A6A	0.061
X2573_FABP7	0.052	X2787_KRT17	0.056	X2481_SORL1	0.059	X4996_CSMD1	-0.06
X4195_CHI3L2	0.052	X4340_MMP7	0.055	X4967_CRABP1	0.059	X4689_ACSS3	-0.056

The Gene List table lists genes that make up each orthogonal component.

For more information regarding PCAs, see: http://psb.stanford.edu/psb-online/proceedings/psb00/raychaudhuri.pdf.

Scatter Plot with Linear Regression

A scatter plot displays values for two variables within a dataset, with a line that best fits the slope of the data.

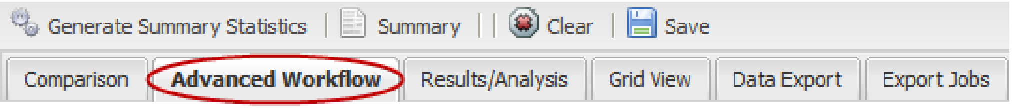
To perform a scatter plot with linear regression analysis:

1. Run tranSMART, then click the **Dataset Explorer** tab.

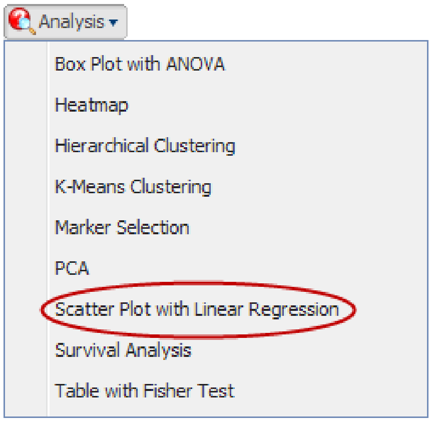
2. Define the cohorts you wish to analyze by dragging one or more concepts from a study into empty subset definition boxes. For more information, see Populating the Study Groups on page 31.

   Only one subset may be specified in this analysis. Information in Subset 2 will be ignored.

3. Click the **Advanced Workflow** tab above Subset 1:



4. Select **Scatter Plot with Linear Regression** from the **Analysis** dropdown menu:

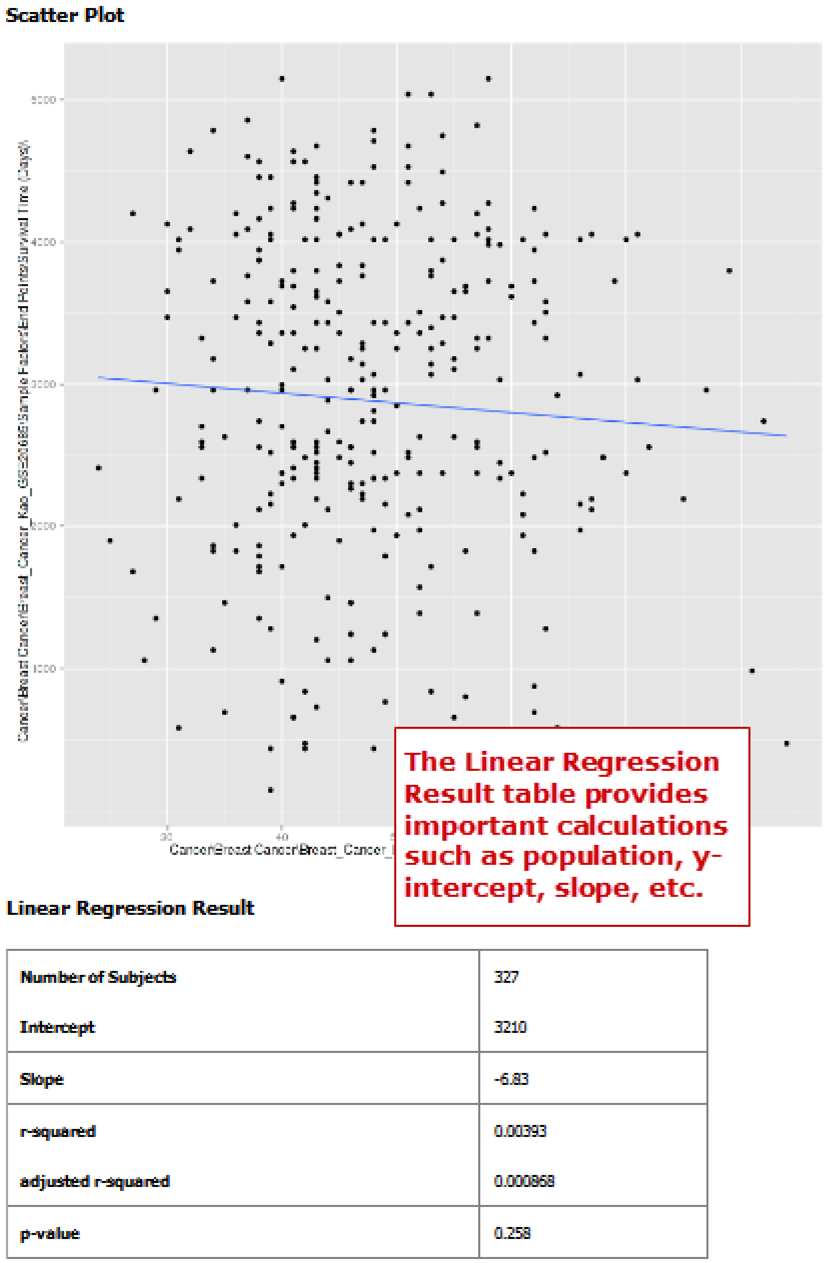


   The Variable Selection section appears. You will need to define what variables in the study are independent, and what variables are dependent. Both variables should be continuous (for example, Age).

5. Define the variables.

6. Click **Run**.

   Your analysis appears below:



Scatter Plot

5000
4000
3000
2000
1000
CancerBreast CancerBreast_Cancer_Kao_GSE20685Sample FactorsEnd PointsSurvival Time (Days)
30
40
CancerBreast CancerBreast_Cancer_

The Linear Regression Result table provides important calculations such as population, y-intercept, slope, etc.

Linear Regression Result

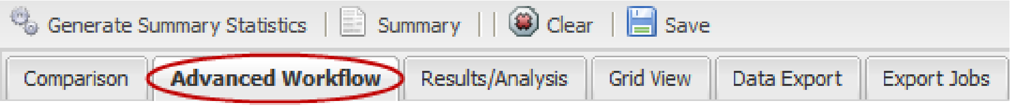
Number of Subjects Intercept	327 3210
Slope	-6.83
r-squared adjusted r-squared	0.00393 0.000868
p-value	0.258

Survival Analysis

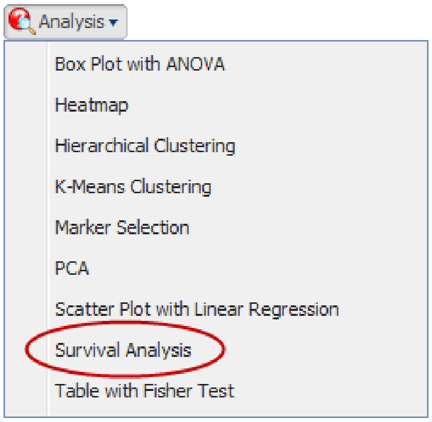
A survival analysis displays time-to-event data.

To perform a survival analysis:

1. Run tranSMART, then click the **Dataset Explorer** tab.
2. Define the cohorts you wish to analyze by dragging one or more concepts from a study into empty subset definition boxes. For more information, see Populating the Study Groups on page 31.
3. Click the **Advanced Workflow** tab above Subset 1:

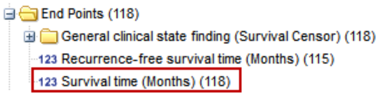


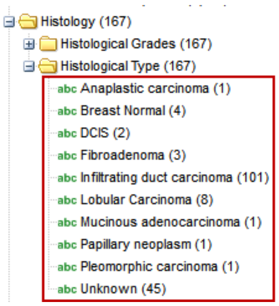
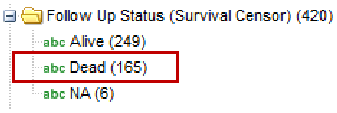
4. Select **Survival Analysis** from the **Analysis** dropdown menu:



The Variable Selection section appears.

5. Define the following variables:

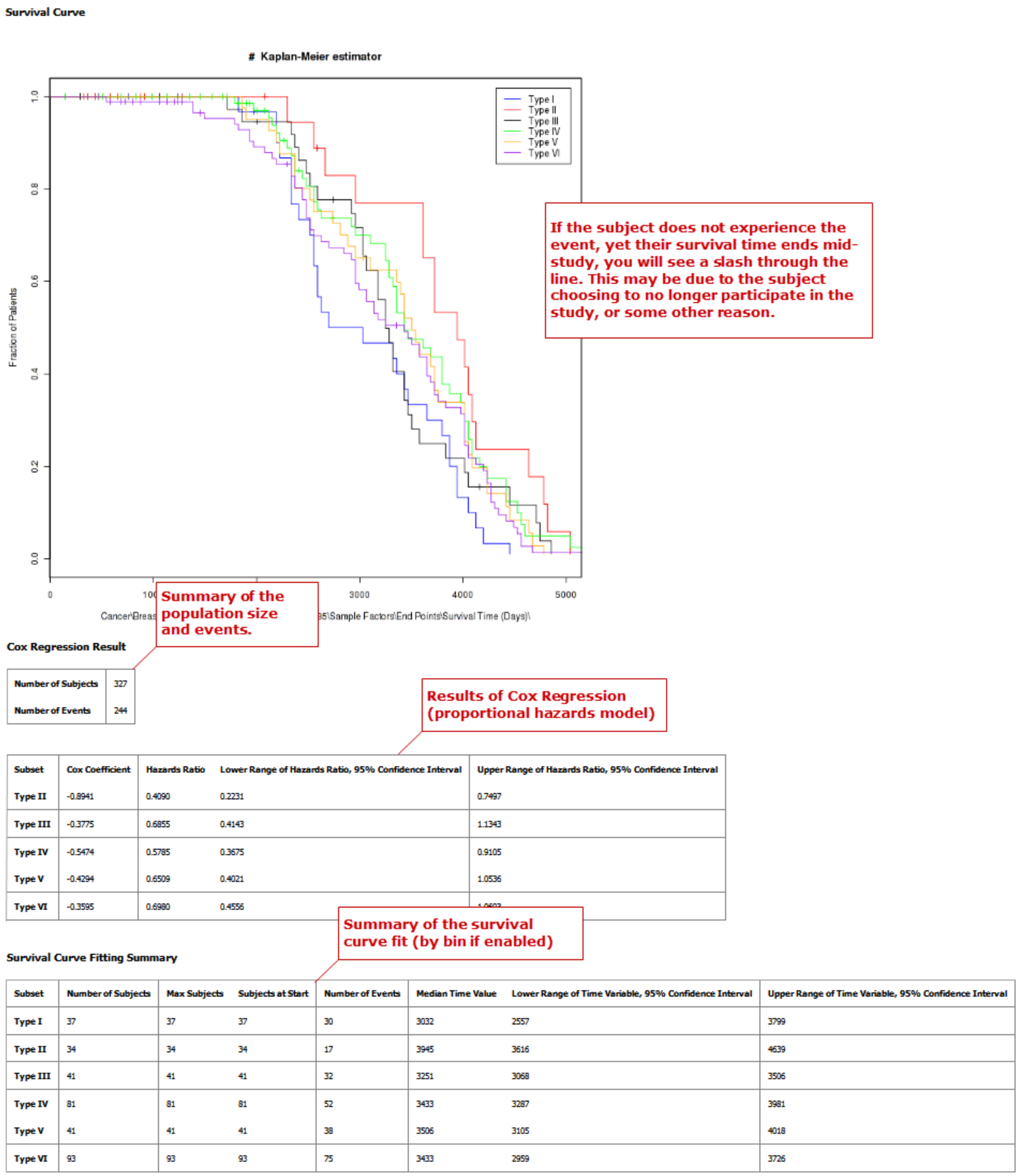
Variable	Required?	Definition	Example
Time	Yes	A numeric field within tranSMART.	Survival at Follow Up (Years)  End Points (118) General clinical state finding (Survival Censor) (118) 123 Recurrence-free survival time (Months) (115) 123 Survival time (Months) (118)

Variable	Required?	Definition	Example
Category	No	A concept that is dragged into this input will dictate the groups into which the data will be split in order to compare their survival times. If this variable is continuous, it requires binning. For details, see [Data Binning Using Survival Analysis]() on page 68.	Cancer Stage  Histology (167) Histological Grades (167) Histological Type (167) abc Anaplastic carcinoma (1) abc Breast Normal (4) abc DCIS (2) abc Fibroadenoma (3) abc Infiltrating duct carcinoma (101) abc Lobular Carcinoma (8) abc Mucinous adenocarcinoma (1) abc Papillary neoplasm (1) abc Pleomorphic carcinoma (1) abc Unknown (45)
Censoring Value	No	Specifies which patients had the event whose time is being measured. For example, if the Time variable selected is **Overall Survival Time (Years)**, an appropriate censoring variable is **Patient Death**.	Dead  Follow Up Status (Survival Censor) (420) abc Alive (249) abc Dead (165) abc NA (6)

In this example, the data binning feature is not used. For future reference, data binning refers to a pre-processing technique used to reduce minor observation errors. Clusters of data are replaced by a value representative of that cluster (the central value). For information on data binning, see [Data Binning Using Survival Analysis]() on page 68.

6. Click **Run**.

   Your analysis appears below:



If the subject does not experience the event, yet their survival time ends mid-study, you will see a slash through the line. This may be due to the subject choosing to no longer participate in the study, or some other reason.

Summary of the population size and events.

Cox Regression Result

Number of Subjects	327
Number of Events	244

Results of Cox Regression (proportional hazards model)

Subset	Cox Coefficient	Hazards Ratio	Lower Range of Hazards Ratio, 95% Confidence Interval	Upper Range of Hazards Ratio, 95% Confidence Interval
Type II	-0.8941	0.4090	0.2231	0.7497
Type III	-0.3775	0.6855	0.4143	1.1343
Type IV	-0.5474	0.5785	0.3675	0.9105
Type V	-0.4294	0.6509	0.4021	1.0536
Type VI	-0.3595	0.6980	0.4556	1.0603

Summary of the survival curve fit (by bin if enabled)

Survival Curve Fitting Summary

Subset	Number of Subjects	Max Subjects	Subjects at Start	Number of Events	Median Time Value	Lower Range of Time Variable, 95% Confidence Interval	Upper Range of Time Variable, 95% Confidence Interval
Type I	37	37	37	30	3032	2557	3799
Type II	34	34	34	17	3945	3616	4639
Type III	41	41	41	32	3251	3068	3506
Type IV	81	81	81	52	3433	3287	3981
Type V	41	41	41	38	3506	3105	4018
Type VI	93	93	93	75	3433	2959	3726

Table with Fisher Test Analysis

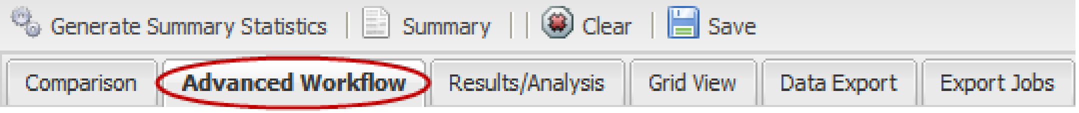
A Fisher Test analysis examines the significance of associated variables.

To perform a table with fisher test analysis:

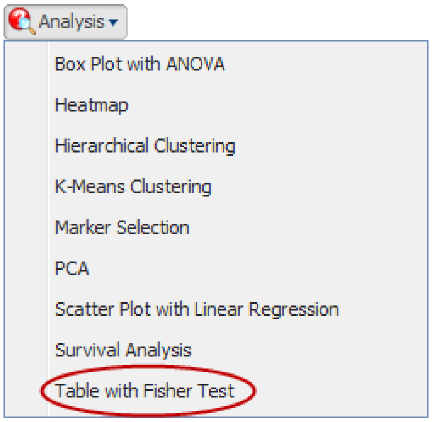
1. Run tranSMART, then click the **Dataset Explorer** tab.
2. Define the cohorts you wish to analyze by dragging one or more concepts from a study into empty subset definition boxes. For more information, see Populating the Study Groups on page 31.

   Only one subset may be specified in this analysis. Information in Subset 2 will be ignored.

3. Click the **Advanced Workflow** tab:



4. Select **Table with Fisher Test** from the **Analysis** dropdown menu:



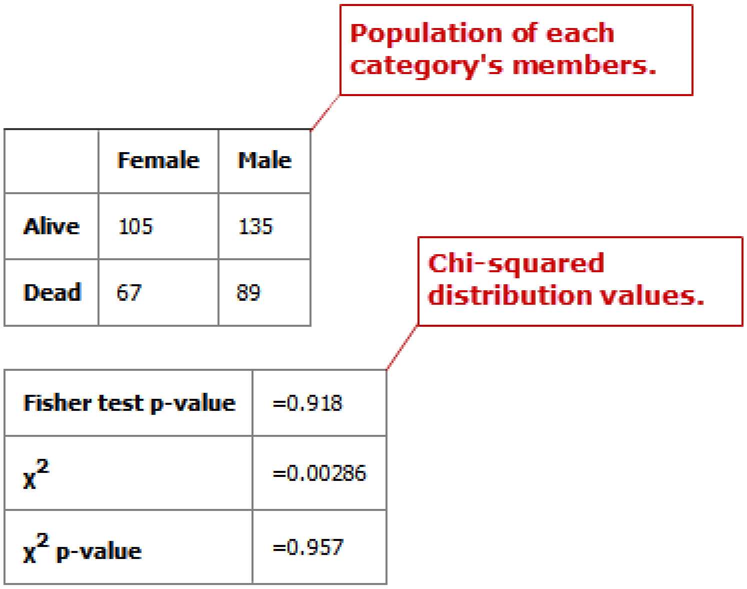
   The Variable Selection section appears. You will need to define what variables in the study are independent, and what variables are dependent. Both variables should be categorical values (for example, Tissue Type).

   In this example, the data binning feature is not used. For future reference, data binning refers to a pre-processing technique used to reduce minor observation errors. Clusters of data are replaced by a value representative of that cluster (the central value). For information on data binning, see Data Binning on page 66.

5. Define the variables.

6. Click **Run**.

   Your analysis appear below:



Population of each category's members.

	Female	Male
Alive	105	135
Dead	67	89

Chi-squared distribution values.

Fisher test p-value	=0.918
x^2	=0.00286
x^2 p-value	=0.957

Data Binning

Data binning refers to a pre-processing technique used to reduce observation errors and to allow continuous variables to become categorical. Clusters of data are replaced by a value representative of that cluster (the central value).

The data displayed after binning represents the data available in the study. If, for example, you have selected to bin based on date range (0-10 years of age), yet there is only data available for subjects eight years old and up, the bin will display the age range as 8-10.

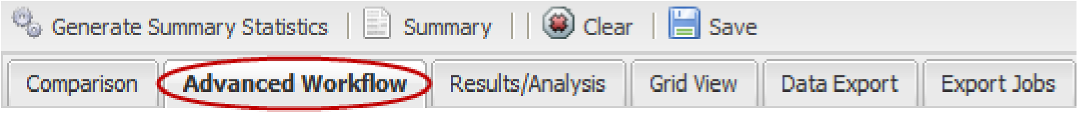
Data Binning Using Box Plot with ANOVA

When conducting a Box Plot with ANOVA analysis, at least one of the variables selected should be a continuous variable (for example, age), and the other should be a categorical value (for example, tumor stage).

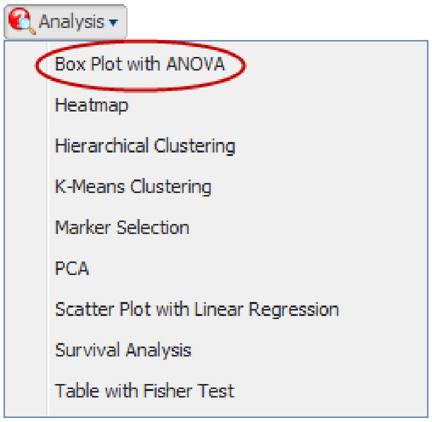
A continuous variable can be viewed as a categorical value using the binning feature, described below. Alternatively, binning can also be used to categorize data. For example, if histological grade with values such as Well Defined, Moderately Well Defined, and Poorly Defined are selected, you can group Moderately Well Defined with Poorly Defined and treat them as one group for the purposes of this analysis.

To use the data binning feature with a box plot analysis:

1. Run tranSMART, then click the **Dataset Explorer** tab.
2. Define the cohorts you wish to analyze by dragging one or more concepts from a study into empty subset definition boxes. For more information, see Populating the Study Groups on page 31.
3. Click the **Advanced Workflow** tab:



4. Select **Box Plot with ANOVA** from the **Analysis** dropdown menu:

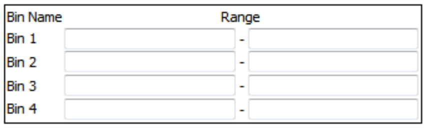
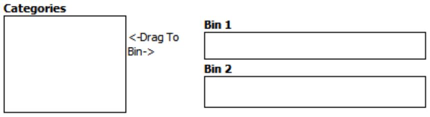


The Variable Selection section appears. You will need to define what variables in the study are independent, and what variables are dependent. At least one of the variables should be continuous (for example, Age), and one should be a categorical value (for example, Tissue Type).

5. Define the variables.
6. Under **Binning**, click **Enable**:

7. Define the following:

Field	Description	Comments
Variable	Select which variable should define the groups (Independent or Dependent) from the dropdown menu.	If the *independent variable* defines the groups, boxes will be plotted horizontally. If the *dependent variable* defines the groups, boxes will be plotted vertically

Field	Description	Comments
Variable Type	Select whether the variable you have defined above is continuous or categorical from the dropdown menu.	A continuous variable can be turned into a categorical variable when you use the binning feature.
Number of Bins	Type the number of bins you would like data to be organized in.	This step may require trial and error based on how you wish to display data.
Bin Assignments	Select how you would like data to be binned from the dropdown menu. **Note:** This feature can only be used when the variable type selected above is continuous.	▪ **Evenly Distribute Population:** assigns bins based on the underlying data. For example, if the majority of the subjects in the study were elderly, bins based on age could look like: [(1-40), (40-80), (81-85), (86-90), (90-92)]. ▪ **Evenly Spaced Bins:** creates bins based on the overall range of the variable. For example, if the majority of the subjects in the study were elderly, bins based on age could look like: [(1-20), (21-40), (41-60), (61-80), (81-100)].
Manual Binning	Select the checkbox if you wish to bin manually. **Note:** This is the only binning method available if you are attempting to bin a categorical variable type.	Complete the binning form that populates as a result of checking the **Manual Binning** box. For continuous data:  Bin Name Range Bin 1 - Bin 2 - Bin 3 - Bin 4 - For categorical data:  Categories <-Drag To Bin-> Bin 1 Bin 2

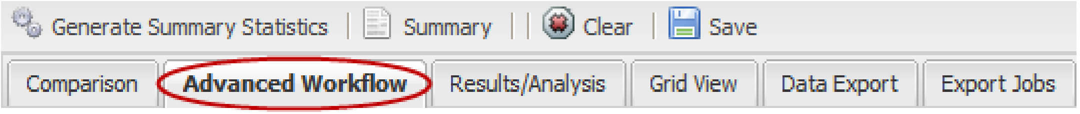
8. Click **Run**.

Data Binning Using Survival Analysis

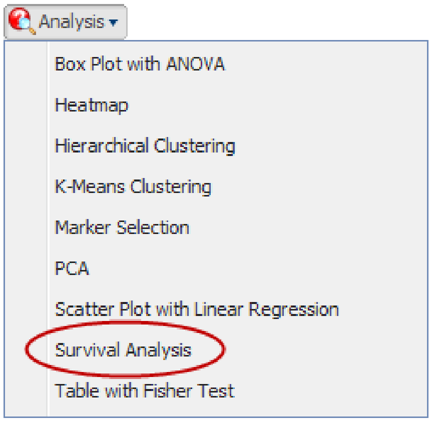
Data binning is used in survival analyses if the variable you wish to use is continuous (for example, age), but needs to be viewed as categorical data. Alternatively, it can be used to regroup categorical data. For example, if histological grade with values such as *Well Defined*, *Moderately Well Defined*, and *Poorly Defined* are selected, you can group *Moderately Well Defined* with *Poorly Defined* and treat them as one group for the purposes of this analysis.

To use the data binning feature with a survival analysis:

1. Run tranSMART, then click the **Dataset Explorer** tab.
2. Define the cohorts you wish to analyze by dragging one or more concepts from a study into empty subset definition boxes. For more information, see Populating the Study Groups on page 31.
3. Click the **Advanced Workflow** tab:



4. Select **Survival Analysis** from the **Analysis** dropdown menu:

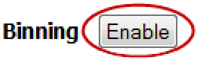


   The Variable Selection section appears.

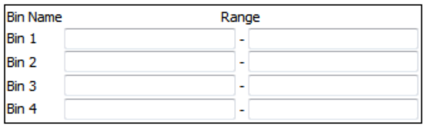
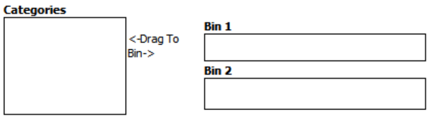
5. Define the variables:

Variable	Required?	Description	Example
Time	Yes	A time variable used in the study.	Survival Time
Category	No	A variable that you wish to use to sort the cohorts. If the variable you wish to use is continuous (for example, age), the binning feature should be used.	Cancer Stage
Censoring Variable	No	A censoring variable (occurs when the value of a measurement/observation is partially known).	Survival (Censor)

6. Under **Binning**, click **Enable**:



7. Define the following:

Field	Description	Comments
Variable Type	Select whether the variable you have defined above is continuous or categorical.	A continuous variable can be treated as a categorical variable when you use the binning feature.
Number of Bins	Type the number of bins you would like data to be organized in.	This step may require trial and error based on how you wish to display data.
Bin Assignments	Select how you would like data to be binned. **Note:** This feature can only be used when the variable type selected above is continuous.	■ Evenly Distribute Population: assigns bins based on the underlying data. For example, if the majority of the subjects in the study were elderly, bins based on age could look like: [(1-40), (40-80), (81-85), (86-90), (90-92)]. ■ Evenly Spaced Bins: creates bins based on the overall range of the variable. For example, if the majority of the subjects in the study were elderly, bins based on age could look like: [(1-20), (21-40), (41-60), (61-80), (81-100)].
Manual Binning	Select the checkbox if you wish to bin manually. **Note:** This is the only binning method available if you are attempting to bin a categorical variable type.	Complete the binning form that populates as a result of checking the **Manual Binning** box. ■ For continuous data:  ■ For categorical data: 

8. Click **Run**.

High Dimensional Data

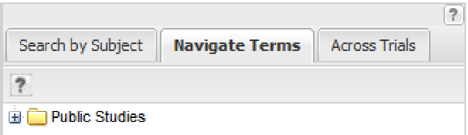
The High Dimensional Data button available within the Advanced Workflow section of Dataset Explorer allows you to specify additional inputs for selected variables. These inputs help filter specific information of value (such as platforms, samples, or timepoints).

> The High Dimensional Data feature must be used when you perform an analysis using high dimensional data (such as SNP, gene expression, RBM, etc.) symbolized by the DNA icon (🧬). Additionally, the High Dimensional Data feature cannot be used without high dimensional data.

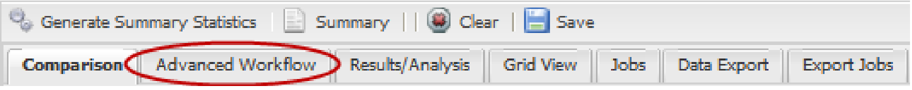
To use the High Dimensional Data feature:

1. Click the tranSMART **Dataset Explorer** tab to display the Dataset Explorer window.
2. In the left pane of the Dataset Explorer window, click the **Navigate Terms** tab.

   The navigation tree appears, showing the categories of available studies:



3. Select the study of interest.
4. Drag the study of interest into a subset definition box in Subset 1.
5. Click the **Advanced Workflow** tab above Subset 1:



6. Select the analysis you wish to perform, and define at least one variable using high dimensional data.
7. Click the **High Dimensional Data** button.

   The Compare Subsets-Pathway Selection dialog appears.
8. Define the available filters listed in the table below.

> Dataset Explorer will attempt to pre-populate default values in the associated fields of the dialog based on the underlying data in the variable selection box.

Filter	Description
Platform	The platform type (for example, SNP, mRNA, etc.) used to collect biomarker data in the study.
GPL Platform	The specific name of the platform used in the study.
Sample	The type of sample tested in the study.
Tissue Type	The type of tissue tested in the study.
Timepoint	The time period when the sample was taken.
Select a Gene/Pathway	The gene, gene signature, or pathway of interest. If you would like to run the analysis on the entire chip, leave this field blank.
Select SNP Type	Select the type of SNP data being used: ■ Genotype – Use for categorical variables. ■ Copy Number – Use for continuous variables. **Note:** This option is only available when you drag SNP data into the variable selection box. **Note:** Both Genotype and Copy Number data may not be available for all studies involving SNP data.
Aggregate Probes?	The checkbox can be selected if the variable chosen is either gene expression data or SNP copy number data. If the checkbox is selected, the algorithm WGCNA (weighted correlation network analysis) is employed. For genes that are comprised of multiple probes, WGCNA selects the probe that best represents the overall expression level or copy number. **Note:** WGCNA was developed by the Department of Human Genetics at UCLA. For more information, see http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/.

9. Click **Apply Selections**.

10. Define any additional required variables, then click **Run**.

Other Features

The sections below illustrate additional features in the Advanced Workflow tab.

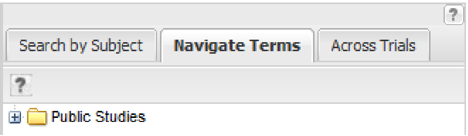
Save to PDF

The Save to PDF feature allows you to save analyses run through the Advanced Workflow function within Dataset Explorer.

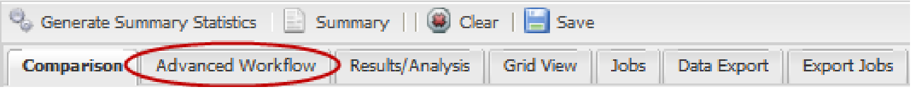
To save advanced workflow analyses as a PDF file:

1. Click the tranSMART **Dataset Explorer** tab to display the Dataset Explorer window.
2. In the left pane of the Dataset Explorer window, click the **Navigate Terms** tab.

   The navigation tree appears, showing the categories of available studies:



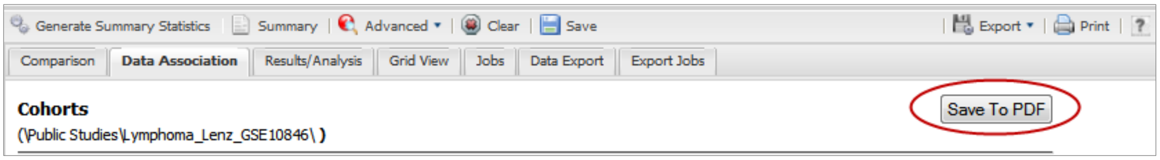
3. Select the study of interest.
4. Drag the study of interest into a subset definition box in Subset 1.
5. Click the **Advanced Workflow** tab above Subset 1:



6. Select the analysis you wish to perform, and define the variables accordingly.
7. Click **Run**.

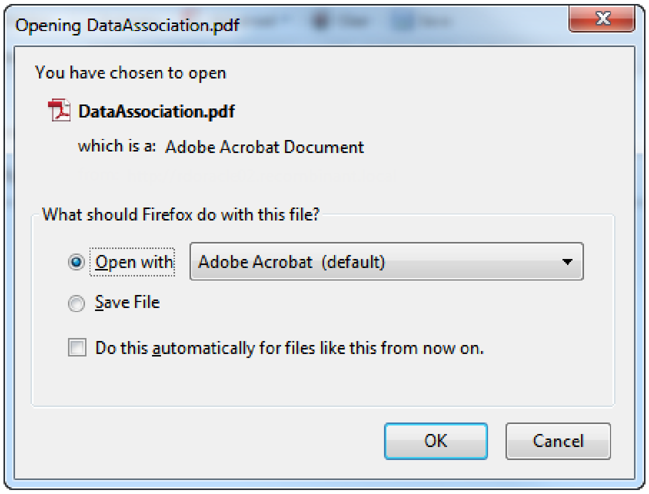
   Your analysis appears below the variable selection panes.

8. Click **Save to PDF**:



Cohorts
(\Public Studies\Lymphoma_Lenz_GSE10846\ )

The following dialog appears:



Opening DataAssociation.pdf

You have chosen to open

DataAssociation.pdf

which is a: Adobe Acrobat Document

What should Firefox do with this file?

● Open with Adobe Acrobat (default)

○ Save File

□ Do this automatically for files like this from now on.

9. Select **Open with** or **Save File**, then click **OK**.

Download Raw R Data

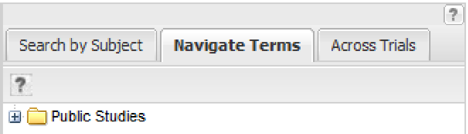
Analyses run through the Advanced Workflow tool within Dataset Explorer use R for computation. You are able to download raw R data for use in an external tool.

For more information on The R Project for Statistical Computing, visit the following site: www.r-project.org.

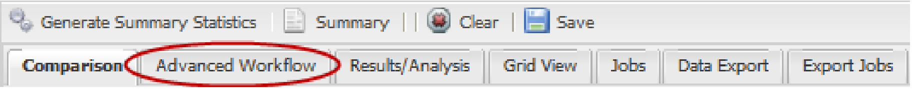
To download advanced workflow analyses as raw R data:

1. Click the tranSMART **Dataset Explorer** tab to display the Dataset Explorer window.
2. In the left pane of the Dataset Explorer window, click the **Navigate Terms** tab.

   The navigation tree appears, showing the categories of available studies:



3. Select the study of interest.
4. Drag the study of interest into a subset definition box in Subset 1.
5. Click the **Advanced Workflow** tab above Subset 1:



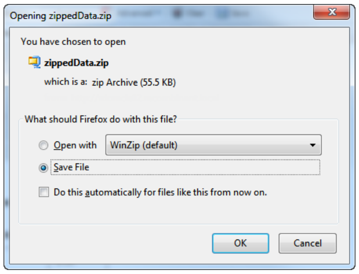
6. Select the analysis you wish to perform, and define the variables accordingly.

7. Click **Run**.

   Your analysis appears below the variable selection panes.

8. Click **Download raw R data**:

   The following dialog appears:

   Opening zippedData.zip

   You have chosen to open

   zippedData.zip

   which is a: zip Archive (55.5 KB)

   What should Firefox do with this file?

   ○ Open with WinZip (default)

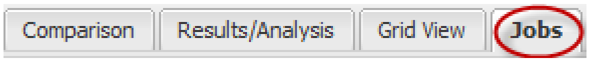
   ● Save File

   □ Do this automatically for files like this from now on.

9. Select whether you would like to open the file or save it to your hard drive, then click **OK**.

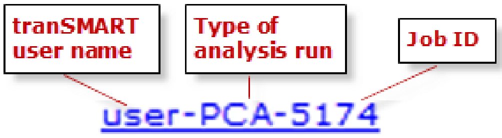
The Jobs Tab

The Jobs tab allows you to review analyses you have run previously.



Each advanced workflow analysis 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:  tranSMART user name Type of analysis run Job ID user-PCA-5174

Column	Description
Status	The status of the analysis. Statuses are explained below: ▪ **Completed** – The job has finished and a visualization or analysis 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 started.

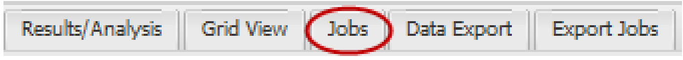
Click the **Refresh** button to view any changes that have been made since the Jobs tab initially populated.

Viewing a Logged Job

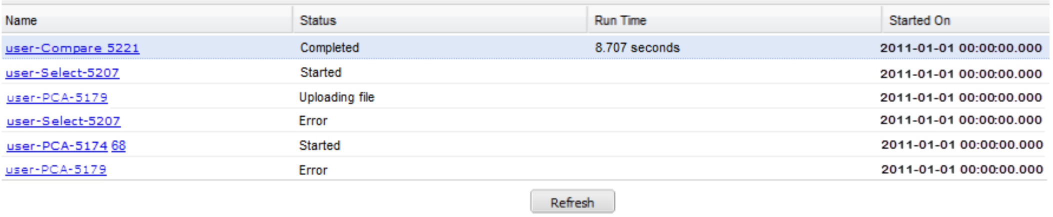
Each advanced analysis that you have run in the previous seven days will be logged in the **Jobs** tab. You may view the visualization or analysis again by selecting it from the list.

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:



Name	Status	Run Time	Started On
user-Compare 5221	Completed	8.707 seconds	2011-01-01 00:00:00.000
user-Select-5207	Started		2011-01-01 00:00:00.000
user-PCA-5179	Uploading file		2011-01-01 00:00:00.000
user-Select-5207	Error		2011-01-01 00:00:00.000
user-PCA-5174 68	Started		2011-01-01 00:00:00.000
user-PCA-5179	Error		2011-01-01 00:00:00.000

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 sample IDs and locate the study that produced the samples in the Dataset Explorer.

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:

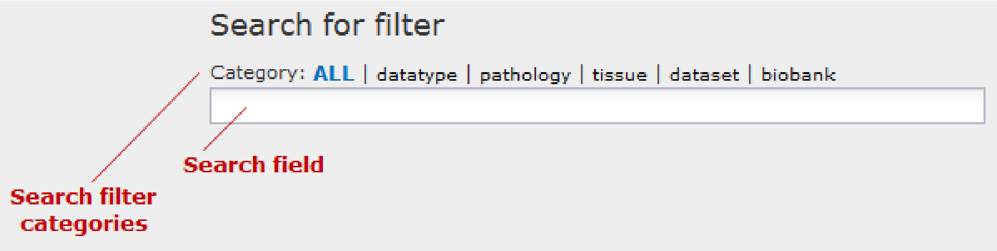
- **Subject Treatment** – The treatment administered to study subjects
- **Data Type** – The type of data associated with the samples
- **Pathology** – The type of disease associated with the samples
- **Tissue** – The physical source of the samples, such as liver or colon tissue
- **Dataset** – The study that generated the samples
- **BioBank** – The sample originated in the BioBank
- **Sample Treatment** – The treatment administered to the sample
- **Source Organism** – The organism from which the sample was taken

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.

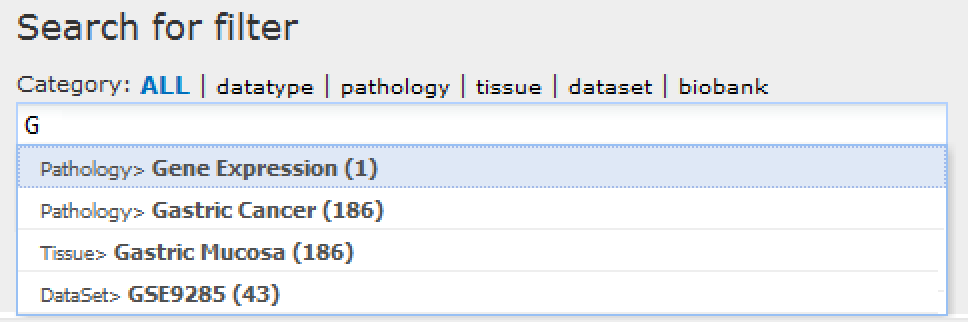
To search for a primary search filter:

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


Search for filter
Category: ALL | datatype | pathology | tissue | dataset | biobank
Search field
Search filter 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:


Search for filter
Category: ALL | datatype | pathology | tissue | dataset | biobank

G

Pathology> Gene Expression (1)
Pathology> Gastric Cancer (186)
Tissue> Gastric Mucosa (186)
DataSet> GSE9285 (43)

   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.

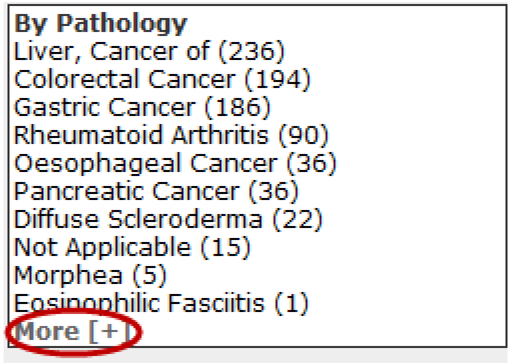
   The search begins immediately, and the results are displayed in a new window (see View and Refine Sample Search Results on page 80).

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:



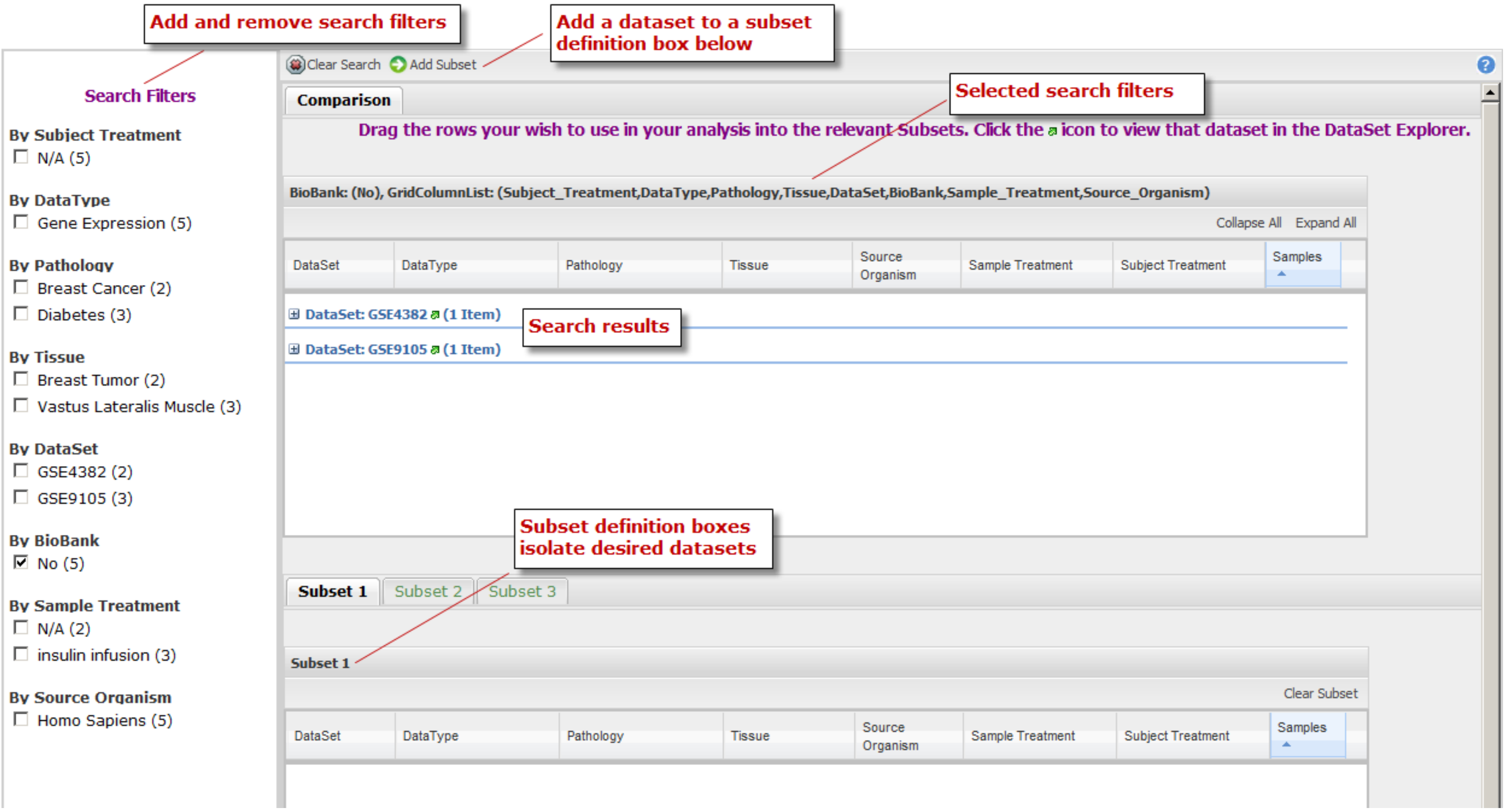
By Pathology
Liver, Cancer of (236)
Colorectal Cancer (194)
Gastric Cancer (186)
Rheumatoid Arthritis (90)
Oesophageal Cancer (36)
Pancreatic Cancer (36)
Diffuse Scleroderma (22)
Not Applicable (15)
Morphea (5)
Eosinophilic Fasciitis (1)
More [+]

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 80).

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:



Add and remove search filters

Add a dataset to a subset definition box below

Selected search filters

Search results

Subset definition boxes isolate desired datasets

Search Filters

By Subject Treatment
☐ N/A (5)

By DataType
☐ Gene Expression (5)

By Pathology
☐ Breast Cancer (2)
☐ Diabetes (3)

By Tissue
☐ Breast Tumor (2)
☐ Vastus Lateralis Muscle (3)

By DataSet
☐ GSE4382 (2)
☐ GSE9105 (3)

By BioBank
☑ No (5)

By Sample Treatment
☐ N/A (2)
☐ insulin infusion (3)

By Source Organism
☐ Homo Sapiens (5)

Drag the rows your wish to use in your analysis into the relevant Subsets. Click the icon to view that dataset in the DataSet Explorer.

BioBank: (No), GridColumnList: (Subject_Treatment,DataType,Pathology,Tissue,DataSet,BioBank,Sample_Treatment,Source_Organism)

DataSet: GSE4382 (1 Item)

DataSet: GSE9105 (1 Item)

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

- Select and remove search filters
- Locate the study that produced the samples in the Dataset Explorer
- 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.

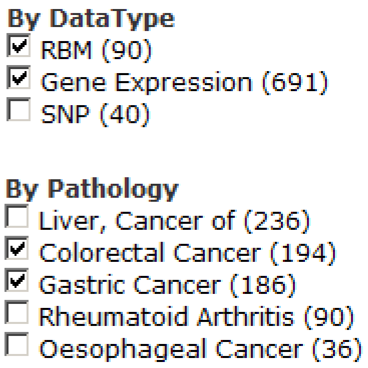
Clicking a filter name rather than the check box next to the name will select that filter and deselect all currently selected filters.

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) 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)
```



By DataType
☑ RBM (90)
☑ Gene Expression (691)
□ SNP (40)

By Pathology
□ Liver, Cancer of (236)
☑ Colorectal Cancer (194)
☑ Gastric Cancer (186)
□ Rheumatoid Arthritis (90)
□ Oesophageal Cancer (36)

Locate the Source of the Samples in Dataset Explorer

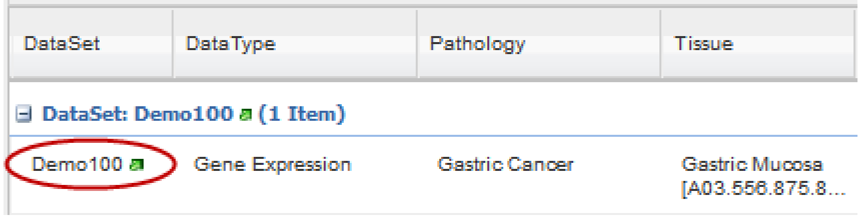
If a dataset of samples was collected for a Dataset Explorer study, you can link back to the study to view information such as the study owner, study description and purpose, demographics of the participants, and other data relevant to the samples.

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 81).

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



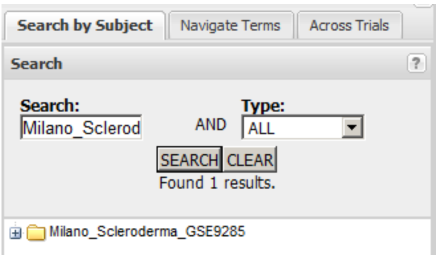
DataSet	Data Type	Pathology	Tissue
DataSet: Demo100 (1 Item)			
Demo100	Gene Expression	Gastric Cancer	Gastric Mucosa [A03.556.875.8...

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:



Search: Milano_Sclerod AND Type: ALL

Found 1 results.

Milano_Scleroderma_GSE9285

3. Open and explore the study of interest.

For information, see Branches and Leaves of the Navigation Tree on page 30.

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.

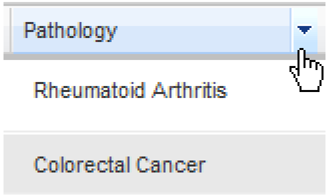
Manage the Sample Search Result List

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:

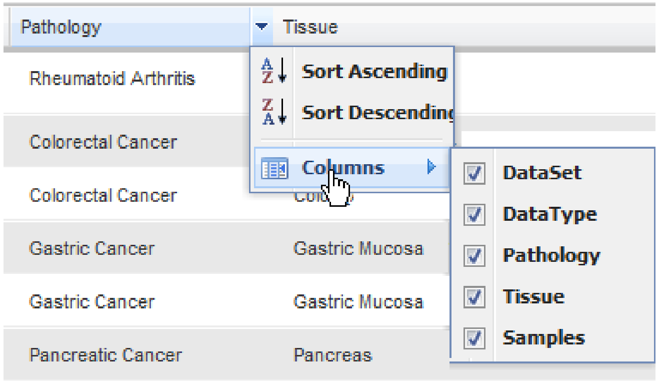


2. Click **Sort Ascending** or **Sort Descending**.

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.

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.

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 deregulated, and that can help you develop a further set of experiments.

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.

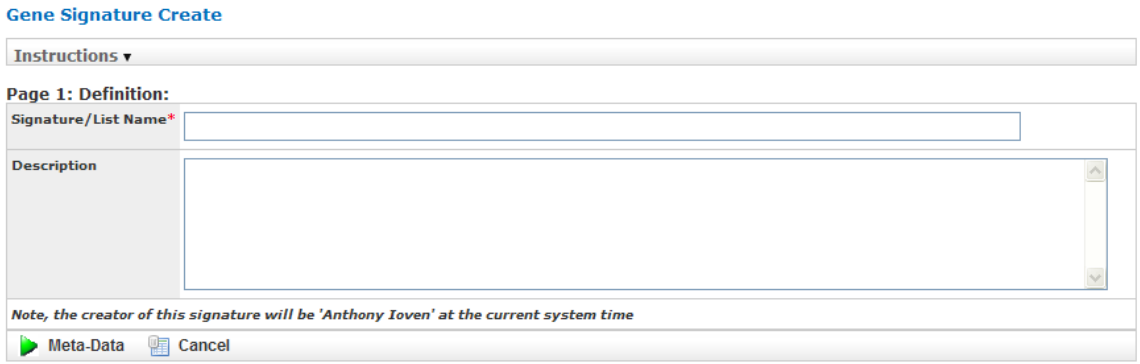
The following table shows the different ways you can specify the genes for your gene signature:

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

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:



Gene Signature Create

Instructions ▾

Page 1: Definition:

Signature/List Name*

Description

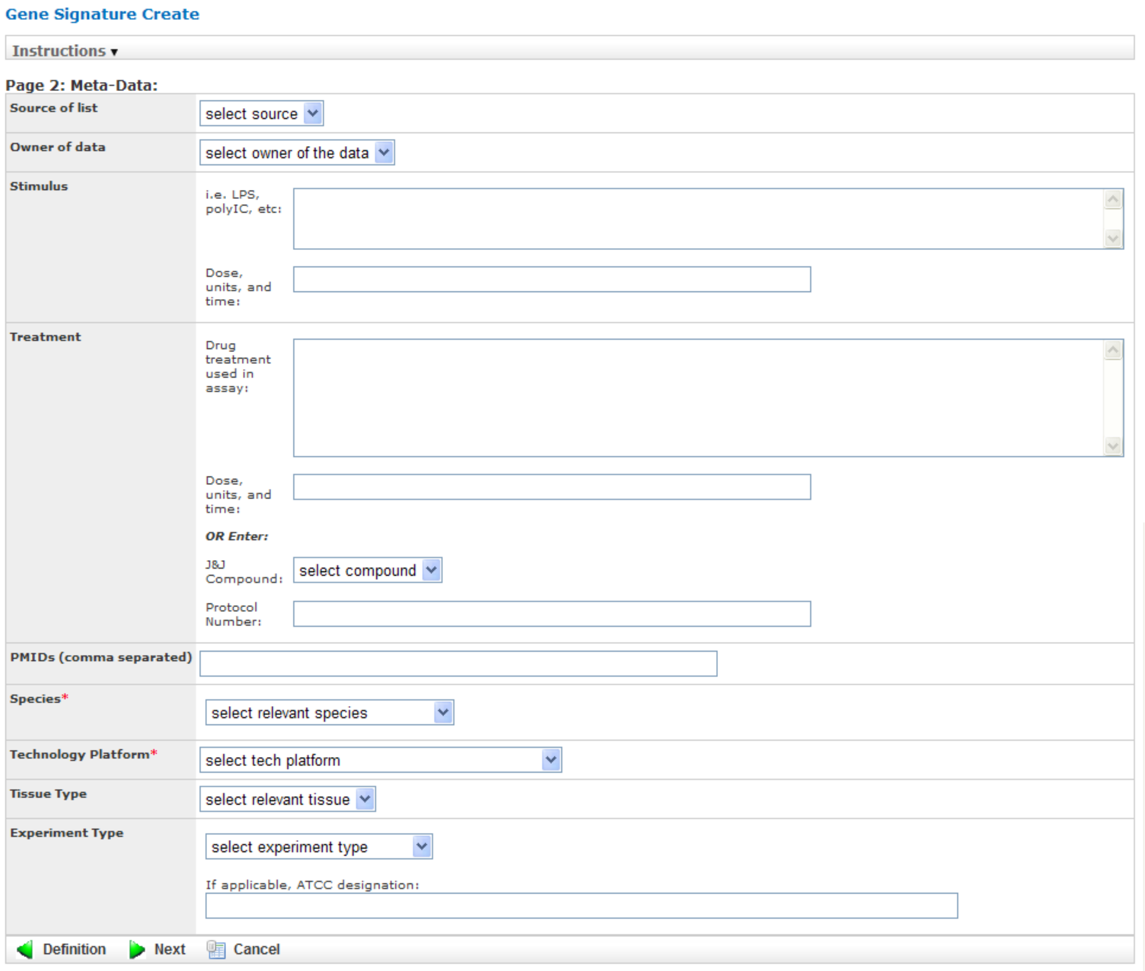
Note, the creator of this signature will be 'Anthony Ioven' at the current system time

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:



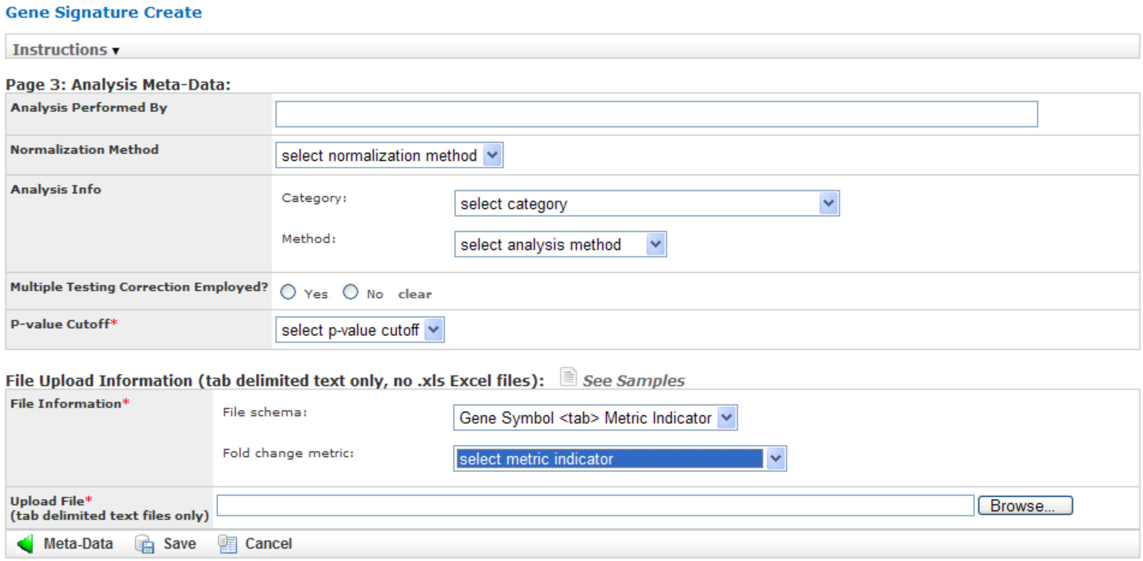
Gene Signature Create

Instructions ▾

Page 2: Meta-Data:

Source of list	select source	
Owner of data	select owner of the data	
Stimulus	i.e. LPS, polyIC, etc:	
	Dose, units, and time:	
Treatment	Drug treatment used in assay:	
	Dose, units, and time:	
	OR Enter:	
	J&J Compound:	select compound
	Protocol Number:	
PMIDs (comma separated)		
Species*	select relevant species	
Technology Platform*	select tech platform	
Tissue Type	select relevant tissue	
Experiment Type	select experiment type	
	If applicable, ATCC designation:	

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:



Gene Signature Create

Instructions ▾

Page 3: Analysis Meta-Data:

Analysis Performed By	
Normalization Method	select normalization method
Analysis Info	Category: select category Method: select analysis method
Multiple Testing Correction Employed?	○ Yes ○ No clear
P-value Cutoff*	select p-value cutoff

File Upload Information (tab delimited text only, no .xls Excel files): *See Samples*

File Information*	File schema: Gene Symbol <tab> Metric Indicator Fold change metric: select metric indicator
Upload File* (tab delimited text files only)	Browse...

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 85, 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.

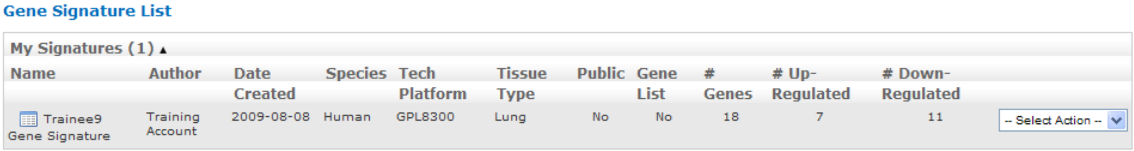
   - In the **Fold change metric** section of **File Information**, select one of the following choices from the dropdown:

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:



Gene Signature List

My Signatures (1) ▴

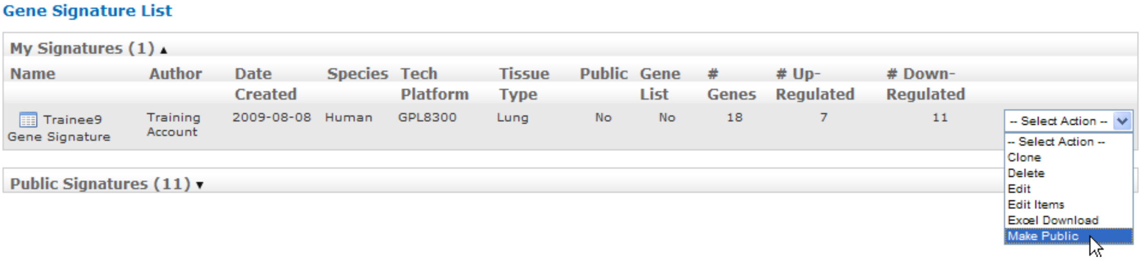
Name	Author	Date Created	Species	Tech Platform	Tissue Type	Public	Gene List	# Genes	# Up-Regulated	# Down-Regulated	
Trainee9 Gene Signature	Training Account	2009-08-08	Human	GPL8300	Lung	No	No	18	7	11	-- Select Action --

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:



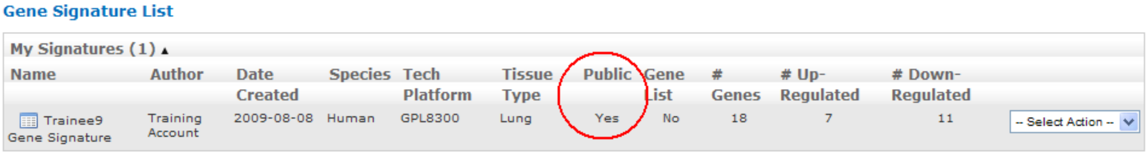
Gene Signature List

My Signatures (1) ▴

Name	Author	Date Created	Species	Tech Platform	Tissue Type	Public	Gene List	# Genes	# Up-Regulated	# Down-Regulated	
Trainee9 Gene Signature	Training Account	2009-08-08	Human	GPL8300	Lung	No	No	18	7	11	-- Select Action --

Public Signatures (11) ▾

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



Gene Signature List

My Signatures (1) ▴

Name	Author	Date Created	Species	Tech Platform	Tissue Type	Public	Gene List	# Genes	# Up-Regulated	# Down-Regulated	
Trainee9 Gene Signature	Training Account	2009-08-08	Human	GPL8300	Lung	Yes	No	18	7	11	-- Select Action --

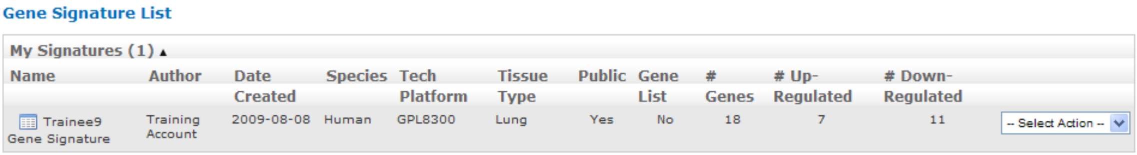
tranSMART users assigned the role `ROLE_ADMIN` have access to both public and private gene signatures.

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:



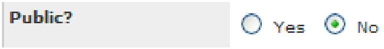
Gene Signature List

My Signatures (1) ▴

Name	Author	Date Created	Species	Tech Platform	Tissue Type	Public	Gene List	# Genes	# Up-Regulated	# Down-Regulated	
Trainee9 Gene Signature	Training Account	2009-08-08	Human	GPL8300	Lung	Yes	No	18	7	11	-- Select Action --

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: Upload New File Only to Override Existing Items ▾ 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:  Public? ○ Yes ● No

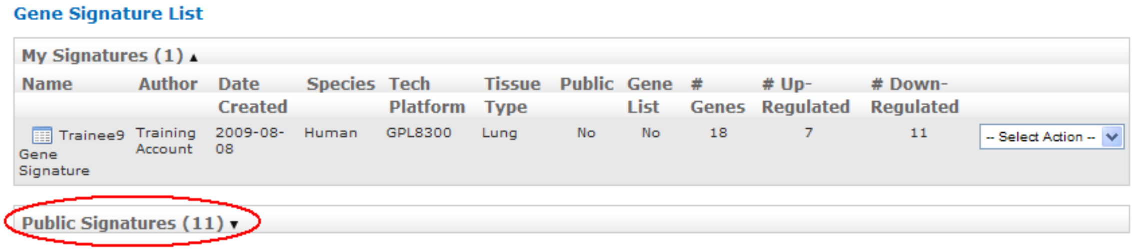
Performing Actions on Other Users' Signatures

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:



Gene Signature List

My Signatures (1) ▴

Name	Author	Date Created	Species	Tech Platform	Tissue Type	Public	Gene List	# Genes	# Up-Regulated	# Down-Regulated	
Trainee9 Gene Signature	Training Account	2009-08-08	Human	GPL8300	Lung	No	No	18	7	11	-- Select Action --

Public Signatures (11) ▾

tranSMART users assigned the role ROLE_ADMIN will see **Other Signatures** instead of **Public Signatures**.

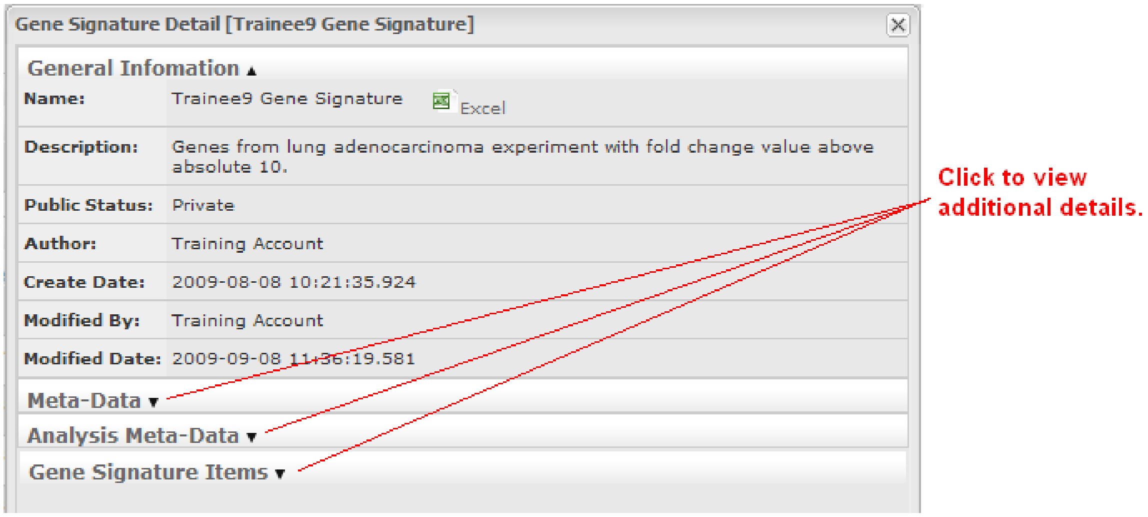
3. Click the **Select Action** dropdown for the gene signature you want to act on.
4. Select the action you want to perform on the gene signature.

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:



Gene Signature Detail [Trainee9 Gene Signature]

General Infomation ▴

Name:	Trainee9 Gene Signature Excel
Description:	Genes from lung adenocarcinoma experiment with fold change value above absolute 10.
Public Status:	Private
Author:	Training Account
Create Date:	2009-08-08 10:21:35.924
Modified By:	Training Account
Modified Date:	2009-09-08 11:36:19.581

Meta-Data ▾

Analysis Meta-Data ▾

Gene Signature Items ▾

Click to view additional details.

Chapter 6

Analysis Data Upload

The **Upload Data** tab at the top of the tranSMART window lets you upload analysis data for a study. It also lets you define information about the analysis (the analysis metadata), such as analysis name, description, data type, sample size, tissue, phenotype, and so on.

Data that you upload through the two-page data upload form is fully integrated with the studies and analyses that have already been loaded in the tranSMART data warehouse. The data is immediately searchable via tranSMART Faceted Search.

Analysis data is uploaded from a tab-separated text file. To ensure that this file is in the proper format for uploading, use one of the provided templates. For information, see File Templates on page 97.

Uploading Analysis Data

Before you begin to upload analysis data, you must have a properly formatted file containing the data in a directory on your local computer or on a network server that is accessible from your computer.

To upload analysis data:

1. At the top of the tranSMART window, click the **Upload Data** tab.

   The first Upload Data page appears. On this page, the fields **Study**, **Analysis Type to Upload**, and **Analysis Name** are required.

2. In **Study**, select the name of the study associated with the analysis.

   To do so, do either of the following:

   □ Type part of the study name in the text box. tranSMART lists all study names that contain those characters in a contiguous string anywhere in the name. When you see the name you want, click it. If you do not see the name, type more characters.

   □ Click the **Browse** button to browse the list of study names. Select the one you want and click **Select**.

   The study metadata appears under the study name. If the selected study is not the one you want, click **Change** to select a different name.

If you do not see the study you want, click the **Email administrator** button at the upper right corner of the page to inquire about the study.

3. In **Analysis Type to Upload**, select the data type of the analysis data.

4. In **Analysis Name**, type a name for the analysis.

   This is the name that will appear in the list of analyses for the study.

5. In `Analysis Description`, type a description of the analysis.

   The description should contain enough information to help a researcher who is scanning the analyses find the ones that are of interest.

6. Click **Enter metadata**.

   The second page of the data upload form appears. The page name will reference the data type you selected on the previous page.

7. In **File**, click the **Browse** button to navigate to the tab-delimited text file that contains the analysis data.

   Optionally, click **Download Template** to open or save the template for the data type that you are uploading.

8. Complete the upload data form by providing information for the remaining fields on the form, as described in the table below:

Analysis Metadata	**Data Type**	**Description**
Disease or Phenotype	All	Type one or more characters in the name of a disease or observation that is relevant to the analysis. tranSMART lists the names that contain the characters in a contiguous string anywhere in the name. Click the name of the relevant disease or observation. If you do not see the name you want, type more characters. To add another disease or observation, click **Add new**.
Population	All	Specify the population of individuals to whom the analysis applies.
Sample Size	All	Specify the number of subjects who were included in the study.
Tissue	All	The type of tissue on which testing was performed.
Cell Type	All	The type of cell on which testing was performed.
(Genotype) Platform: Vendor	All	The genotype platform vendor. To add another vendor and platform, click **Add new**.

Analysis Metadata	Data Type	Description
(Genotype) Platform	All	The specific genotype platform involved in the study. Platform names will vary with the vendor you select.
Genome Version	All	The human genome version on which the analysis data is based.
Expression Platform: Vendor	eQTL	The gene expression platform vendor. To add another vendor and platform, click **Add new**.
Expression Platform	eQTL	The specific gene expression platform involved in the study. Platform names will vary with the vendor you select.
Model Name	All	The name of the model that was used to perform the analysis.
Model Description	All	The description of the model.
Statistical Test	All	The statistical test that was used to analyze the data.
P-value cutoff <=	All	The p-value threshold applied to the data being uploaded. The uploaded data contains records with a p-value that is equal to or less than the specified p-value cutoff.
Research Unit	All	The research unit that performed the analysis or to whom the analysis is relevant..

9. Click **Upload** to upload the analysis data and metadata.

File Templates

Analysis data must be contained in a properly formatted, tab-separated text file.

tranSMART includes templates for the supported analysis data types (GWAS, Metabolic GWAS, eQTL). Be sure to use these templates for the analysis data to upload.

The data must contain values in the `rsid` and `p-value` columns.

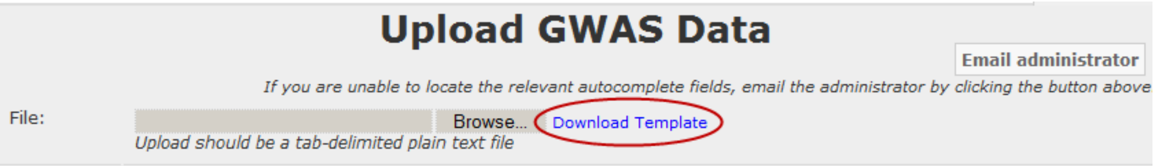
You can download the templates from the data upload form in tranSMART. To do so:

1. At the top of the tranSMART window, click the **Upload Data** tab.

   The first page of the upload data form appears. The templates are available from the second page, so you must provide information on the required fields of this page to proceed to the next.

2. In **Study**, select any study name.
3. In **Analysis Type to Upload**, select the data type for the analysis data that you will upload.
4. In **Analysis Name**, type any name.
5. Click **Enter metadata**.

   The second page of the data upload form appears.
6. Click **Download Template** to the right of the File field:

   Upload GWAS Data

   If you are unable to locate the relevant autocomplete fields, email the administrator by clicking the button above

   File:

   Upload should be a tab-delimited plain text file

   You will be prompted to open or save the template for the data type you specified in step 3.
7. Click **Save**, specify a location for the file to be saved, and click **Save**.
8. Close the Download dialog box.
9. Click the **Cancel** button on the upload data form.

Handling an Upload Failure

If the analysis data does not appear in tranSMART after an upload attempt, and you are sure that all procedures described in this chapter were performed correctly, contact the person who is responsible for loading data into tranSMART. Ask that person to consult the chapter “Loading GWAS and eQTL Analysis Data” in the tranSMART ETL Guide for troubleshooting and manual data loading procedures.

Appendix A

Glossary

AGGREGATE PROBES

Used in Dataset Explorer, the Aggregate Probes checkbox allows you to group probes used in high-dimensional data samples to form a total quantity that analyses will be performed on.

ANALYSIS OF VARIANCE (ANOVA)

Analysis of Variance (ANOVA) is a statistical method used in Dataset Explorer to make concurrent comparisons between two or more means in a box plot.

ANALYSIS VIEW

Used in the Search tool, the Analysis View option displays the statistically significant analyses from your search filter(s).

ANTI-REGULATION

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*.

ARRAY DATA

See: Microarray

ARRAYEXPRESS

Database of gene expression and other microarray data at the European Bioinformatics Institute (EBI).

See http://www.ebi.ac.uk/arrayexpress for details.

BINOMIAL DISTRIBUTION

Graph that displays the discrete probability distribution of obtaining *n* successes out of N Bernoulli trials.

See http://mathworld.wolfram.com/BinomialDistribution.html for details.

BIOMARKER

Short for Biological Marker, a biomarker is a key molecular or cellular event that links a specific environmental exposure to a health outcome.

BOX PLOT

Also known as a Box and Whisker Plot, a box plot is a histogram-like method of displaying data. Box plots are useful when conveying location and variation information in datasets.

CATEGORICAL VARIABLE

Also known as a nominal value, a categorical variable is one that has two or more categories, but with no intrinsic ordering to the categories. An example of a categorical value is hair color – there is no way to order these variables from highest to lowest.

CENSORING VALUE

Used in Survival Analyses. The Censoring Value specifies which patients had the event whose time is being measured. For example, if the Time variable selected is Overall Survival Time (Years), an appropriate censoring variable is Patient Death.

CHI SQUARED

Let the probabilities of various classes in a distribution be p_1, p_2, ..., p_k, with observed frequencies m_1, m_2, ..., m_k. The quantity

$$\chi_s^2 = \sum_{i=1}^{k} \frac{(m_i - N\, p_i)^2}{N\, p_i}$$

is therefore a measure of the deviation of a sample from expectation, where $N$is the sample size.

COHORT

A group of subjects who have shared a specific event or characteristic.

CONTINUOUS VARIABLE

Continuous variables have an infinite number of values between two points. For example, age or temperature.

CO-REGULATION

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*.

CORRELATION ANALYSIS

A type of Regression Analysis, correlation analysis measures the correlation coefficient – the linear association between two variables. Values of the correlation coefficient are always between -1 and +1. A correlation coefficient of +1 indicates that two variables are perfectly related in a positive linear sense, while a correlation coefficient of -1 indicates that two variables are perfectly related in a negative linear sense.

COX COEFFICIENT

The Cox coefficient refers to the coefficients in a Cox regression model (also known as the proportional hazards model for survival-time). The analysis investigates the effects of one or more variables upon the time a specified event takes to happen. The cox coefficient relates to a hazard; a positive coefficient indicates a worse prognosis, while a negative coefficient indicates a protective effect of the variable.

DATA BINNING

Defers to a data pre-processing technique used to reduce observation errors and to allow continuous variables to become categorical. Clusters of data are replaced by a value representative of that cluster (often but not necessarily, the central value).

DATA WAREHOUSE

A database used for reporting and analysis.

DATASET

Collection of data, most commonly presented in a tabular form where each column represents a specific variable, and each row represents a value for that variable.

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 that you select.

DEPENDENT VARIABLE

In an experiment, the dependent variable is the response that is measured.

DIFFERENTIAL MODULATION

DOWN-REGULATION

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

ENTREZ GENE

Reference sequences for a wide range of species. For details, see http://www.ncbi.nlm.nih.gov/gene/.

ENTREZ GLOBAL

Federated search engine that allows users to search various health sciences databases at the National Center for Biotechnology Information (NCBI) website.

See www.ncbi.nlm.nih.gov/Entrez/ for details.

FOLD CHANGE RATIO

A number describing how much a quantity changes going from an initial to a final value. An initial value of 50 and a final value of 100 corresponds to a fold change of 2 (a two-fold increase).

GENE

Stretches of DNA and RNA that code for a polypeptide or for an RNA chain – contains hereditary molecular information.

GENE CHIP

See: Microarray

GENE EXPRESSION

The flow of genetic information from gene to protein; the process, or the regulation of the process, by which the effects of a gene are manifested; the manifestation of a heritable trait in an individual carrying the gene or genes that determine it.

GENE EXPRESSION OMNIBUS

GEO is an international public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomics data submitted by the research community. For more information, see http://www.ncbi.nlm.nih.gov/geo.

GENE SET ENRICHMENT ANALYSIS (GSEA)

Computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (for example, phenotypes).

See http://www.broadinstitute.org/gsea/index.jsp for details.

GENE SIGNATURE

A group of genes whose combined expression pattern is uniquely characteristic of a medical condition or other clinical outcome of interest.

GENE SYMBOL

A unique abbreviation of a gene name consisting of italicized uppercase Latin letters and Arabic numbers. we use Entrez as the full list of genes (related to but not identical to HUGO)

See http://www.genenames.org/ for details.

GENECARDS

Database that offers information about human genes (and mouse homologues).

See http://www.genecards.org for details.

GOOGLE SCHOLAR

Google application that provides a search of scholarly literature across multiple disciplines and sources.

See http://scholar.google.com for details.

GPL PLATFORM

A Platform record is composed of a summary description of the array or sequencer and, for array-based Platforms, a data table defining the array template. Each Platform record is assigned a unique and stable GEO accession number (GPLxxx). A Platform may reference many Samples that have been submitted by multiple submitters.

HEATMAP

Display of differential expression. Individual values contained in the matrix are represented by colors.

HIERARCHICAL CLUSTERING

Hierarchical clustering is a type of clustering analysis whose goal is to organize data so that the objects in the same cluster are more similar to each other than to those in other clusters.

HIGH DIMENSIONAL DATA

Datasets where the intersection of a subject and measurement is comprised of hundreds or thousands of points. For example, in a low dimensional data measurement such as height the intersection of subject and measurement is one number (ex. 180 cm) whereas in a high dimensional data measurement such as gene expression in a lymph node the measurement is 50,000 individual probe expression values.

HISTOGRAM

A visual representation of the distribution of data values within a dataset.

HOMOLOGY

The basis for comparative biology – where organs/structures from one organism are compared to a similar organ/structure in a different organism.

IN VITRO STUDY

Those that are conducted using components of an organism that have been isolated from their usual biological surroundings.

IN VIVO STUDIES

Experimentation using a whole, living organism.

INDEPENDENT VARIABLE

In an experiment, the independent variable is the variable that is manipulated.

JOB

In Valhalla, a job refers to a command you have given Dataset Explorer to process or export data. Jobs and job-related events can be found within the **Jobs** tab in Dataset Explorer.

KENDALL CORRELATION

Kendall's rank correlation provides a distribution-free test of independence and a measure of the strength of dependence between two variables.

K-MEANS CLUSTERING

The K-Means clustering heatmap clusters genes and/or samples into a specified number of clusters. The result is *k* clusters, each centered around a randomly-selected data point.

LINE GRAPH

Line graphs illustrate the temporal relationship between two major variables.

MARKER SELECTION

Marker Selection is a display of the top differentially expressed genes between two specified cohorts. .

MESH ONTOLOGY

MeSH is the National Library of Medicine's controlled vocabulary thesaurus. It consists of sets of terms naming descriptors in a hierarchical structure that permits searching at various levels of specificity.

MICROARRAY

A two-dimensional array on a chip or solid surface that assays large amounts of DNA material.

MRNA ANALYSIS

Assays that quantify the expression levels of all mRNA molecules in an experiment.

NAVIGATION TREE

The Window’s Explorer-like, hierarchical representation of study data that has been loaded into Dataset Explorer.

NCBI

The National Center for Biotechnology Information.

See http:// www.ncbi.nlm.nih.gov/ for details.

NUMERIC-NODE

Used in Dataset Explorer, numeric-nodes are indicated by the (**123**) symbol, numeric nodes indicate that the data values associated with the concept are only numeric (for example, age values, date values, etc.). For more information, see Continuous Variable.

ONTOLOGY

A hierarchical description of the concepts and relationships that can exist for an agent or a community of agents.

ORTHOGONAL COMPONENT

When performing statistical analysis, independent variables that affect a particular dependent variable are said to be orthogonal if they are uncorrelated, since the covariance forms an inner product.

PATHOLOGY

The study of diagnosis and disease.

PATHWAY

A group of genes interacting to form an aggregate biological function.

PEARSON CORRELATION

Obtained by dividing the covariance of the two variables by the product of their standard deviations

PRINCIPAL COMPONENT ANALYSIS

A Principal Component Analysis (PCA) is commonly used as a tool in exploratory data analysis. Data is split into orthogonal components, and the genes/probes that contribute the most variance to the components are displayed.

PROBE SET

A probe set is a collection of probes designed to interrogate a given sequence.

PROBE SET ID

A probe set ID is used to refer to a probe set, which looks like the following:

```
12345_at or 12345_a_at or 12345_s_at or 12345_x_at
```

The last three characters (`_at`) identify the probe set strand.

P-VALUE

The number corresponding probability that the occurrences of your experiment and analysis did not happen by chance. P-value cutoffs are often 0.05 or 0.01 – when the value is under the threshold, the result is said to be statistically significant.

R

R is a language and environment for statistical computing and graphics.

See http://www.r-project.org for details.

RBM DATA

Rules Based Medicine. They provide an array measurement of metabolites

REGRESSION ALGORITHMS

Algorithms that are particularly suited for mining data sets that have high dimensionality (many attributes), including transactional and unstructured data.

RHO-VALUE

Also known as Spearman's rho, the rho-value is a non-parametric measure of statistical dependence between two variables. See: Spearman Correlation.

R-VALUE

The value assigned to a correlation coefficient.

SCATTER PLOT

Type of graph that uses Cartesian coordinates to display values for two variables for a set of data.

SEARCH FILTER

A biomedical concept used to define search criteria in the Search tool.

SEARCH STRING

A sequence of biomedical concepts used to define search criteria in the Search tool.

SLOPE

The steepness of the line of best fit in a graph (Δy/Δx).

SNP DATA

Single Nucleotide Polymorphism. DNA sequence data marking variation occurring when a single nucleotide — A, T, C or G — in the genome.

SPEARMAN CORRELATION

The Spearman's rank-order correlation is the nonparametric version of the Pearson product-moment correlation. Spearman's correlation coefficient, (, also signified by rho-value) measures the strength of association between two ranked variables.

STATISTICAL SIGNIFICANCE

Results of analyses on data that are statistically significant indicate a confidence level that the results did not happen by chance.

STUDY GROUP

The subjects in a study grouped together due to some common attribute of interest (for example, a study can have two study groups: normal and control).

SUBSET

A smaller grouping of participants in a study. See cohort.

SURVIVAL ANALYSIS

Assessment of the amount of time that a person or population lives after a particular intervention or condition.

T STATISTIC

Ratio of the departure of an estimated parameter from its notional value and its standard error.

TABLE WITH FISHER TEST

Examines the significance of associated categorical variables.

TEA ANALYSES

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

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.

TEA SCORE

TEXT-NODE

Indicated by the (**abc**) symbol, text nodes indicate that the data values associated with the concept are only textual (for example, race or gender). For more information, see Categorical Variable.

TISSUE TYPE

The specific type of tissue that has been used in the experiment (for example, breast tissue, lung tissue, etc.)

UP-REGULATION

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

X-AXIS

The horizontal axis of a two-dimensional Cartesian coordinate system.

Y-AXIS

The vertical axis of a two-dimensional Cartesian coordinate system.