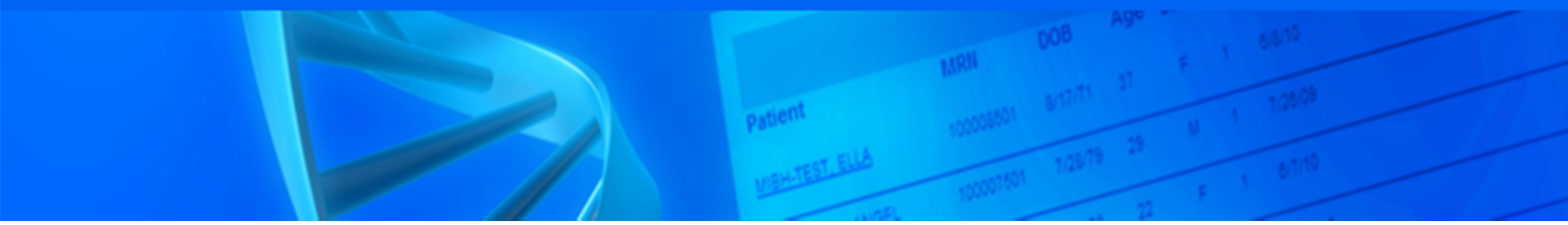
SANOFI

tranSMART

Release Candidate 2 (RC2)

User Manual

Version E01, July 2014



Edition	Date	Changes
E01	July, 2014	Initial version

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

tranSMART

Contents

Welcome

This tutorial describes the main features associated with the tranSMART version called Release Candidate 2 (RC2) deployed at Sanofi.

tranSMART overview

tranSMART is an emerging open source solution which is being adopted by major non-profit research organizations, biopharma companies, academics, hospitals, to **integrate and analyze multidisciplinary data in the context of translational research activities**:

- Clinical data: demographics, clinical observations, trial outcome and preclinical data
- Experimental / molecular data from different models (human, animal, cell line): gene expression, genotyping, histology, proteomics, metabolomics, etc.
- From proprietary, collaborative or public studies

tranSMART enables scientists to test correlations between the different data types, mainly phenotypic and molecular data and to investigate biomarkers of drug toxicity, efficacy, patient responder profile, etc.

For any questions related to tranSMART, please contact:

Annick.Peleraux@sanofi.com
HeikeDagmar.Schuermann@sanofi.com

Connection to tranSMART

Your tranSMART account will be enabled by the tranSMART Team after you have received appropriate training on the application.

Login to tranSMART is done automatically through your company network username and password.

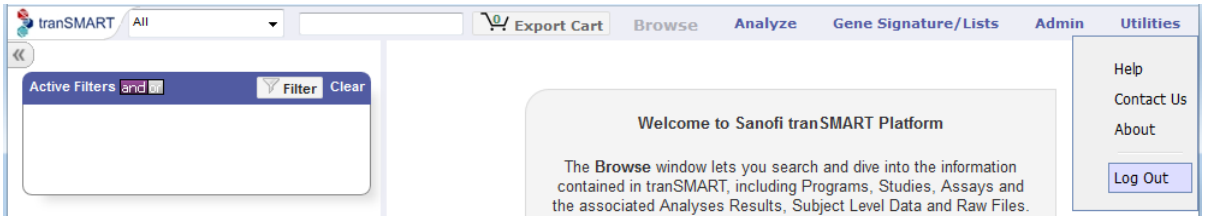
Use the address provided below to log in to the tranSMART (production instance):

https://transmart.sanofi.com/transmart/

Note: A distinct tranSMART-Test instance is available for practicing. You can contact the tranSMART Team to get an access.

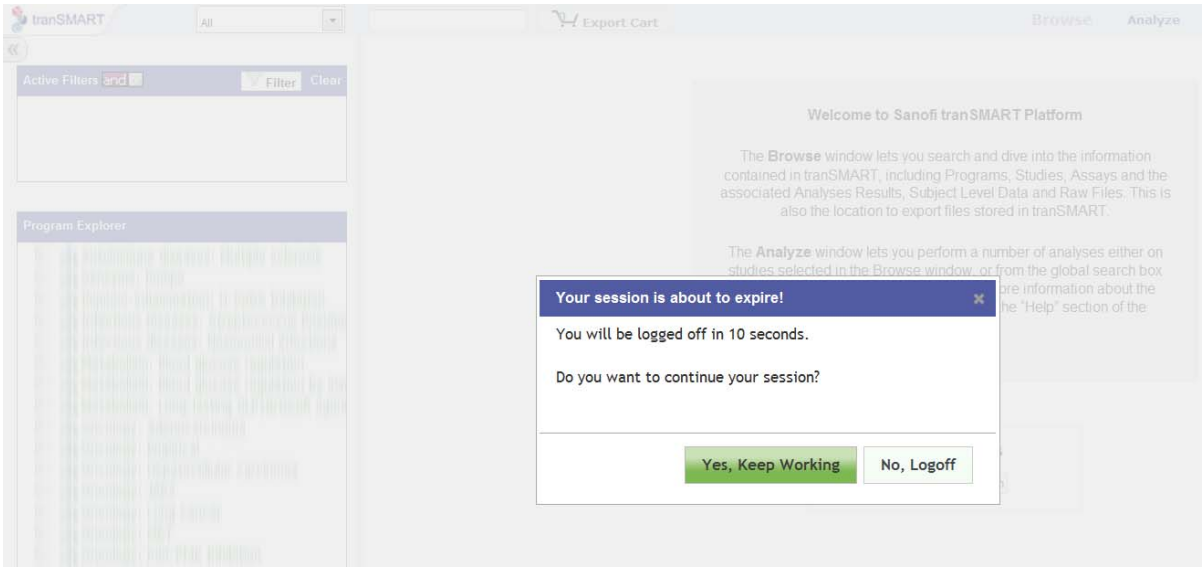
Disconnection

To disconnect tranSMART, click Utilities, then Log Out.



Disconnection in case of inactivity

tranSMART disconnects after a long moment of inactivity.



Click "Yes, Keep Working" to continue working with tranSMART.

Click "No, Logoff" to close tranSMART.

Application and Data Differences

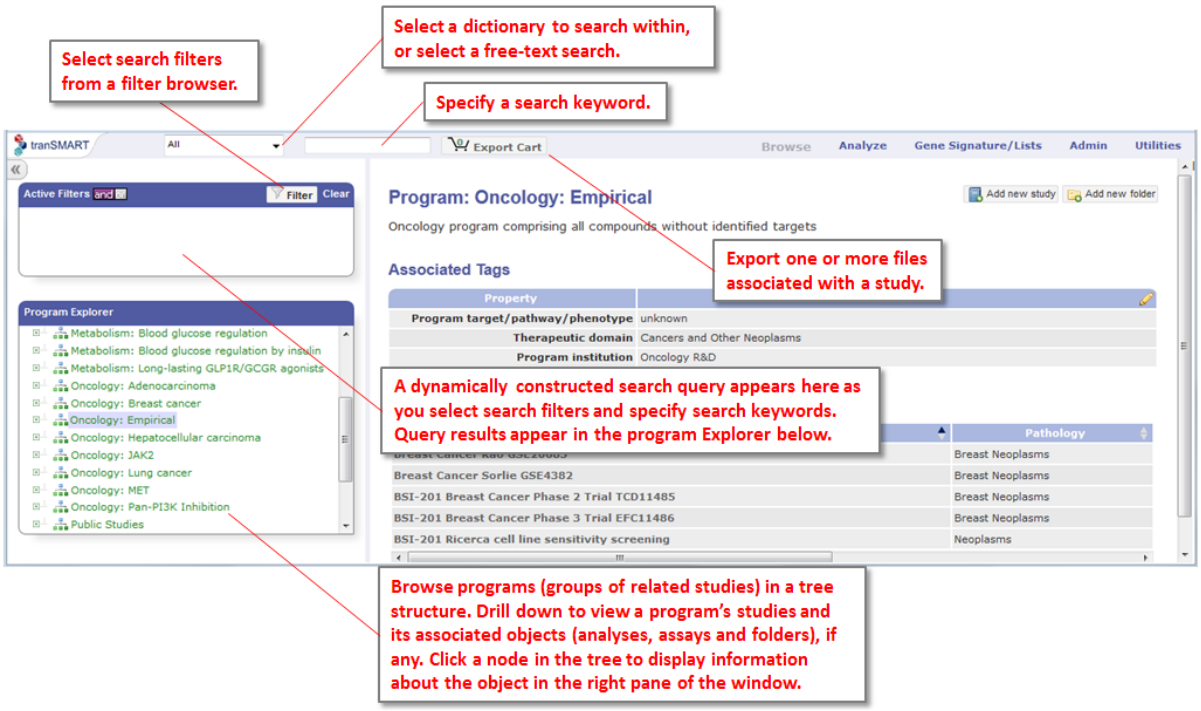
Due to periodic updates to the application and data on tranSMART, the figures and references to specific data in this tutorial may be different than what you see on your screen.

Lesson 1 – Quick Tour of the Browse Module

The tranSMART Browse window lets you browse and search for programs and studies, and for associated objects such as analyses, assays, and file folders.

To open the Browse window, click **Browse** in the tranSMART toolbar:

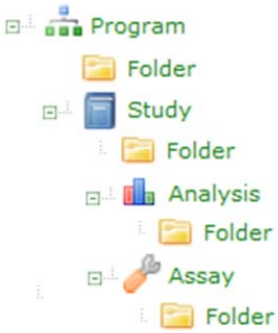
Lesson Goal: Become acquainted with the tranSMART Browse window.



Program Explorer Tree

The Program Explorer displays the results of the search query in the Active Filters box. As the search query changes, the contents of the Program Explorer changes along with it.

The following illustration shows the hierarchy of objects in the Program Explorer tree. Note that each node in the tree is associated with an icon that represents the type of object at that node:



Program is the top-level component of the hierarchy whose purpose is to group related studies together. Most of the time a program is defined by a molecular target, but it may also be a disease or a pathway.

Study is a collection of subjects on which one or several assays were performed. It can be a clinical trial, a preclinical study, or a discovery experiment.

Assay is an investigative procedure for qualitatively or quantitatively assessing the amount or functional activity of an entity. An assay is defined by a unique experimental protocol.

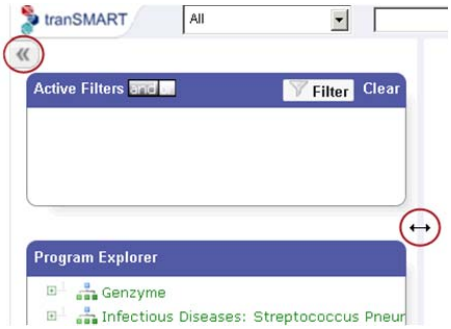
Analysis is a result obtained by analyzing data from a study. In most cases, an analysis is a signature, i.e. a list of molecular entities affected by a particular experimental condition or phenotype.

Folders contain one or several files with information about the associated program, study, analysis, or assay.

Click an object name to view information about the object in the right pane of the Browse window.

Widen the Window

To widen the right pane of the window to view more of its contents, reduce the size of the left pane either by clicking the leftward-pointing arrow button or sliding to the left the vertical bar that separates the two panes.

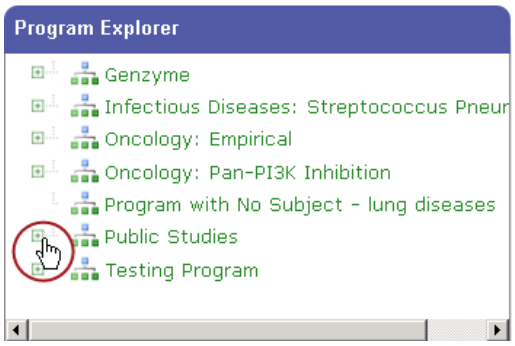


Lesson 2 – Browse Studies

Lesson Goal: Browse programs and studies in the Program Explorer.

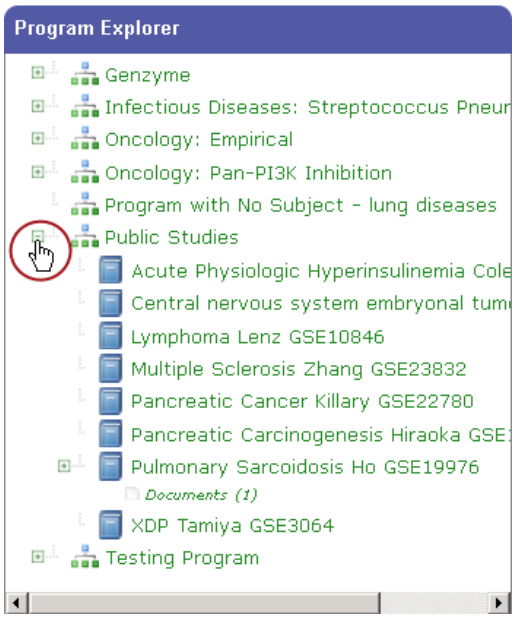
Scenario: You want to view assay information about a public breast cancer study that you are aware of, conducted by T. Sorlie.

1. In the tranSMART Browse window, open the nested **Public Studies** program in the Program Explorer by clicking the plus-sign icon (⊞) to the left of the program name:

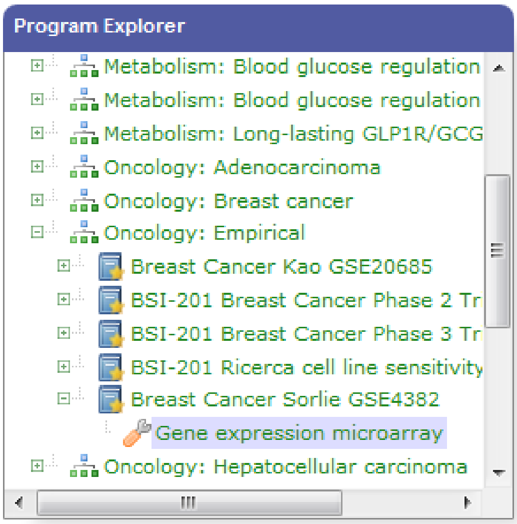


   The list of public studies is short, but you don't see the Sorlie breast cancer study.

2. Close the **Public Studies** program by clicking the minus-sign icon (⊟) to the left of the program name:



3. Open the nested **Oncology: Empirical** program to see if the breast cancer study is there:



The study is there with an assay associated with it (the study has a sub-node with a wrench icon:).

You decide to see in the PubMed site the assay information you want.

4. Click the study name, **Breast cancer Sorlie GSE4382**.

The study's metadata appears in the right pane of the Browse window.

5. In the **Study PubMed ID** field, click the link to the study's PubMed information:

Study identifier	GSE4382
Pathology	Breast Neoplasms
Study compound	Not Applicable
Study phase	Phase II
Study objective	Discover biomarkers
Study design	Interventional
Study biomarker type	Patient selection biomarker
Study design factors	Disease vs. disease
Study link	NA
Number of followed subjects	167
Organism	Homo sapiens
Study access type	Public
Study Institution	Other
Country	UNITED STATES
Study date	2004
Study PubMed ID	12829800
Study publication DOI	Unknown
Study publication author list	Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, Demeter J, Perou CM, LÃ¿Â¿Ã¿Â¿Ã¿Â¿Ã¿Â¸nning PE, Brown PO, BÃ¿Â¿Ã¿Â¿Ã¿Â¿Ã¿Â¸rresen-Dale AL, Botstein D
Study publication title	Repeated observation of breast tumor subtypes in independent gene expression data sets.
Study publication status	Published

6. When finished viewing the PubMed information, close the PubMed page and return to the tranSMART Browse window.

Lesson 3 – Build a Search Query Using Filters

When you add a filter to the Active Filters box, the Program Explorer will display only those programs and associated objects that match the search filter.

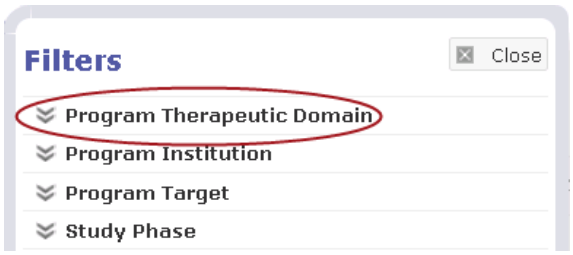
Lesson Goals: (1) Build a search query using the Filter button, and (2) Join multiple search filters using logical operators.

Scenario: You want to search for studies of cancer patients with either of two study objectives: to discover biomarkers, or to demonstrate clinical benefit.

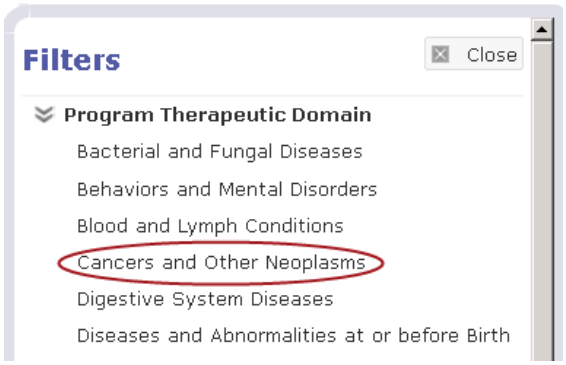
1. In the tranSMART Browse window, click the **Filter** button:

The Filters browser opens, displaying a list of filter categories.

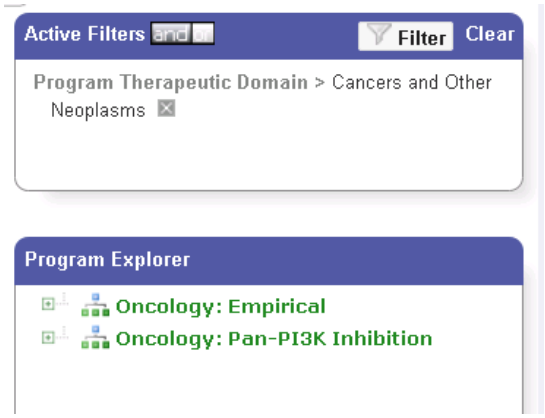
2. In the Filters browser, click the **Program Therapeutic Domain** category.



3. Click **Cancers and other Neoplasms**.



4. Notice that the filter you just selected appears in the Active Filters box, and that the contents of the Program Explorer changes to reflect the filter:

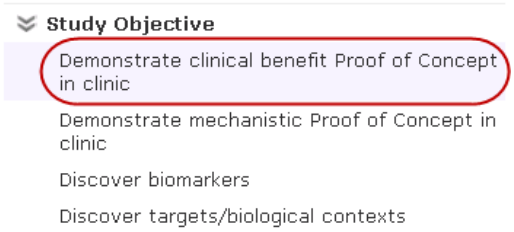


5. Click the plus-sign icon (⊞) to the left of the two program names to view all the studies in the programs that match the filter.

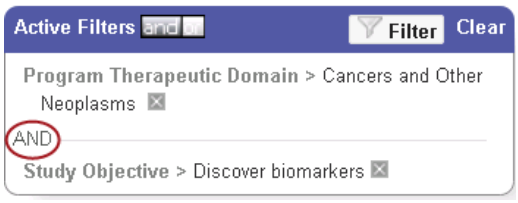
   Now you need to select filters for the study objective.

6. Click the **Filter** button.

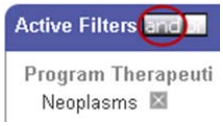
7. In the Filters browser, click **Study Objective**. If necessary, scroll down to find it.
8. Click **Demonstrate clinical benefit Proof of Concept in clinic**:



9. Notice that a Study Objective filter appears in Active Filters, and that it is joined to the Program Therapeutic Domain filter by an AND logical operator:



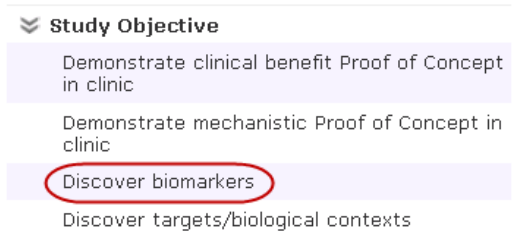
Filters from different filter categories are joined by the logical operator selected in the AND/OR button at the top of the Active Filters box – in this case, AND:



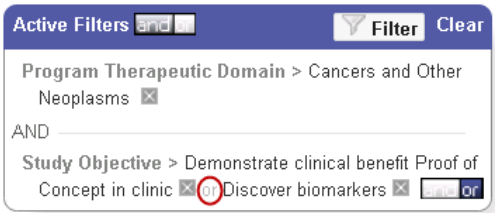
10. Notice also that there are now fewer studies in the Program Explorer. Those are the only studies that match both the Program Therapeutic Doman filter AND the Study Objective filter.

    Now you want to add a filter to the query for the second study objective.

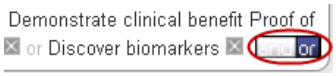
11. Click **Discover biomarkers** in the Study Objective list:



12. Notice that the two study-objective filters are joined by a logical OR operator.



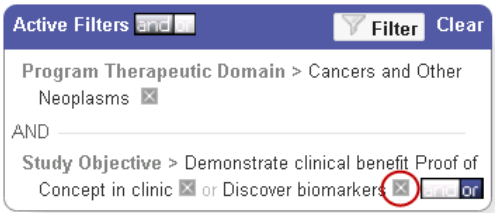
Filters within the same filter category are joined by the logical operator selected in the AND/OR button for the category – in this case, OR:



13. Notice also that there are now more studies in the Program Explorer. That is because the new query matches *either* the clinical benefit objective *or* the discover biomarkers objective.

    You have changed your mind and decide to investigate only the studies with the clinical benefit objective. To display just those studies in the Program Explorer, remove the discover biomarkers objective from the query.

14. In **Active Filters**, click the x icon (⊠) to the right of the **Discover biomarkers** filter:



    The **Discover biomarkers** filter is removed from the Active Filters box, and the studies displayed in Program Explorer is refreshed.

15. Click **Clear** to remove all filters from the Active Filters box in preparation for the next lesson:

    When Active Filters is empty, the Program Explorer contains all programs and their associated objects.

Lesson 4 – Build a Search Query Using Keywords

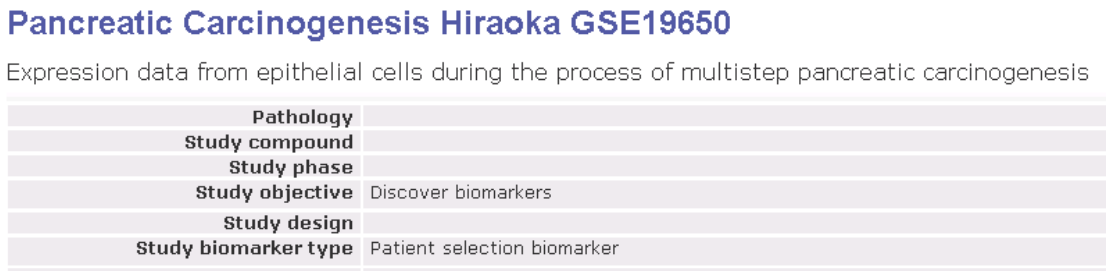
You can build a search query by selecting filters from the Filters browser (as you did in the previous lesson), by typing search keywords in the keyword box (as you will do in this lesson), or by any combination of selected filters and search keywords.

There are two kinds of keyword searches:

- Controlled-vocabulary searches.

  With controlled-vocabulary searches, tranSMART looks for the text you type within a pre-defined dictionary, and displays a list of terms (called the *autocomplete* list) that begin with the typed text. You then select a term in the autocomplete list to use as a search filter.

  Metadata fields that contain controlled vocabulary appear as shaded rows in the Browse window. In the illustration below, **Discover biomarkers** and **Patient selection biomarker** are examples of controlled-vocabulary metadata.

- Free-text searches.

  With free-text searches, tranSMART looks for the text you type within metadata fields that do not contain controlled vocabulary; for example, titles and descriptions of studies and related objects. tranSMART also looks for free-text keywords in text files that are attached to objects.

  With controlled-vocabulary searches, tranSMART includes in the autocomplete list only those terms that begin with the text you type.

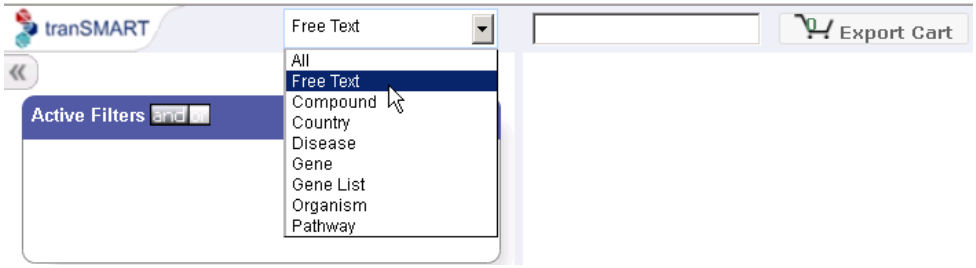
  With free-text searches, tranSMART can find the text you type anywhere within the free-text metadata. For example, tranSMART would find **Hiraoka** in the title of the Pancreatic Carcinogenesis metadata above. It would also find **pancreatic carcinogenesis** in the description.

Keyword searches are not case-sensitive.

Lesson Goals: (1) Build a search string using keywords, and (2) Change the default logical operator that is inserted between filter categories.

Scenario: Find all studies that involved the compound iniparib. To broaden the scope of the search, you want tranSMART to search for all instances of the compound, whether the name appears in controlled-vocabulary or in free-text strings.

1. At the top of the Browse window, click the item **Free Text** in the keyword category dropdown box:



2. In the entry box to the right of the keyword category dropdown box, type the word **iniparib** and press the `Enter` key:

   Pressing `Enter` initiates a free-text search. Now you want to initiate a controlled-vocabulary search.

3. In the dictionary dropdown box, click the item **Compound**.

   This dropdown box contains major categories of keywords (dictionaries), as shown in step **Erreur ! Source du renvoi introuvable.** above. Selecting a category narrows the scope of the autocomplete suggestions. If you select **All**, the autocomplete suggestions can include values from all categories.

4. In the keyword entry box, type the letters **ini** (the first three letters of iniparib):

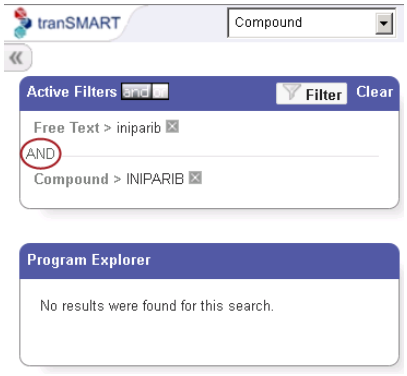
   When you stop typing, the tranSMART autocomplete feature displays a list of terms that begin with the letters you typed. However, iniparib is not there. You need to type more characters to reduce the number of matches.

5. Type the letter **p** after the letters you already typed:

By making the keyword more specific, you narrow the number of terms that match the keyword. Iniparib now appears in the autocomplete list.

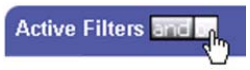
6. Click **iniparib**.

   The Active Filters box now shows two filters joined by the AND logical operator:



   You see from the message displayed in the Program Explorer that there are no matches for *both* the free-text AND the controlled-vocabulary search filters. But you want to broaden the scope of the search by having tranSMART search for matches in either free-text strings OR controlled-vocabulary strings.

7. At the top of the Active Filters box, click **OR**:

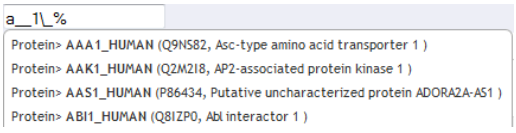


   The AND logical operator between the filters in the different search categories changes to OR:

   The Program Explorer now contains the objects that match *either* of the search category filters.

8. Click **Clear** to remove all filters from the Active Filters box in preparation for the next lesson.

Tips



Use wildcard characters to find information:

% (percent symbol) Match one or more characters
_ (underscore) Match any single character
**** (backslash) Escape character—do not treat the next character as a wildcard

Lesson 5 – Exporting files

Lesson Goal: Download a data file attached to an object in the tranSMART Browse window.

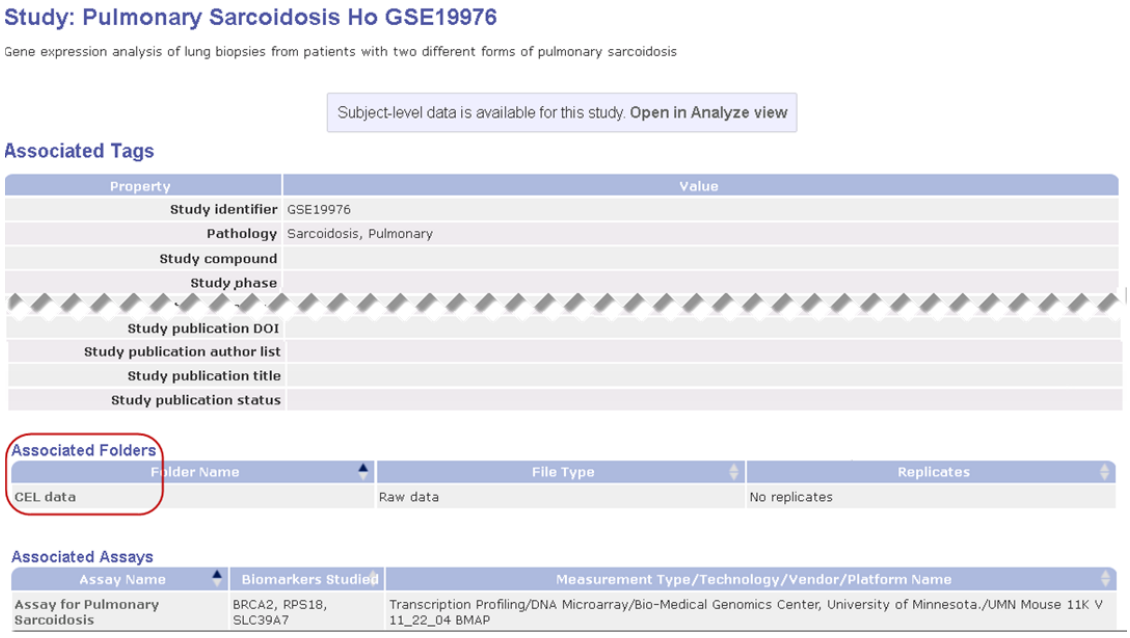
Scenario: You want to analyze the gene expression data generated from a public study of pulmonary sarcoidosis conducted by researcher Ling-Pei Ho and others.

1. In the tranSMART Browse window, open the **'Immuno-inflammation : Pulmonary Sarcoidosis'** program in the Program Explorer by clicking the plus-sign icon ( ) to the left of the program name.

   You see the study **Pulmonary Sarcoidosis Ho GSE19976**. You also notice that it has a file attached. That may be the file of gene expression data that you are looking for.

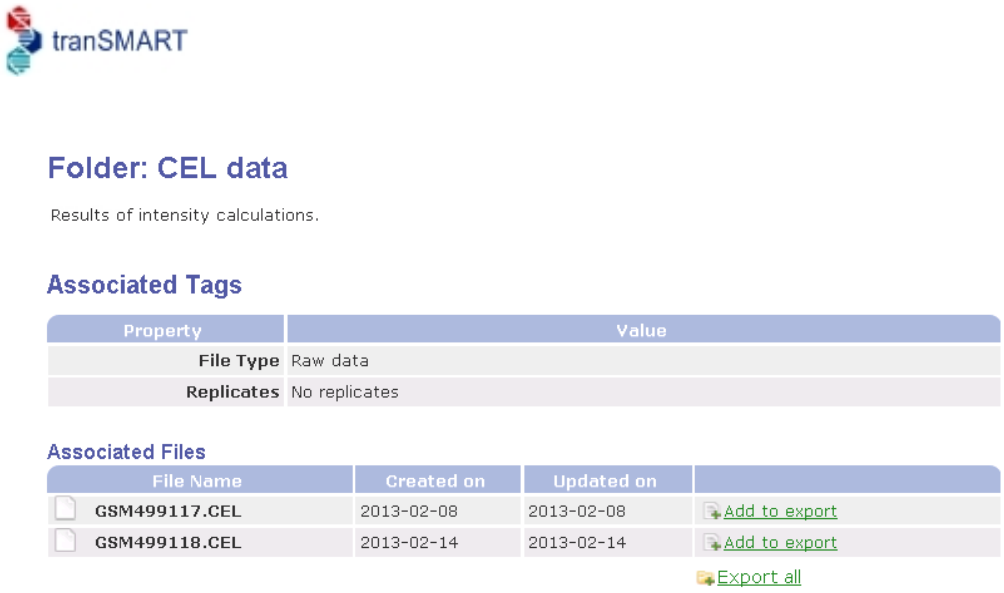
2. Click **Pulmonary Sarcoidosis Ho GSE19976**.

   Information about the study appears in the right pane of the Browse window. You see that the study has one associated folder.



3. Click the folder name, **CEL data**.

   Information about the folder appears, including a list of the files in the folder:



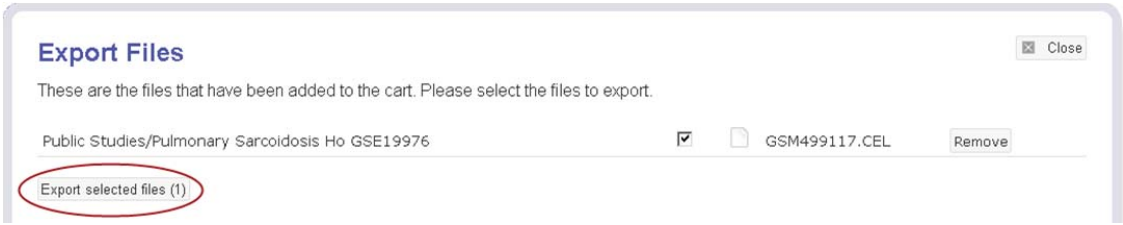
4. Click **Add to export** button to the far right of the file **GSM499117.CEL**.

   The file is not exported immediately. Instead, it is added to the Export Cart. As you continue to work in the Browse window, you can add files from other studies to the cart, and then export the files in the cart in one download operation.

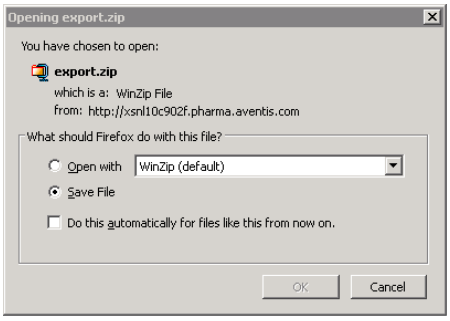
   In this case, you want to download just one file from this study only.

5. Click the **Export Cart** button at the top of the Browse window:

6. In the Export Files dialog box, click **Export selected files (1)**:



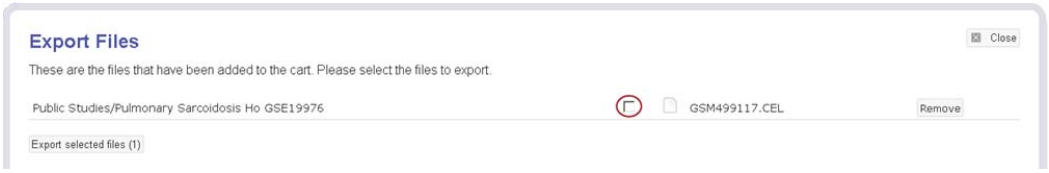
7. When prompted, select **Save File** and click **OK**:



Note the following:

- tranSMART formats the file as a zip file, assigns it the name `export.zip`, and downloads the file to the `Downloads` directory on your computer.
- If a file named `export.zip` already exists in the directory, tranSMART changes the name to `export-1.zip` (or `export-2.zip`, `export-3.zip`, etc., depending on how many files have been exported previously).
- If multiple files are selected for export on the Export Files dialog box, all are downloaded in one zip file.

8. Do one of the following in the Export Files dialog box:
   - To keep the file in the Export Cart in case you want to export it again later, but to remove it from the list of files to be downloaded in the next export operation, clear the check box next to the file name:

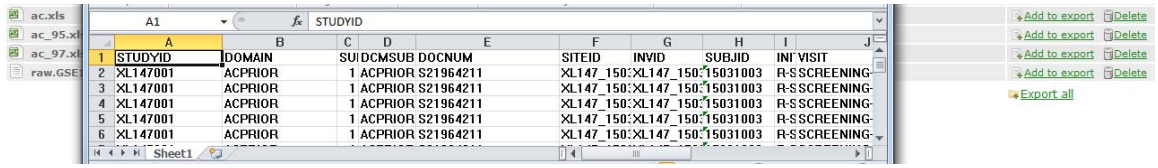


   - To remove the file from the Export Cart, click the **Remove** button next to the file name:

9. Click **Close** in the upper right corner of the Export Files dialog box to close it.

Tips

If you want to visualize the content of a file, click directly the file and open it with the appropriate software.



Lesson 6 – Create and Edit Objects

This lesson is for tranSMART 'Advanced Users' only. Users with basic permissions will not see the object's editing tools that are described in this lesson.

Lesson Goal: Learn to use Browse window tools to add and edit programs, studies, and associated objects.

This lesson is intended to give you a quick tour of the tools you will use to create and edit objects in the Browse window.

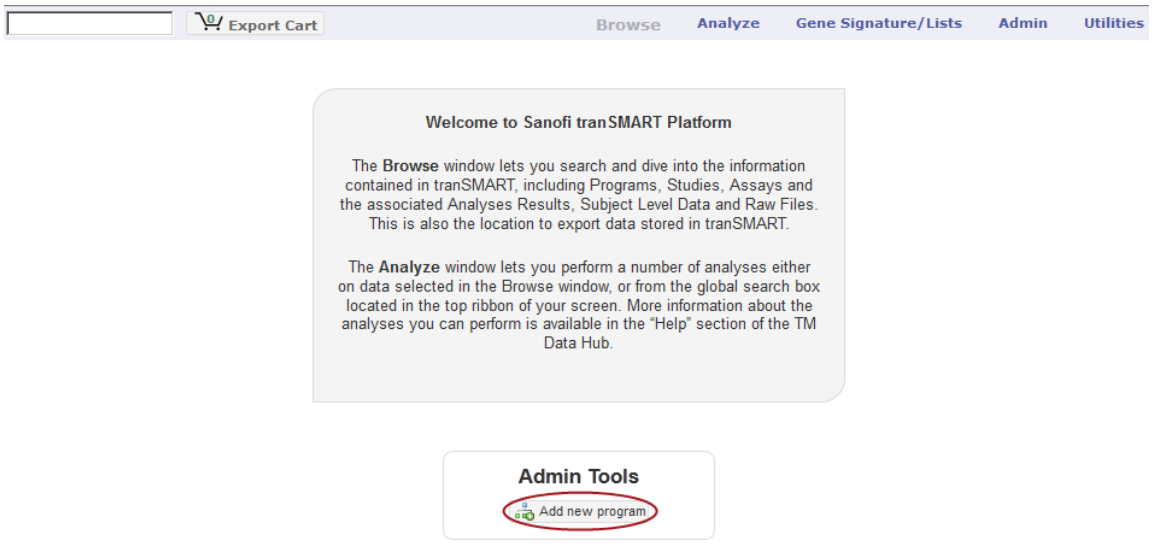
Create Objects

Creating a Program

- To create a program, open the tranSMART Browse window.

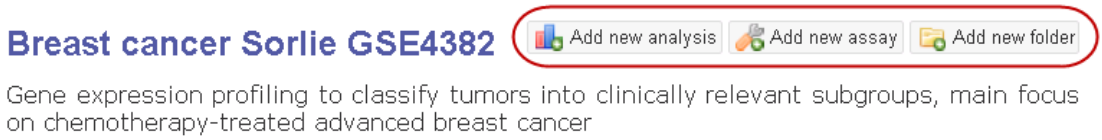
  If you are already working in the Browse window when you want to create a program, click the tranSMART icon at the upper-left corner of the screen to return to this view.

- Click **Add new program**.



Creating Other Objects

The upper right corner of the tranSMART window shows the objects you can create for the currently selected object in the Program Explorer. For example, if a study is selected, you can create an analysis, assay, and folder for the study:



The following table shows the objects you can create for a currently selected object:

Selected Object in Program Explorer	Objects You Can Create
Program	▪ Study ▪ Folder
Study	▪ Analysis ▪ Assay ▪ Folder
Analysis	▪ Folder
Assay	▪ Folder
Folder	▪ Sub-folder

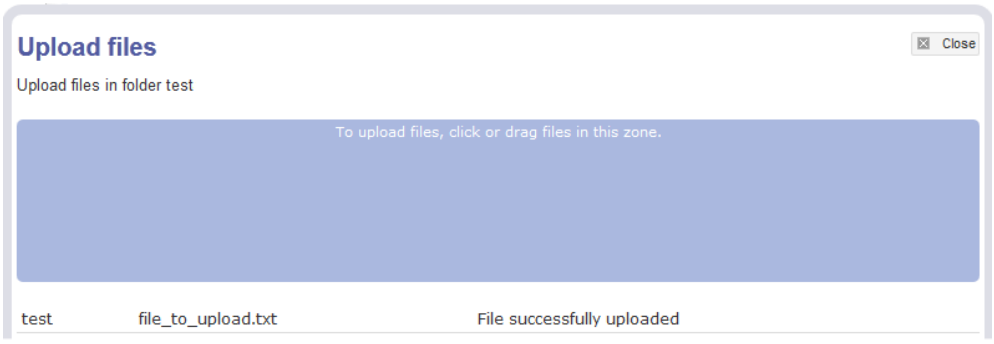
Adding Files to Folders

Folders allow you to attach files to an object; for example, you might add a folder to contain files pertaining to the analysis of a study, or a gene list for an analysis.

You can upload any type of file to a folder. However, the free-text search feature will only search files in format that can be text-indexed, such as Microsoft Word documents, text files, and electronically generated PDFs.

To upload a file to an existing folder, click the Upload files button in the upper right corner of the tranSMART window.

To upload files, click the zone and add the file or drag and drop the file directly in the zone.

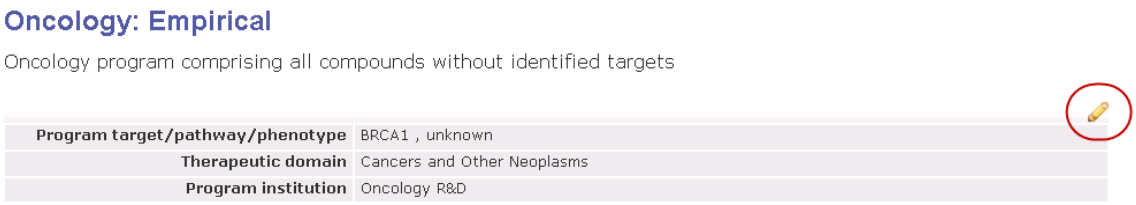


File successfully uploaded.

Edit Objects

You can edit any object's controlled-vocabulary fields and its description, except the study ID. However, program name and study name should not be modified once subject level data have been loaded.

To edit an object, click the pencil icon above the upper right corner of the shaded rows:



Assign Values to Fields

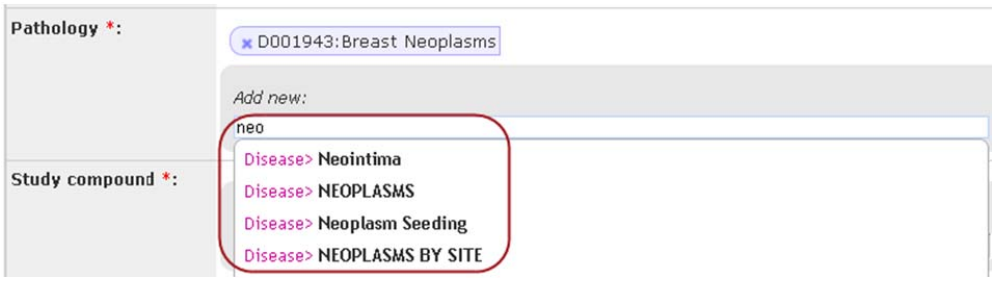
When creating or editing objects, keep the following points in mind:

- You must assign values to fields with a red asterisk after the field name:

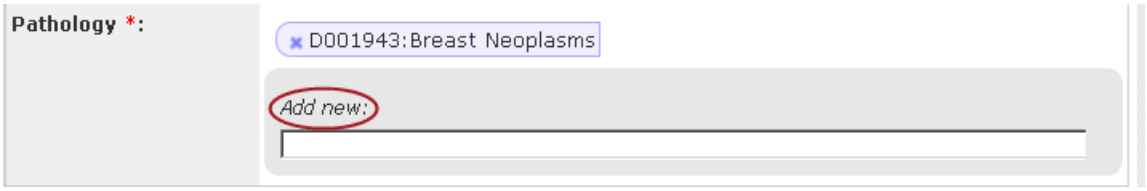
- Study identifier must be unique and cannot be modified later on. It is limited to 100 characters. Avoid using specific characters.

- Shaded fields are autocomplete fields. Type one or more characters at the beginning of the value that you want to assign, and tranSMART will display a list of text strings that begin with those characters.

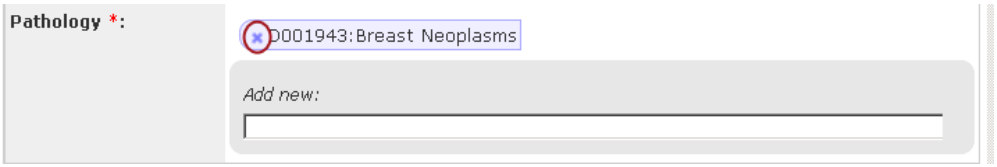
  Alternatively, insert the cursor in the text field and press the Down arrow key to view an alphabetical list of suggested field values.



- Most fields allow multiple values to be assigned. These fields contain the label ***Add new*** above the field.



- To remove a value from a field, click the x icon (✖) to the left of the field name:



Save or Cancel Changes

- **Save** and **Cancel** buttons appear at the bottom of Create… and Edit… dialog boxes:

  Click the appropriate button to save or cancel your changes.

- To close the dialog box, click **Close** at the upper right corner.

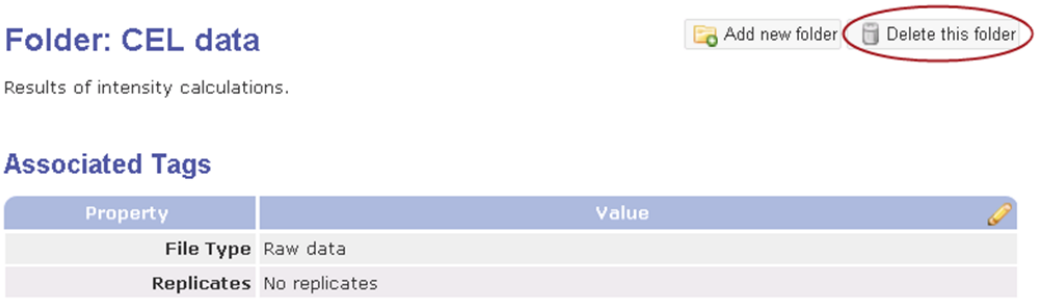
  If you click **Close** before you click **Save**, the dialog box will close without saving your changes and without a warning message.

Delete Objects

In the Browse window, you can delete analyses, assays, folders, and files. Programs and studies can only be deleted from the database directly.

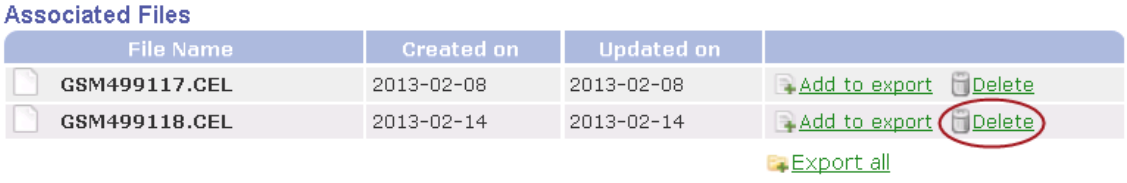
Deleting an Analysis, Assay, or Folder

With the analysis, assay, or folder displayed in the Browse window, click the **Delete this ...** button; for example, to delete a folder, click **Delete this Folder** as shown below:



Deleting Files

With a list of files displayed in the Browse window, click the Delete button at the rightmost end of the row that contains the name of the file to delete. For example:

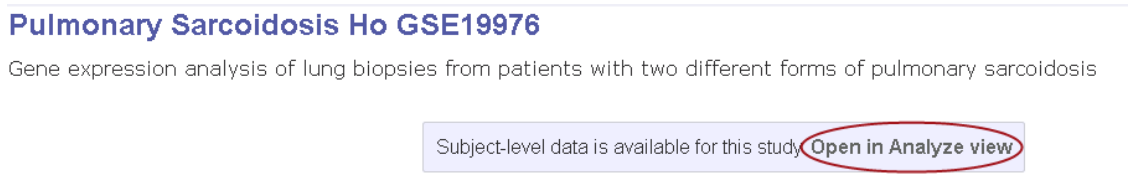


Lesson 7 – Open a study in Analyze View

Lesson Goal: Open a study in the tranSMART Analyze window to perform your own analyses.

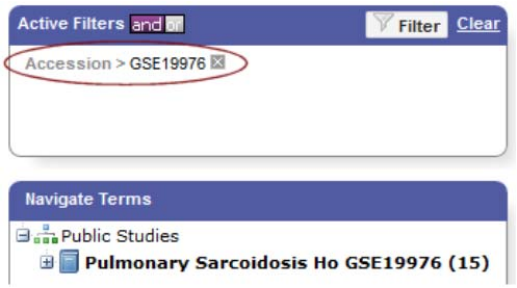
Under the Browse tab, in Program Explorer, select the Study and

click **Open in Analyze view**:



You will only see the message **Subject-level data is available for this study** and the link **Open in Analyze view** if the study's subject-level data has been loaded into tranSMART.

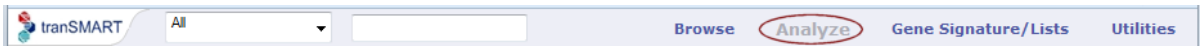
tranSMART displays the Analyze window and opens the study you had just been viewing in the Browse window. Note that the study ID has been added to the Active Filters. You can now define cohorts, run advanced analyses, and perform any other operation that the Analyze window provides.



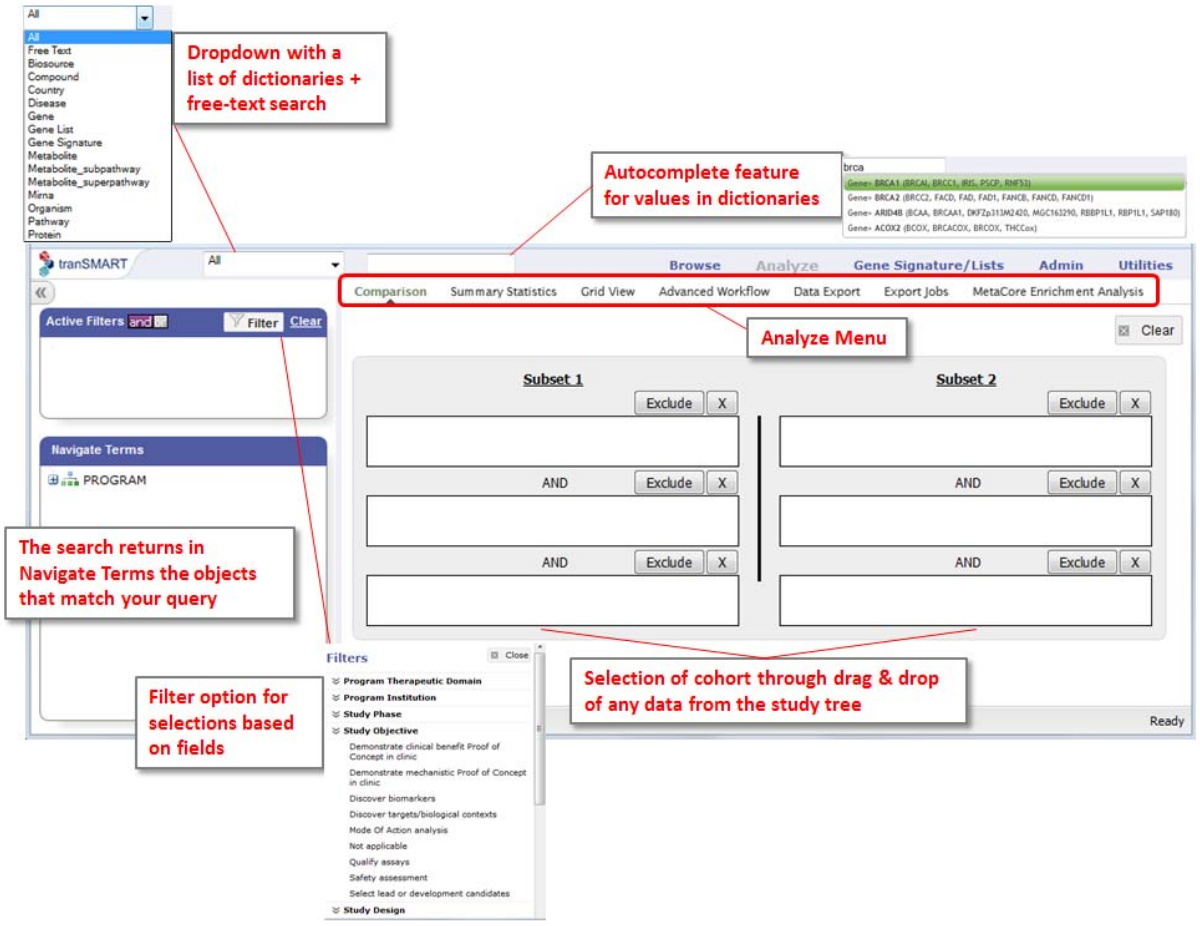
Lesson 8 – Quick tour of the Analyze Module

The tranSMART Analyze window lets you compare data generated for subjects in two different study groups, based on criteria and points of comparison that you specify. Functionalities available under the Analyze tab are useful to help you test a hypothesis around the criteria and points of comparison of your interest.

To open the Analyze window, click **Analyze** in the tranSMART toolbar:



Lesson Goal: Become acquainted with the tranSMART Analyze window.



Overview of the main features available under Analyze

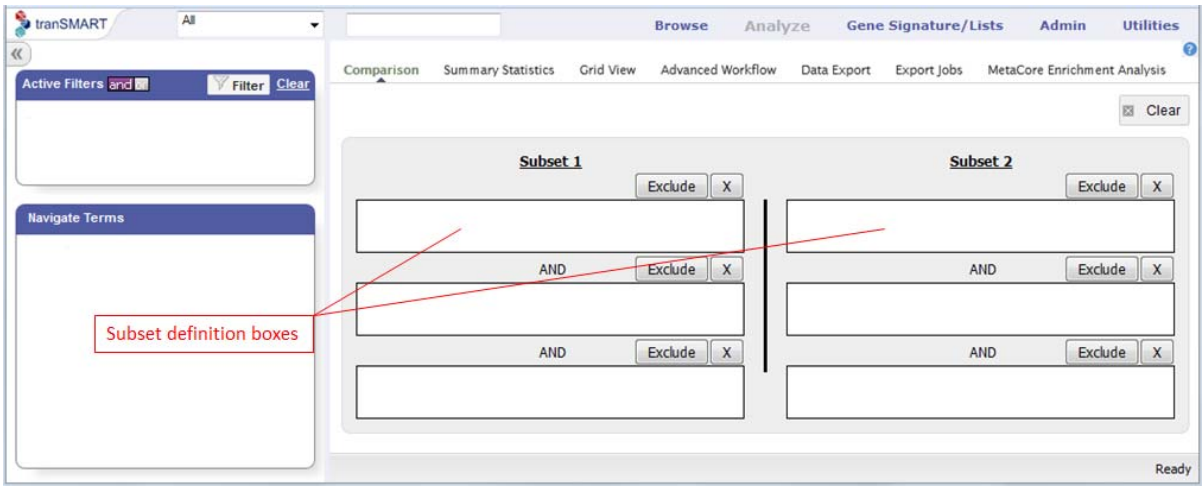
The Analyze module interface is divided into two panels:

Left panel

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

Right panel

- Lets you define the criteria that 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 of the study.
  You define the criteria for the subsets in the subset definition boxes shown below. Subjects who do not satisfy the criteria you define are excluded from the subsets.
- Provides summary data about the subjects being compared, and several different views of the comparison data.



The following table describes the buttons and tabs in the right panel of Analyze:

Button or Tab	Description
Comparison tab	This allows you to further refine the subsets for the comparison.
Summary Statistics tab	Displays tables and charts that describe demographic information about the subjects in the subsets, and also analyses of criteria included in the subset definitions. Other variables from the tree can be described easily by dragging and dropping the node.
Grid View tab	Displays the comparison and analysis data in grid format.
Advanced Workflow tab	Lets you view subset data in the following ways: ■ As a Heatmap ■ As Hierarchical Clustering ■ As K-Means Clustering ■ As Marker Selection ■ As a principal component analysis (PCA) ■ As Survival Analysis ■ As Box Plot with ANOVA ■ As Line Graph ■ As Scatter Plot ■ As Table with Fisher Test ■ As Correlation Analysis
Data Export tab	Export selected data for further analysis in an external tool.
Export Jobs tab	Displays links to previously selected data in Data Export.
MetaCore Enrichment Analysis tab	MetaCore Enrichment Analysis helps understand experimental findings (omics data) in the context of validated biological pathways. Currently this analysis scores and ranks the most relevant MetaBase pathways for a list of genes identified from a gene expression dataset.
Clear button	Clears all the criteria in the subset definition boxes.

Lesson 9 – Browsing and Searching Data in Analyze Module

Lesson Goal: Learn to browse and search data in Analyze Module.

Diversity of data

The following data types can be loaded into tranSMART and displayed in the Analyze window:

- Clinical data: demographics, endpoints, laboratory test results …
- Preclinical data: weight, food intake, laboratory test results …
- Other low dimensional data: histology, mutations …
- mRNA data (Microarray, Sequencing)
- miRNA data (PCR, Sequencing)
- SNP data (Microarray)
- Proteomics data (Mass Spec)
- Metabolomics data (Mass Spec)

Browsing data using Navigate Terms Panel

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

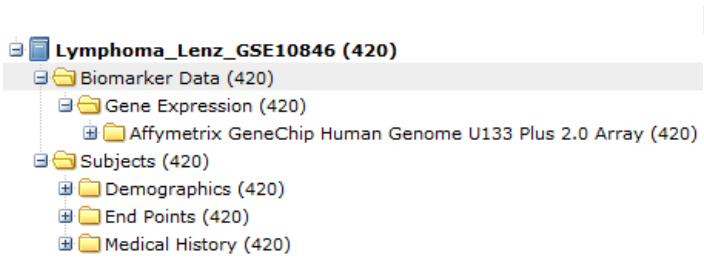
Studies that are grayed out are private studies that you are not authorized to access. To display the details box for a study, right-click the study name and click **Show**. You can display the details box for a study whether or not the study is grayed out.

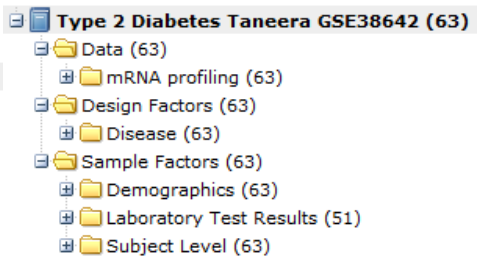
Branches and Leaves of the Navigation Tree

The Analyze module navigation tree looks and works much like the Microsoft Windows Explorer. Windows Explorer is a hierarchy of folders, sub-folders, and files. Analyze module 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 Analyze module, all levels of the tree, both branches and leaves, are referred to as nodes.

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





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

Node	Description
Biomarker Data or Data	Measurements of biomarkers such as RBM antigens, gene expressions, antibodies and antigens in ELISA tests, and SNPs.
Design Factors	Main study parameters (for example: Treatment Arms, Disease Groups, …).
Sample Factors	Other parameters (for example: Demographics, Endpoints, …).

There are three types of leaves in the tree as indicated by different icons:

- **abc** refers to a concept that is non-numeric – for example, gender.
- **123** refers to a concept that is numeric – for example, age, height, weight.
-  refers to a concept of high dimension – for example, gene expression data.

Selecting a cohort (subset of subjects)

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. Analyze module 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 subjects of the study.

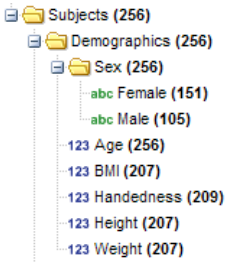
Selecting Criteria for the Cohort

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

Visual Aids to Help You Select the Criteria

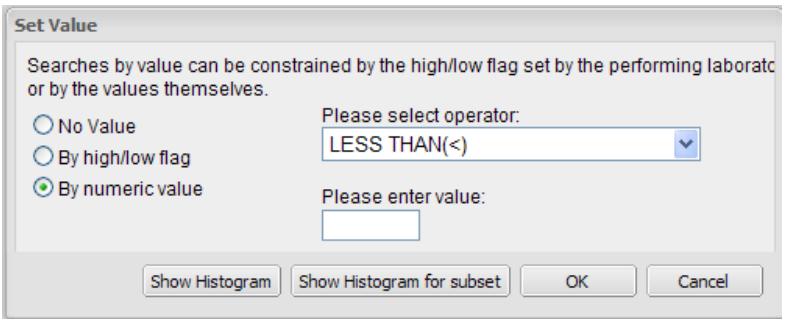
Each concept node in the navigation tree displays the following information about the concept:

- The numbers in parentheses at each node of the tree indicate the number of subjects to whom that node applies. For example, in the figure below, there are a total of 256 subjects in the study, 151 females and 105 males. Further, height and weight measurements were taken for only 207 of the subjects.
- Some nodes have the icon **abc** before them, and others have the icon **123**.
  - **abc** refers to a concept that is non-numeric – for example, gender.
  - **123** refers to a concept that is numeric – for example, age, height, weight.



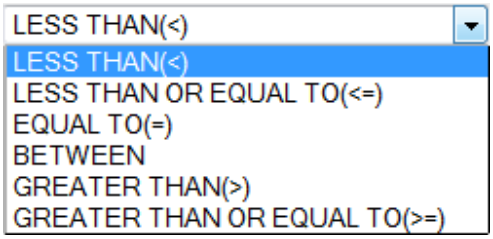
Specifying a Numeric Value

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

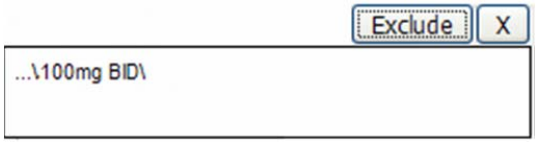


Use the Set Value dialog to specify how you want to constrain the numeric values to use in the subset definition. To do so, first select one of the following choices:

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

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

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

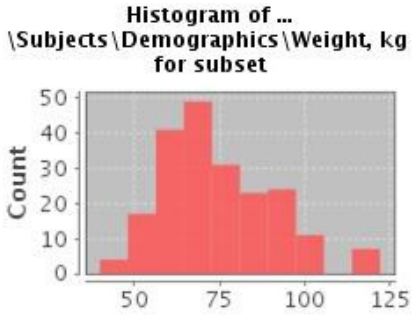


Viewing the Numeric Values Associated with a Concept

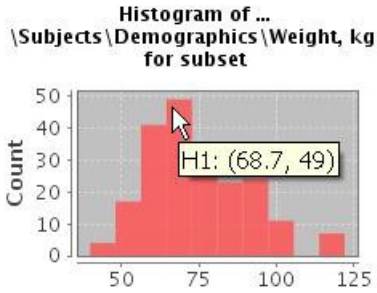
Note the buttons **Show Histogram** and **Show Histogram for subset** in the Set Value dialog. The histograms show how the numeric values associated with the concept that you placed in the subset box are distributed among the subjects across both subsets, or in the particular subset you are currently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

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

represents just under 50 subjects. You may want to use these weight parameters to help you determine the value to set for the weight concept.



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



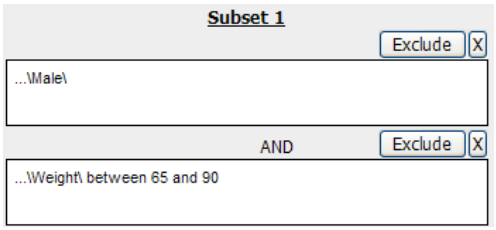
Joining Multiple Criteria for a Subset Definition

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

The rules for joining multiple criteria are as follows:

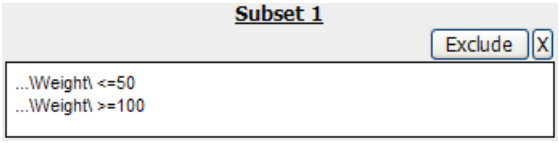
- Criteria in separate subset definition boxes are joined by an AND operator.

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



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

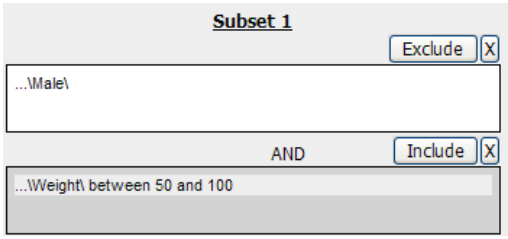
For example, to use the extreme ends of the weight scale for your weight criterion, you might add the following to a definition box:



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

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

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

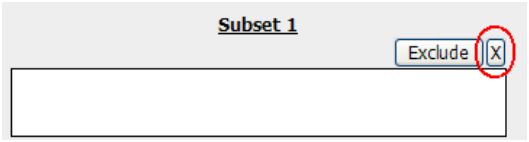


Note that when you click the **Exclude** button, the button label changes to **Include**, allowing you to join the criteria in the box with an AND operator later if you choose.

Modifying or Deleting Criteria

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

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



Lesson 10 – Generating Summary Statistics

Lesson Goal: Learn to generate Summary Statistics.

Steps to generate summary statistics on selected cohort(s)

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

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\ )

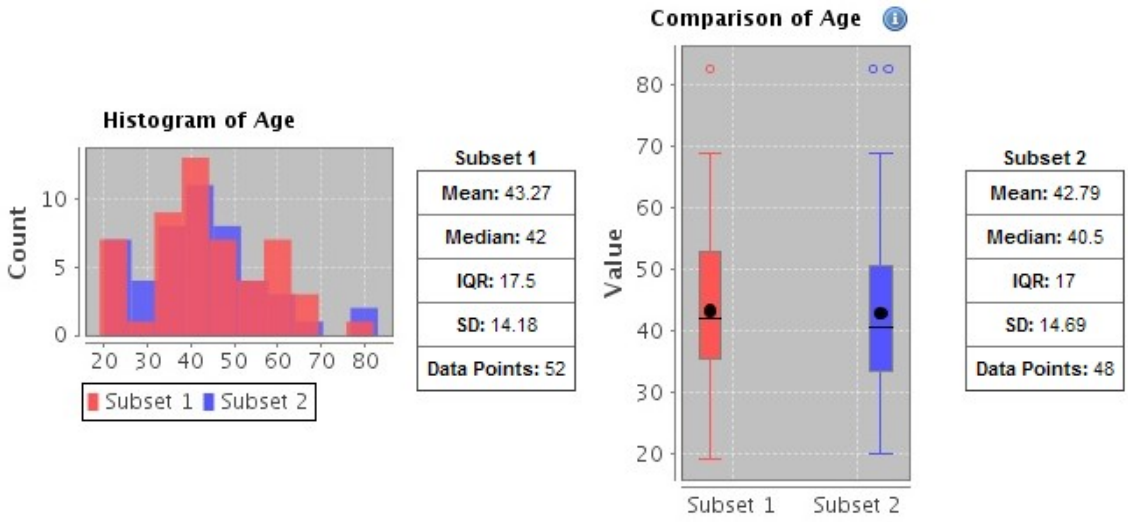
- A table showing the number of subjects in each subset who match the subset criteria. Example:

Subject Totals

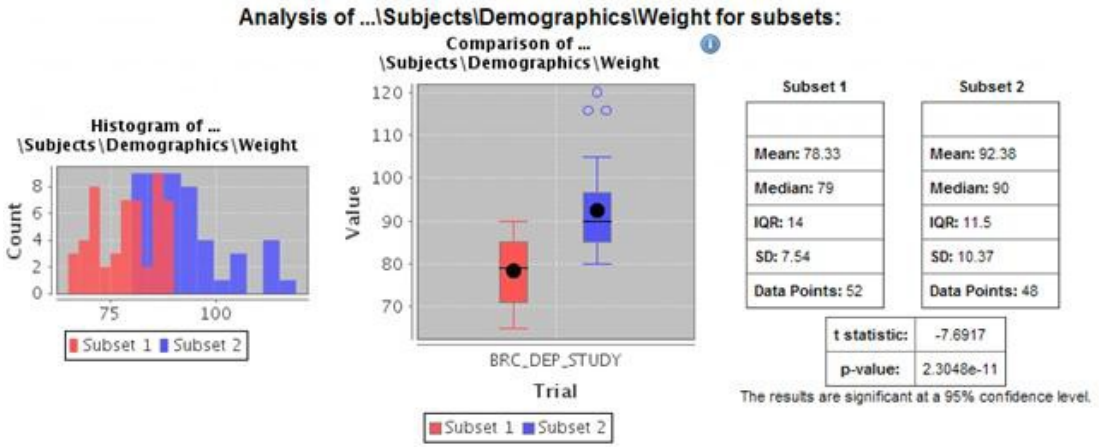
Subset 1	Both	Subset 2
52	25	48

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):



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



Significance Tests

The above figure includes the results of significance testing that Analyze module 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.

Analyze module 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/proper/commons-math/apidocs/org/apache/commons/math3/stat/inference/TTest.html#tTest(double[], double[])

- 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/proper/commons-math/apidocs/org/apache/commons/math3/stat/inference/ChiSquareTest.html#chiSquare(long[][])

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

- If two identical subsets are defined. In this case, the significance test results are not meaningful.

- If all subjects in the first subset have one set of values for the categorical value, and all subjects in the second subset have other categorical values. For example, suppose you set Subset 1 to contain only males and Subset 2 to contain only females. Also, suppose that Subset 1 has 15 subjects and Subset 2 has 20. If you then try to show statistics by gender, a table like the following would result:

	Subset 1	Subset 2
Female	0	20
Male	15	0

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

Defining Points of Comparison

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

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

To apply a point of comparison to the subsets:

1. You must already have defined the subsets and have generated summary statistics for the subsets, as described in the previous section.

2. Drag the concept that you want to introduce as the point of comparison from the

navigation tree, and drop it anywhere in the Results/Analysis view.

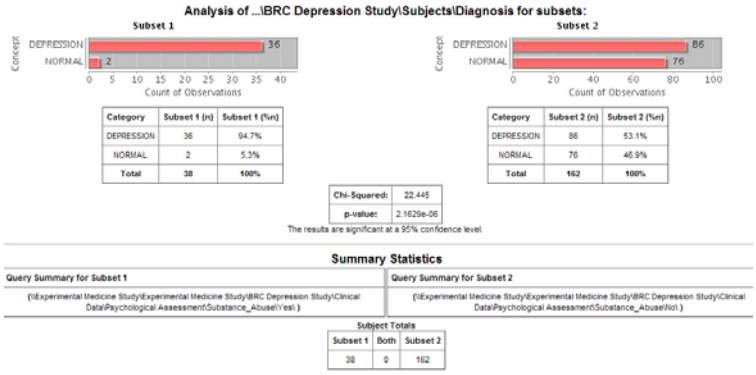
As soon as you drop the point of comparison into the Results/Analysis view, tranSMART begins to compare the subsets based on that point of comparison. When finished, tranSMART displays a side-by-side summary of how the subjects in each subset match or respond to the point of comparison.

Results of a Comparison

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

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

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



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

Lesson 11 – Generating a Grid View

Lesson Goal: Learn to generate a Grid View.

Steps to generate Grid View on selected cohort(s)

When you finish defining criteria for the groups to compare – the subsets – click the **Summary Statistics** tab, then click the **Grid View** tab.

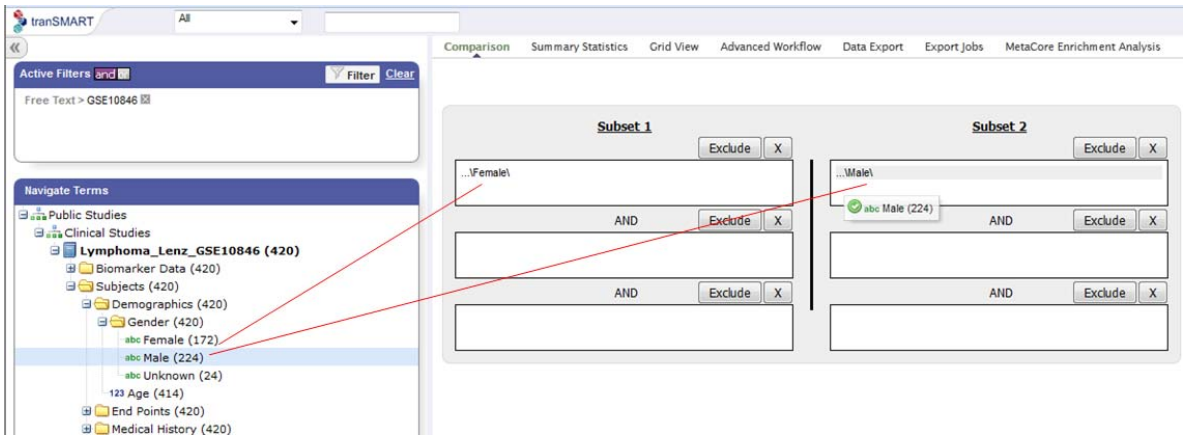
The Grid View displays the selected data in grid format.

Example

1. Go to the Analyze tab, select the search criteria **Free Text**, write in the box: **GSE10846** and 'Enter key' on your keyboard.

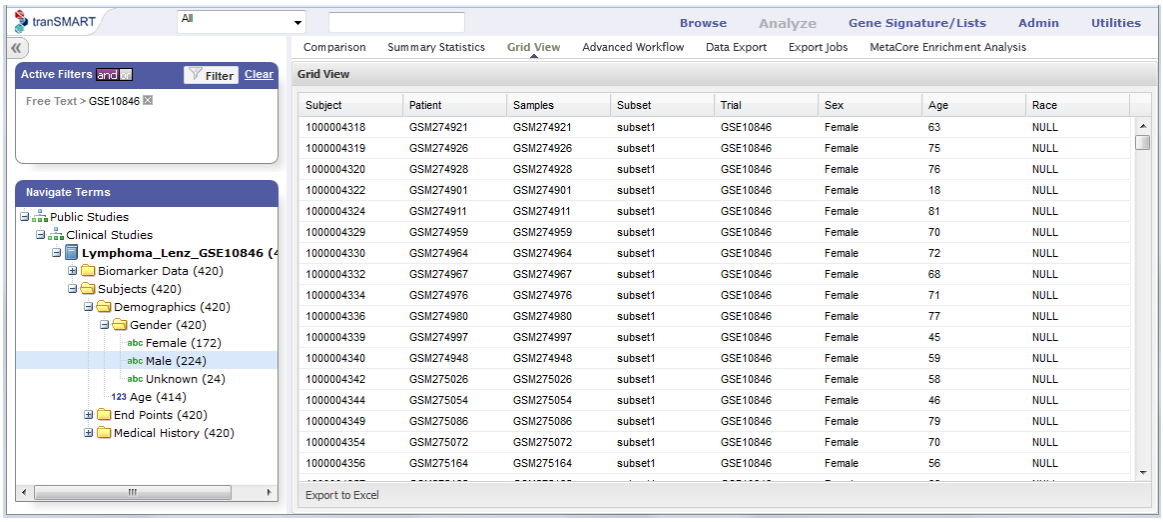
2. In the Navigate Terms area, under the Program, open the Study '**Lymphoma_Lenz_GSE10846**'.

3. In the Comparison tab, drag and drop the Study 'Public Studies\Clinical Studies \Lymphoma_Lenz_GSE10846\Subjects\Demographics\Gender**Female**' as Subset 1 cohort and 'Public Studies\Clinical Studies \Lymphoma_Lenz_GSE10846\Subjects\Demographics\Gender**Male**' as Subset 2 cohort.

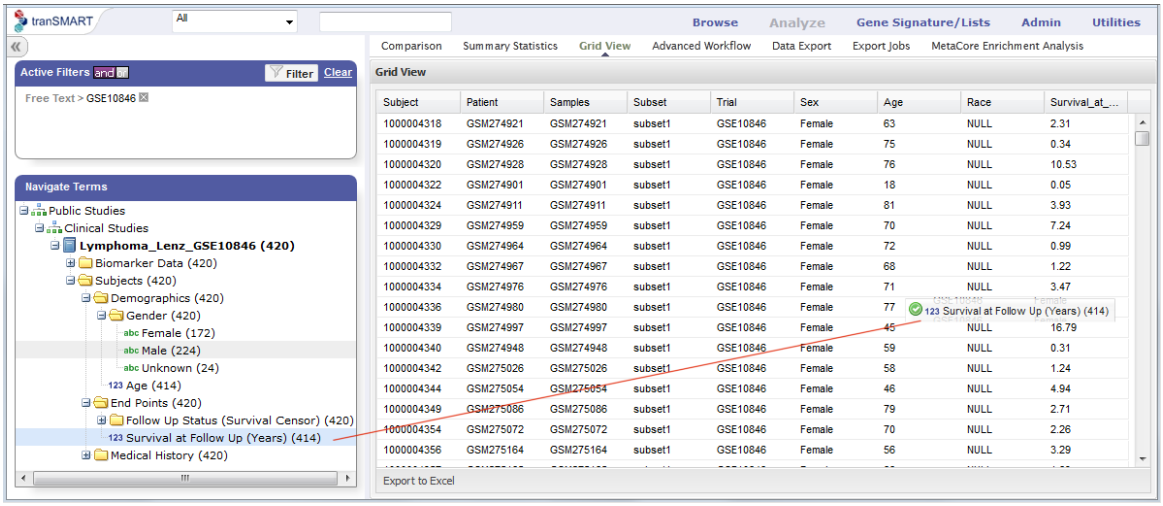


4. Click the **Summary Statistics** tab.

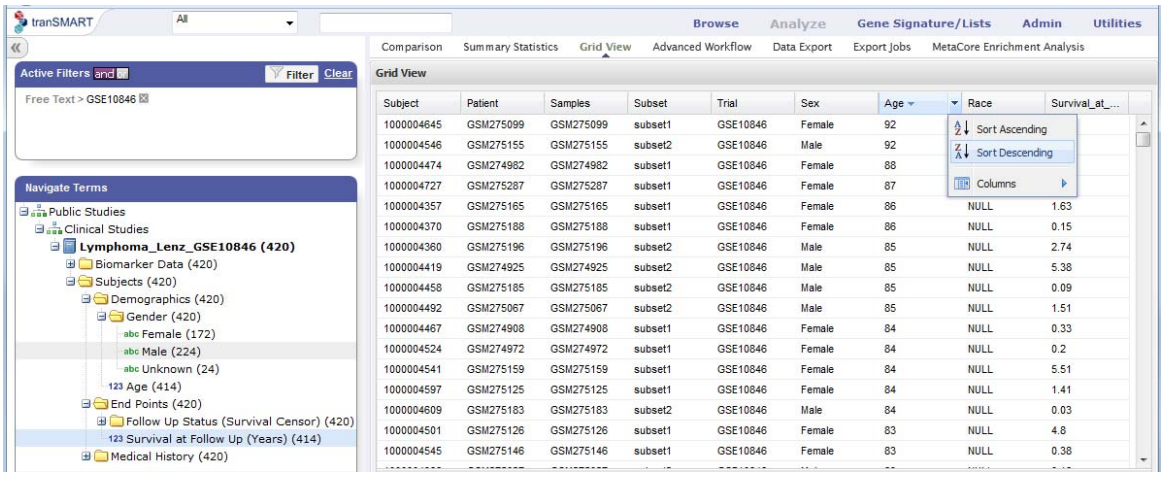
5. Click the **Grid View** tab.



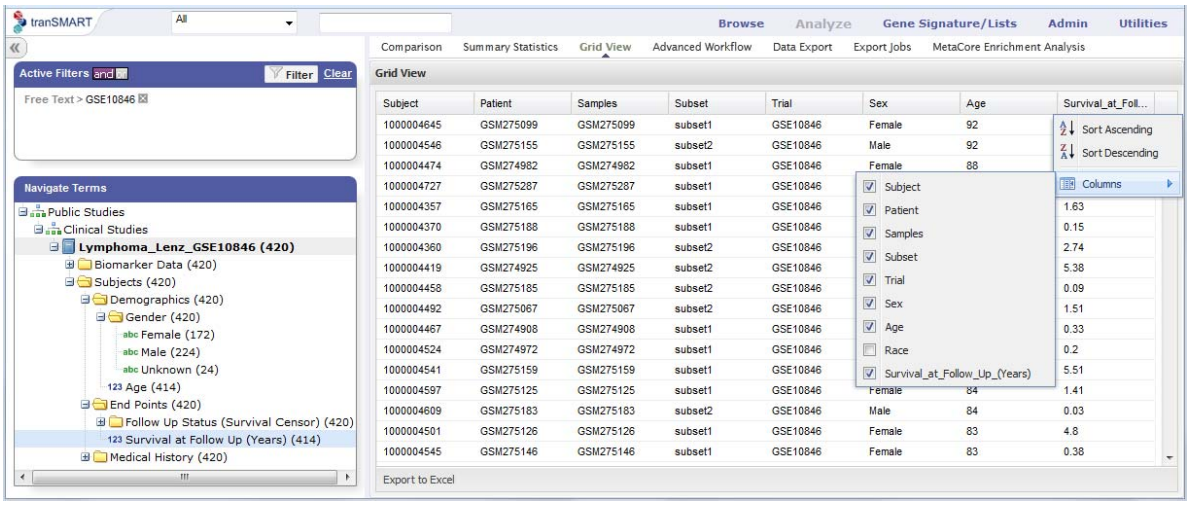
6. Drag and drop the '\Public Studies\Clinical Studies\Lymphoma_Lenz_GSE10846\Subjects\End Points**Survival at Follow Up (Years)**' node in the Grid View table.



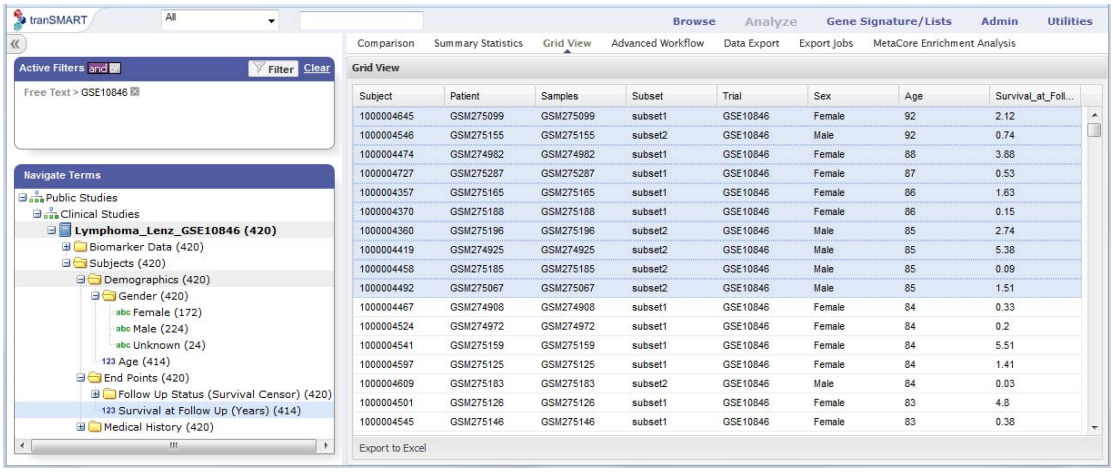
7. Click the **Age** title and sort descending this column.



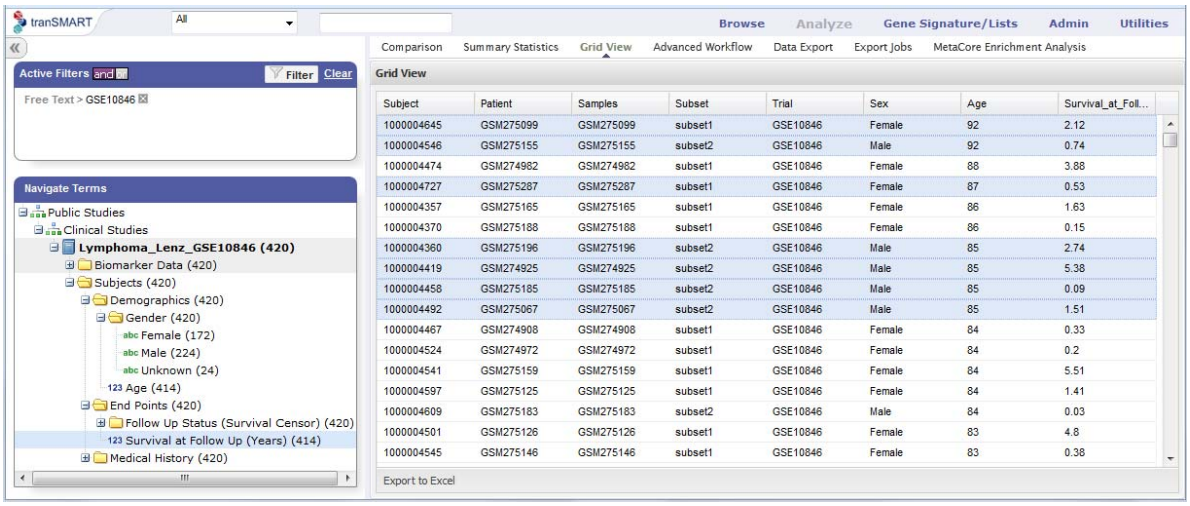
8. Click the **Race** drop down title and deselect this column.



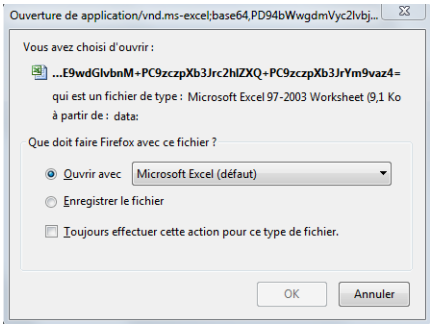
9. Select only the 10 first rows by clicking the first row in the 'Grid View' table, SHIFT in your keyboard and the tenth row.



10. Deselect the rows having ‘Age = 86’ and ‘Age = 88’, using CTRL in your keyboard.



11. Click the **Export to Excel** button at the bottom of the table.

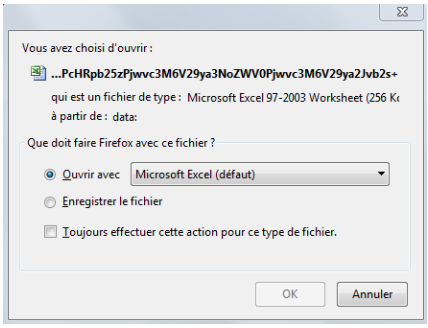


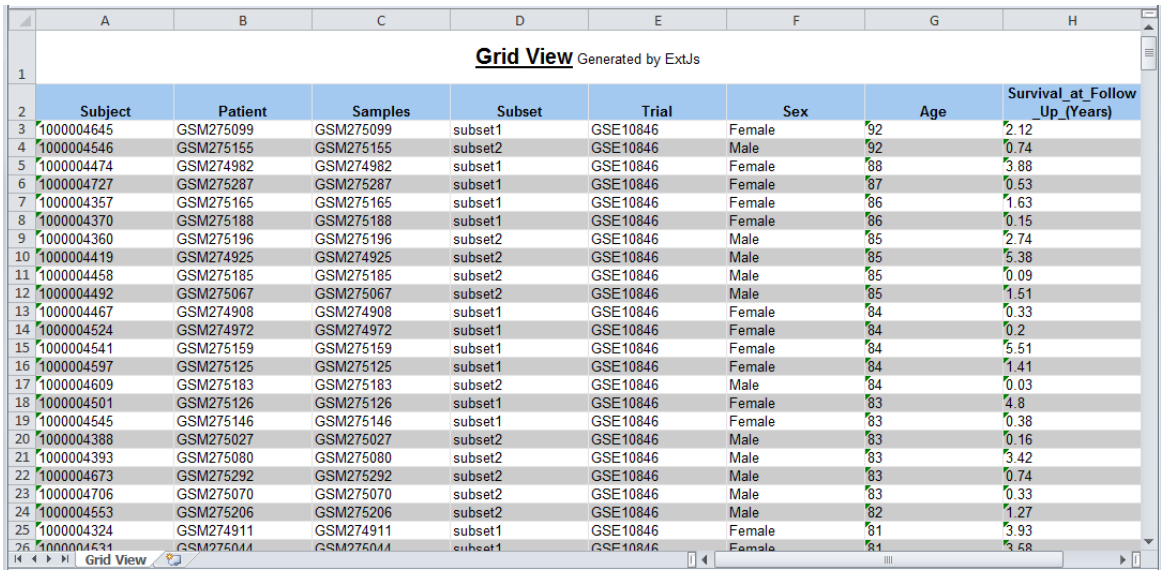
11. Visualize data in the Excel file.

	A	B	C	D	E	F	G	H
1	Grid View Generated by ExtJs							
2	Subject	Patient	Samples	Subset	Trial	Sex	Age	Survival_at_Follow_Up_(Years)
3	1000004645	GSM275099	GSM275099	subset1	GSE10846	Female	92	2.12
4	1000004546	GSM275155	GSM275155	subset2	GSE10846	Male	92	0.74
5	1000004727	GSM275287	GSM275287	subset1	GSE10846	Female	87	0.53
6	1000004360	GSM275196	GSM275196	subset2	GSE10846	Male	85	2.74
7	1000004419	GSM274925	GSM274925	subset2	GSE10846	Male	85	5.38
8	1000004458	GSM275185	GSM275185	subset2	GSE10846	Male	85	0.09
9	1000004492	GSM275067	GSM275067	subset2	GSE10846	Male	85	1.51

12. Return to Grid View. Use CTRL to deselect highlighted rows.

13. To export all data in Grid View (no highlighted rows), click the **Export to Excel** button at the bottom of the table.





Grid View Generated by ExtJs

Subject	Patient	Samples	Subset	Trial	Sex	Age	Survival_at_Follow_Up_(Years)
1000004645	GSM275099	GSM275099	subset1	GSE10846	Female	92	2.12
1000004546	GSM275155	GSM275155	subset2	GSE10846	Male	92	0.74
1000004474	GSM274982	GSM274982	subset1	GSE10846	Female	88	3.88
1000004727	GSM275287	GSM275287	subset1	GSE10846	Female	87	0.53
1000004357	GSM275165	GSM275165	subset1	GSE10846	Female	86	1.63
1000004370	GSM275188	GSM275188	subset1	GSE10846	Female	86	0.15
1000004360	GSM275196	GSM275196	subset2	GSE10846	Male	85	2.74
1000004419	GSM274925	GSM274925	subset2	GSE10846	Male	85	5.38
1000004458	GSM275185	GSM275185	subset2	GSE10846	Male	85	0.09
1000004492	GSM275067	GSM275067	subset2	GSE10846	Male	85	1.51
1000004467	GSM274908	GSM274908	subset1	GSE10846	Female	84	0.33
1000004524	GSM274972	GSM274972	subset1	GSE10846	Female	84	0.2
1000004541	GSM275159	GSM275159	subset1	GSE10846	Female	84	5.51
1000004597	GSM275125	GSM275125	subset1	GSE10846	Female	84	1.41
1000004609	GSM275183	GSM275183	subset2	GSE10846	Male	84	0.03
1000004501	GSM275126	GSM275126	subset1	GSE10846	Female	83	4.8
1000004545	GSM275146	GSM275146	subset1	GSE10846	Female	83	0.38
1000004388	GSM275027	GSM275027	subset2	GSE10846	Male	83	0.16
1000004393	GSM275080	GSM275080	subset2	GSE10846	Male	83	3.42
1000004673	GSM275292	GSM275292	subset2	GSE10846	Male	83	0.74
1000004706	GSM275070	GSM275070	subset2	GSE10846	Male	83	0.33
1000004553	GSM275206	GSM275206	subset2	GSE10846	Male	82	1.27
1000004324	GSM274911	GSM274911	subset1	GSE10846	Female	81	3.93
1000004531	GSM275044	GSM275044	subset1	GSE10846	Female	81	3.68

Grid View

Lesson 12 – Generating a Heatmap

Lesson Goals: Become acquainted with the heatmap analyses: (1) Heatmap, (2) Hierarchical Clustering, (3) K-Means Clustering, (4) Marker Selection

Heatmap analyses require **several rows of data** (either several probes, transcripts, peptides etc… or a pathway representing several genes or proteins).

Heatmap

A heatmap is a graphical representation of data where the individual values contained in a matrix are represented as colors.

To generate a Heatmap:

1. Run tranSMART, under Analyze tab, select the study of interest.
2. In the Comparison tab, define one or two cohorts whose data points will be represented in the Heatmap visualization.
3. Click the **Advanced Workflow** tab.
4. Click **Analysis > Heatmap**.
5. In the Variable Selection box, drag and drop the high dimensional data node.
6. Click the **High Dimensional Data** button and add a pathway (or other object with multiple rows).
7. Click the **Run** button.

Example

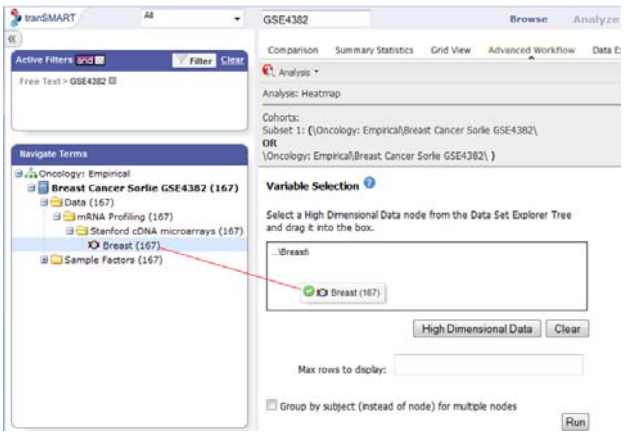
1. Go to the Analyze tab, select the search criteria **Free Text**, write in the box: **GSE4382** and 'Enter key' on your keyboard.

2. In the Navigate Terms area, under the Program, open the Study '**Breast Cancer Sorlie GSE4382**'.

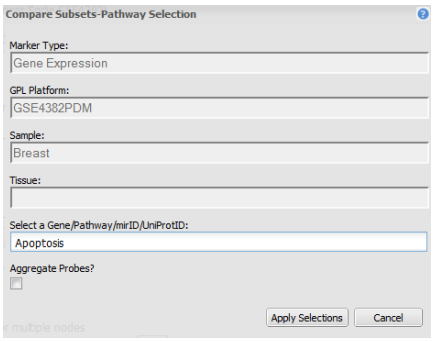
3. In the Comparison tab, drag and drop the Study '**Breast Cancer Sorlie GSE4382**' as Subset 1 cohort.

4. Go to the **Advanced Workflow**, then select **Heatmap** from the Analysis dropdown menu.

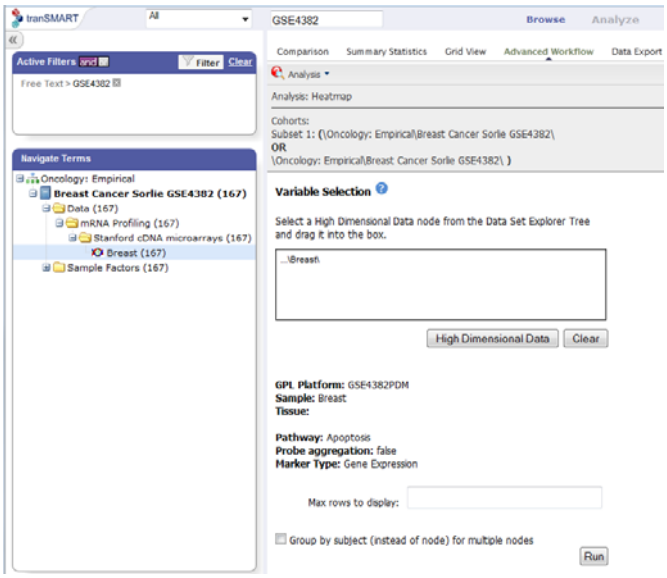
5. Drag and drop the node '\Breast Cancer Sorlie GSE4382\Data\...**Breast**' under Variable Selection.



6. Click the **High Dimensional Data** button, add the pathway **Apoptosis** and click **Apply Selections**.

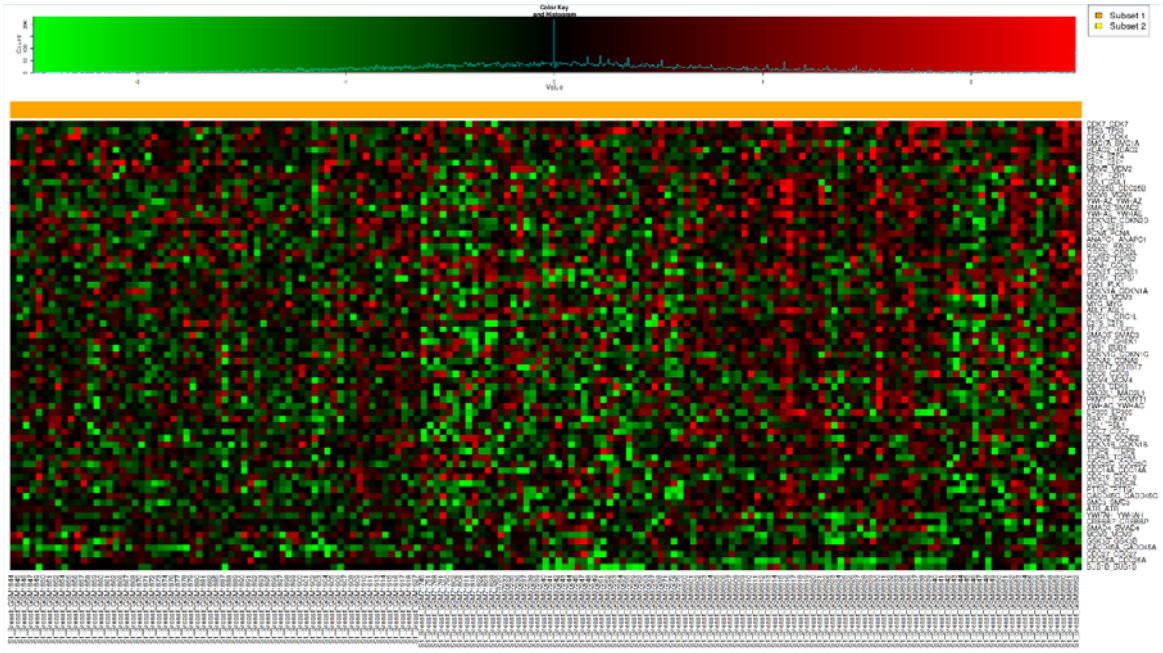


For using Aggregate Probes, see Appendix A – Additional Material>Aggregate Probes.



7. Click the **Run** button.

8. Click the Heatmap generated and visualize the image.



9. Click the link **Download raw R data** to export data.

Tips

For Heatmap, if no specific pathway / gene / mirID / UniProtID needs to be selected, click directly the Run button to by-pass the High Dimensional Data pop-up!

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.

To generate a Hierarchical Clustering Heatmap:

1. Run tranSMART, under Analyze tab, select the study of interest.
2. In the Comparison tab, define one or two cohorts whose data points will be represented in the Heatmap visualization.
3. Click the **Advanced Workflow** tab.
4. Click **Analysis > Hierarchical Clustering**.
5. In the Variable Selection box, drag and drop the high dimensional data node.
6. Click the **High Dimensional Data** button and add a pathway (or other object with multiple rows).
7. Click the **Run** button.

Example

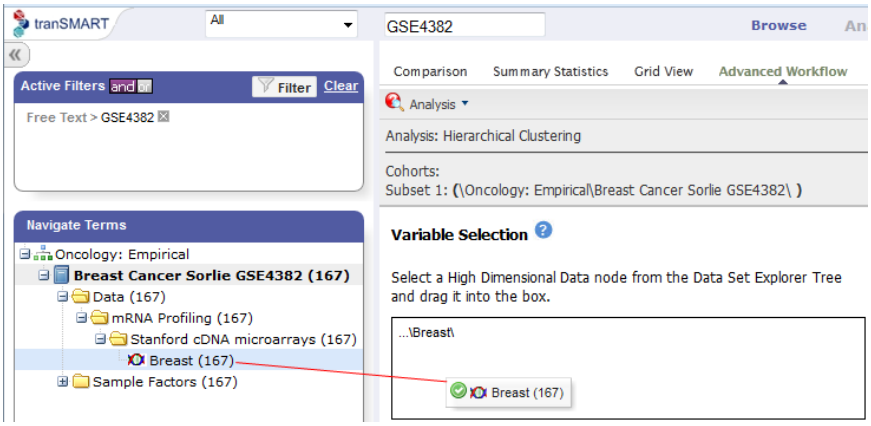
1. Go to the Analyze tab, select the search criteria **Free Text**, write in the box: **GSE4382** and 'Enter key' on your keyboard.

2. Under the Program, open the Study '**Breast Cancer Sorlie GSE4382**'.

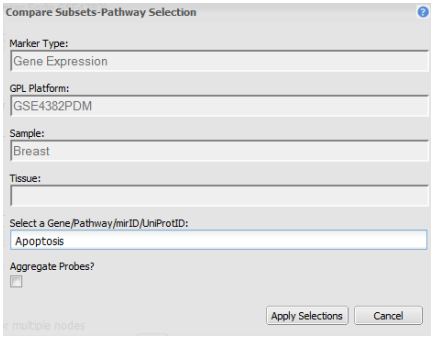
3. In the Comparison tab, drag and drop the Study '**Breast Cancer Sorlie GSE4382**' as Subset 1 cohort.

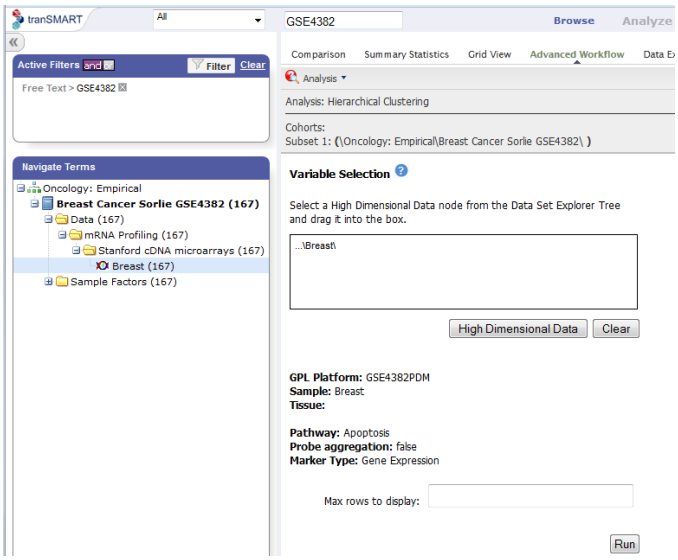
4. Go to the **Advanced Workflow**, then select **Hierarchical Clustering** from the Analysis dropdown menu.

5. Drag and drop the node '\Breast Cancer Sorlie GSE4382\Data\...**Breast**' under Variable Selection.



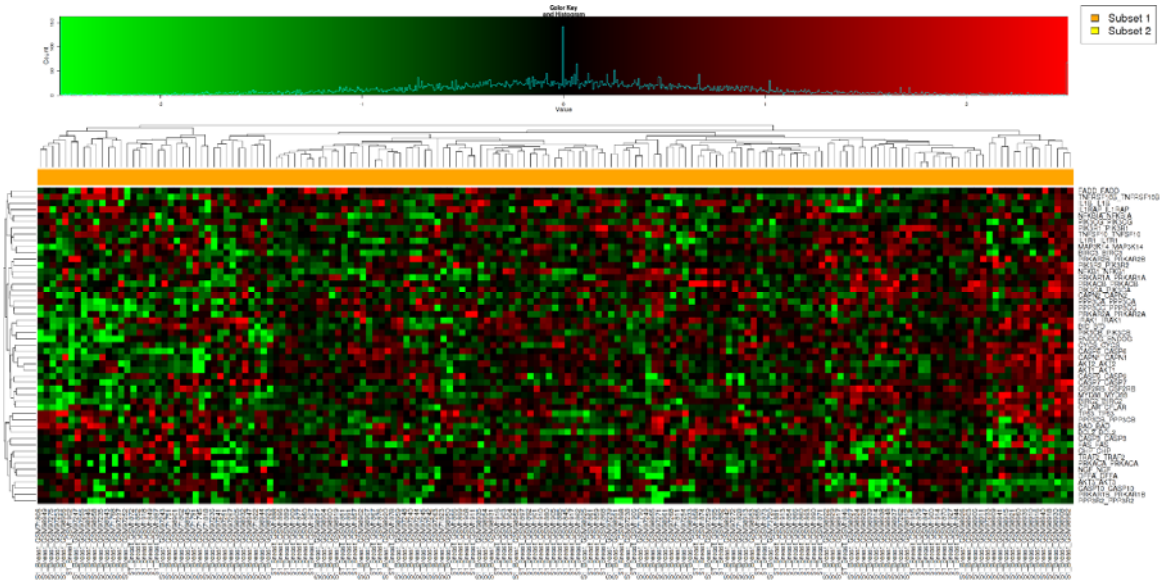
6. Click the **High Dimensional Data** button, add the pathway **Apoptosis** and click **Apply Selections**.





7. Click the **Run** button.

8. Click the Heatmap generated and visualize the image.



9. Click the link **Download raw R data** to export data.

K-Means Clustering

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

To generate a K-Means clustering Heatmap:

1. Run tranSMART, under Analyze tab, select the study of interest.
2. In the Comparison tab, define one or two cohorts whose data points will be represented in the Heatmap visualization.
3. Click the **Advanced Workflow** tab.
4. Click **Analysis > K-Means Clustering**.
5. In the Variable Selection box, drag and drop the high dimensional data node.
6. Click the **High Dimensional Data** button and add a pathway (or other object with multiple rows).
7. Click the **Run** button.

Example

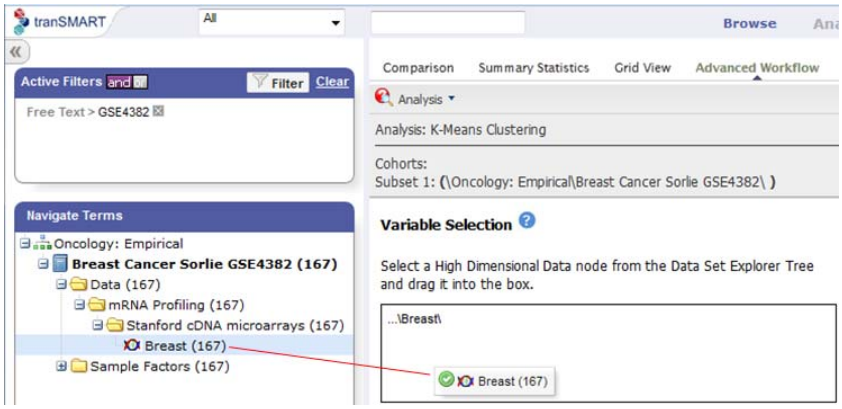
1. Go to the Analyze tab, select the search criteria **Free Text**, write in the box: **GSE4382** and 'Enter key' on your keyboard.

2. Under the Program, open the Study '**Breast Cancer Sorlie GSE4382**'.

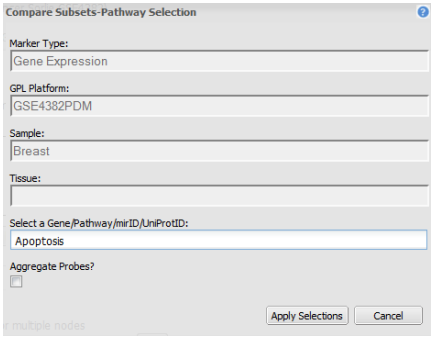
3. In the Comparison tab, drag and drop the Study '**Breast Cancer Sorlie GSE4382**' as Subset 1 cohort.

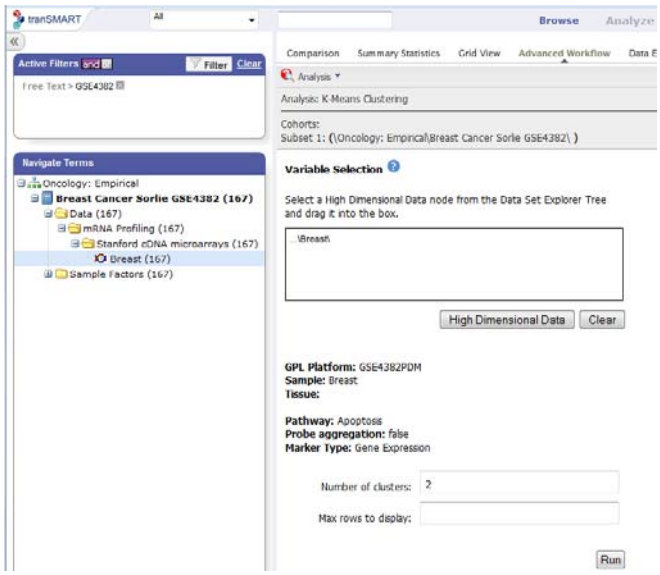
4. Go to the **Advanced Workflow**, then select **K-Means Clustering** from the Analysis dropdown menu.

5. Drag and drop the node '\Breast Cancer Sorlie GSE4382\Data\...**Breast**' under Variable Selection.



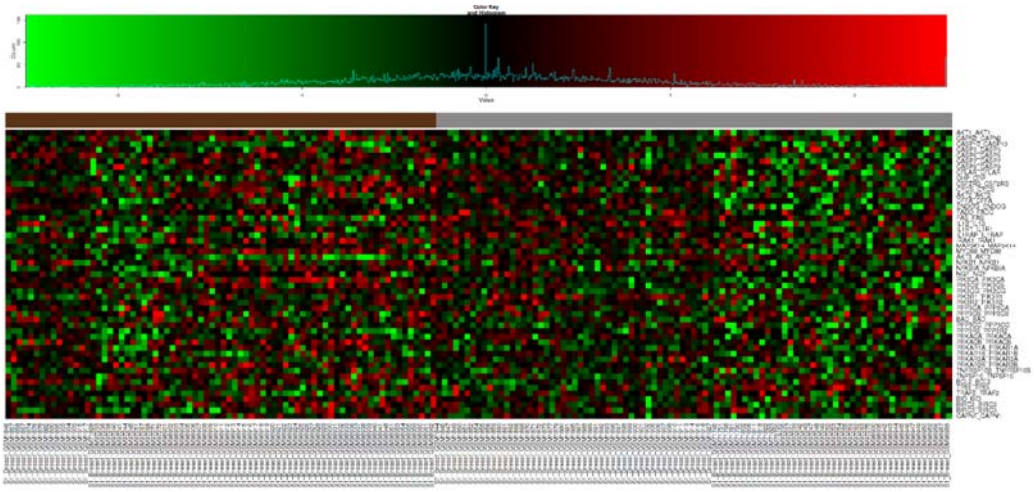
6. Click the **High Dimensional Data** button, add the pathway **Apoptosis** and click **Apply Selections**.





7. Click the **Run** button.

8. Click the Heatmap generated and visualize the image.



9. Click the link **Download raw R data** to export data.

Marker Selection

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

To generate a Marker Selection Heatmap:

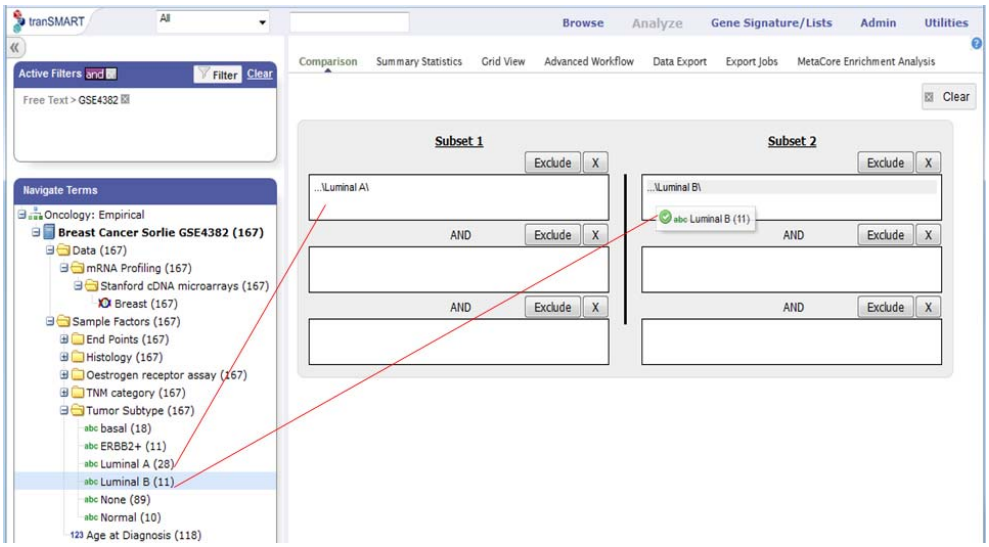
1. Run tranSMART, under Analyze tab, select the study of interest.
2. In the Comparison tab, define the two cohorts whose data points will be represented in the Heatmap visualization.
3. Click the **Advanced Workflow** tab.
4. Click **Analysis > Marker Selection**.
5. In the Variable Selection box, drag and drop the high dimensional data node.
6. Click the **High Dimensional Data** button and add a pathway (or other object with multiple rows).
7. Click the **Run** button.

Example

1. Go to the Analyze tab, select the search criteria **Free Text**, write in the box: **GSE4382** and 'Enter key' on your keyboard.

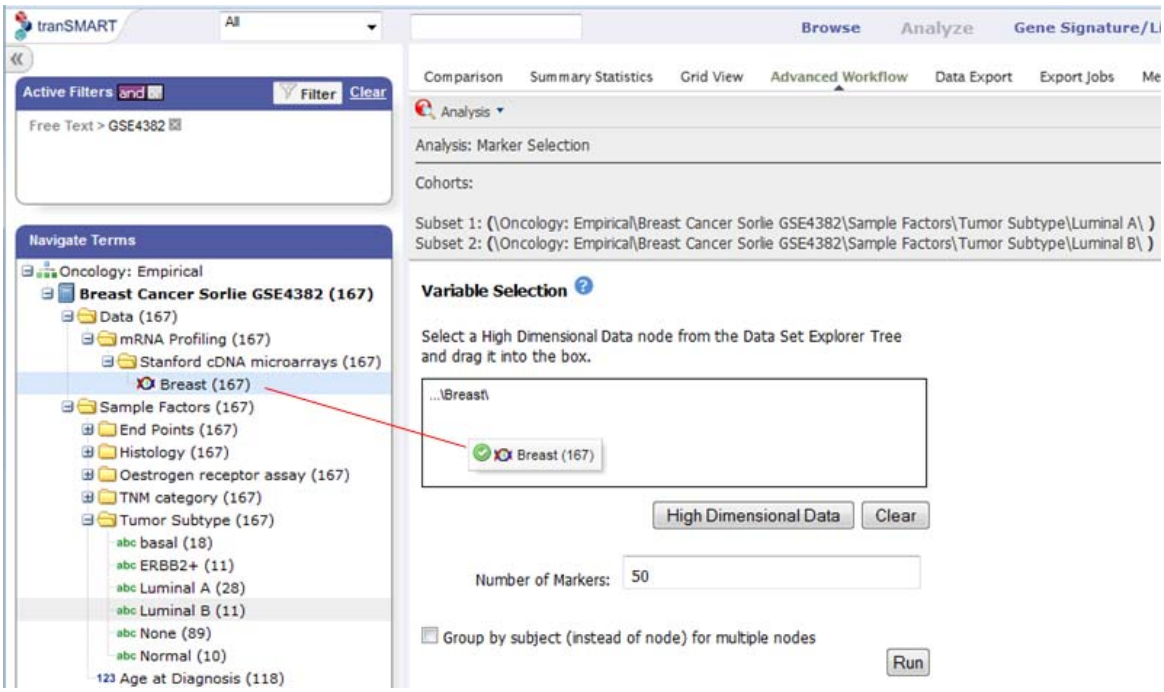
2. Under the Program, open the Study '**Breast Cancer Sorlie GSE4382**'.

3. In the Comparison tab, drag and drop the Study '\Breast Cancer Sorlie GSE4382\Sample Factors\Tumor Subtype**Luminal A**' as Subset 1 cohort and '\Breast Cancer Sorlie GSE4382\Sample Factors\Tumor Subtype**Luminal B**' as Subset 2 cohort.

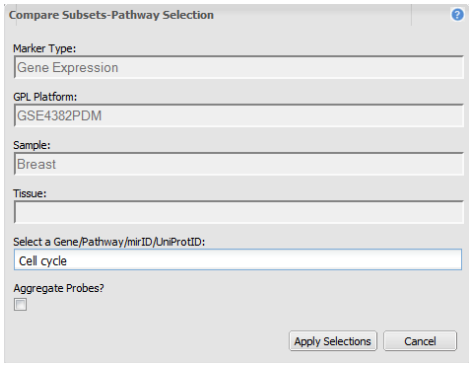


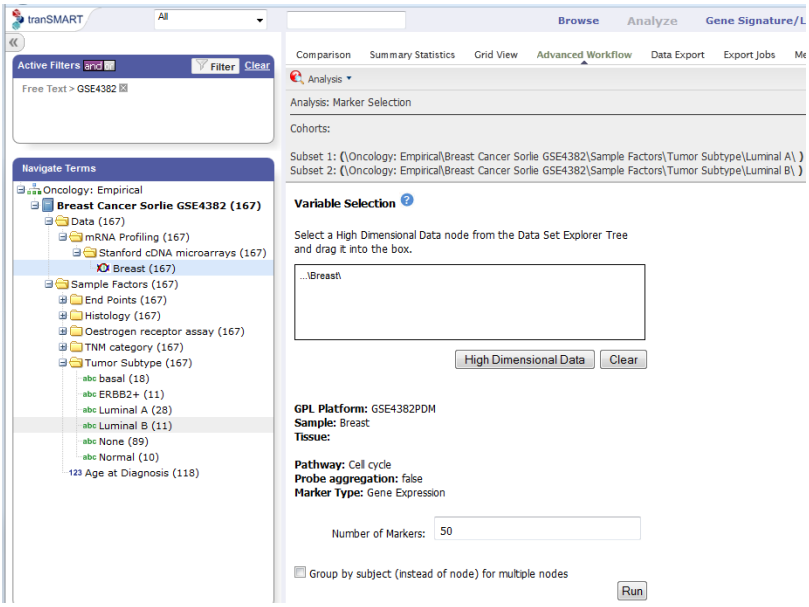
4. Go to the **Advanced Workflow**, then select **Marker Selection** from the Analysis dropdown menu.

5. Drag and drop the node '\Breast Cancer Sorlie GSE4382\Data\...**Breast**' under Variable Selection.

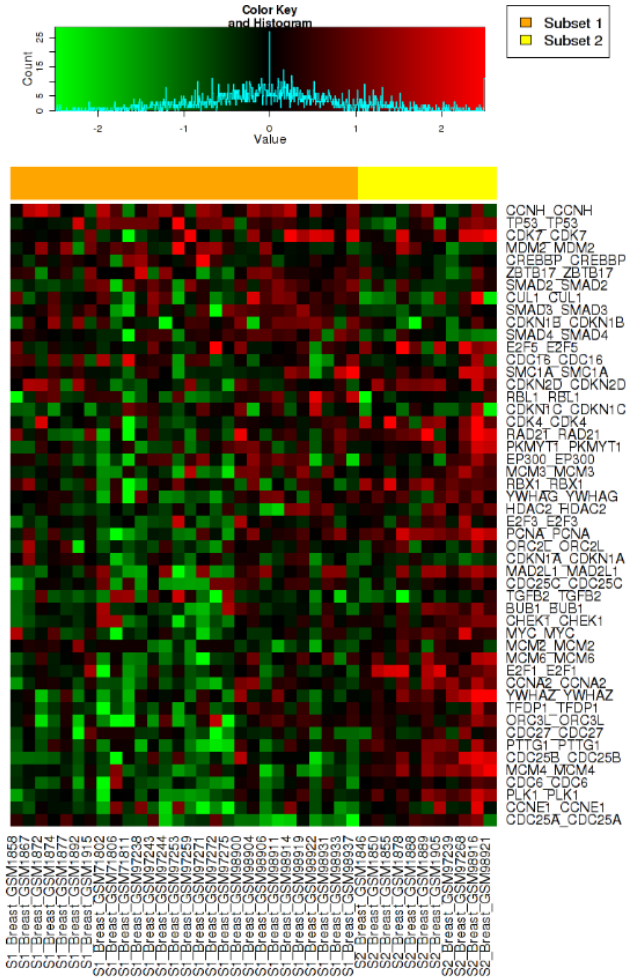


6. Click the **High Dimensional Data** button, add the pathway **Cell cycle** and click **Apply Selections**.





7. Click the **Run** button.



8. Click the Heatmap generated and visualize the image.

9. Click the link **Download raw R data** to export data.

Tips

For Marker Selection, if no pathway/gene/mirID/UniProdID needs to be selected, click directly the Run button to by-pass the High Dimensional Data pop-up!

Lesson 13 – Generating a PCA

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

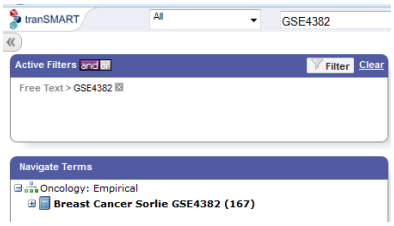
Lesson Goal: Become acquainted with the PCA analysis.

To generate a PCA visualization:

1. Run tranSMART, under Analyze tab, select the study of interest.
2. In the Comparison tab, define one or two cohorts whose data points will be represented in the PCA visualization.
3. Click the **Advanced Workflow** tab.
4. Click **Analysis > Principal Component Analysis**.
5. In the Variable Selection box, drag and drop the high dimensional data node.
6. Click the **High Dimensional Data** button and add a pathway (or other object with multiple rows).
7. Click the **Run** button.

Example

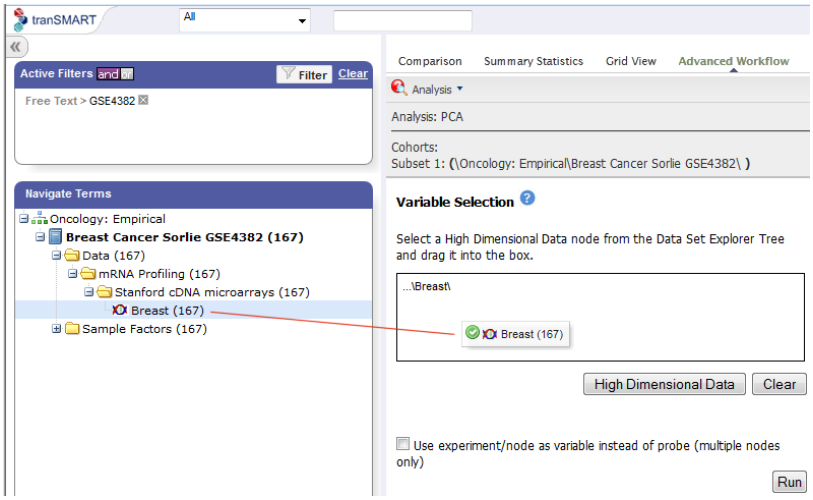
1. Go to the Analyze tab, select the search criteria **Free Text**, write in the box: **GSE4382** and 'Enter key' on your keyboard.
2. Under the Program, open the Study '**Breast Cancer Sorlie GSE4382**'.



3. In the Comparison tab, drag and drop the Study '**Breast Cancer Sorlie GSE4382**' as Subset 1 cohort.

4. Go to the **Advanced Workflow**, then select **PCA** from the Analysis dropdown menu.

5. Drag and drop the node '\Breast Cancer Sorlie GSE4382\Data\...**Breast**' under Variable Selection.

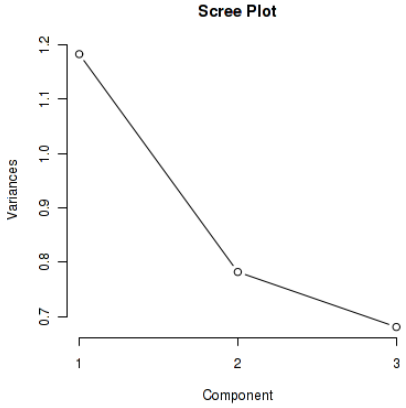


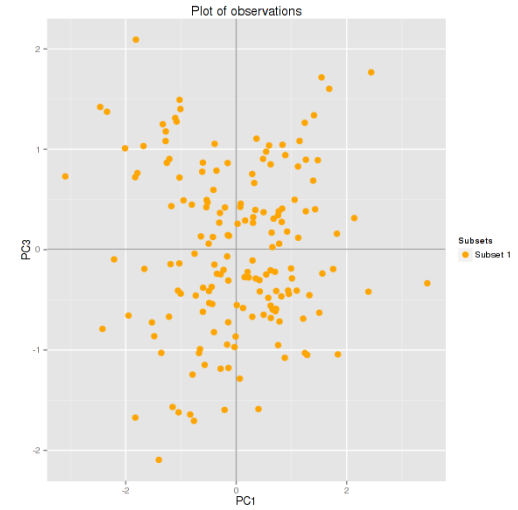
6. Click the **High Dimensional Data** button, add the pathway **Vitamin B6 metabolism** and click **Apply Selections**.

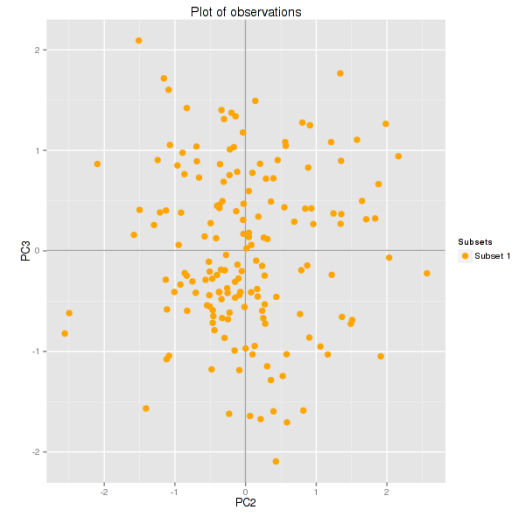
7. Click the **Run** button.

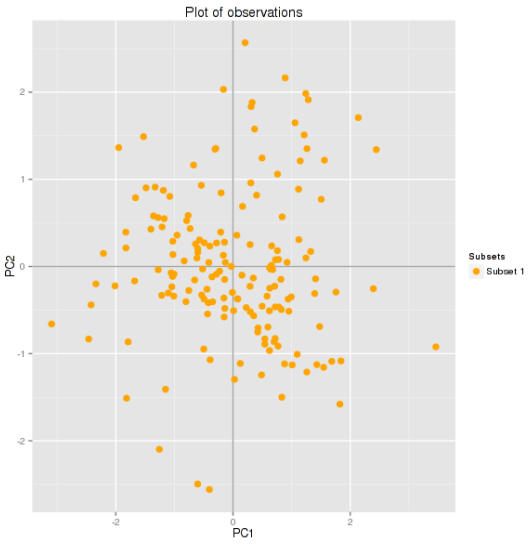
Component Summary

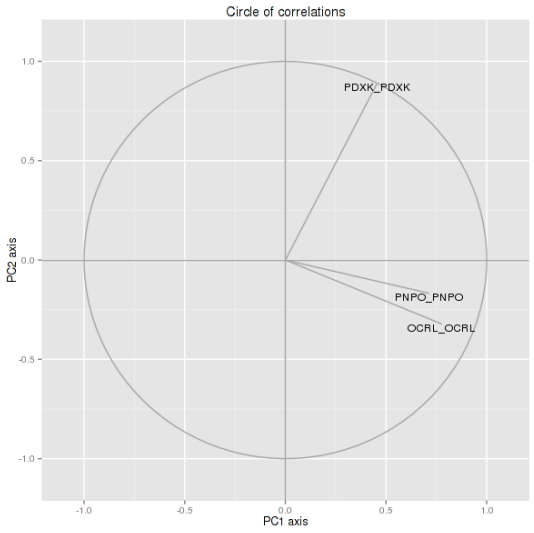
Primary Component	Eigen Value	Percent Variance
PC1	1.18313	44.70749
PC2	0.78225	29.55932
PC3	0.681	25.73319

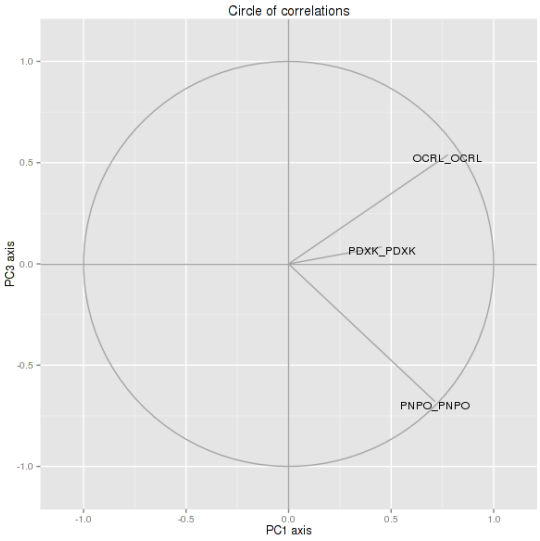


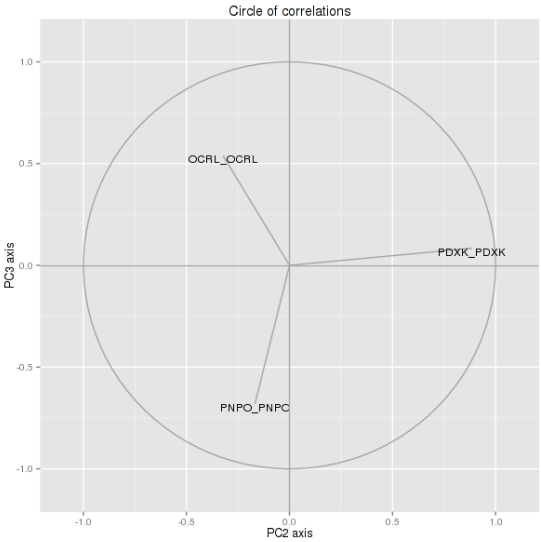

Plot of observations
PC3
PC1
Subsets
Subset 1


Plot of observations
PC3
PC2
Subsets
Subset 1


Plot of observations
PC2
PC1
Subsets
Subset 1


Circle of correlations
PDXK_PDXK
PNPO_PNPO
OCRL_OCRL
PC2 axis
PC1 axis
1.0
0.5
0.0
-0.5
-1.0


Circle of correlations
OCRL_OCRL
PDXK_PDXK
PNPO_PNPO
PC3 axis
PC1 axis
1.0
0.5
0.0
-0.5
-1.0


Circle of correlations
OCRL_OCRL
PDXK_PDXK
PNPO_PNPC
PC3 axis
PC2 axis
1.0
0.5
0.0
-0.5
-1.0

Gene list by proximity to Component

Component 1		Component 2		Component 3	
OCRL_OCRL	0.691	PDXK_PDXK	0.919	PNPO_PNPO	-0.77
PNPO_PNPO	0.613	OCRL_OCRL	-0.352	OCRL_OCRL	0.631
PDXK_PDXK	0.383	PNPO_PNPO	-0.177	PDXK_PDXK	0.094

8. Click the link **Download raw R data** to export data.

Tips

Reminder:

High dimensional data analytics				
1 or 2 cohorts				2 cohorts
Heatmap	Hierarchical Clustering	K-means clustering	PCA	Marker Selection

Analytics restricted to 1 cohort					
using high or low dim variables					using low dim variables
Survival Analysis	Box Plot with ANOVA	Line graph	Scatter plot with linear regression	Table with Fisher Test	Correlation analysis

Lesson 14 – Generating a Survival Analysis

Survival analysis is generally defined as a set of methods for analyzing data where the outcome variable is the time until the occurrence of an event of interest. The time to event or survival time can be measured in days, weeks, years, etc. The Kaplan–Meier estimator is an estimator for estimating the survival function from lifetime data.
It is often used to measure the fraction of patients living for a certain amount of time after treatment.

Lesson Goal: Become acquainted with the Survival analysis.

To generate a survival analysis:

1. Run tranSMART, under Analyze tab, select the study of interest.
2. In the Comparison tab, define one cohort you wish to analyze by dragging one or more concepts from a study into empty subset 1 definition boxes for the survival analysis.
3. Click the **Advanced Workflow** tab.
4. Click **Analysis > Survival Analysis**.
5. In the Time box, drag and drop the survival time node.
6. In the Category box, drag and drop the categorical variables.

Optional: Drag and drop a categorical censoring variable in the censoring variable box. 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.

7. Click the **Run** button.

Example

1. Go to the Analyze tab, select the search criteria **Free Text**, write in the box: **GSE4382** and 'Enter key' on your keyboard.

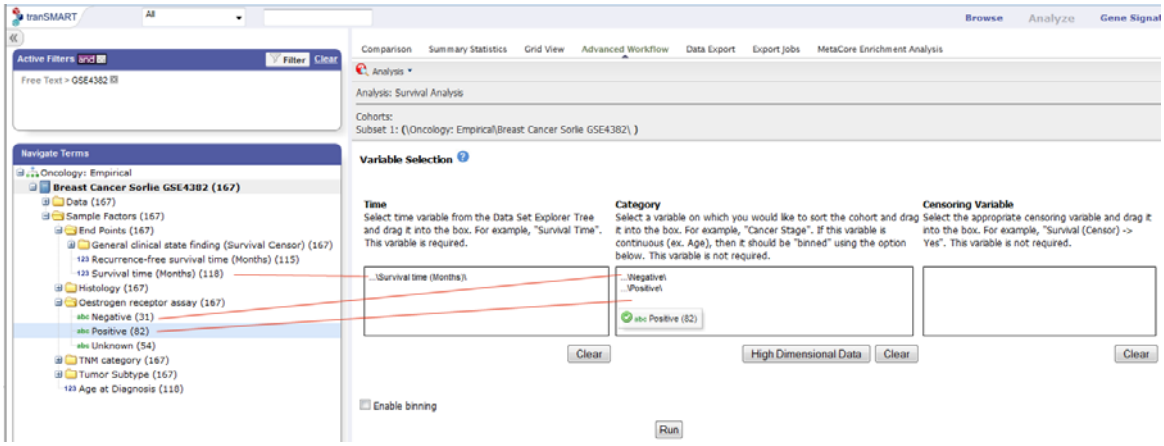
2. Under the Program, open the Study '**Breast Cancer Sorlie GSE4382**'.

3. In the Comparison tab, drag and drop the Study '**Breast Cancer Sorlie GSE4382**' as Subset 1 cohort.

4. Click the **Advanced Workflow** tab, then select **Survival Analysis** from the Analysis dropdown menu.

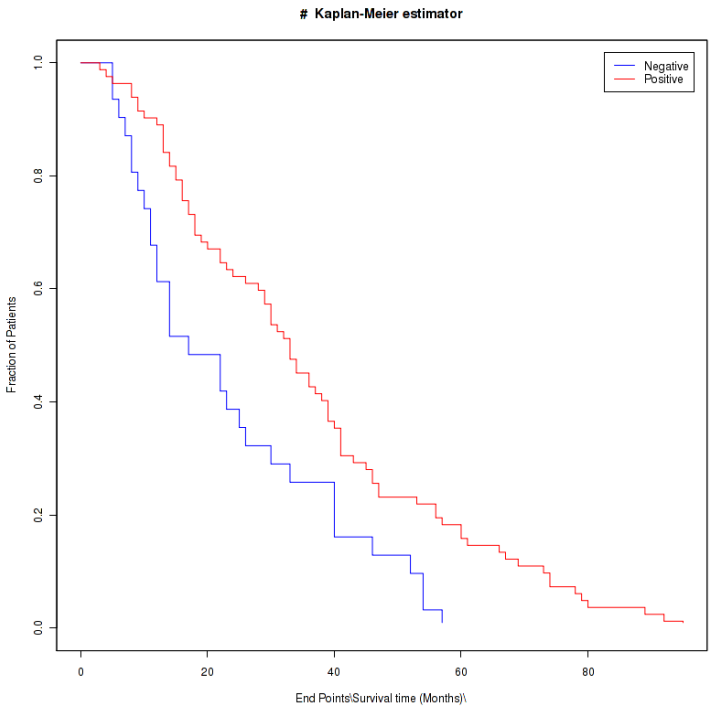
5. Drag and drop the node '\Breast Cancer Sorlie GSE4382\Sample Factors\End Points**Survival time (Months)**' under the Time box.

6. Drag and drop the '\Breast Cancer Sorlie GSE4382\Sample Factors\Oestrogen receptor assay**Negative**' and '\Breast Cancer Sorlie GSE4382\Sample Factors\Oestrogen receptor assay**Positive**' under the Category box.



7. Click the **Run** button.

Survival Curve



Cox Regression Result

Number of Subjects	113
Number of Events	113
Likelihood ratio test	8.53 on 1 df, p=0.003485
Wald test	9.34 on 1 df, p=0.002238
Score (logrank) test	9.69 on 1 df, p=0.001854

Subset	Cox Coefficient	Hazards Ratio	Lower Range of Hazards Ratio, 95% Confidence Interval	Upper Range of Hazards Ratio, 95% Confidence Interval
Positive	-0.6706	0.5114	0.3327	0.7862

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
Negative	31	31	31	31	17	12	33
Positive	82	82	82	82	33	29	39

8. Click the link **Download raw R data** to export data.

Lesson 15 – Generating a Box Plot

A Box Plot with ANOVA analysis displays the distribution of a numerical variable as a box and whisker plot in several groups of subjects, and performs the corresponding analysis of variance.

Lesson Goal: Become acquainted with the Box Plot with ANOVA analysis.

To perform a boxplot with ANOVA analysis:

1. Run tranSMART, under Analyze tab, select the study of interest.

2. In the Comparison tab, define one cohort you wish to analyze by dragging one or more concepts from a study into empty subset 1 definition boxes for the Box Plot analysis.

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

4. Click **Analysis > Box Plot with ANOVA**.

   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, then boxes will be plotted horizontally. If the *dependent variable*, on the other hand, defines the groups, boxes will be plotted vertically.

5. In the Independent Variable box, drag and drop a numerical (or categorical) variable.

6. In the Dependent Variable box, drag and drop a categorical (or numerical) variable.

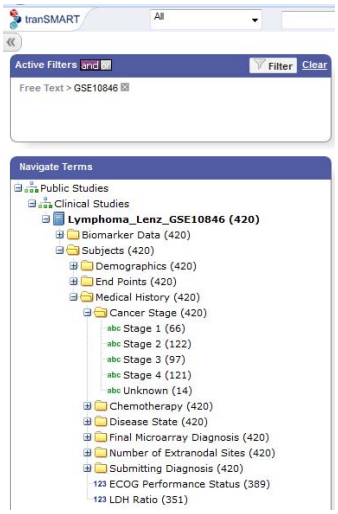
   If two numerical variables have been chosen, one must be categorized using binning. See Appendix A – Additional Material>Data Binning.

7. Click the **Run** button.

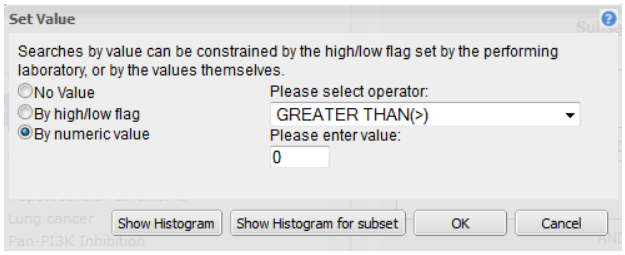
Example

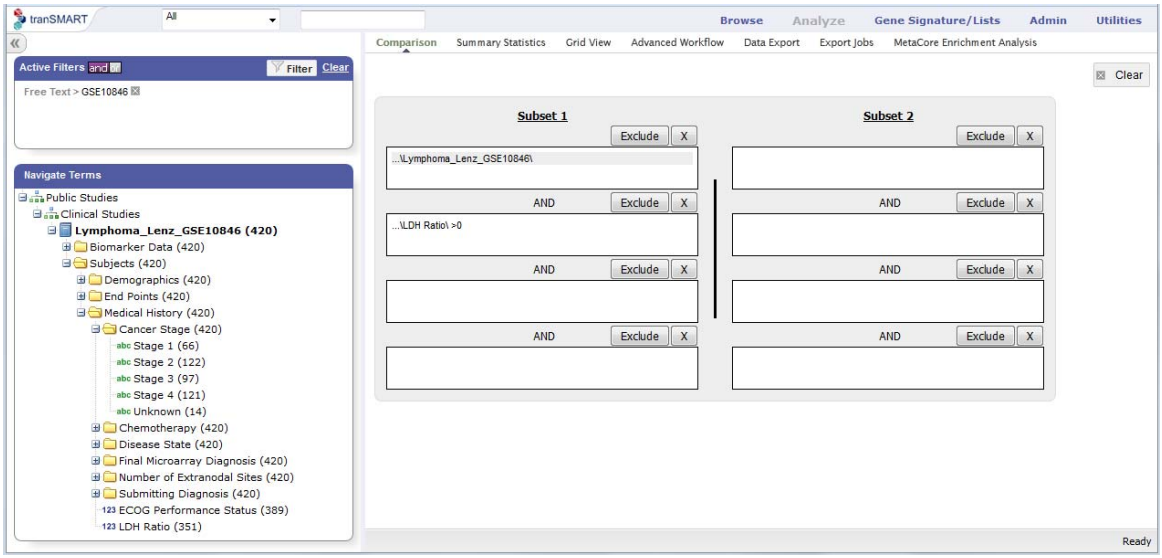
1. Go to the Analyze tab, select the search criteria **Free Text**, write in the box: **GSE4382** and 'Enter key' on your keyboard.

2. Under the Program, open the nested tree of the Study '**Lymphoma_Lenz_GSE10846**'.



3. In the Comparison tab, drag and drop the Study '**Lymphoma_Lenz_GSE10846**' into a subset definition box in Subset 1 and the node **LDH Ratio** >0 in another Subset 1 box.

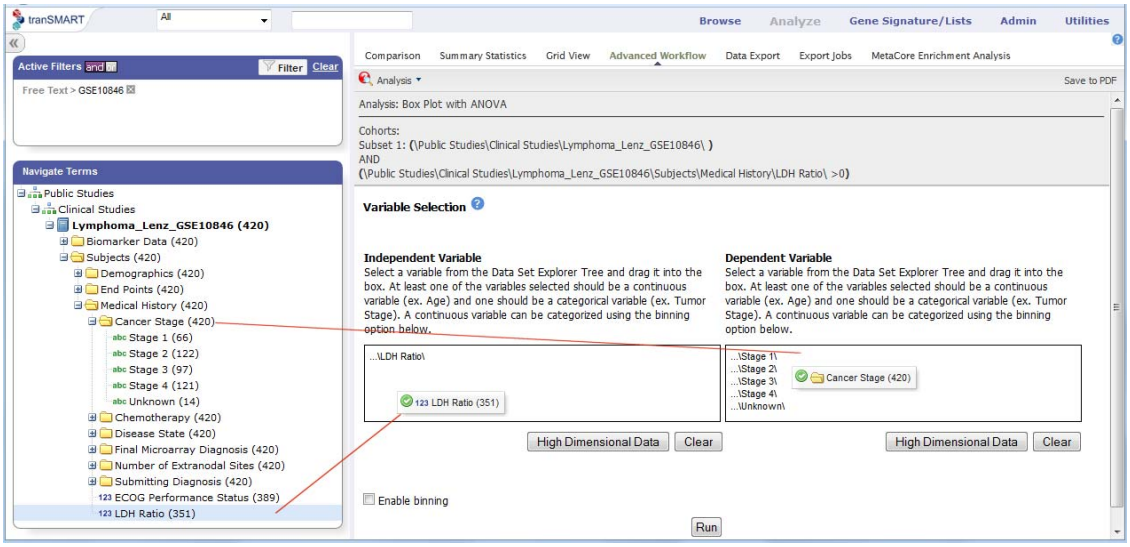




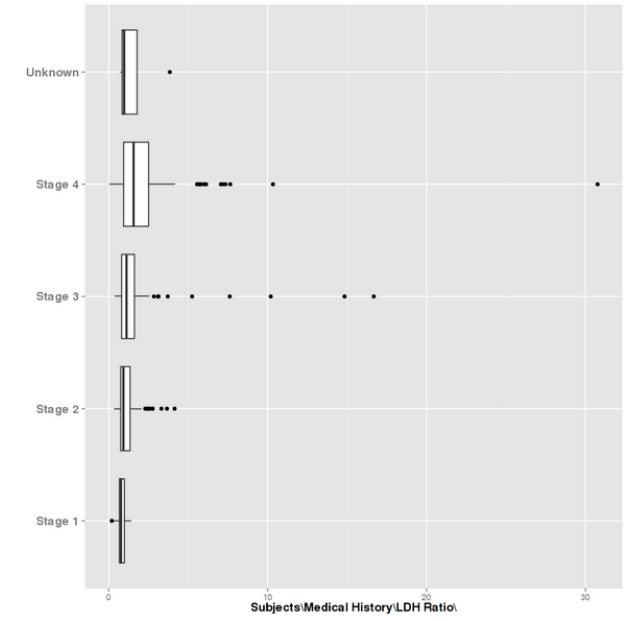
4. Click the **Advanced Workflow** tab, then select **Box Plot with ANOVA** from the Analysis dropdown menu.

5. Drag and drop the node **LDH Ratio** into the Independent Variable box.

6. Drag and drop the node **Cancer Stage** into the Dependent Variable box.



7. Click the **Run** button.



ANOVA Result

p-value	7.5e-05
F value	6.24

Group	Mean	n
Stage 1	0.821	54
Stage 2	1.177	103
Stage 3	1.836	82
Stage 4	2.476	101
Unknown	1.645	4

Pairwise t-Test p-Values

	Stage 1	Stage 2	Stage 3	Stage 4
Stage 2	3.58e-01	NA	NA	NA
Stage 3	1.22e-02	5.36e-02	NA	NA
Stage 4	2.54e-05	6.75e-05	6.20e-02	NA
Unknown	4.89e-01	6.90e-01	8.71e-01	4.79e-01

Download raw R data

8. Click the link **Download raw R data** to export data.

Lesson 16 – Generating a Line Graph

In a line graph analysis, you can plot serial data (time series, dose response or series of conditions) for 1 or several subsets of subjects in a study (for example, treatment groups).

Lesson Goal: Become acquainted with the Line Graph analysis.

To perform a Line Graph:

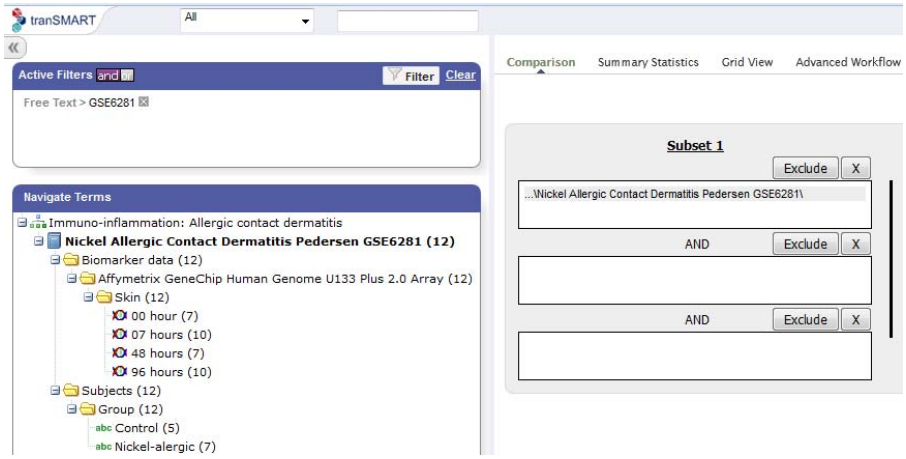
1. Run tranSMART, under Analyze tab, select the study of interest.
2. In the Comparison tab, define one cohort you wish to analyze by dragging one or more concepts from a study into empty subset 1 definition boxes for the Line Graph.
3. Click the **Advanced Workflow** tab.
4. Click **Analysis > Line Graph**.
5. In the Time/Measurement Concepts box, drag and drop one or multiple numerical or high dimensional nodes (serial measurements).
6. In the Group Concepts box, drag and drop one or multiple categorical nodes. One node Numerical or High Dimensional can be dragged and dropped with binning to be categorical.
7. Click the **Run** button.

Example

1. Go to the Analyze tab, select the search criteria **Free Text**, write in the box: **GSE6281** and 'Enter key' on your keyboard.

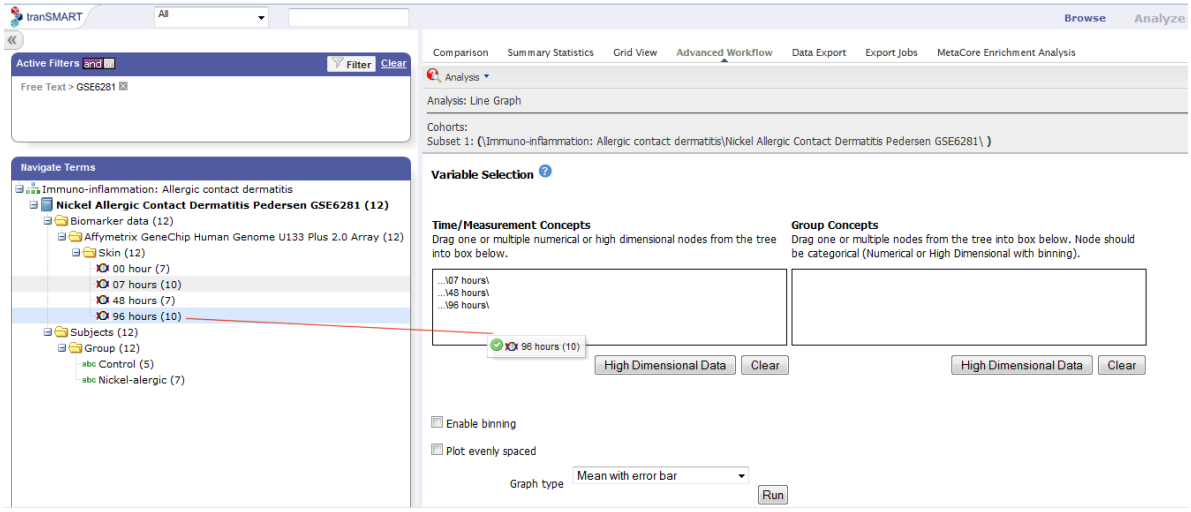
2. Under the Program, open the nested tree of the Study '**Nickel Allergic Contact Dermatitis Pedersen GSE6281**'.

3. In the Comparison tab, drag and drop the Study '**Nickel Allergic Contact Dermatitis Pedersen GSE6281**' as Subset 1 cohort.

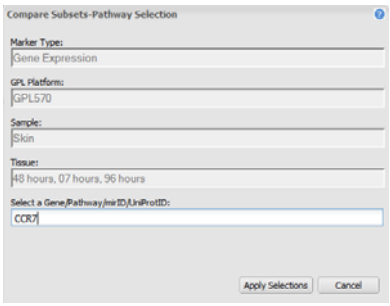


4. Click the **Advanced Workflow** tab, then select **Line Graph** from the Analysis dropdown menu.

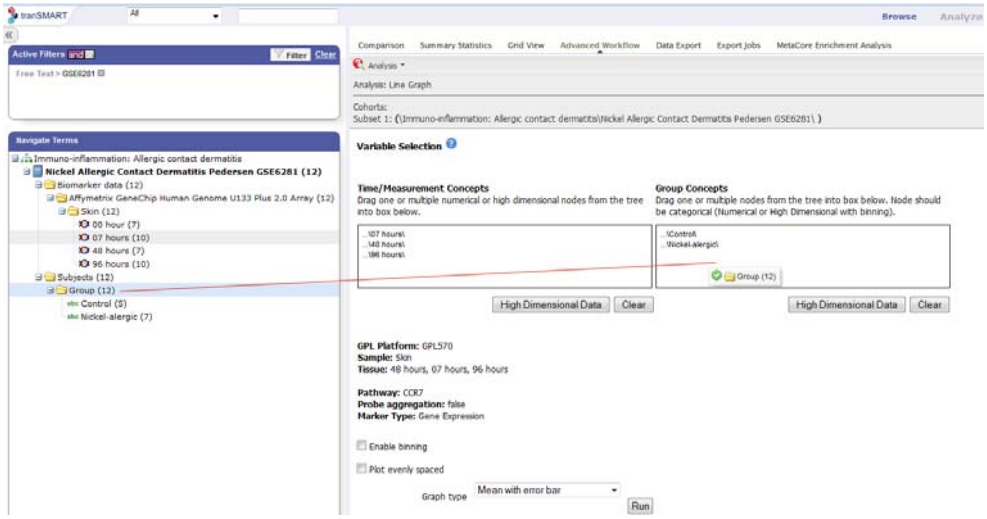
5. Drag and drop the nodes '\Immuno-inflammation: Allergic contact dermatitis\Nickel Allergic Contact Dermatitis Pedersen GSE6281\Biomarker data\Affymetrix GeneChip Human Genome U133 Plus 2.0 Array\Skin**07 hours**', '\...**48 hours**' and '\...**96 hours**' under the Time/Measurement Concepts box.



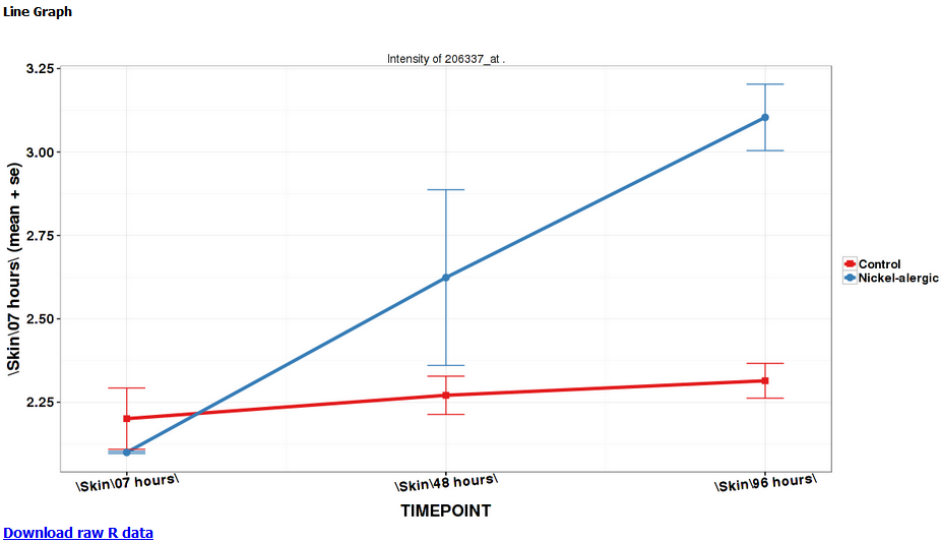
6. Click the **High Dimensional Data** button, add **CCR7** then **Apply Selections**.



7. Drag and drop the node '\Nickel Allergic Contact Dermatitis Pedersen GSE6281\Subjects**Group**' under the Group Concepts box.



8. Click the **Run** button.



9. Click the link **Download raw R data** to export data.

Lesson 17 – Generating a Scatter Plot

A scatter plot displays values for two numerical variables within a dataset, and performs linear regression analysis.

Lesson Goal: Become acquainted with the Scatter Plot analysis.

To perform a scatter plot with linear regression analysis:

1. Run tranSMART, under Analyze tab, select the study of interest.
2. In the Comparison tab, define one cohort you wish to analyze by dragging one or more concepts from a study into empty subset 1 definition boxes for the Scatter Plot.
3. Click the **Advanced Workflow** tab.
4. Click **Analysis > Scatter Plot**.
5. In the Independent Variable box, drag and drop a continuous variable.
6. In the Dependent Variable box, drag and drop another continuous variable.
7. Click the **Run** button.

Example

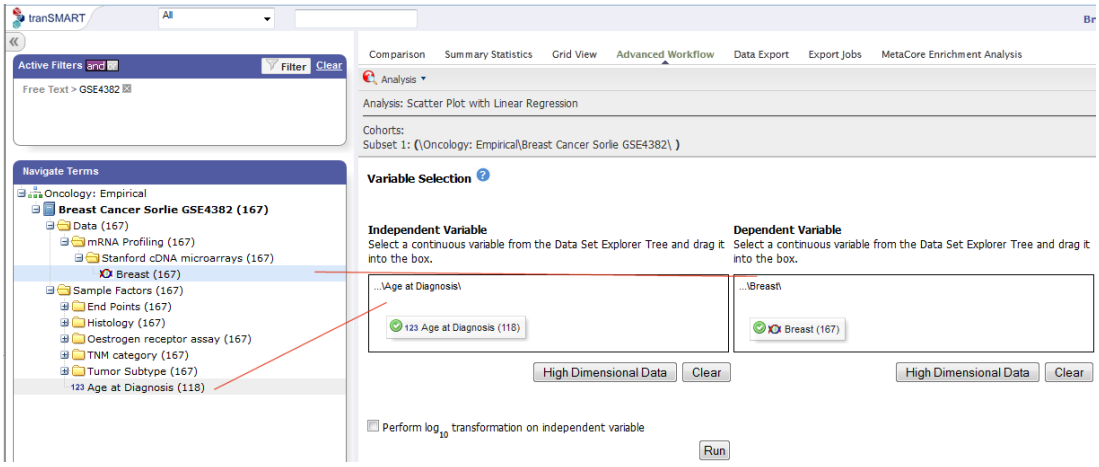
1. Go to the Analyze tab, select the search criteria **Free Text**, write in the box: **GSE4382** and 'Enter key' on your keyboard.
2. Under the Program, open the Study '**Breast Cancer Sorlie GSE4382**'.

3. In the Comparison tab, drag and drop the Study '**Breast Cancer Sorlie GSE4382**' as Subset 1 cohort.

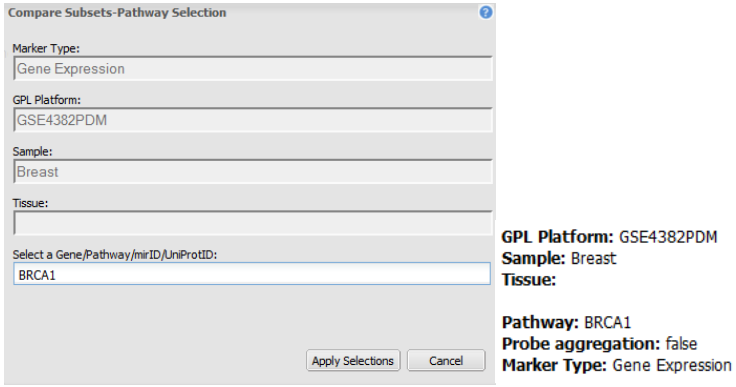
4. Click the **Advanced Workflow** tab, then select **Scatter Plot with Linear Regression** from the Analysis dropdown menu.

5. Drag and drop the node '\Breast Cancer Sorlie GSE4382\Sample Factors**Age at Diagnosis**' under the Independent Variable box.

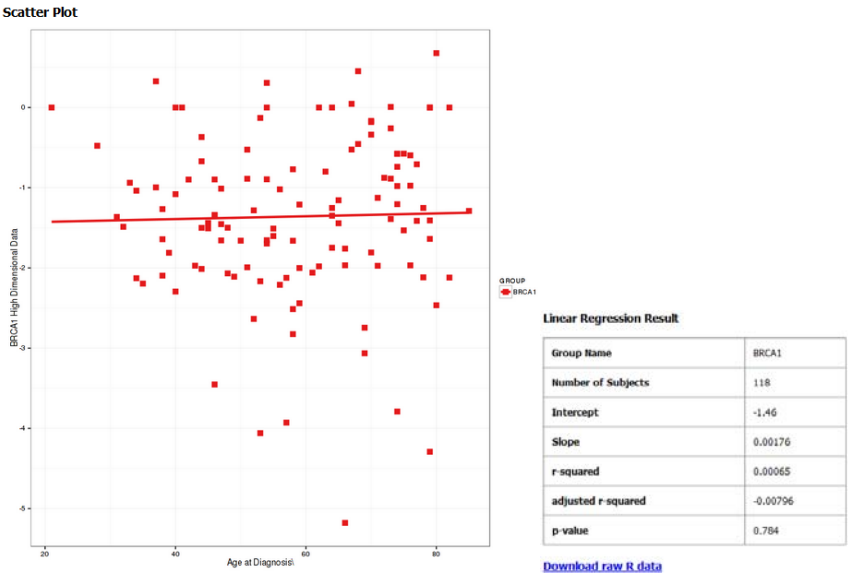
6. Drag and drop the '\Breast Cancer Sorlie GSE4382\Data\mRNA Profiling\Stanford cDNA microarrays**Breast**' under the Dependent Variable box.



7. Click the **High Dimensional Data** button, add **BRCA1** then **Apply Selections**.



8. Click the **Run** button.



9. Click the link **Download raw R data** to export data.

Lesson 18 – Generating a Table with Fisher Test

A Fisher Test analysis compares the classification of subjects based on 2 categorical variables. It displays contingency tables and Fisher test results.

Lesson Goal: Become acquainted with the Fisher Test analysis.

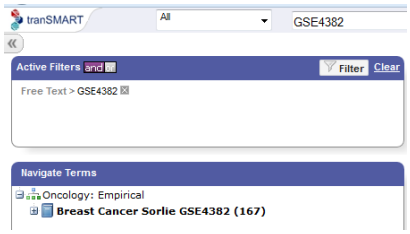
To perform a Fisher Test:

1. Run tranSMART, under Analyze tab, select the study of interest.
2. In the Comparison tab, define one cohort you wish to analyze by dragging one or more concepts from a study into empty subset 1 definition boxes for the.
3. Click the **Advanced Workflow** tab.
4. Click **Analysis > Table with Fisher Test**.
5. In the Independent Variable box, drag and drop a categorical variable.
6. In the Dependent Variable box, drag and drop another categorical variable.
7. Click the **Run** button.

Example

1. Go to the Analyze tab, select the search criteria **Free Text**, write in the box: **GSE4382** and 'Enter key' on your keyboard.

2. Under the Program, open the Study '**Breast Cancer Sorlie GSE4382**'.

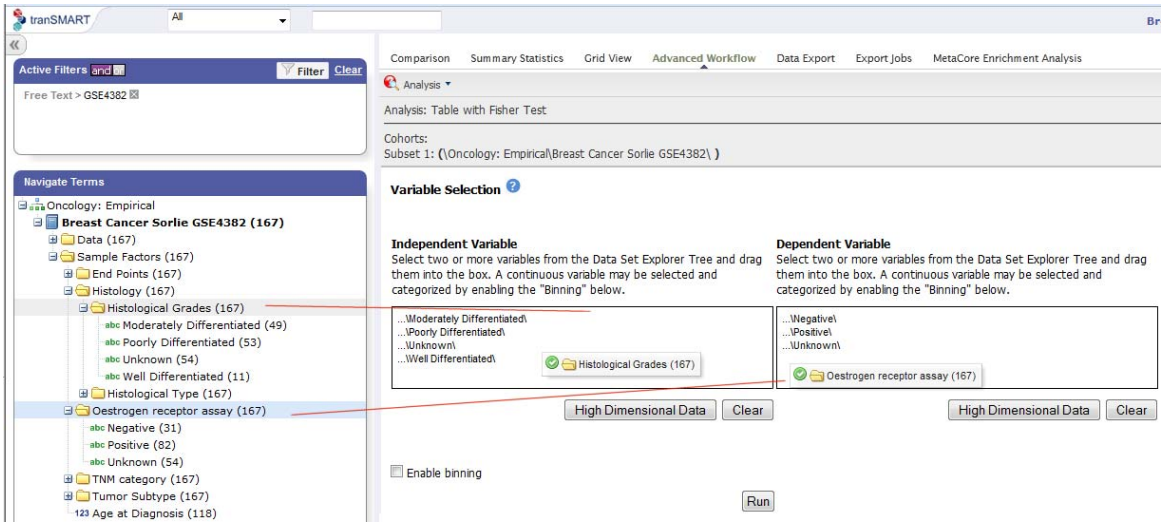


3. In the Comparison tab, drag and drop the Study '**Breast Cancer Sorlie GSE4382**' as Subset 1 cohort.

4. Click the **Advanced Workflow** tab, then select **Table with Fisher Test** from the Analysis dropdown menu.

5. Drag and drop the node '\Breast Cancer Sorlie GSE4382\Sample Factors\Histology**Histological Grades**' under the Independent Variable box.

6. Drag and drop the '\Breast Cancer Sorlie GSE4382\Sample Factors**Oestrogen receptor assay**' under the Dependent Variable box.



7. Click the **Run** button.

	Negative	Positive	Unknown
Moderately Differentiated	8	40	1
Poorly Differentiated	20	32	1
Unknown	1	1	52
Well Differentiated	2	9	0

Fisher test p-value	5e-04
x^2	159
x^2 **p-value**	1.16e-31

Download raw R data

8. Click the link **Download raw R data** to export data.

Lesson 19 – Generating a 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.

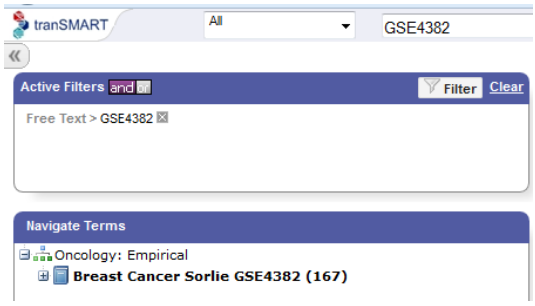
Lesson Goal: Become acquainted with the Correlation Analysis.

To perform a correlation analysis:

1. Run tranSMART, under Analyze tab, select the study of interest.
2. In the Comparison tab, define one cohort you wish to analyze by dragging one or more concepts from a study into empty subset 1 definition boxes.
3. Click the **Advanced Workflow** tab.
4. Click **Analysis > Correlation analysis**.
5. In the Variable Selection box, drag and drop two or more numerical nodes.
6. Choose the Correlation Type: Spearman, Pearson or Kendall.
7. Click the **Run** button.

Example

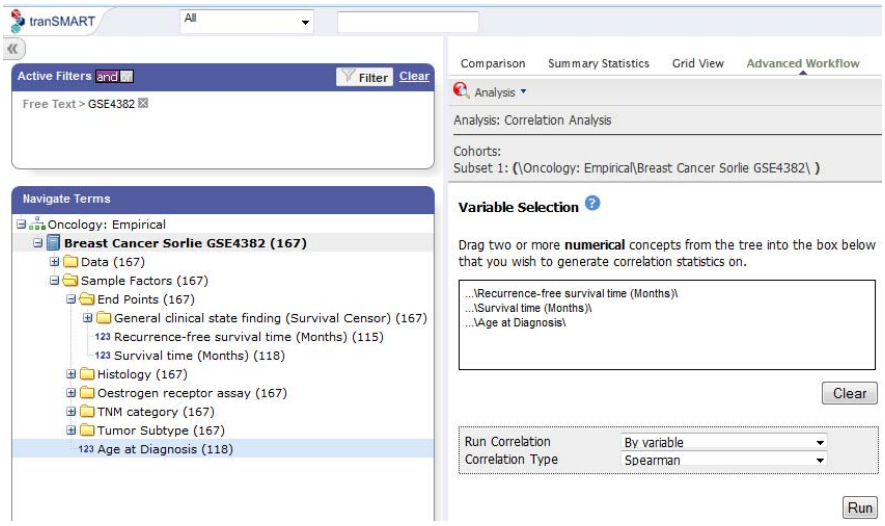
1. Go to the Analyze tab, select the search criteria **Free Text**, write in the box: **GSE4382** and 'Enter key' on your keyboard.
2. Under the Program, open the Study '**Breast Cancer Sorlie GSE4382**'.



3. In the Comparison tab, drag and drop the Study '**Breast Cancer Sorlie GSE4382**' as Subset 1 cohort.

4. Click the **Advanced Workflow** tab, then select **Correlation analysis** from the Analysis dropdown menu.

5. Drag and drop the nodes '\Breast Cancer Sorlie GSE4382\Sample Factors\End Points**Recurrence-free survival time (Months)**', '\Breast Cancer Sorlie GSE4382\Sample Factors\End Points**Survival time (Months)**' and '\Breast Cancer Sorlie GSE4382\Sample Factors**Age at Diagnosis**' in the box.

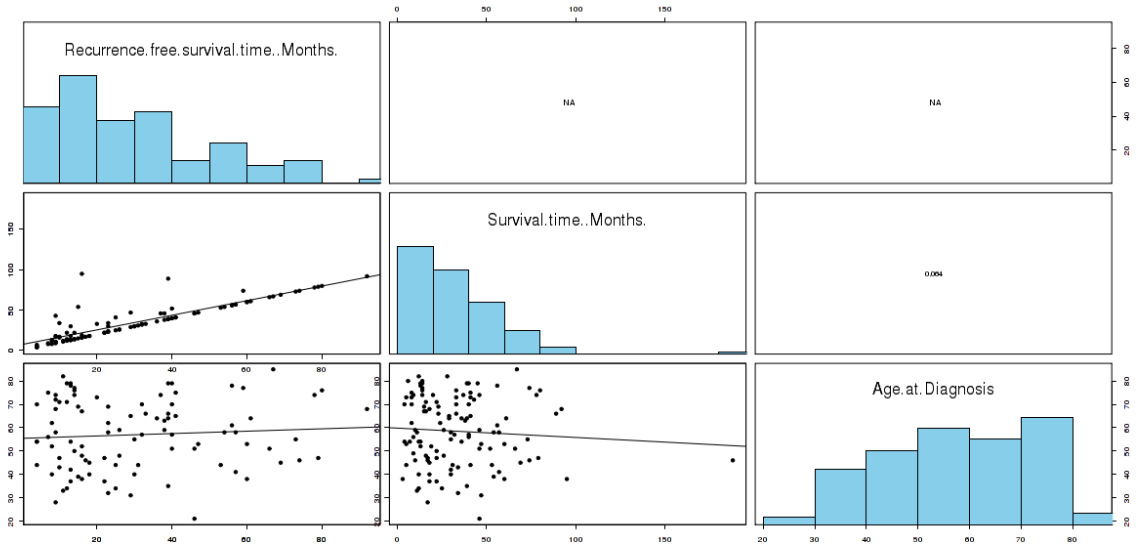


6. Keep the default Correlation Type Spearman.

7. Click the **Run** button.

Correlation Table (p-values on top right half, correlation coefficient on bottom left)

	Recurrence free survival time Months	Survival time Months	Age at Diagnosis
Recurrence free survival time Months	1.000000	0.000000	0.647422
Survival time Months	0.887228	1.000000	0.668716
Age at Diagnosis	0.042536	-0.039800	1.000000

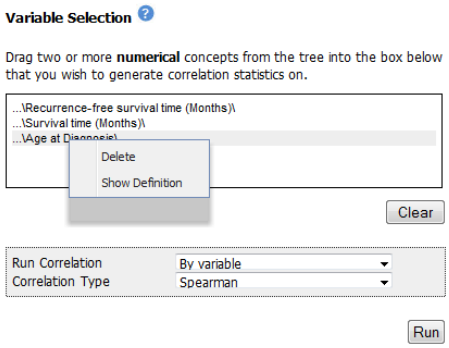


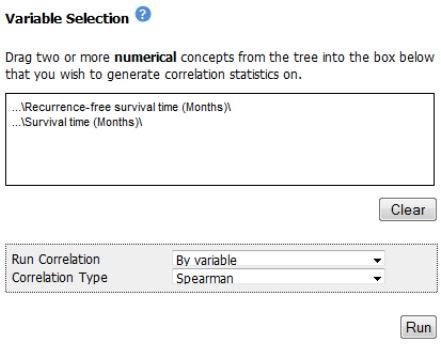
Download raw R data

8. Click the link **Download raw R data** to export data.

Tips

If you have added a node by mistake in the variable selection box, right-click the node and remove it.





Lesson 20 – Exporting data

Subject-level data (including raw data and processed data) can be exported under the Analyze tab. Export pertains only to data related to the predefined cohort(s). tranSMART allows to export all data or only nodes of interest.

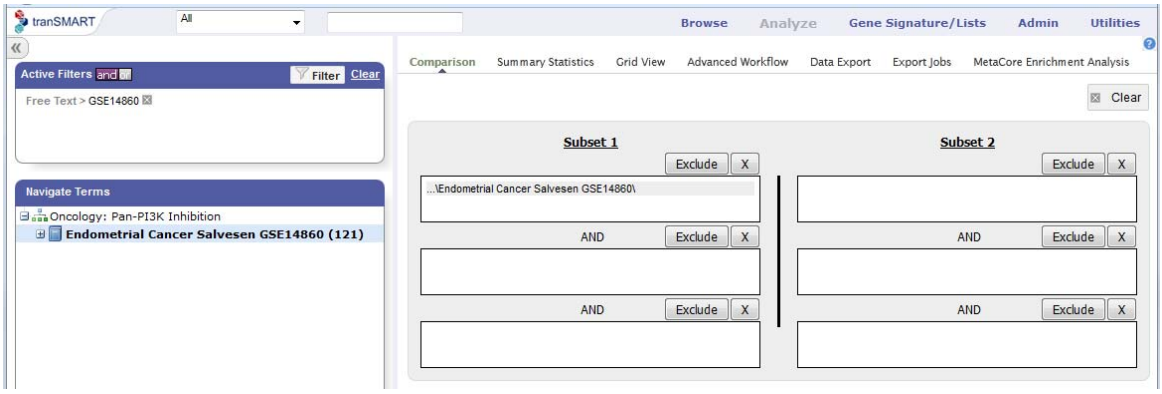
Under Analyze, in the Data Export tab, data available for a study are categorized – e.g., clinical and low dimensional data, gene expression, SNP, etc. (one category for each high dimensional assay type).

Lesson Goals: (1) Export data for a Study using the check boxes to indicate the data types and file formats that are desired (2) Filter data to export by dragging and dropping some node onto each data type row (3) Use cohorts to export a subset of data (4) Export of SNP data.

Scenario (1): You want to export all 'Clinical and Low Dimensional' data and 'mRNA (microarray)' data of the Study 'Endometrial Cancer Salvesen GSE14860'.

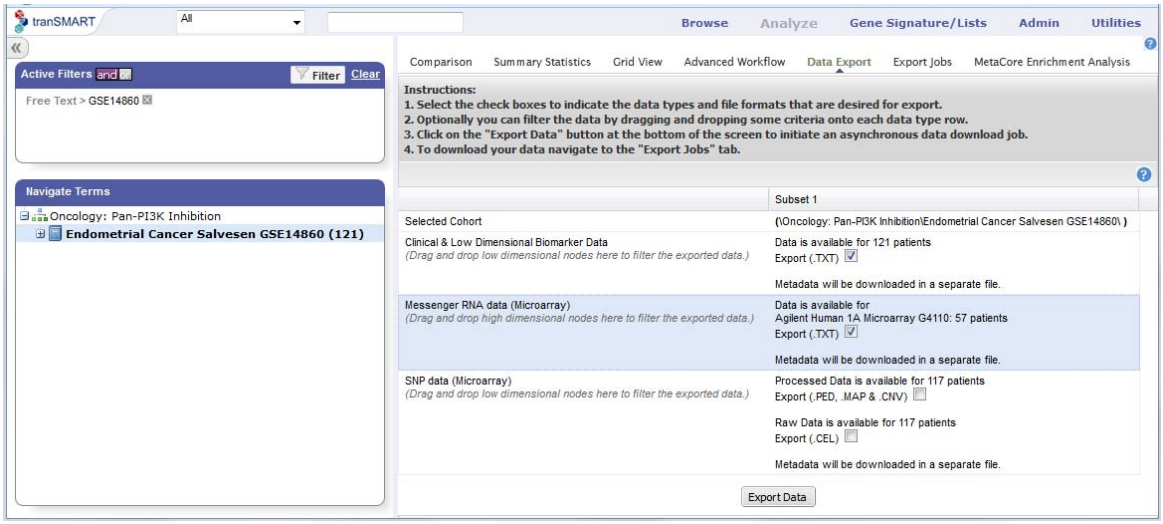
1. Go to the Analyze tab, select the search criteria **Free Text**, write in the box: **GSE14860** and 'Enter key' on your keyboard.

2. In the Comparison tab, drag and drop the Study '**Endometrial Cancer Salvesen GSE14860'** as Subset 1 cohort.

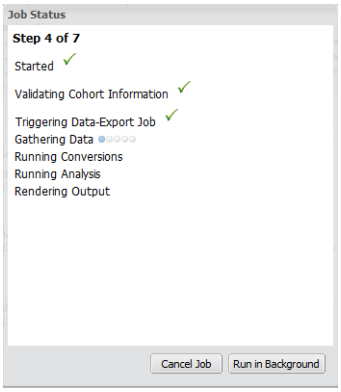


3. Go to the **Data Export** tab.

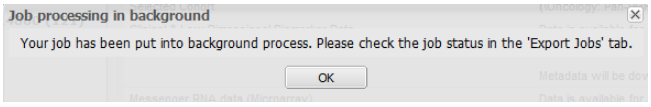
4. Activate the check boxes of 'Clinical and Low Dimensional Biomarker data' and 'Messenger RNA data (microarray)'.



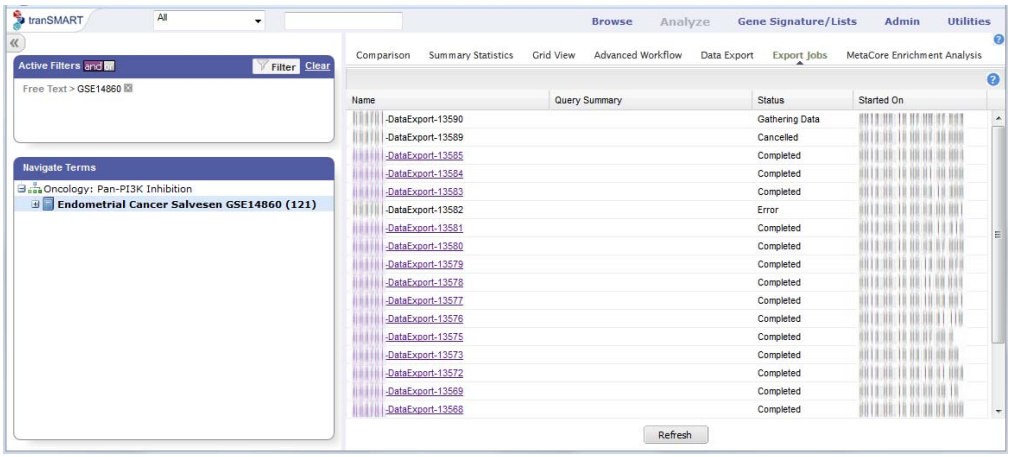
5. Click the **Export Data** button.



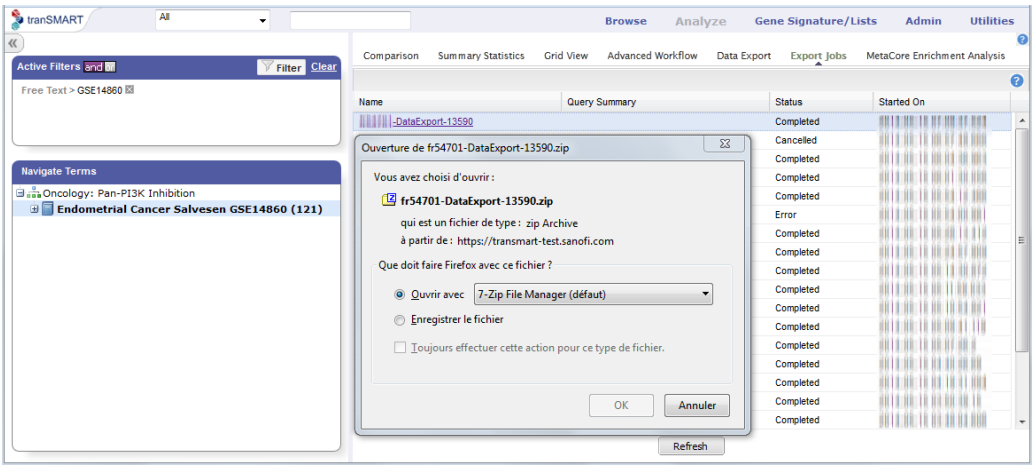
6. Click the button **Run in background** when the pop-up window **Job Status** appears. Then **OK**.

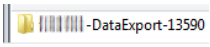


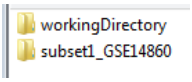
7. Click the **Export Jobs** tab.

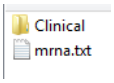


8. Click the **Refresh** button if the status of the job is not completed.
9. Click the link when the status is completed.

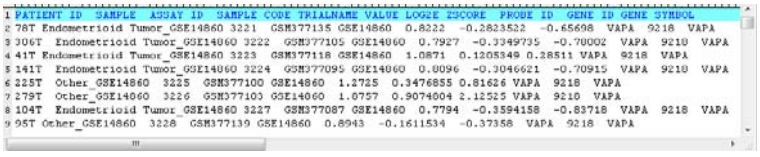


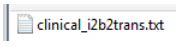




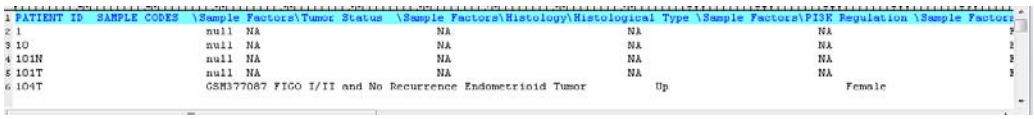


mrna.txt: Gene expression data (intensities, log2-transformed intensities, z-scores) exported for 'endometrioid' and 'other' tumors.





clinical_i2b2trans.txt: All clinical & LDD data exported.



Scenario (2): You want to filter your export of 'Clinical and Low Dimensional Biomarker data' and 'Messenger RNA data (microarray) data' for the Study 'Endometrial Cancer Salvesen GSE14860', by dragging and dropping some criteria onto each data type row.

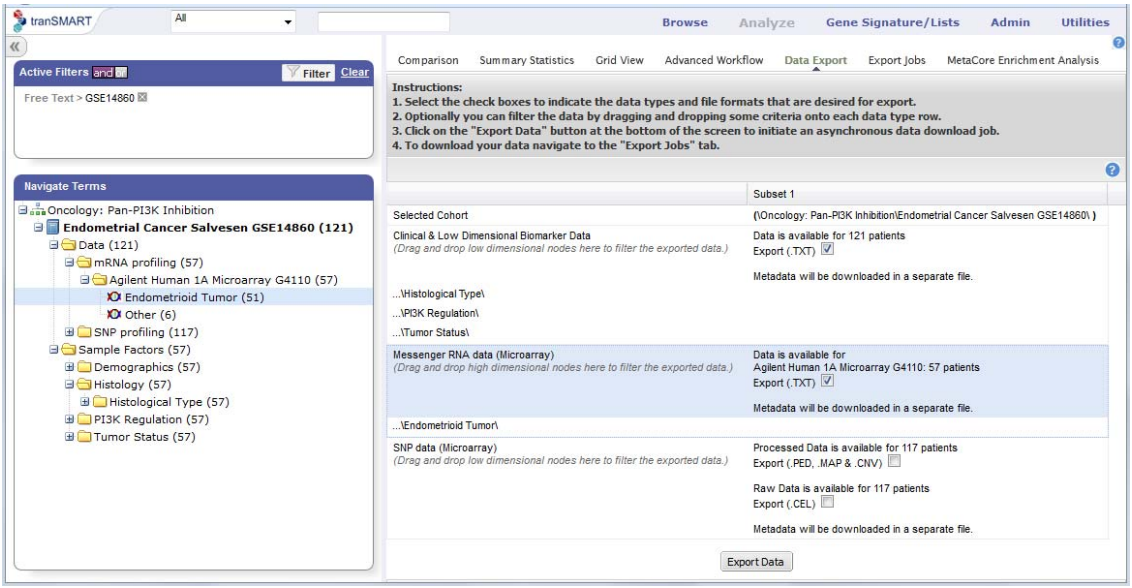
10. Return to the **Data Export** tab

11. Keep the check boxes of 'Clinical and Low Dimensional Biomarker data' and 'Messenger RNA data (microarray) data' activated.

12. In the Navigate Terms area, open the nested nodes of the Study 'Endometrial Cancer Salvesen GSE14860'.

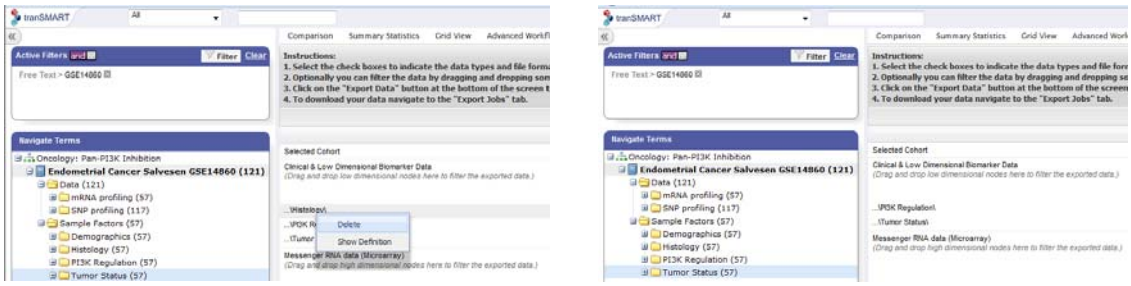
13. Drag and drop the nodes '\Endometrial Cancer Salvesen GSE14860\Sample Factors\Histology**Histology Type**', '\...**PI3K Regulation**' and '\...**Tumor Status**' in the 'Clinical & Low Dimensional Biomarker Data' area.

14. Drag and drop the node '\Endometrial Cancer Salvesen GSE14860\Data\mRNA profiling\Agilent Human 1A Microarray G4110**Endometrioid Tumor**' in the 'Messenger RNA data (microarray) data' area.

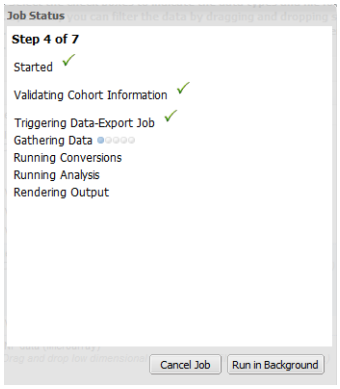


Tips

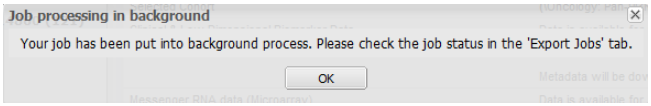
If you have added a node by mistake in the export filter area, right-click the node and remove it.



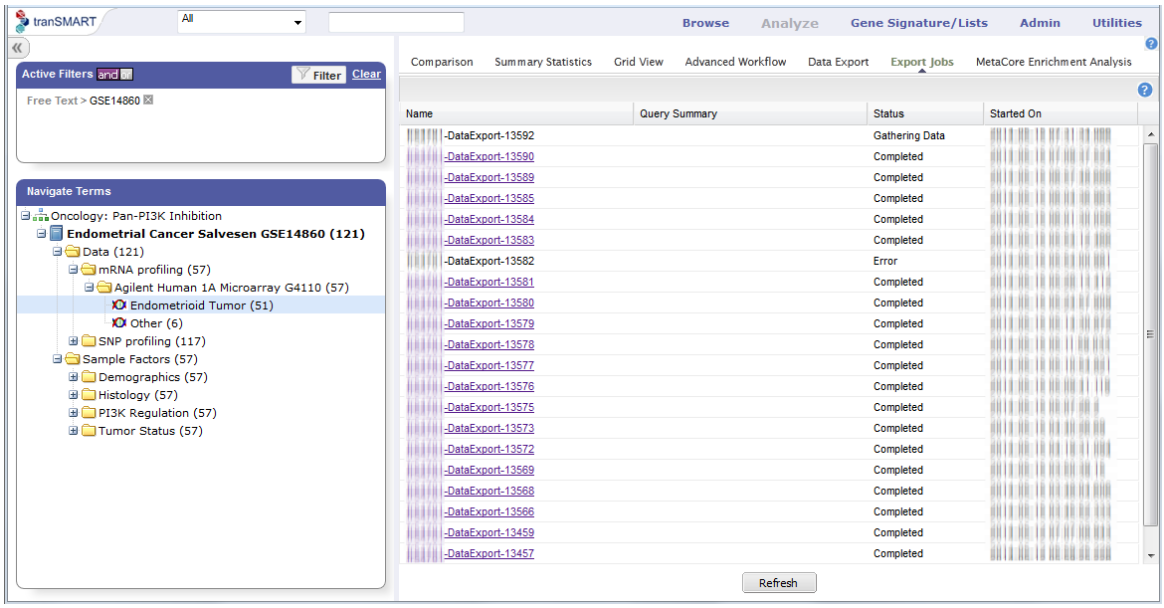
15. Click the **Export Data** button.



16. Click the button **Run in background** when the pop-up window **Job Status** appears. Then **OK**.

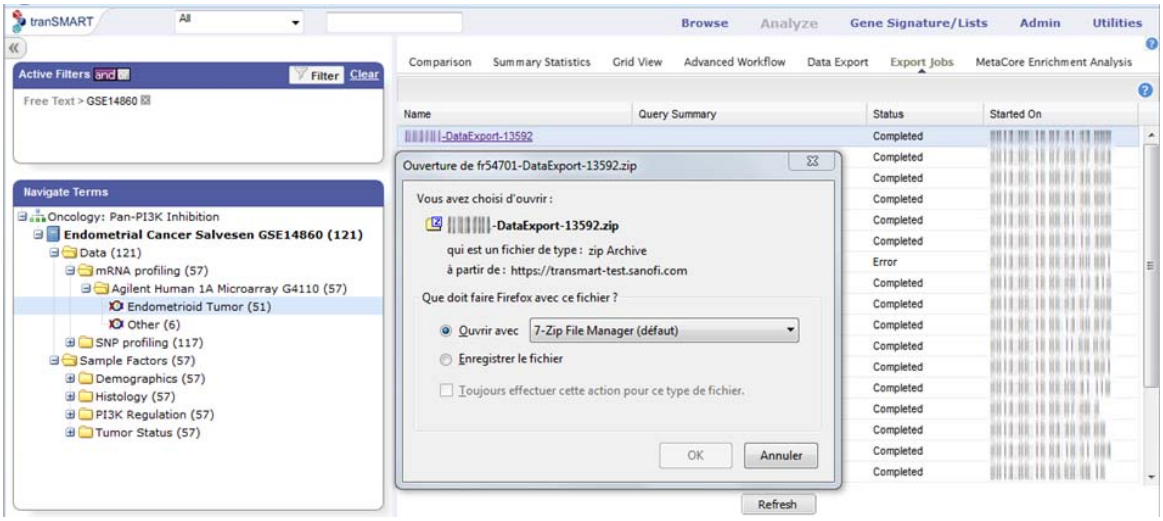


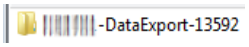
17. Click the **Export Jobs** tab.

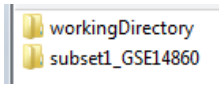


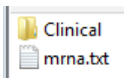
18. Click the **Refresh** button if status of the job is not completed.

19. Click the link when the status is completed.

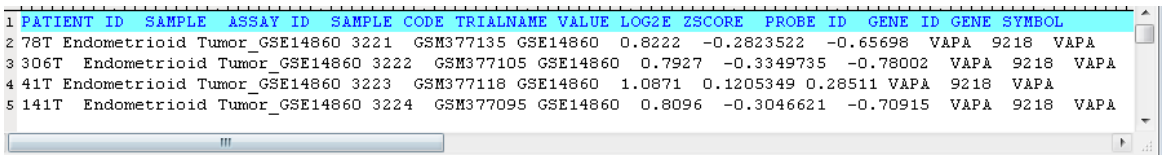






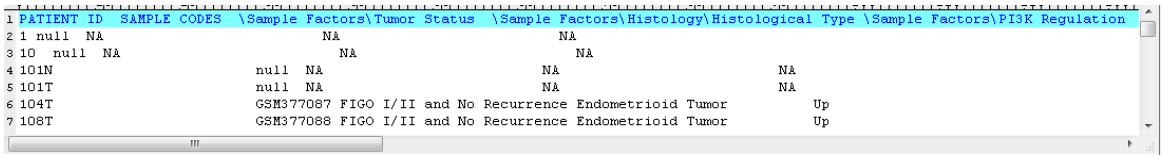


Gene expression data exported only for Endometrioid Tumor.



PATIENT ID	SAMPLE	ASSAY ID	SAMPLE CODE	TRIALNAME	VALUE	LOG2E	ZSCORE	PROBE ID	GENE ID	GENE SYMBOL
78T	Endometrioid Tumor_GSE14860	3221	GSM377135	GSE14860	0.8222	-0.2823522	-0.65698	VAPA	9218	VAPA
306T	Endometrioid Tumor_GSE14860	3222	GSM377105	GSE14860	0.7927	-0.3349735	-0.78002	VAPA	9218	VAPA
41T	Endometrioid Tumor_GSE14860	3223	GSM377118	GSE14860	1.0871	0.1205349	0.28511	VAPA	9218	VAPA
141T	Endometrioid Tumor_GSE14860	3224	GSM377095	GSE14860	0.8096	-0.3046621	-0.70915	VAPA	9218	VAPA

clinical_i2b2trans.txt



PATIENT ID	SAMPLE CODES	\Sample Factors\Tumor Status	\Sample Factors\Histology\Histological Type	\Sample Factors\PI3K Regulation
1	null	NA	NA	NA
10	null	NA	NA	NA
101N	null	NA	NA	NA
101T	null	NA	NA	NA
104T	GSM377087	FIGO I/II and No Recurrence	Endometrioid Tumor	Up
108T	GSM377088	FIGO I/II and No Recurrence	Endometrioid Tumor	Up

Only the nodes Histology Type, PI3K Regulation, Tumor Status are exported.

Scenario (3): Use cohorts to export a subset of data for the Study 'Endometrial Cancer Salvesen GSE14860'.

20. Return to the Comparison tab and clear the Subset 1.

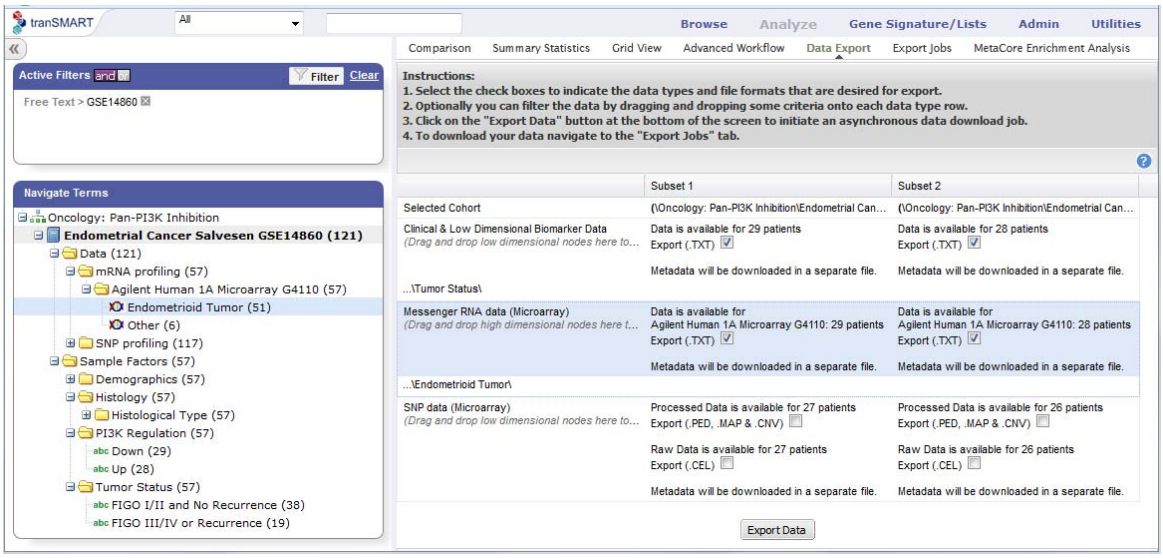
21. Drag and drop the node '\Endometrial Cancer Salvesen GSE14860\Sample Factors\PI3K Regulation**Down**' as cohort Subset 1 and '\Endometrial Cancer Salvesen GSE14860\Sample Factors\PI3K Regulation**Up**' as cohort Subset 2.

22. Go to the **Data Export** tab.

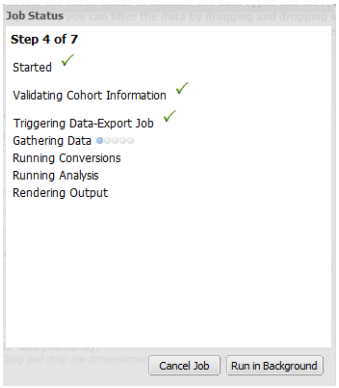
23. Drag and drop the node '\Endometrial Cancer Salvesen GSE14860\Sample Factors**Tumor Status**' in the 'Clinical & Low Dimensional Biomarker Data' area.

24. Drag and drop the node '\Endometrial Cancer Salvesen GSE14860\Data\mRNA profiling\Agilent Human 1A Microarray G4110**Endometrioid Tumor**' in the 'Messenger RNA data (microarray) data' area.

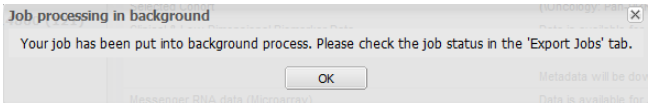
25. Activate the check boxes of 'Clinical & Low Dimensional Biomarker Data' and 'Messenger RNA data (microarray) data' for the two cohorts Subset 1 and Subset 2.



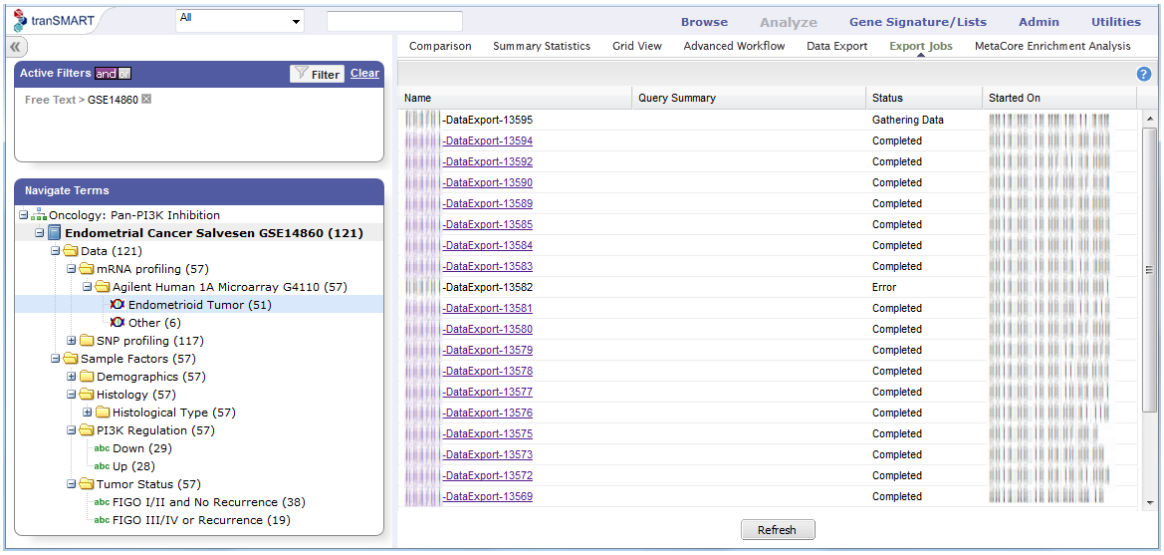
26. Click the **Export Data** button.



27. Click the button **Run in background** when the pop-up window **Job Status** appears. Then **OK**.

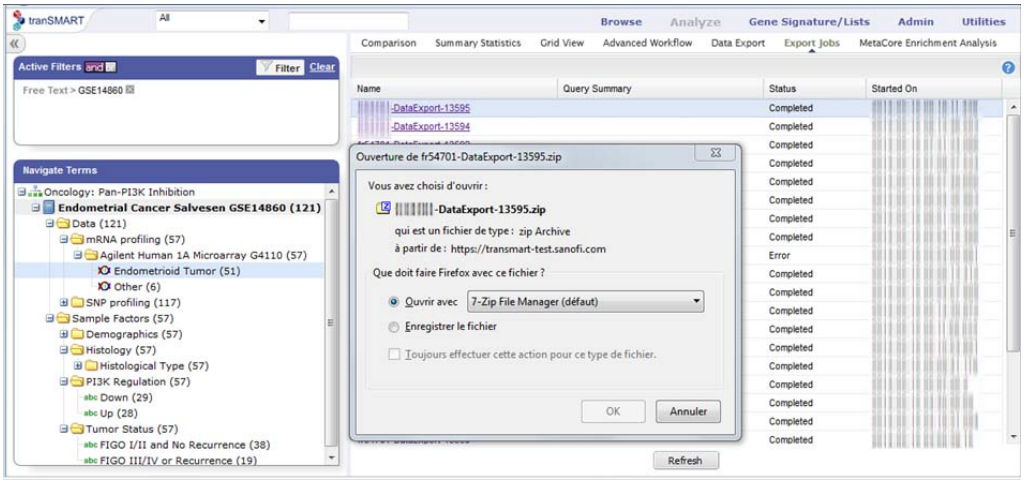


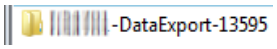
28. Click the **Export Jobs** tab.

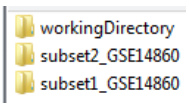


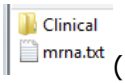
29. Click the **Refresh** button if status of the job is not completed.

30. Click the link when the status is completed.



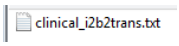




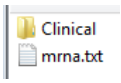


(subset 1)

```
PATIENT ID  SAMPLE  ASSAY ID  SAMPLE CODE TRIALNAME VALUE LOG2E ZSCORE  PROBE ID  GENE ID GENE SYMBOL
306T  Endometrioid Tumor_GSE14860 3222  GSM377105 GSE14860  0.7927  -0.3349735  -0.78002  VAPA  9218  VAPA
436T  Endometrioid Tumor_GSE14860 3229  GSM377123 GSE14860  1.5345  0.6178222 1.44793 VAPA  9218  VAPA
172T  Endometrioid Tumor_GSE14860 3231  GSM377098 GSE14860  0.8830  -0.179378 -0.41619  VAPA  9218  VAPA
445T  Endometrioid Tumor_GSE14860 3233  GSM377127 GSE14860  1.3190  0.3995302 0.93749 VAPA  9218  VAPA
```

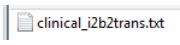



```
PATIENT ID  SAMPLE CODES  \Sample Factors\Tumor Status
110T                      GSM377090 FIGO I/II and No Recurrence
111T                      GSM377091 FIGO I/II and No Recurrence
113T                      GSM377092 FIGO I/II and No Recurrence
116T                      GSM377141 FIGO I/II and No Recurrence
```



(subset 2)

```
PATIENT ID  SAMPLE  ASSAY ID  SAMPLE CODE TRIALNAME VALUE LOG2E ZSCORE  PROBE ID  GENE ID GENE SYMBOL
78T Endometrioid Tumor_GSE14860 3221  GSM377135 GSE14860  0.8222  -0.2823522  -0.65698  VAPA  9218  VAPA
41T Endometrioid Tumor_GSE14860 3223  GSM377118 GSE14860  1.0871  0.1205349 0.28511 VAPA  9218  VAPA
141T  Endometrioid Tumor_GSE14860 3224  GSM377095 GSE14860  0.8096  -0.3046621  -0.70915  VAPA  9218  VAPA
104T  Endometrioid Tumor_GSE14860 3227  GSM377087 GSE14860  0.7794  -0.3594158  -0.83718  VAPA  9218  VAPA
```




```
PATIENT ID  SAMPLE CODES  \Sample Factors\Tumor Status
104T                      GSM377087 FIGO I/II and No Recurrence
108T                      GSM377088 FIGO I/II and No Recurrence
10T2                      GSM377089 FIGO I/II and No Recurrence
129T                      GSM377094 FIGO I/II and No Recurrence
```

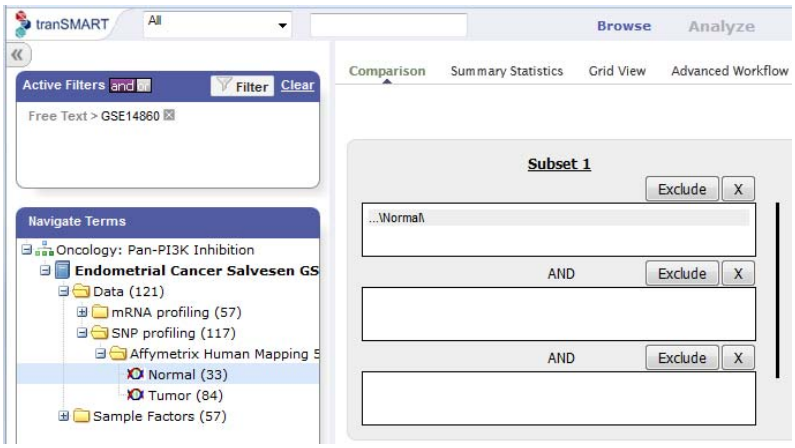
Scenario (4): You want to export SNP data.

Be careful when you export SNP data. Even if you drag and drop one SNP node, the export will include all SNP nodes.

If each SNP node concerns a different cohort, you can use the cohort subset box to export only one SNP node.

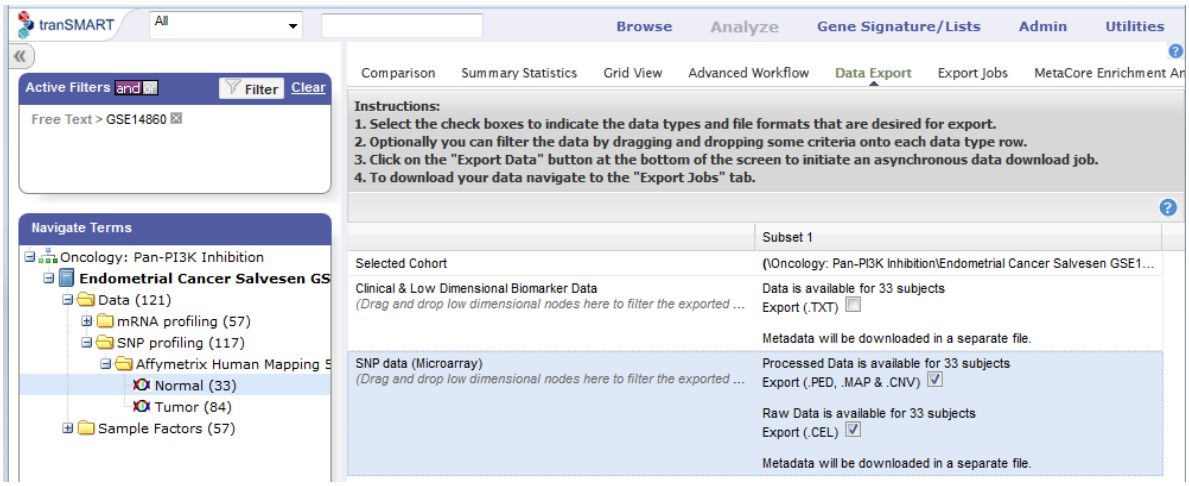
31. Return to the Comparison tab and clear Subset 1 and Subset 2.

32. Drag and drop the node \Endometrial Cancer Salvesen GSE14860\Data\SNP profiling\...**Normal**' as cohort Subset 1.



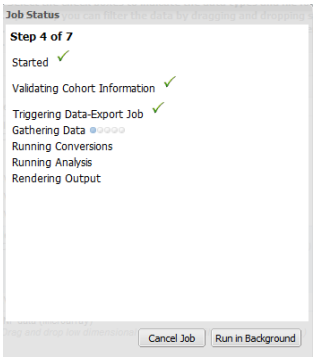
33. Go to the **Data Export** tab.

34. Activate the check boxes for Processed (.PED, .MAP & .CNV files) and Raw data (.CEL files).

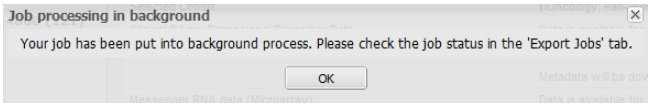


Remark: Data available for 33 subjects, only the Normal SNP node.

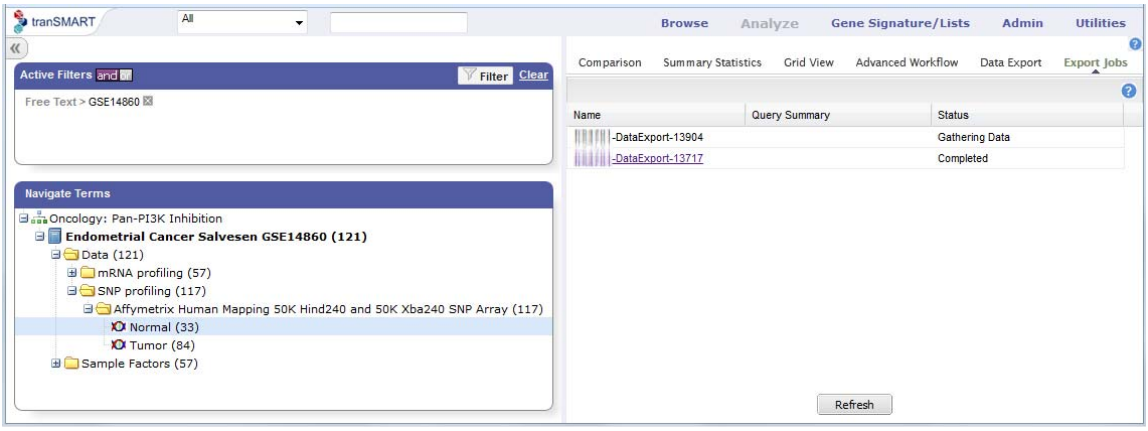
35. Click the **Export Data** button.



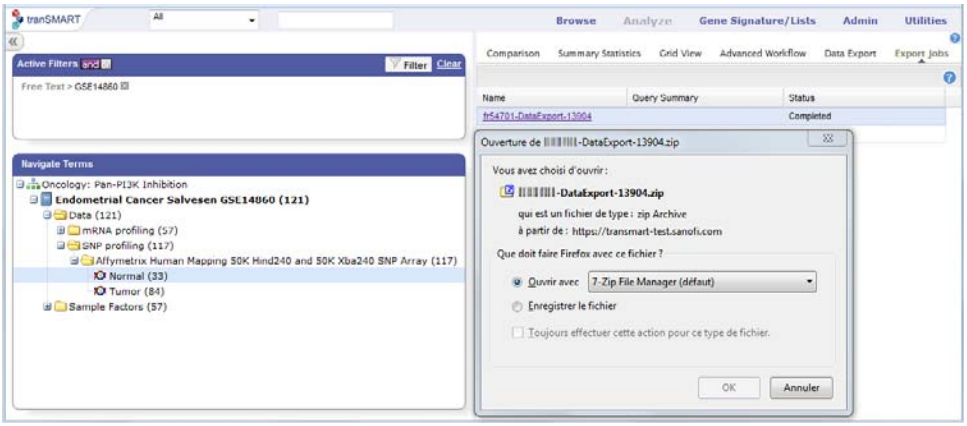
36. Click the button **Run in background** when the pop-up window **Job Status** appears. Then **OK**.



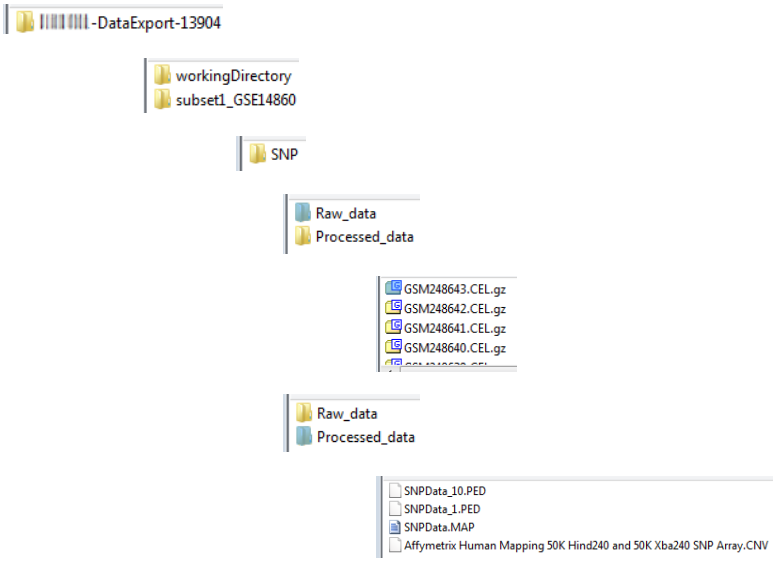
37. Click the **Export Jobs** tab.



38. Click the **Refresh** button if status of the job is not completed.

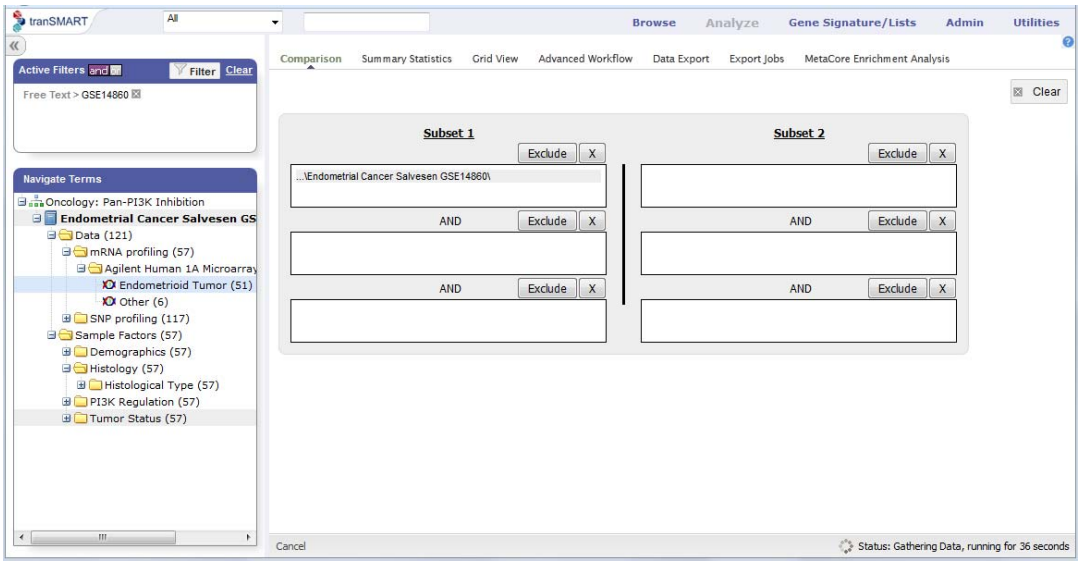


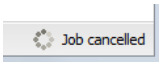
39. Click the link when the status is completed.

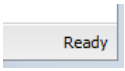


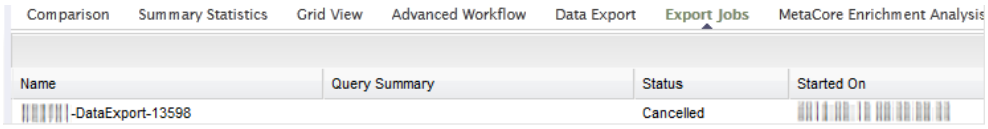
Tips

The current job can be cancelled by clicking the Cancel button displayed at the bottom of the page in the Comparison tab.


tranSMART
All
Browse
Analyze
Gene Signature/Lists
Admin
Utilities
Active Filters
and
Filter
Clear
Free Text > GSE14860
Navigate Terms
Oncology: Pan-PI3K Inhibition
Endometrial Cancer Salvesen GS
Data (121)
mRNA profiling (57)
Agilent Human 1A Microarray
Endometrioid Tumor (51)
Other (6)
SNP profiling (117)
Sample Factors (57)
Demographics (57)
Histology (57)
Histological Type (57)
PI3K Regulation (57)
Tumor Status (57)
Comparison
Summary Statistics
Grid View
Advanced Workflow
Data Export
Export Jobs
MetaCore Enrichment Analysis
Clear
Subset 1
Subset 2
Exclude
X
...\Endometrial Cancer Salvesen GSE14860\
AND
Cancel
Status: Gathering Data, running for 36 seconds


Job cancelled


Ready


Comparison
Summary Statistics
Grid View
Advanced Workflow
Data Export
Export Jobs
MetaCore Enrichment Analysis
Name
Query Summary
Status
Started On
-DataExport-13598
Cancelled

Lesson 21 – MetaCore Enrichment Analysis

MetaCore Enrichment Analysis helps understand experimental findings (omics data) in the context of validated biological pathways. Currently this analysis scores and ranks the most relevant MetaBase pathways for a list of genes identified from a gene expression dataset.

Lesson Goal: Become acquainted with the MetaCore Enrichment Analysis.

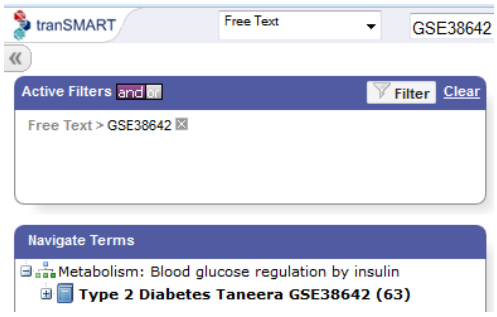
To perform MetaCore Enrichment Analysis:

1. Run tranSMART, under Analyze tab, select the study of interest.
2. In the Comparison tab, define the cohorts.
3. Click the **MetaCore Enrichment Analysis** tab.
4. In the Variable Selection box, drag and drop the high dimensional data node.
5. Click the **High Dimensional Data** button and add a list of genes (pathway, gene list or gene signature).
6. Click the **Run** button.
7. Click the links to visualize the pathway maps in MetaCore.

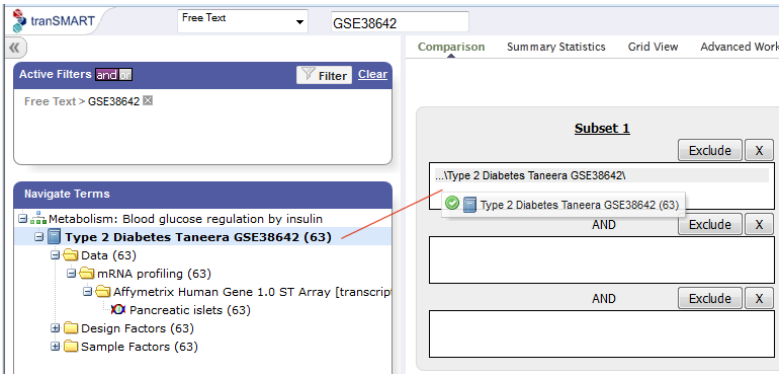
Example

1. Go to the Analyze tab, select the search criteria **Free Text**, write in the box: **GSE38642** and 'Enter key' on your keyboard.

2. Under the Program, open the Study '**Type 2 Diabetes Taneera GSE38642**'.

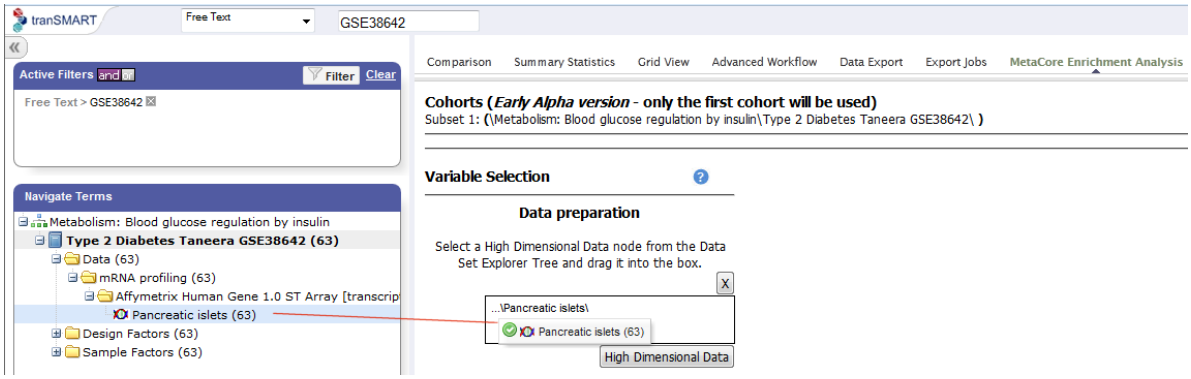


3. In the Comparison tab, drag and drop the Study '**Type 2 Diabetes Taneera GSE38642**' as Subset 1 cohort.

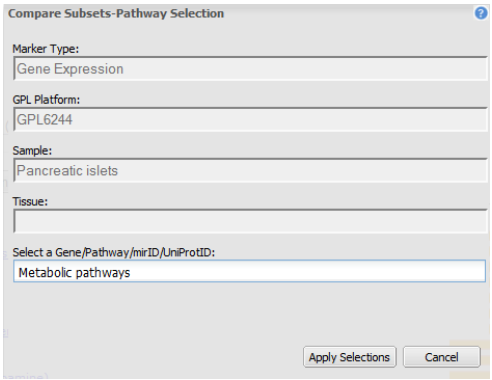


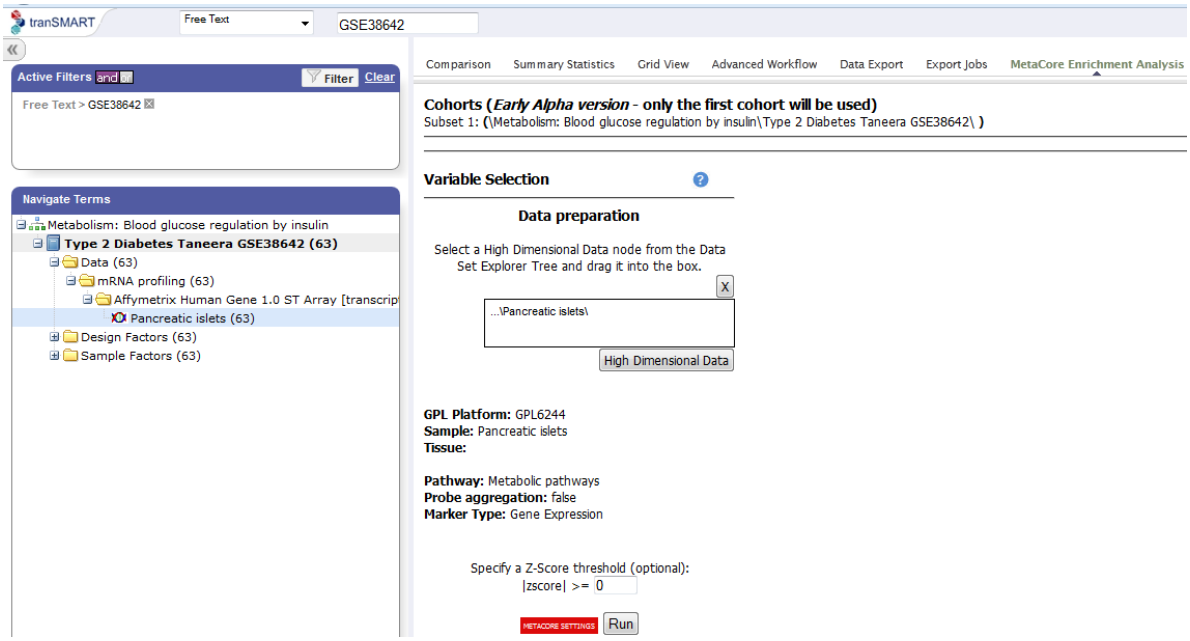
4. Go to the **MetaCore Enrichment Analysis** tab.

5. Drag and drop the node '\Type 2 Diabetes Taneera GSE38642\...**Pancreatic islets**' under Variable Selection.

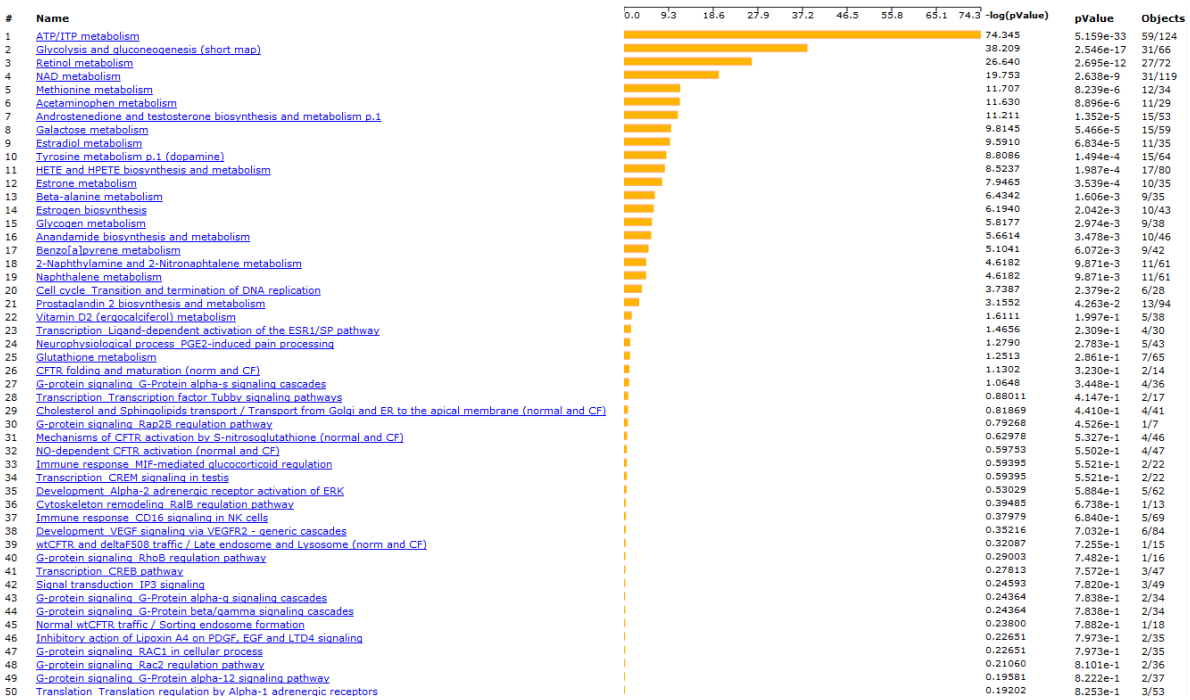


6. Click the **High Dimensional Data** button, add the KEGG pathway '**Metabolic pathways**' and click **Apply Selections**.



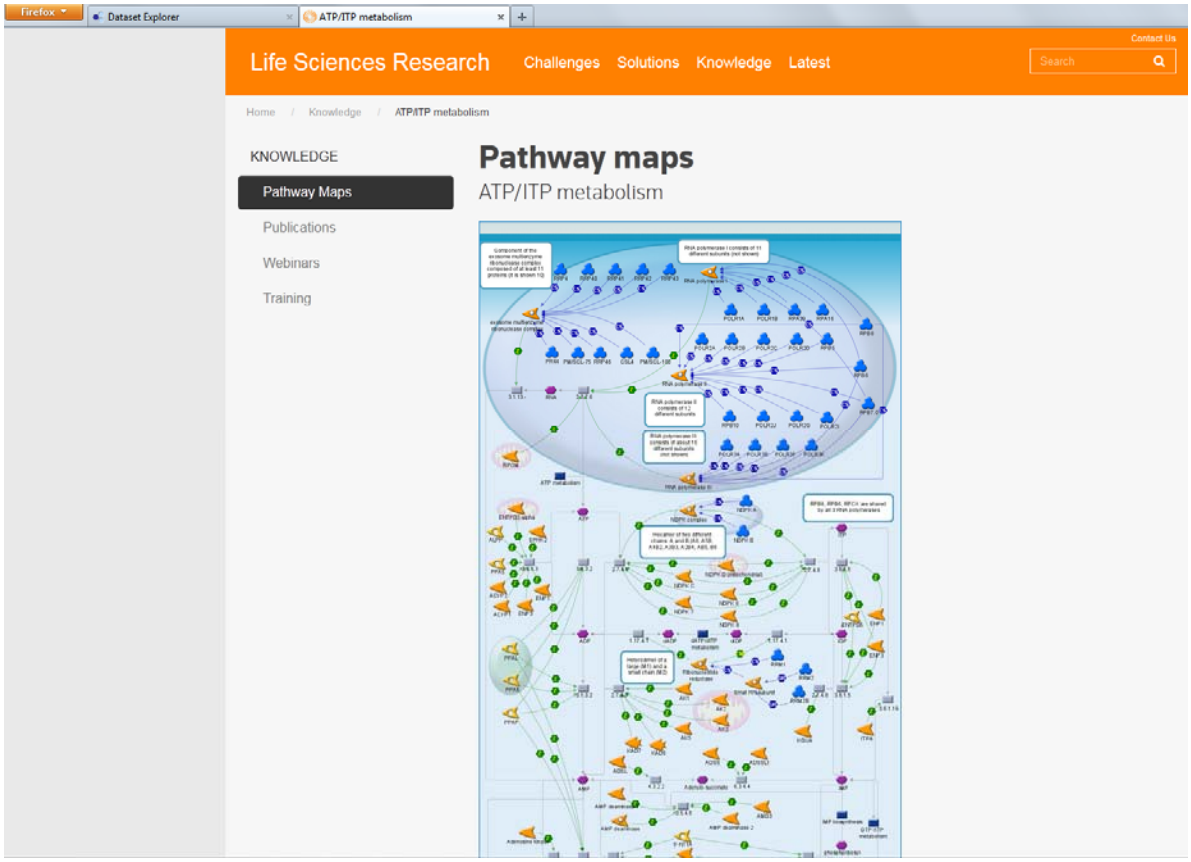


7. Click the **Run** button.



#	Name	-log(pValue)	pValue	Objects
1	ATP/ITP metabolism	74.345	5.159e-33	59/124
2	Glycolysis and gluconeogenesis (short map)	38.209	2.546e-17	31/66
3	Retinol metabolism	26.640	2.695e-12	27/72
4	NAD metabolism	19.753	2.638e-9	31/119
5	Methionine metabolism	11.707	8.239e-6	12/34
6	Acetaminophen metabolism	11.630	8.896e-6	11/29
7	Androstenedione and testosterone biosynthesis and metabolism p.1	11.211	1.352e-5	15/53
8	Galactose metabolism	9.8145	5.466e-5	15/59
9	Estradiol metabolism	9.5910	6.834e-5	11/35
10	Tyrosine metabolism p.1 (dopamine)	8.8086	1.494e-4	15/64
11	HETE and HPETE biosynthesis and metabolism	8.5237	1.987e-4	17/80
12	Estrone metabolism	7.9465	3.539e-4	10/35
13	Beta-alanine metabolism	6.4342	1.606e-3	9/35
14	Estrogen biosynthesis	6.1940	2.042e-3	10/43
15	Glycogen metabolism	5.8177	2.974e-3	9/38
16	Anandamide biosynthesis and metabolism	5.6614	3.478e-3	10/46
17	Benzo[a]pyrene metabolism	5.1041	6.072e-3	9/42
18	2-Naphthylamine and 2-Nitronaphtalene metabolism	4.6182	9.871e-3	11/61
19	Naphthalene metabolism	4.6182	9.871e-3	11/61
20	Cell cycle_Transition and termination of DNA replication	3.7387	2.379e-2	6/28
21	Prostaglandin 2 biosynthesis and metabolism	3.1552	4.263e-2	13/94
22	Vitamin D2 (ergocalciferol) metabolism	1.6111	1.997e-1	5/38
23	Transcription_Ligand-dependent activation of the ESR1/SP pathway	1.4656	2.309e-1	4/30
24	Neurophysiological process_PGE2-induced pain processing	1.2790	2.783e-1	5/43
25	Glutathione metabolism	1.2513	2.861e-1	7/65
26	CFTR folding and maturation (norm and CF)	1.1302	3.230e-1	2/14
27	G-protein signaling_G-Protein alpha-s signaling cascades	1.0648	3.448e-1	4/36
28	Transcription_Transcription factor Tubby signaling pathways	0.88011	4.147e-1	2/17
29	Cholesterol and Sphingolipids transport / Transport from Golgi and ER to the apical membrane (normal and CF)	0.81869	4.410e-1	4/41
30	G-protein signaling_Rap2B regulation pathway	0.79268	4.526e-1	1/7
31	Mechanisms of CFTR activation by S-nitrosoglutathione (normal and CF)	0.62978	5.327e-1	4/46
32	NO-dependent CFTR activation (normal and CF)	0.59753	5.502e-1	4/47
33	Immune response_MIF-mediated glucocorticoid regulation	0.59395	5.521e-1	2/22
34	Transcription_CREM signaling in testis	0.59395	5.521e-1	2/22
35	Development_Alpha-2 adrenergic receptor activation of ERK	0.53029	5.884e-1	5/62
36	Cytoskeleton remodeling_RalB regulation pathway	0.39485	6.738e-1	1/13
37	Immune response_CD16 signaling in NK cells	0.37979	6.840e-1	5/69
38	Development_VEGF signaling via VEGFR2 - generic cascades	0.35216	7.032e-1	6/84
39	wtCFTR and deltaF508 traffic / Late endosome and Lysosome (norm and CF)	0.32087	7.255e-1	1/15
40	G-protein signaling_RhoB regulation pathway	0.29003	7.482e-1	1/16
41	Transcription_CREB pathway	0.27813	7.572e-1	3/47
42	Signal transduction_IP3 signaling	0.24593	7.820e-1	3/49
43	G-protein signaling_G-Protein alpha-q signaling cascades	0.24364	7.838e-1	2/34
44	G-protein signaling_G-Protein beta/gamma signaling cascades	0.24364	7.838e-1	2/34
45	Normal wtCFTR traffic / Sorting endosome formation	0.23800	7.882e-1	1/18
46	Inhibitory action of Lipoxin A4 on PDGF, EGF and LTD4 signaling	0.22651	7.973e-1	2/35
47	G-protein signaling_RAC1 in cellular process	0.22651	7.973e-1	2/35
48	G-protein signaling_Rac2 regulation pathway	0.21060	8.101e-1	2/36
49	G-protein signaling_G-Protein alpha-12 signaling pathway	0.19581	8.222e-1	2/37
50	Translation_Translation regulation by Alpha-1 adrenergic receptors	0.19202	8.253e-1	3/53

8. Click a pathway **link** to view the pathway map in MetaCore.



Lesson 22 – Gene Signature / Gene List

The tranSMART gene signature wizard guides you through the process of creating a gene list or a gene signature (gene list with fold-changes). 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 in the Browse tab 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. User-defined gene lists and gene signatures can also be used from the Analyze tab to restrict high dimensional analyses (Heatmap, Clustering analyses, Marker Selection, PCA) to this specific list of genes.

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

Lesson Goal: Learn to create a Gene Signature and Gene List.

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.

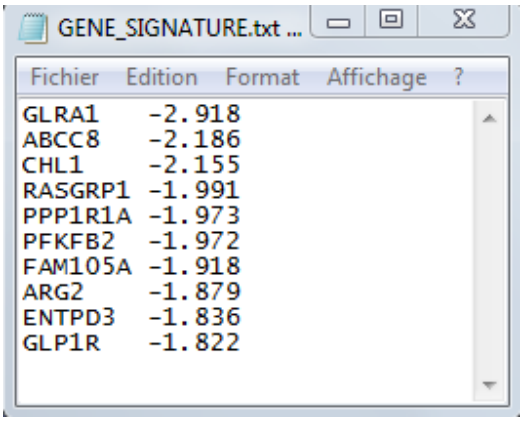
- **0**. No change.

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_at1 A_24_P93251-1

Example

A text file is available on your computer with the following gene symbols and Fold-Change Metrics:



No column headings.

Tips

Create the text file from Excel using **Save as type** option **Text (Tab delimited) (*.txt)**.

Step 2. Creating the Gene Signature

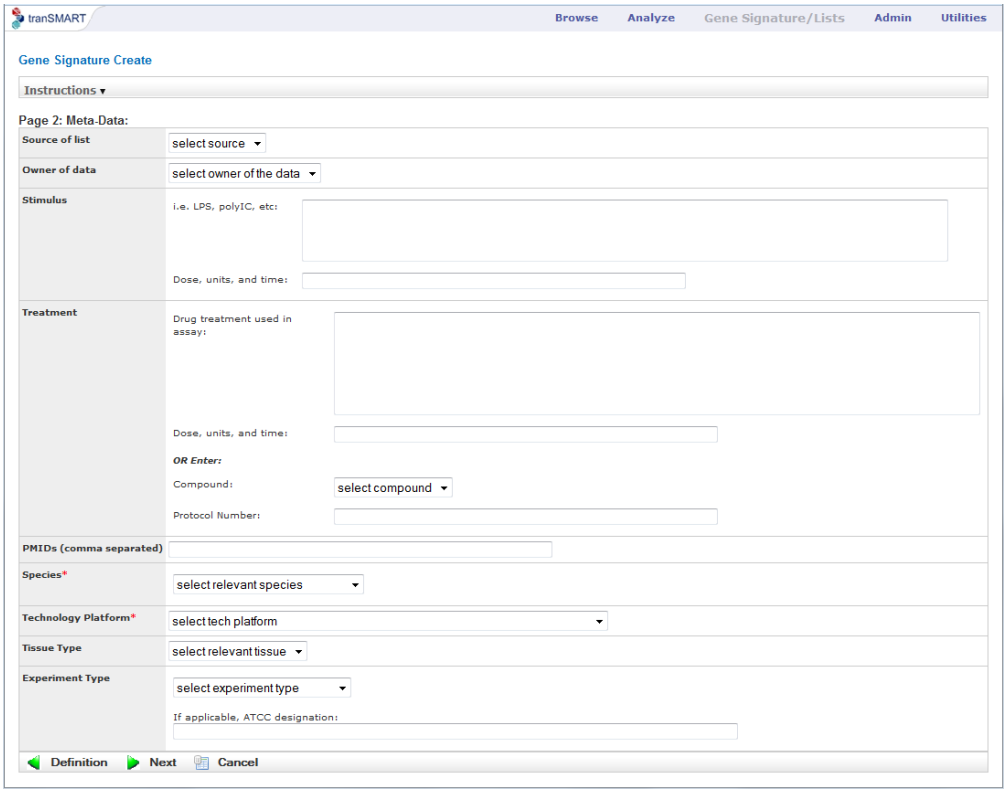
1. In tranSMART, click the **Gene Signature / Lists** tab.
2. Click the **New Signature** button.

The first page of the gene signature wizard appears:

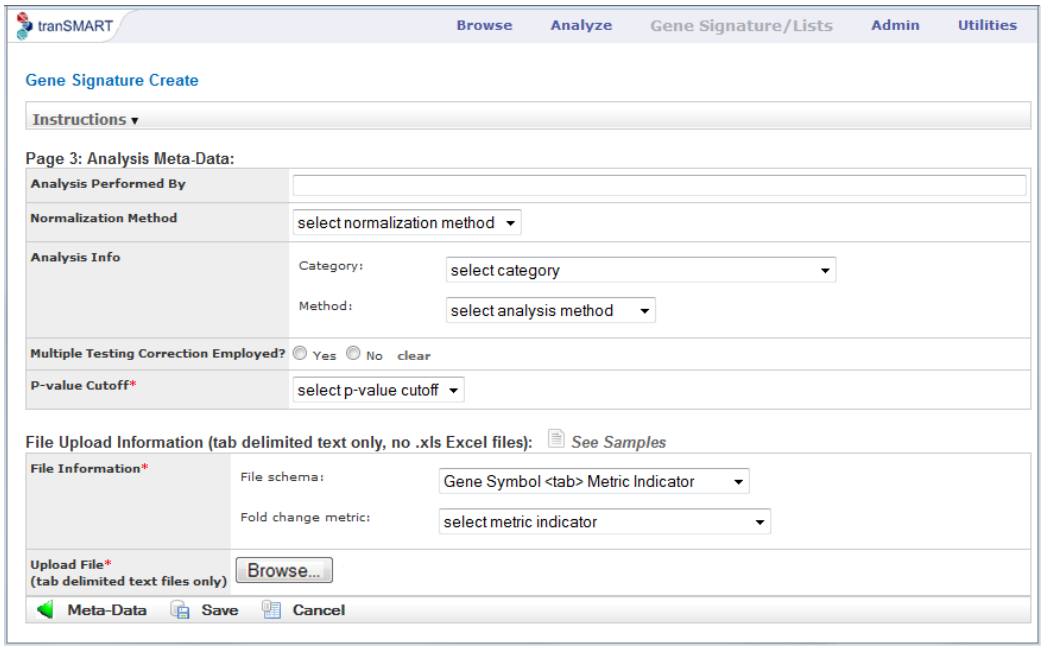
Required fields on gene signature wizard pages are marked with a red asterisk (*).

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

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



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



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

6. In the section **File Upload Information**, describe the text file you created in the section Step 1. Adding the Genes to a Text File on page 93, using the required fields **File Information** and **Upload File**:

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

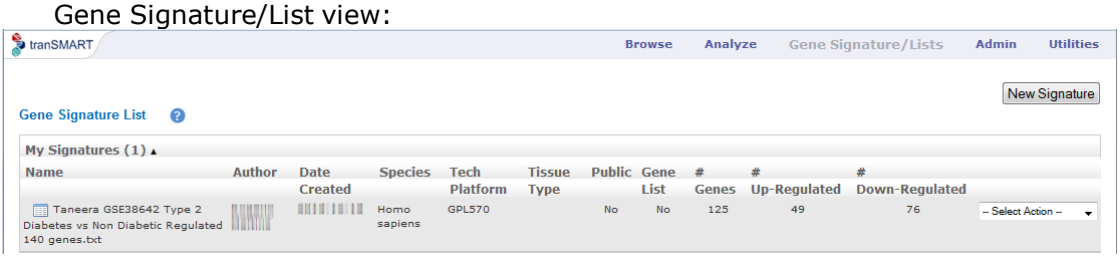
   - In the **Fold change metric** section of **File Information**, select one of the following choices from the dropdown:

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

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

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

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



Making a New Gene Signature Public

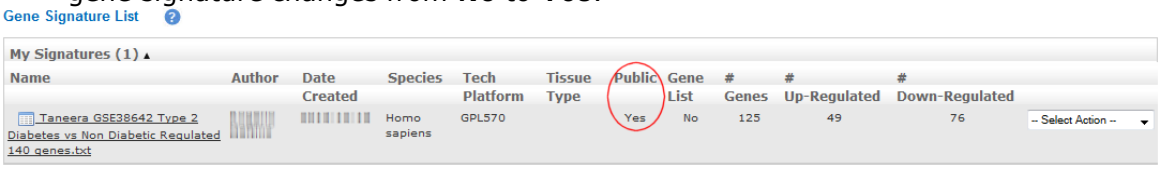
By default, a newly created gene signature is private.

To make a gene signature public:

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

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

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

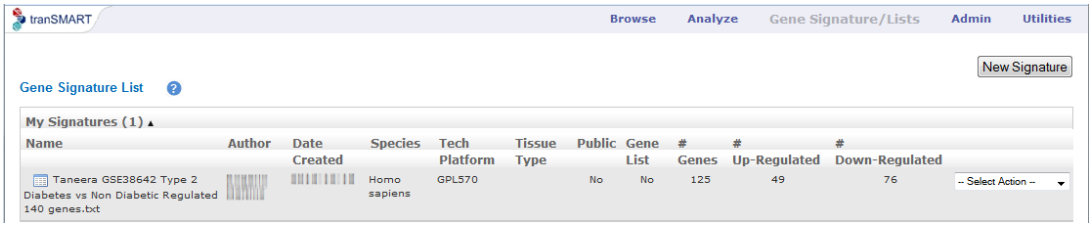


tranSMART users assigned the role `ROLE_ADMIN` have access to both public and private gene signatures.

Performing Actions on Your Gene Signatures

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

1. In tranSMART, click the **Gene Signature / Lists** tab.
   The **Gene Signature List** appears, containing all the genes you have created:



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

Action	Description
Clone	Create an exact duplicate of the gene signature definition (*except* for the text file containing the gene symbols and fold change values), and display the definition in the gene signature wizard. Cloning a gene signature helps you create a new gene signature with a similar definition to an existing one. However, it is expected you will import a different set of genes into the gene signature.
Delete	Delete the gene signature.
Edit	Open the gene signature in the gene signature wizard for editing. The gene signature wizard displays all the information in the gene signature, including the reference to the text file containing the list of genes and fold change values. If you want to choose a different text file, click the following label: 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.
Download .GMT file	Download .the GMT files. A GMT file format is a tab delimited file format that describes gene sets. In the GMT format, each row represents a gene set.
Make Public	Make a private gene signature public. **Note:** To make a public gene signature private, edit the gene signature and set the **Public?** field to **No** on the first page of the gene signature wizard: 

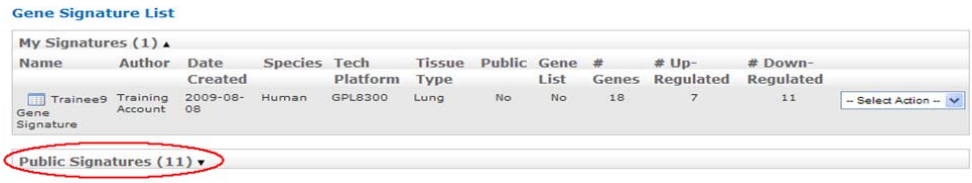
Performing Actions on Other Users' Signatures

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

Role	Authorized Actions
ROLE_ADMIN	All actions on all gene signatures, both public and private.
ROLE_SPECTATOR ROLE_STUDY_OWNER ROLE_DATASET_EXPLORER_ADMIN	Only **Clone** and **Excel Download**, and only on public gene signatures.

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

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

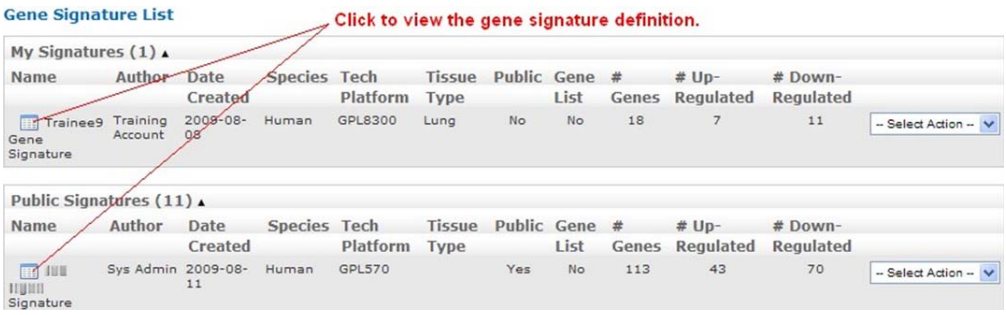


tranSMART users assigned the role `ROLE_ADMIN` will see OtherSignatures instead of Public Signatures.

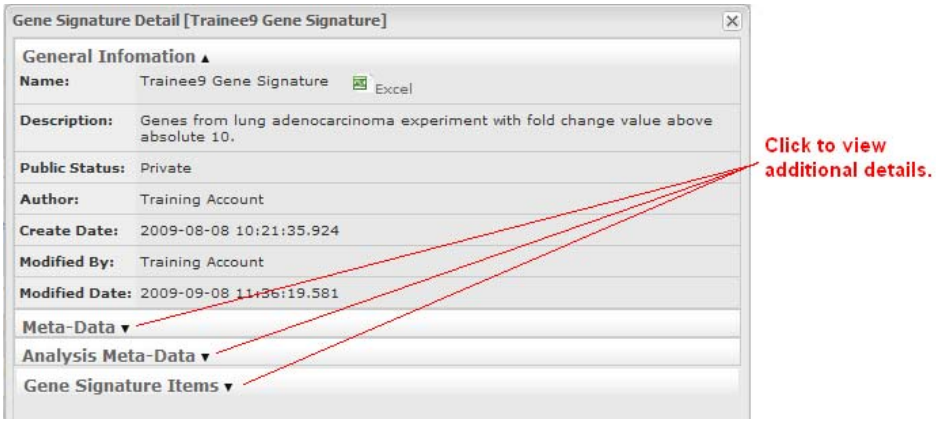
3. Click the **--Select Action--** dropdown for the gene signature you want to act on.
4. Select the action you want to perform on the gene signature.

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.

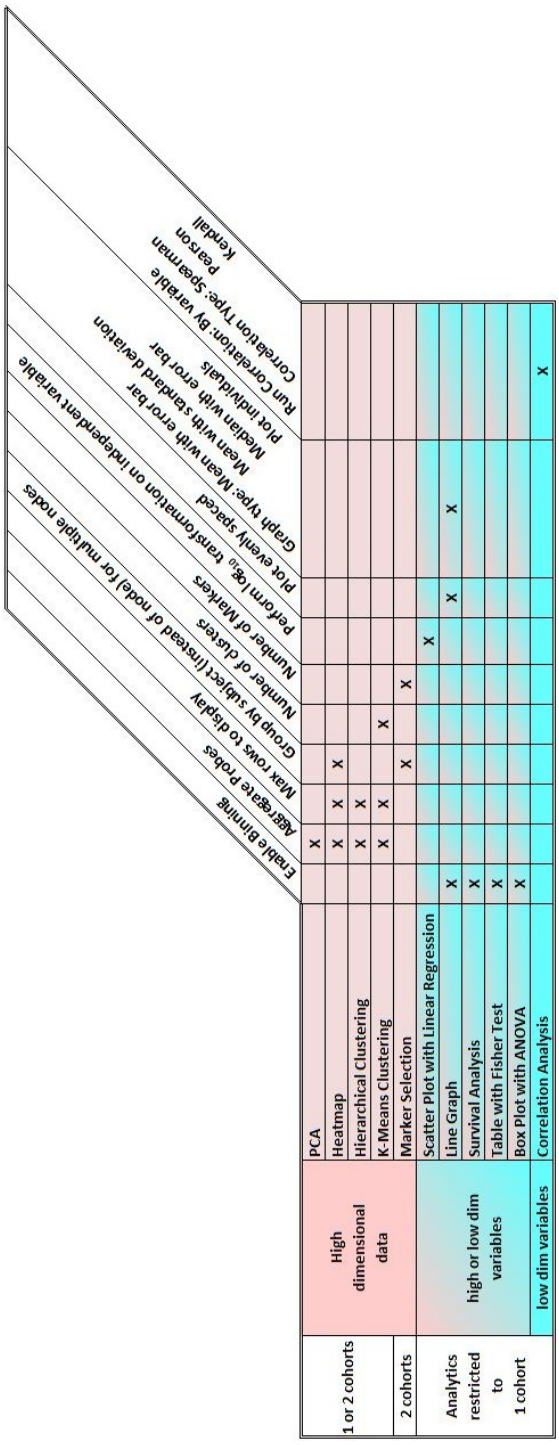
To view a gene signature definition, click the **Detail** icon ( ) next to the gene signature name:



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



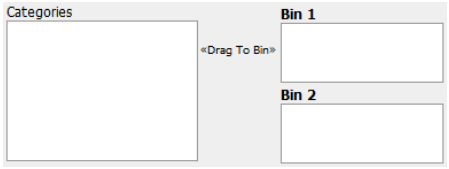
Appendix A – Additional Material



Data Binning

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

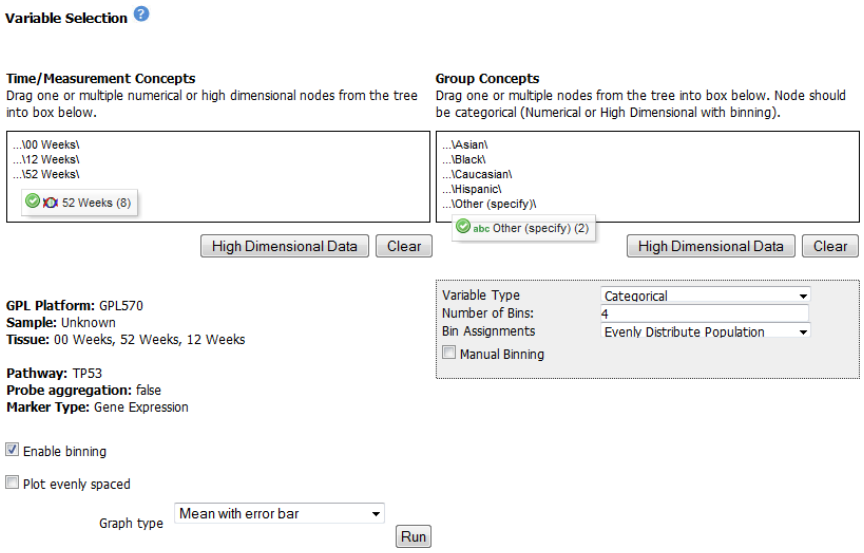
The data displayed after binning represents the data available in the study. If, for example, you have selected to bin based on date range (0-10 years of age), yet there is only data available for subjects eight years old and up, the bin will display the age range as 8-10.

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, even 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: 

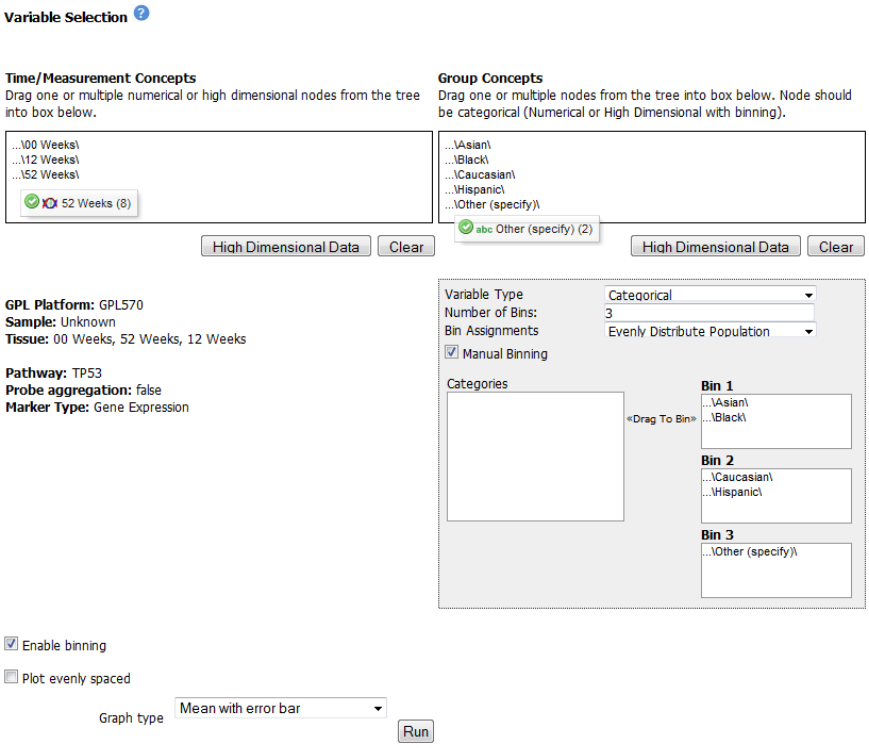
Example in Line Graph with ‘Enable Binning’ checked

Variable Type: Categorical

Without Manual Binning:

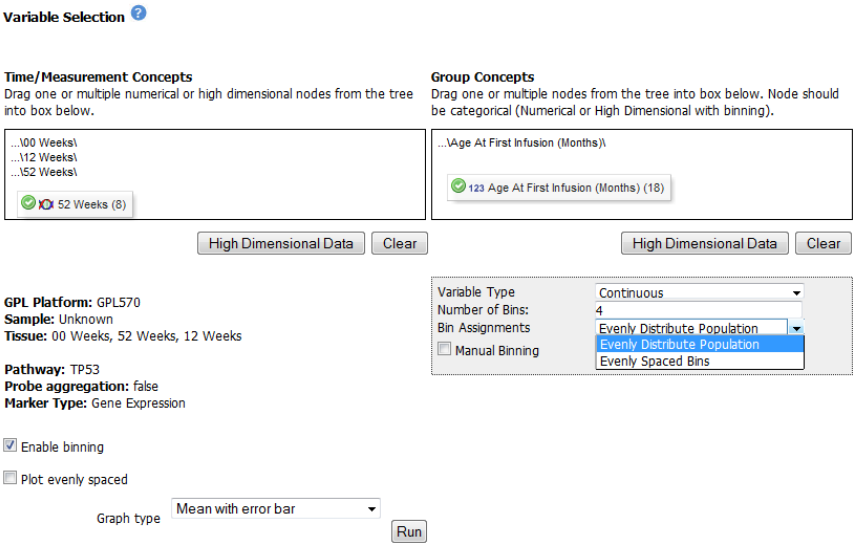


With Manual Binning:

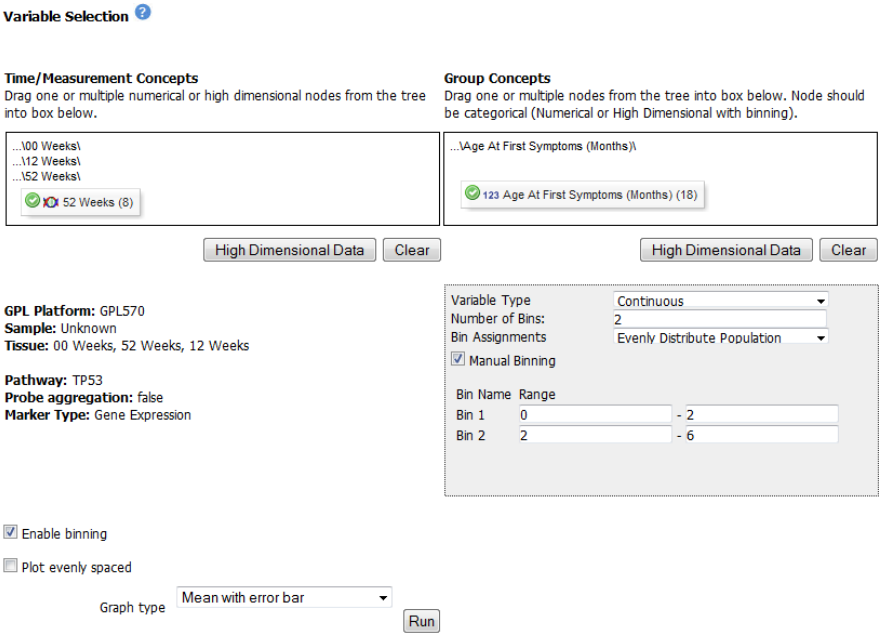


Variable Type: Continuous

Without Manual Binning:



With Manual Binning:



Aggregate Probes

Only usable for high dimensional data in Heatmap, Hierarchical Clustering, K-Means Clustering and PCA. 'Aggregate Probes' reduces the number of variables by retaining only 1 variable per dictionary entity, for example 1 probe per gene for gene expression, 1 transcript per gene for RNAseq data, 1 peptide per protein for mass spec proteomic data, etc. 'Aggregate Probes' does not have any effect if the dataset contains only 1 variable of each entity.

If the checkbox is selected, the algorithm WGCNA (weighted correlation network analysis) is employed. The variable retained is the one with the maximum mean value across the defined cohort(s).

Note: WGCNA was developed by the Department of Human Genetics at UCLA. For more information, see http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/.

Example

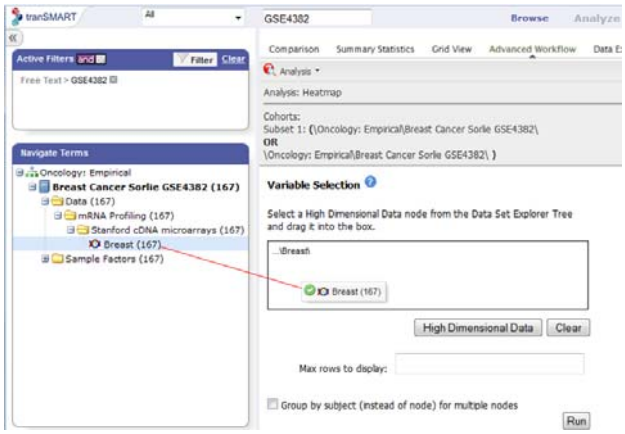
1. Go to the Analyze tab, select the search criteria **Free Text**, write in the box: **GSE4382** and 'Enter key' on your keyboard.

2. In the Navigate Terms area, under the Program, open the Study '**Breast Cancer Sorlie GSE4382**'.

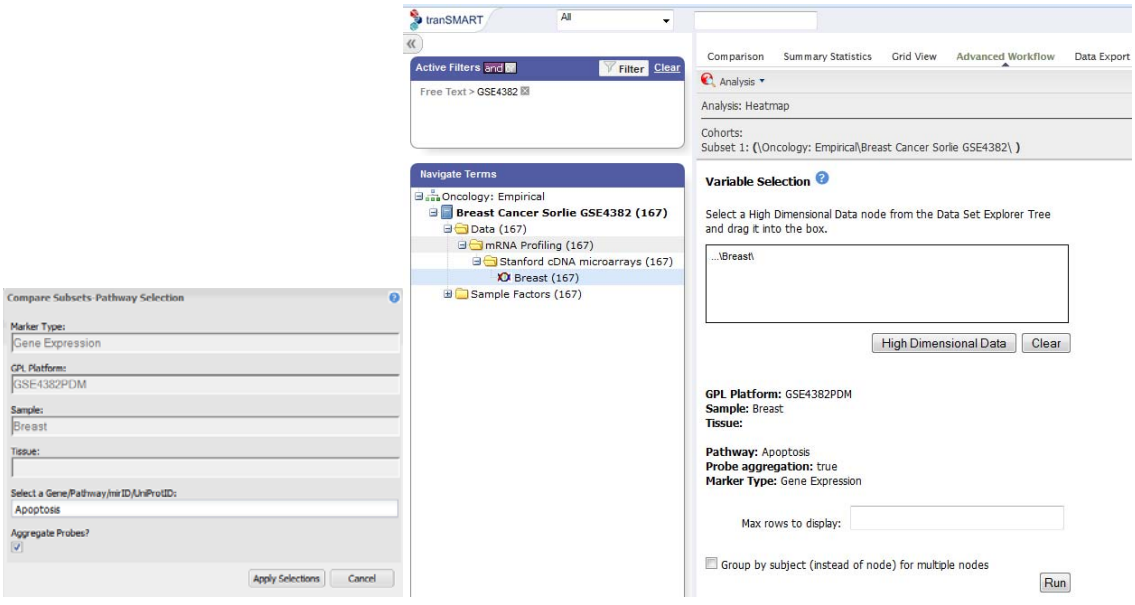
3. In the Comparison tab, drag and drop the Study '**Breast Cancer Sorlie GSE4382**' as Subset 1 cohort.

4. Go to the **Advanced Workflow**, then select **Heatmap** from the Analysis dropdown menu.

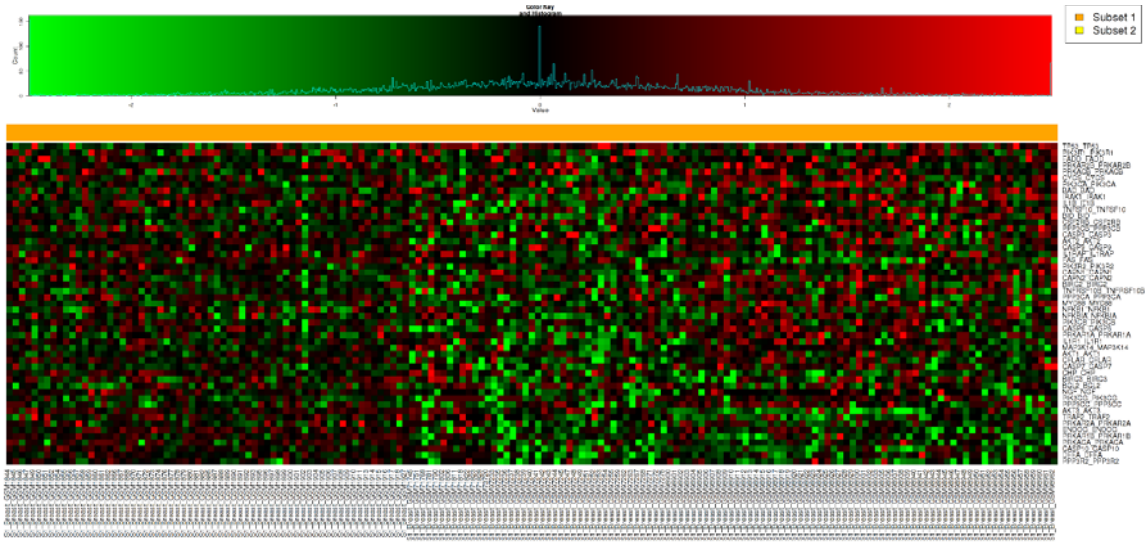
5. Drag and drop the node '\Breast Cancer Sorlie GSE4382\Data\...**Breast**' under Variable Selection.



6. Click the **High Dimensional Data** button, add the pathway **Apoptosis** , activate the checkbox **Aggregate Probes** and click **Apply Selections**.

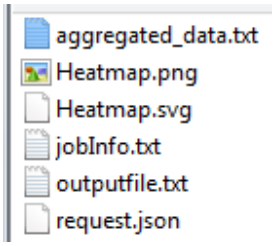


7. Click the **Run** button.



8. Click the link **Download raw R data** to export and visualize data.

An additional file named aggregated_data.txt has been generated.



Appendix B – Glossary of Terms

AGGREGATE PROBES

Used in the Analyze module, 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 the Analyze module to make concurrent comparisons between two or more means in a box plot.

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 measurable indicator of some biological state or condition.

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.

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 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.

DEPENDENT VARIABLE

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

ENTREZ GENE

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

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 correspond 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 a short sequence of 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 matrix data. Individual values are represented by colors.

HIERARCHICAL CLUSTERING

Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. The 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 tranSMART, a job refers to a command you have given the Analyze module to process or export data. Jobs and job-related events can be found within the **Jobs** tab in Analyze module.

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 samples into a specified number of clusters. The result is *k* clusters, each centered around a randomly-selected data point.

LINE GRAPH

Line graph illustrates the changes in one numerical variable in a series of measurements (time series, dose response or series of conditions).

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 the Analyze module.

NCBI

The National Center for Biotechnology Information.

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

NUMERIC-NODE

Used in the Analyze module, 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.

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 variables 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 in Affymetrix microarrays, 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 to the 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 proteins and metabolites.

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 numerical variables for a set of data.

SEARCH FILTER

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

SEARCH STRING

A sequence of 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 differs.

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

A T-test is a statistical comparison of two population means. The test statistic in the t-test is known as the t-statistic.

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.

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.)

X-AXIS

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

Y-AXIS

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