



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.

Contents

Welcome	4
Lesson 1 – Quick Tour of the Browse Module.....	6
Lesson 2 – Browse Studies	9
Lesson 3 – Build a Search Query Using Filters	11
Lesson 4 – Build a Search Query Using Keywords.....	15
Lesson 5 – Exporting files.....	19
Lesson 6 – Create and Edit Objects.....	22
Lesson 7 – Open a study in Analyze View	27
Lesson 8 – Quick tour of the Analyze Module.....	28
Lesson 9 – Browsing and Searching Data in Analyze Module	31
Lesson 10 – Generating Summary Statistics	38
Lesson 11 – Generating a Grid View	42
Lesson 12 – Generating a Heatmap	47
Lesson 13 – Generating a PCA	60
Lesson 14 – Generating a Survival Analysis.....	65
Lesson 15 – Generating a Box Plot	68
Lesson 16 – Generating a Line Graph	71
Lesson 17 – Generating a Scatter Plot	74
Lesson 18 – Generating a Table with Fisher Test.....	77
Lesson 19 – Generating a Correlation Analysis	79
Lesson 20 – Exporting data	82
Lesson 21 – MetaCore Enrichment Analysis.....	96
Lesson 22 – Gene Signature / Gene List	100
Appendix A – Additional Material.....	111
Appendix B – Glossary of Terms	118



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.

A screenshot of the tranSMART web application. At the top, there's a navigation bar with links for "All", "Export Cart", "Browse", "Analyze", "Gene Signature/Lists", "Admin", and "Utilities". A dropdown menu under "Utilities" contains links for "Help", "Contact Us", "About", and "Log Out". On the left, there's a sidebar titled "Active Filters" with a "Filter" and "Clear" button. The main content area displays a "Welcome to Sanofi tranSMART Platform" message and a brief description of the "Browse" window. A tooltip for the "Browse" link is visible, explaining its function.

Disconnection in case of inactivity

tranSMART disconnects after a long moment of inactivity.

A screenshot of the tranSMART interface showing an inactivity warning. A blue modal dialog box appears in the center, stating "Your session is about to expire!" with a close button (X). Below it, a message says "You will be logged off in 10 seconds." and a question "Do you want to continue your session?". At the bottom are two buttons: "Yes, Keep Working" (green) and "No, Logoff" (white).

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.

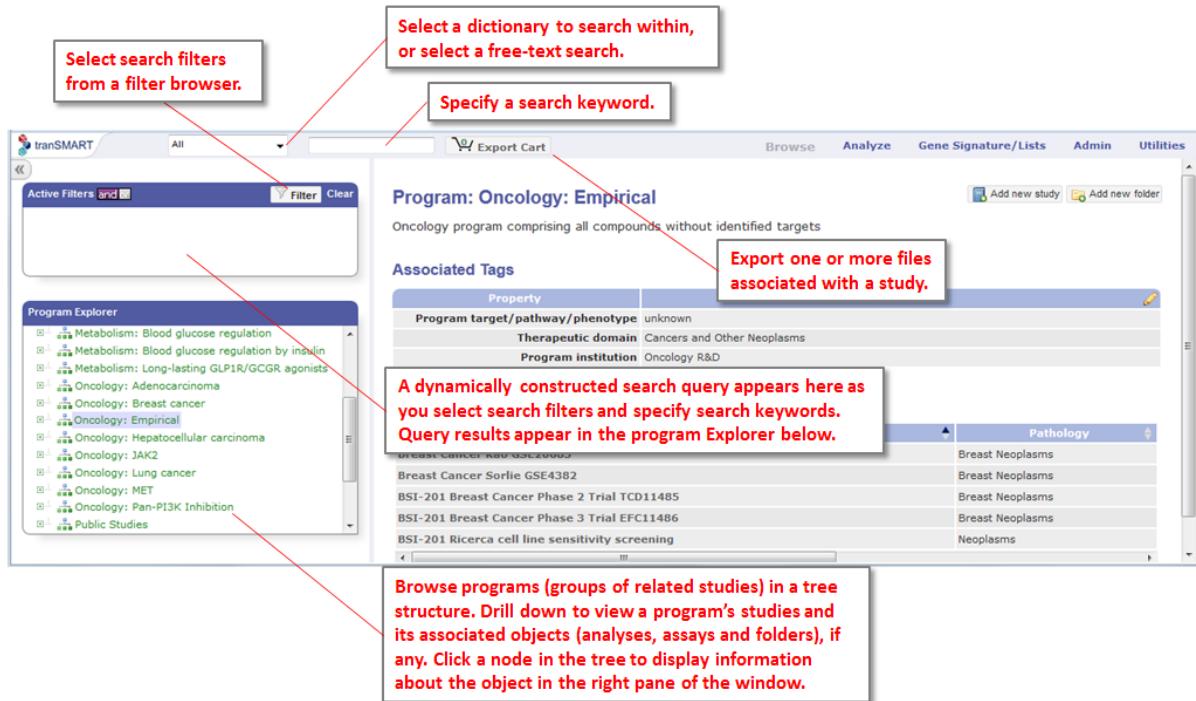
Lesson 1 – Quick Tour of the Browse Module

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.



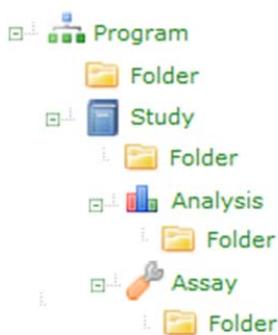
The screenshot shows the tranSMART Browse window with several annotations:

- Select search filters from a filter browser.** Points to the "Active Filters" section at the top left.
- Select a dictionary to search within, or select a free-text search.** Points to the search bar area.
- Specify a search keyword.** Points to the search bar area.
- A dynamically constructed search query appears here as you select search filters and specify search keywords. Query results appear in the program Explorer below.** Points to the main content area where a table of study details is shown.
- Export one or more files associated with a study.** Points to the "Add new study" and "Add new folder" buttons at the top right.
- Browse programs (groups of related studies) in a tree structure. Drill down to view a program's studies and its associated objects (analyses, assays and folders), if any. Click a node in the tree to display information about the object in the right pane of the window.** Points to the "Program Explorer" tree on the left side.

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.

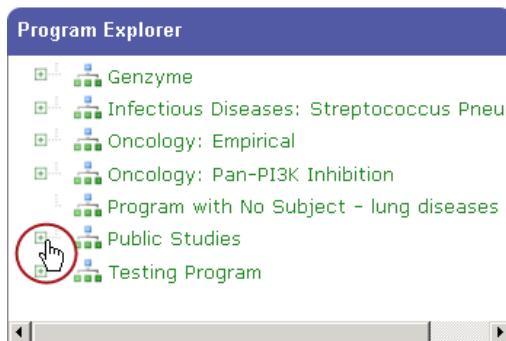


Lesson 2 – Browse Studies

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

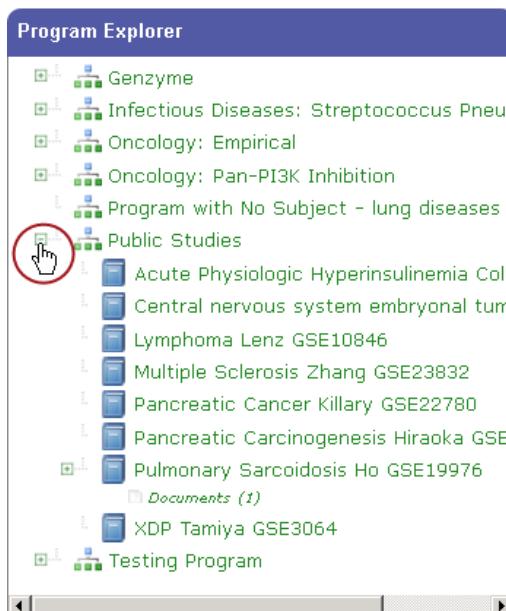
Scenario: You want to view assay information about a public breast cancer study that you are aware of, conducted by T. Sorlie.

1. In the transSMART 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:

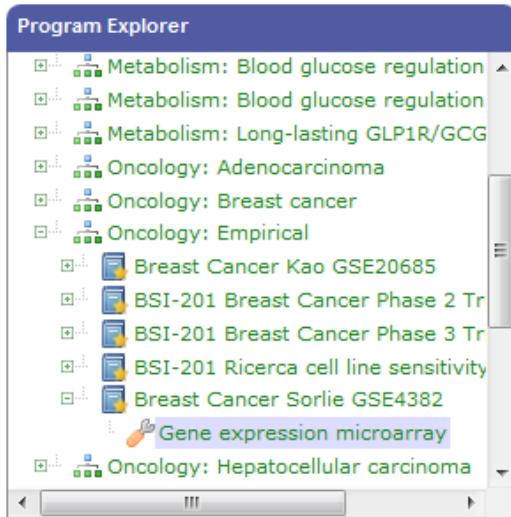


The list of public studies is short, but you don't see the Sorlie breast cancer study.

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, Lai A, Alizadeh A, Oning PE, Brown PO, Bitterman J, Brusen 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 transSMART 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.

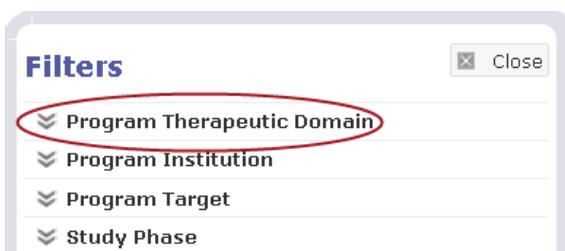
Scenario: You want to search for studies of cancer patients with either of two study objectives: to discover biomarkers, or to demonstrate clinical benefit.

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

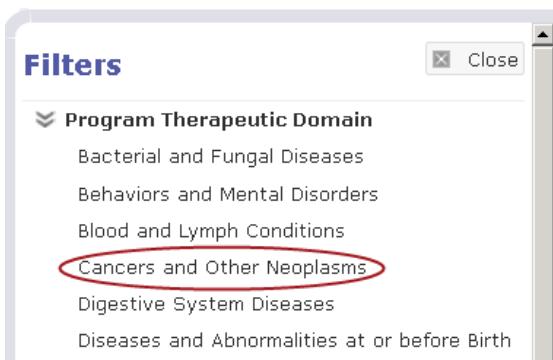


The Filters browser opens, displaying a list of filter categories.

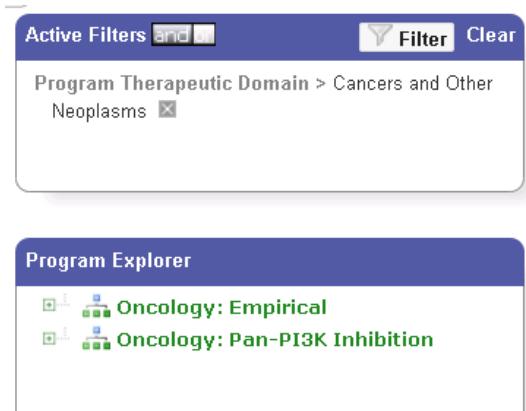
2. In the Filters browser, click the **Program Therapeutic Domain** category.



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**:

Study Objective

- Demonstrate clinical benefit Proof of Concept in clinic
- Demonstrate mechanistic Proof of Concept in clinic
- Discover biomarkers
- Discover targets/biological contexts

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:

Active Filters **AND** **Filter** **Clear**

Program Therapeutic Domain > Cancers and Other Neoplasms

AND

Study Objective > Discover biomarkers

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:

Active Filters **AND**

Program Therapeutic
Neoplasms

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:

Study Objective

- Demonstrate clinical benefit Proof of Concept in clinic
- Demonstrate mechanistic Proof of Concept in clinic
- Discover biomarkers**
- Discover targets/biological contexts



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

The screenshot shows the 'Active Filters' section of the transSMART interface. It has a blue header bar with 'Active Filters', 'and' (with a dropdown arrow), 'Filter' (with a magnifying glass icon), and 'Clear'. Below this, there are two filter categories: 'Program Therapeutic Domain > Cancers and Other Neoplasms' (with a close button) and 'AND' (with a dropdown arrow). Under 'AND', there are two more filter categories: 'Study Objective > Demonstrate clinical benefit Proof of Concept in clinic' (with a close button) and 'Discover biomarkers' (with a close button). A red circle highlights the 'or' button between the two 'Discover biomarkers' entries.

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:

A close-up screenshot of the 'Discover biomarkers' filter entry. It shows the text 'Discover biomarkers' followed by a close button, an 'or' button (highlighted with a red circle), and another 'or' button. This indicates that the 'Discover biomarkers' filter is part of an OR group.

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 screenshot shows the 'Active Filters' section again. The 'Discover biomarkers' filter entry is now removed, leaving only the 'Program Therapeutic Domain > Cancers and Other Neoplasms' filter. The 'Clear' button is highlighted with a red circle.

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:

The screenshot shows the 'Active Filters' section with the 'Clear' button highlighted with a red circle. All previous filters ('Program Therapeutic Domain > Cancers and Other Neoplasms' and the removed 'Discover biomarkers') are no longer present.

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.

Pancreatic Carcinogenesis Hiraoka GSE19650

Expression data from epithelial cells during the process of multistep pancreatic carcinogenesis

Pathology	
Study compound	
Study phase	
Study objective	Discover biomarkers
Study design	
Study biomarker type	Patient selection biomarker

- 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.

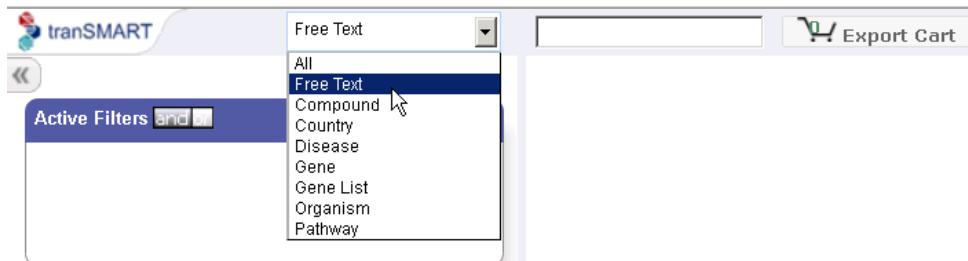
With free-text searches, tranSMART can find the text you type anywhere within the free-text metadata. For example, tranSMART would find **Hiraoka** in the title of the Pancreatic Carcinogenesis metadata above. It would also find **pancreatic carcinogenesis** in the description.

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 transSMART to search for all instances of the compound, whether the name appears in controlled-vocabulary or in free-text strings.

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



- 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.

- 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.

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



When you stop typing, the transSMART 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.

- 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:

The screenshot shows the tranSMART interface. At the top, there's a navigation bar with icons for Home, Log Out, and Help. Below it is a search bar with the placeholder "Search". To the right of the search bar is a dropdown menu set to "Compound". Underneath the search bar is the "Active Filters" section. It has a blue header with "Active Filters" and a radio button set to "and". Below this are two search results: "Free Text > iniparib" and "Compound > INIPARIB". Both results have a small "X" icon to their right. A red circle highlights the word "AND" between the two results. To the right of the Active Filters is a "Program Explorer" panel with a blue header. It displays the message "No results were found for this search.".

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:

This screenshot shows the "Active Filters" section with the radio button set to "or". Both "Free Text > iniparib" and "Compound > INIPARIB" results are listed, indicating that the search is now inclusive of both types of filters.

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.



Use wildcard characters to find information:

This screenshot shows the search results for the query "a_1\%". The results list several proteins: AAA1_HUMAN, AAK1_HUMAN, AAS1_HUMAN, and ABI1_HUMAN, each with its corresponding protein ID and description.



% (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.

Study: Pulmonary Sarcoidosis Ho GSE19976

Gene expression analysis of lung biopsies from patients with two different forms of pulmonary sarcoidosis

Subject-level data is available for this study. Open in Analyze view

Associated Tags

Property	Value
Study identifier	GSE19976
Pathology	Sarcoidosis, Pulmonary
Study compound	
Study phase	
Study publication DOI	
Study publication author list	
Study publication title	
Study publication status	

Associated Folders

Folder Name	File Type	Replicates
CEL data	Raw data	No replicates

Associated Assays

Assay Name	Biomarkers Studied	Measurement Type/Technology/Vendor/Platform Name
Assay for Pulmonary Sarcoidosis	BRCA2, RPS18, SLC39A7	Transcription Profiling/DNA Microarray/Bio-Medical Genomics Center, University of Minnesota./UMN Mouse 11K V 11_22_04_BMAP

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

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

Folder: CEL data

Results of intensity calculations.

Associated Tags

Property	Value
File Type	Raw data
Replicates	No replicates

Associated Files

File Name	Created on	Updated on	
GSM499117.CEL	2013-02-08	2013-02-08	Add to export
GSM499118.CEL	2013-02-14	2013-02-14	Add to export

[Export all](#)

- 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.

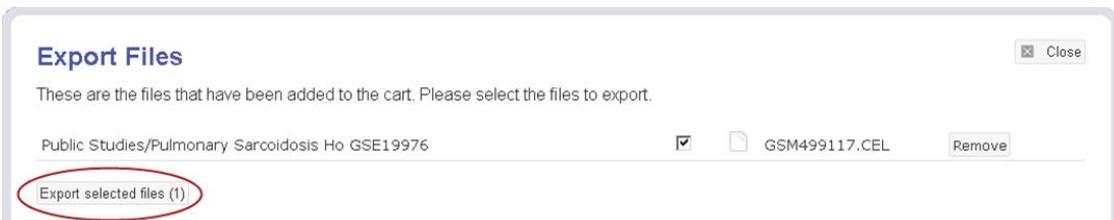
In this case, you want to download just one file from this study only.

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



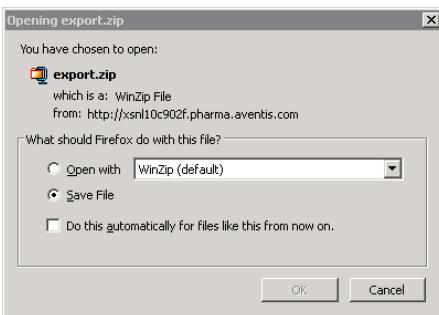
Study: Pulmonary Sarcoidosis Ho GSE19976
Gene expression analysis of lung biopsies from patients with two different forms

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



Export Files
These are the files that have been added to the cart. Please select the files to export.
Public Studies/Pulmonary Sarcoidosis Ho GSE19976 GSM499117.CEL

- When prompted, select **Save File** and click **OK**:

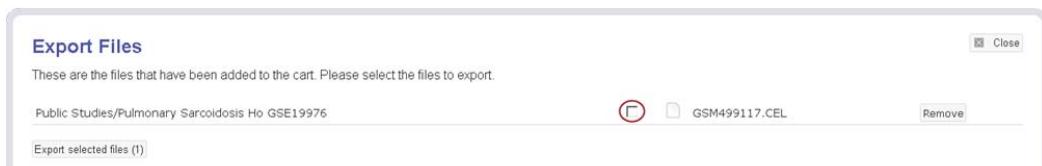


Note the following:

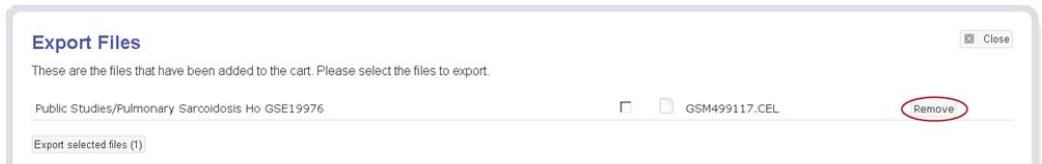
- tranSMART formats the file as a zip file, assigns it the name `export.zip`, and downloads the file to the `Downloads` directory on your computer.
- If a file named `export.zip` already exists in the directory, tranSMART changes the name to `export-1.zip` (or `export-2.zip`, `export-3.zip`, etc., depending 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:

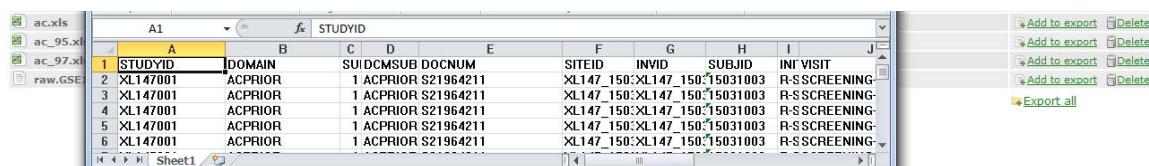


9. Click **Close** in the upper right corner of the Export Files dialog box to close it.



Tips

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



A1	STUDYID	DOMAIN	SUID	DCMSUB	DOCNUM	SITEID	INVID	SUBJID	INTVISIT	R-SCREENING
1	XL147001	ACPRIOR	1	ACPRIOR	S21964211	XL147_150	XL147_150	15031003	R-SCREENING	
2	XL147001	ACPRIOR	1	ACPRIOR	S21964211	XL147_150	XL147_150	15031003	R-SCREENING	
3	XL147001	ACPRIOR	1	ACPRIOR	S21964211	XL147_150	XL147_150	15031003	R-SCREENING	
4	XL147001	ACPRIOR	1	ACPRIOR	S21964211	XL147_150	XL147_150	15031003	R-SCREENING	
5	XL147001	ACPRIOR	1	ACPRIOR	S21964211	XL147_150	XL147_150	15031003	R-SCREENING	
6	XL147001	ACPRIOR	1	ACPRIOR	S21964211	XL147_150	XL147_150	15031003	R-SCREENING	

Lesson 6 – Create and Edit Objects

This lesson is for transSMART ‘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.

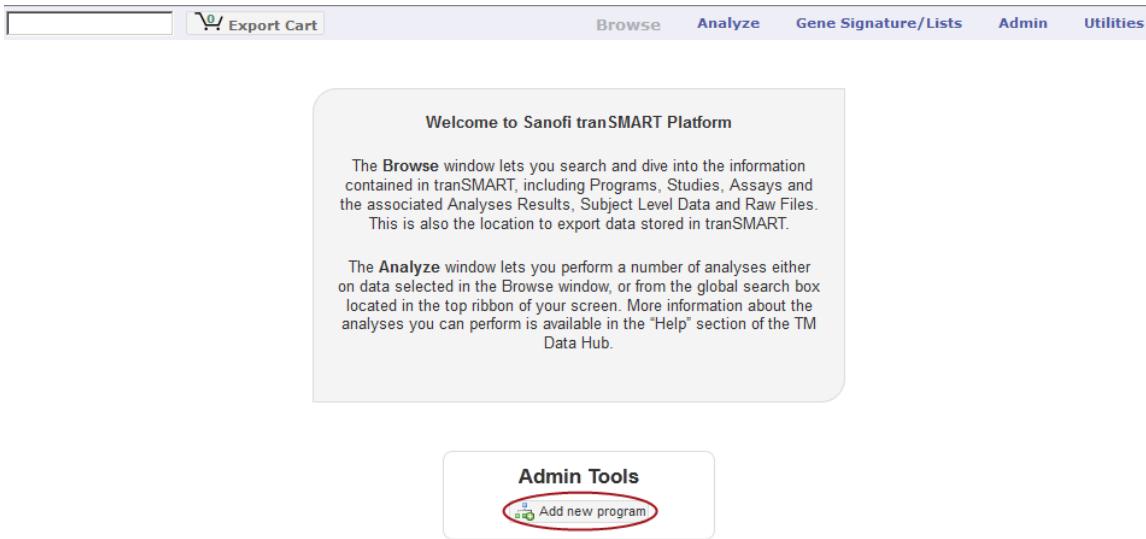
Create Objects

Creating a Program

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

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

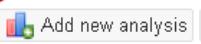
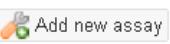
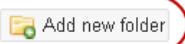
- Click **Add new program**.



The screenshot shows the transSMART platform interface. At the top, there is a navigation bar with tabs: Export Cart, Browse, Analyze, Gene Signature/Lists, Admin, and Utilities. Below the navigation bar, a large central area displays a welcome message: "Welcome to Sanofi transSMART Platform". This area contains two sections of text: one about the Browse window and another about the Analyze window. At the bottom of this central area, there is a section titled "Admin Tools" which contains a button labeled "Add new program". The "Add new program" button is circled with a red oval.

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:

Breast cancer Sorlie GSE4382   

Gene expression profiling to classify tumors into clinically relevant subgroups, main focus on chemotherapy-treated advanced breast cancer

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

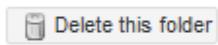
Selected Object in Program Explorer	Objects You Can Create
Program	<ul style="list-style-type: none">▪ Study▪ Folder
Study	<ul style="list-style-type: none">▪ Analysis▪ Assay▪ Folder
Analysis	<ul style="list-style-type: none">▪ Folder
Assay	<ul style="list-style-type: none">▪ Folder
Folder	<ul style="list-style-type: none">▪ Sub-folder

Adding Files to Folders

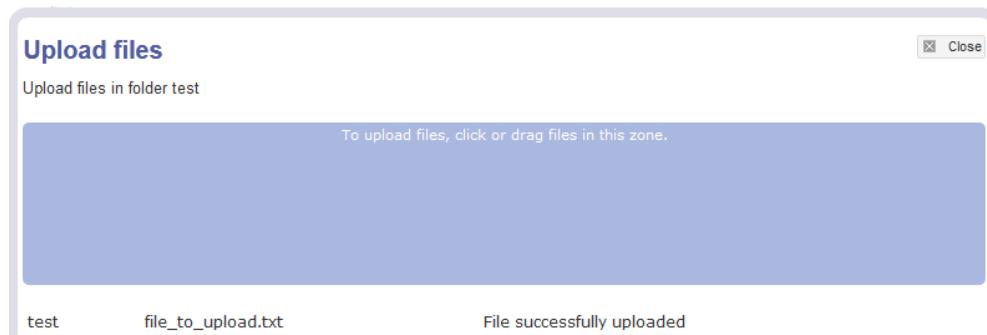
Folders allow you to attach files to an object; for example, you might add a folder to contain files pertaining to the analysis of a study, or a gene list for an analysis.

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:

Oncology: Empirical

Oncology program comprising all compounds without identified targets

Program target/pathway/phenotype	BRCA1 , unknown	
Therapeutic domain	Cancers and Other Neoplasms	
Program institution	Oncology R&D	

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:

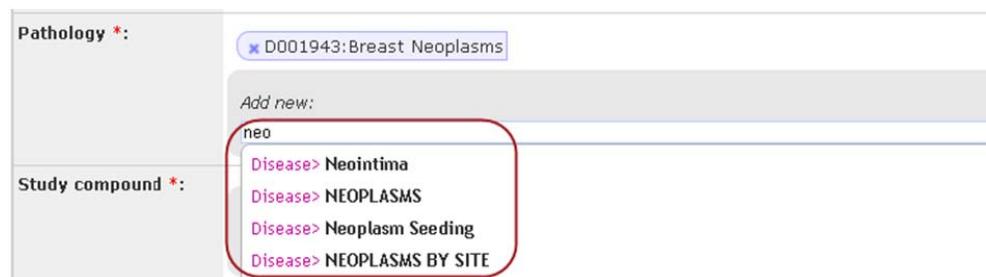


- Study identifier must be unique and cannot be modified later on. It is limited to 100 characters. Avoid using specific characters.



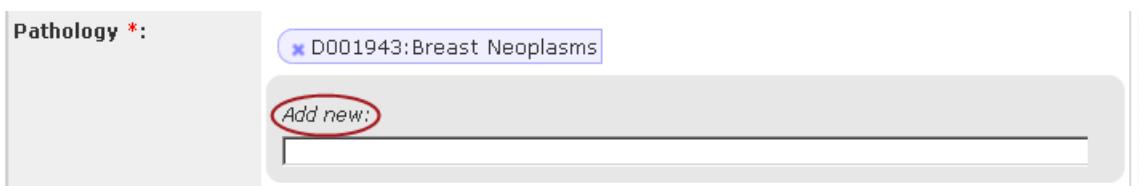
- Shaded fields are autocomplete fields. Type one or more characters at the beginning of the value that you want to assign, and tranSMART will display a list of text strings that begin with those characters.

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



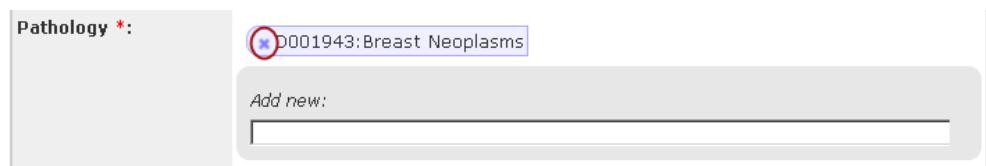
The screenshot shows a 'Pathology *:' field containing the value 'D001943:Breast Neoplasms'. To the right, a dropdown menu titled 'Add new:' lists several suggestions starting with 'neo': 'Disease> Neointima', 'Disease> NEOPLASMS', 'Disease> Neoplasm Seeding', and 'Disease> NEOPLASMS BY SITE'. The first suggestion, 'Disease> Neointima', is highlighted with a red oval.

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



The screenshot shows a 'Pathology *:' field containing the value 'D001943:Breast Neoplasms'. Above the field, the text 'Add new:' is displayed in bold. Below the field is a text input box.

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



The screenshot shows a 'Pathology *:' field containing the value 'D001943:Breast Neoplasms'. To the left of the value, there is a small blue circle with a white 'x' icon, indicating it can be removed. Below the field is a text input box.

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:

Folder: CEL data

Add new folder  Delete this folder

Results of intensity calculations.

Associated Tags

Property	Value
File Type	Raw data
Replicates	No replicates

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:

Associated Files

File Name	Created on	Updated on	
GSM499117.CEL	2013-02-08	2013-02-08	 
GSM499118.CEL	2013-02-14	2013-02-14	 



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**:

Pulmonary Sarcoidosis Ho GSE19976

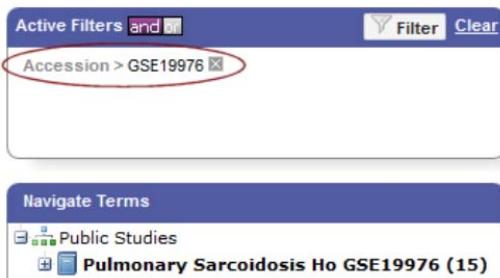
Gene expression analysis of lung biopsies from patients with two different forms of pulmonary sarcoidosis

Subject-level data is available for this study. [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.

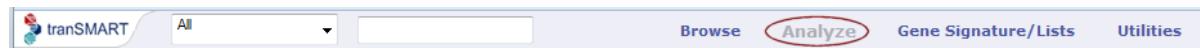


The screenshot shows the tranSMART Analyze interface. At the top, there is a header bar with 'Active Filters' and a search bar. Below this, the 'Active Filters' panel shows a single filter: 'Accession > GSE19976'. The 'Navigate Terms' panel below it lists 'Public Studies' and 'Pulmonary Sarcoidosis Ho GSE19976 (15)'. Both the 'Accession > GSE19976' filter in the Active Filters panel and the 'Pulmonary Sarcoidosis Ho GSE19976 (15)' term in the Navigate Terms panel are circled with red ovals.

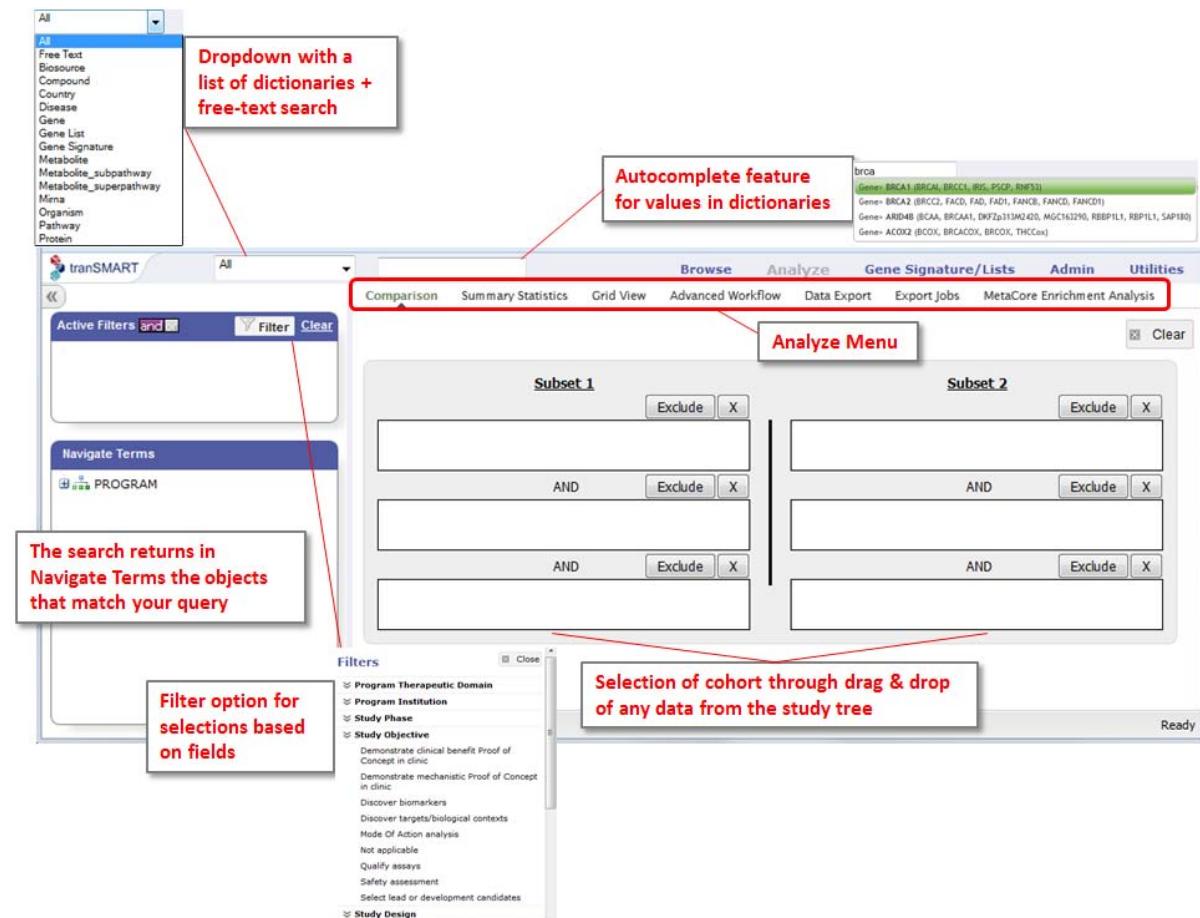
Lesson 8 – Quick tour of the Analyze Module

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.



The screenshot shows the tranSMART Analyze window with several annotations highlighting its features:

- Dropdown with a list of dictionaries + free-text search:** Points to the dropdown menu in the top-left corner, which includes options like All, Free Text, Biosource, Compound, Country, Disease, Gene, Gene List, Gene Signature, Metabolite, Metabolite_subpathway, Metabolite_superpathway, Mims, Organism, Pathway, and Protein. A red box highlights the "All" option.
- Autocomplete feature for values in dictionaries:** Points to a search bar containing the text "brca". Below it, a dropdown menu lists gene names: ERCA1, BRCA1, ERCC1, IUS, PSCP, RNF31; ERCA2, BRCA2, FAD, FAD1, FANCB, FANCD1; ARID4B, ERCA1, ERCC1, DRF2p13M24D0, MGC163210, RBBP1L1, SAP180; and ACOD2, BC0X, BRCA0X, BC0X, THC0X.
- Analyze Menu:** Points to the "Analyze" tab in the top navigation bar.
- Subset 1 and Subset 2:** Points to the two main sections of the analysis interface, each containing a list of items and an "Exclude" button.
- Comparison, Summary Statistics, Grid View, Advanced Workflow, Data Export, Export Jobs, MetaCore Enrichment Analysis:** Points to the tabs in the top navigation bar.
- Active Filters and Filter:** Points to the filter section on the left side of the interface.
- Navigate Terms:** Points to the "PROGRAM" category under Navigate Terms.
- The search returns in Navigate Terms the objects that match your query:** Points to the search results in the Navigate Terms section.
- Filter option for selections based on fields:** Points to the "Filters" section at the bottom left.
- Selection of cohort through drag & drop of any data from the study tree:** Points to the "Ready" section at the bottom right.

Overview of the main features available under Analyze

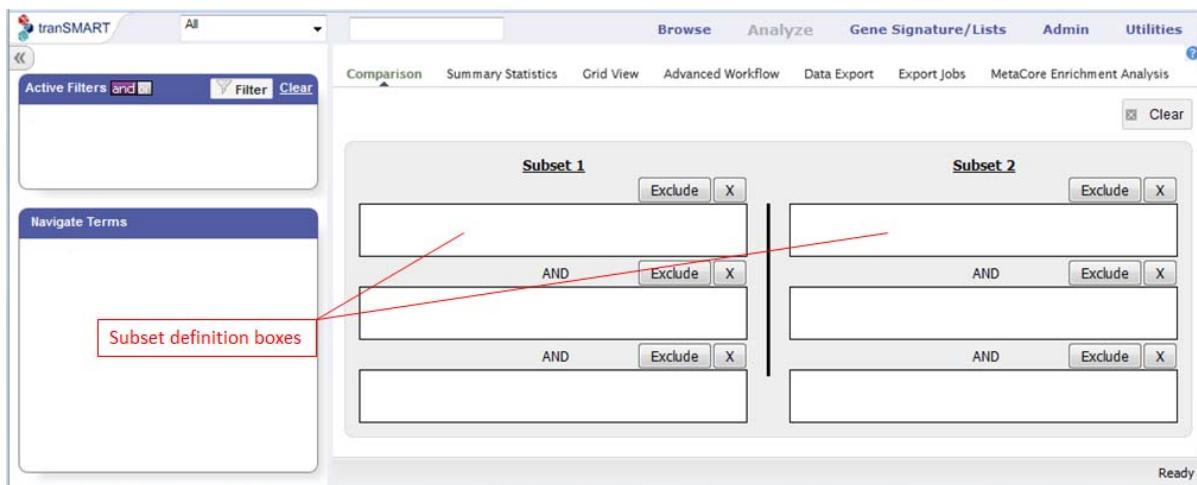
The Analyze module interface is divided into two panels:

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: <ul style="list-style-type: none"> ■ 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.

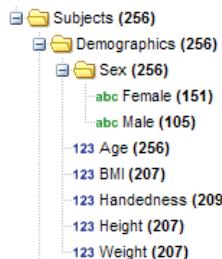
Selecting Criteria for the Cohort

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.



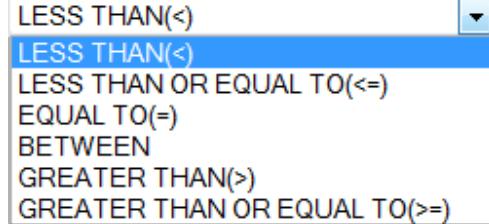
Specifying a Numeric Value

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	<p>Values are not constrained. All the numeric data associated with the concept are factored into the subset definition.</p> <p>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.</p>
By high/low flag	<p>If the testing laboratory has grouped the numeric values into High/Low/Normal ranges, select the range to factor into the subset definition.</p> <p>When you select By high/low flag, the Please select range field appears. Select the range you want and click OK.</p>

Selection	Description
	<p>By numeric value Values are constrained by an exact value or a range of values. After you select By numeric value:</p> <p>Select one of the following numeric operators in the Please select operator dropdown:</p>  <p>In Please enter value, type the numeric value that the operator applies to.</p> <p>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.</p> <p>Click OK.</p> <p>See the next section for information on viewing the numeric values associated with the concept and that you can select from.</p>



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:

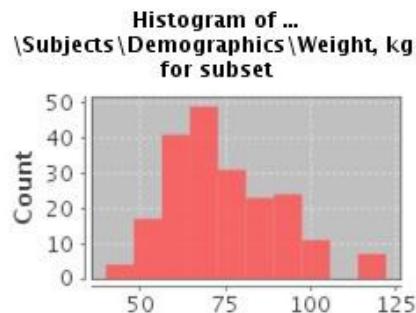


Viewing the Numeric Values Associated with a Concept

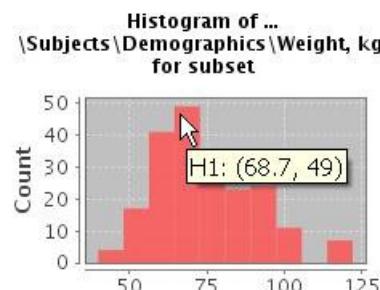
Note the buttons **Show Histogram** and **Show Histogram for subset** in the Set Value dialog. The histograms show how the numeric values associated with the concept that you placed in the subset box are distributed among the subjects across both subsets, or in the particular subset you are currently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

A screenshot of a software interface showing a "Subset 1" dialog box. It contains two stacked subset definition boxes. The top box has an "Exclude" button and contains the criterion "...Male". The bottom box also has an "Exclude" button and contains the criterion "...Weight between 65 and 90". Between the two boxes is an "AND" operator.

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

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

Subset 1

<code>...Weight <=50</code> <code>...Weight >=100</code>	<input type="button" value="Exclude"/> 
---	--

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

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

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

Subset 1

<code>...Male\</code>	<input type="button" value="Exclude"/> 
AND	
<code>...!Weight between 50 and 100</code>	<input type="button" value="Include"/> 

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:

Subset 1

	<input type="button" value="Exclude"/> 
--	---

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:

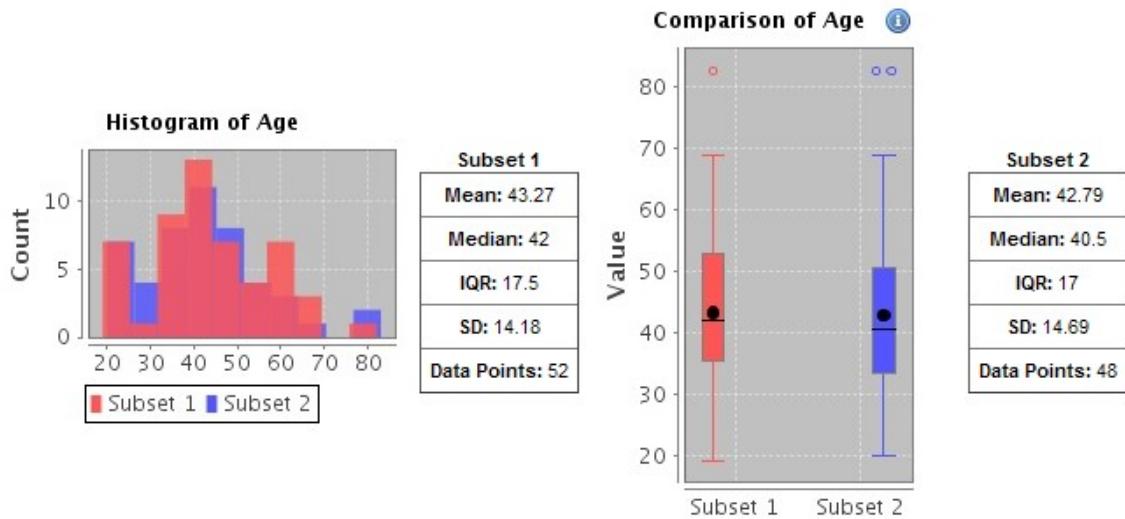
Query Summary for Subset 1	Query Summary for Subset 2
(\Public Studies\Public Studies\Lymphoma_Staudt_GSE10846\Biomarker Data\Gene Expression\) AND (\Public Studies\Public Studies\Lymphoma_Staudt_GSE10846\Subjects \Demographics\Gender\Female\)	(\Public Studies\Public Studies\Lymphoma_Staudt_GSE10846\Biomarker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

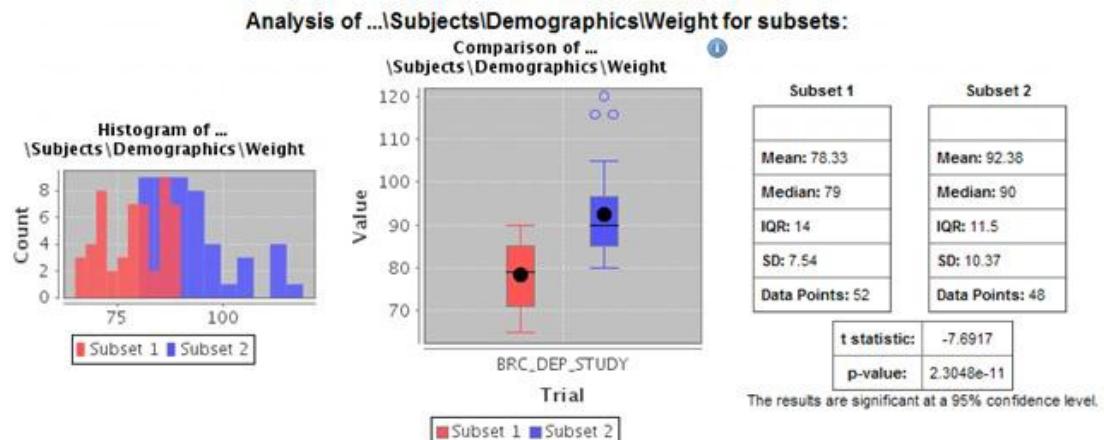
Subject Totals		
Subset 1	Both	Subset 2
52	25	48

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\[\]\[\]\)](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\[\]\[\]\)](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.

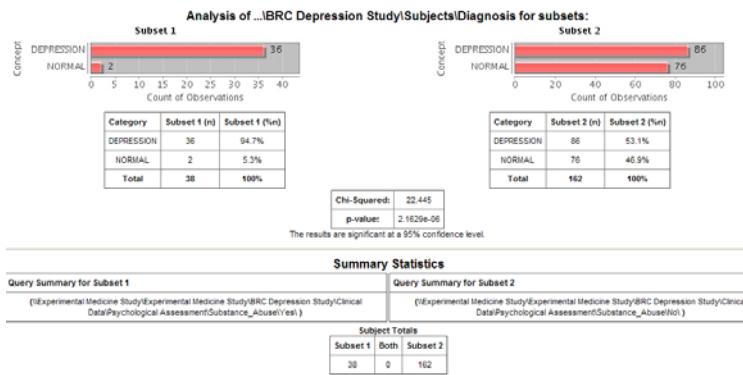
As soon as you drop the point of comparison into the Results/Analysis view, tranSMART begins to compare the subsets based on that point of comparison. When finished, tranSMART displays a side-by-side summary of how the subjects in each subset match or respond to the point of comparison.

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.

The screenshot shows the tranSMART interface with the 'Comparison' tab selected. On the left, the 'Navigate Terms' sidebar shows the study 'Lymphoma_Lenz_GSE10846' expanded. Two red arrows point from the 'Female' and 'Male' nodes under 'Demographics\Gender' in the tree view to their respective entries in the 'Subset 1' and 'Subset 2' sections of the comparison builder. The 'Subset 1' section contains the criterion '...Female' and the 'Subset 2' section contains '...Male'.

4. Click the **Summary Statistics** tab.
5. Click the **Grid View** tab.

Active Filters and Filter Clear

Free Text > GSE10846

Grid View

Subject	Patient	Samples	Subset	Trial	Sex	Age	Race
1000004318	GSM274921	GSM274921	subset1	GSE10846	Female	63	NULL
1000004319	GSM274926	GSM274926	subset1	GSE10846	Female	75	NULL
1000004320	GSM274928	GSM274928	subset1	GSE10846	Female	76	NULL
1000004322	GSM274901	GSM274901	subset1	GSE10846	Female	18	NULL
1000004324	GSM274911	GSM274911	subset1	GSE10846	Female	81	NULL
1000004329	GSM274959	GSM274959	subset1	GSE10846	Female	70	NULL
1000004330	GSM274964	GSM274964	subset1	GSE10846	Female	72	NULL
1000004332	GSM274967	GSM274967	subset1	GSE10846	Female	68	NULL
1000004334	GSM274976	GSM274976	subset1	GSE10846	Female	71	NULL
1000004336	GSM274980	GSM274980	subset1	GSE10846	Female	77	NULL
1000004339	GSM274997	GSM274997	subset1	GSE10846	Female	45	NULL
1000004340	GSM274948	GSM274948	subset1	GSE10846	Female	59	NULL
1000004342	GSM275026	GSM275026	subset1	GSE10846	Female	58	NULL
1000004344	GSM275054	GSM275054	subset1	GSE10846	Female	46	NULL
1000004349	GSM275086	GSM275086	subset1	GSE10846	Female	79	NULL
1000004354	GSM275072	GSM275072	subset1	GSE10846	Female	70	NULL
1000004356	GSM275164	GSM275164	subset1	GSE10846	Female	56	NULL

Export to Excel

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.

Active Filters and Filter Clear

Free Text > GSE10846

Grid View

Subject	Patient	Samples	Subset	Trial	Sex	Age	Race	Survival_at...
1000004318	GSM274921	GSM274921	subset1	GSE10846	Female	63	NULL	2.31
1000004319	GSM274926	GSM274926	subset1	GSE10846	Female	75	NULL	0.34
1000004320	GSM274928	GSM274928	subset1	GSE10846	Female	76	NULL	10.53
1000004322	GSM274901	GSM274901	subset1	GSE10846	Female	18	NULL	0.05
1000004324	GSM274911	GSM274911	subset1	GSE10846	Female	81	NULL	3.93
1000004329	GSM274959	GSM274959	subset1	GSE10846	Female	70	NULL	7.24
1000004330	GSM274964	GSM274964	subset1	GSE10846	Female	72	NULL	0.99
1000004332	GSM274967	GSM274967	subset1	GSE10846	Female	68	NULL	1.22
1000004334	GSM274976	GSM274976	subset1	GSE10846	Female	71	NULL	3.47
1000004336	GSM274980	GSM274980	subset1	GSE10846	Female	77	NULL	123 Survival at Follow Up (Years) (414)
1000004339	GSM274997	GSM274997	subset1	GSE10846	Female	45	NULL	16.79
1000004340	GSM274948	GSM274948	subset1	GSE10846	Female	59	NULL	0.31
1000004342	GSM275026	GSM275026	subset1	GSE10846	Female	58	NULL	1.24
1000004344	GSM275054	GSM275054	subset1	GSE10846	Female	46	NULL	4.94
1000004349	GSM275086	GSM275086	subset1	GSE10846	Female	79	NULL	2.71
1000004354	GSM275072	GSM275072	subset1	GSE10846	Female	70	NULL	2.26
1000004356	GSM275164	GSM275164	subset1	GSE10846	Female	56	NULL	3.29

Export to Excel

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

Subject	Patient	Samples	Subset	Trial	Sex	Age	Race	Survival_at...
1000004645	GSM275099	GSM275099	subset1	GSE10846	Female	92		
1000004546	GSM275155	GSM275155	subset2	GSE10846	Male	92		
1000004474	GSM274982	GSM274982	subset1	GSE10846	Female	88		
1000004727	GSM275287	GSM275287	subset1	GSE10846	Female	87		
1000004357	GSM275165	GSM275165	subset1	GSE10846	Female	86		
1000004370	GSM275188	GSM275188	subset1	GSE10846	Female	86		
1000004360	GSM275196	GSM275196	subset2	GSE10846	Male	85		
1000004419	GSM274925	GSM274925	subset2	GSE10846	Male	85		
1000004458	GSM275185	GSM275185	subset2	GSE10846	Male	85		
1000004492	GSM275067	GSM275067	subset2	GSE10846	Male	85		
1000004467	GSM274908	GSM274908	subset1	GSE10846	Female	84		
1000004524	GSM274972	GSM274972	subset1	GSE10846	Female	84		
1000004541	GSM275159	GSM275159	subset1	GSE10846	Female	84		
1000004597	GSM275125	GSM275125	subset1	GSE10846	Female	84		
1000004609	GSM275183	GSM275183	subset2	GSE10846	Male	84		
1000004501	GSM275126	GSM275126	subset1	GSE10846	Female	83		
1000004545	GSM275146	GSM275146	subset1	GSE10846	Female	83		

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

Subject	Patient	Samples	Subset	Trial	Sex	Age	Survival_at_Foll...
1000004645	GSM275099	GSM275099	subset1	GSE10846	Female	92	
1000004546	GSM275155	GSM275155	subset2	GSE10846	Male	92	
1000004474	GSM274982	GSM274982	subset1	GSE10846	Female	88	
1000004727	GSM275287	GSM275287	subset1	GSE10846			
1000004357	GSM275165	GSM275165	subset1	GSE10846			
1000004370	GSM275188	GSM275188	subset1	GSE10846			
1000004360	GSM275196	GSM275196	subset2	GSE10846			
1000004419	GSM274925	GSM274925	subset2	GSE10846			
1000004458	GSM275185	GSM275185	subset2	GSE10846			
1000004492	GSM275067	GSM275067	subset2	GSE10846			
1000004467	GSM274908	GSM274908	subset1	GSE10846			
1000004524	GSM274972	GSM274972	subset1	GSE10846			
1000004541	GSM275159	GSM275159	subset1	GSE10846			
1000004597	GSM275125	GSM275125	subset1	GSE10846			
1000004609	GSM275183	GSM275183	subset2	GSE10846			
1000004501	GSM275126	GSM275126	subset1	GSE10846			
1000004545	GSM275146	GSM275146	subset1	GSE10846			

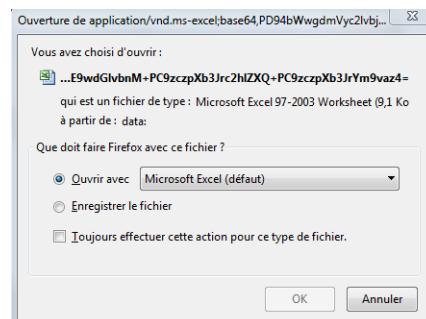
9. Select only the 10 first rows by clicking the first row in the 'Grid View' table, SHIFT in your keyboard and the tenth row.

Subject	Patient	Samples	Subset	Trial	Sex	Age	Survival_at_Foll...
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

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

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
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

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

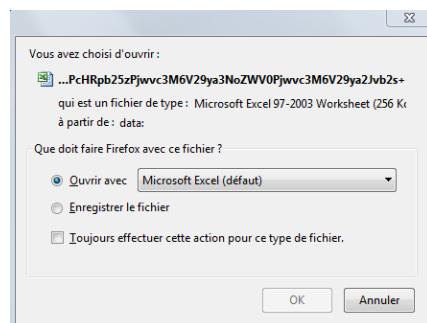


11. Visualize data in the Excel file.

	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.



	A	B	C	D	E	F	G	H
1	Grid View Generated by ExtJS							
2	Subject	Patient	Samples	Subset	Trial	Sex	Age	Survival_at_FollowUp_(Years)
3	1000004645	GSM275099	GSM275099	subset1	GSE10846	Female	92	2.12
4	1000004546	GSM275155	GSM275155	subset2	GSE10846	Male	92	0.74
5	1000004474	GSM274982	GSM274982	subset1	GSE10846	Female	88	3.88
6	1000004727	GSM275287	GSM275287	subset1	GSE10846	Female	87	0.53
7	1000004357	GSM275165	GSM275165	subset1	GSE10846	Female	86	1.63
8	1000004370	GSM275188	GSM275188	subset1	GSE10846	Female	86	0.15
9	1000004360	GSM275196	GSM275196	subset2	GSE10846	Male	85	2.74
10	1000004419	GSM274925	GSM274925	subset2	GSE10846	Male	85	5.38
11	1000004458	GSM275185	GSM275185	subset2	GSE10846	Male	85	0.09
12	1000004492	GSM275067	GSM275067	subset2	GSE10846	Male	85	1.51
13	1000004467	GSM274908	GSM274908	subset1	GSE10846	Female	84	0.33
14	1000004524	GSM274972	GSM274972	subset1	GSE10846	Female	84	0.2
15	1000004541	GSM275159	GSM275159	subset1	GSE10846	Female	84	5.51
16	1000004597	GSM275125	GSM275125	subset1	GSE10846	Female	84	1.41
17	1000004609	GSM275183	GSM275183	subset2	GSE10846	Male	84	0.03
18	1000004501	GSM275126	GSM275126	subset1	GSE10846	Female	83	4.8
19	1000004545	GSM275146	GSM275146	subset1	GSE10846	Female	83	0.38
20	1000004388	GSM275027	GSM275027	subset2	GSE10846	Male	83	0.16
21	1000004393	GSM275080	GSM275080	subset2	GSE10846	Male	83	3.42
22	1000004673	GSM275292	GSM275292	subset2	GSE10846	Male	83	0.74
23	1000004706	GSM275070	GSM275070	subset2	GSE10846	Male	83	0.33
24	1000004553	GSM275206	GSM275206	subset2	GSE10846	Male	82	1.27
25	1000004324	GSM274911	GSM274911	subset1	GSE10846	Female	81	3.93
26	1000004531	GSM275044	GSM275044	subset1	GSE10846	Female	81	5.8

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.

The screenshot shows the tranSMART interface with the search term 'GSE4382' entered in the search bar. The 'Active Filters' section shows 'Free Text > GSE4382'. Below the search bar, a 'Navigate Terms' sidebar lists 'Oncology: Empirical' and 'Breast Cancer Sorlie GSE4382 (167)'.

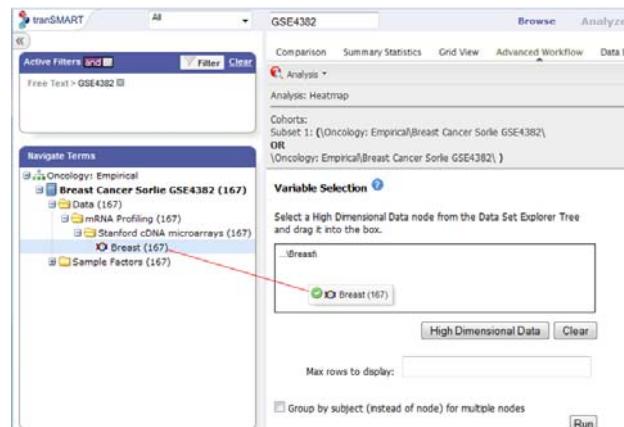
2. In the Navigate Terms area, under the Program, open the Study '**Breast Cancer Sorlie GSE4382**'.

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

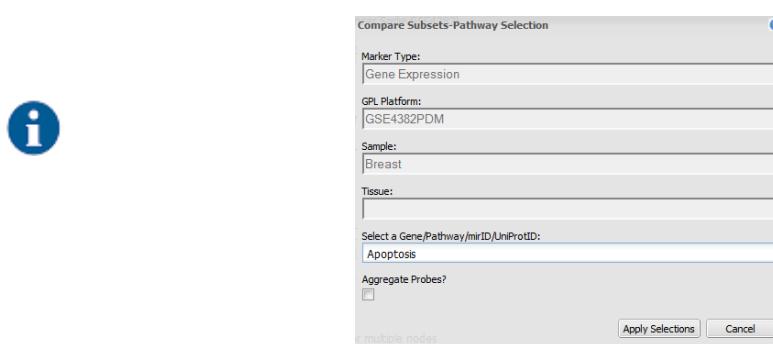


4. Go to the **Advanced Workflow**, then select **Heatmap** from the Analysis dropdown menu.

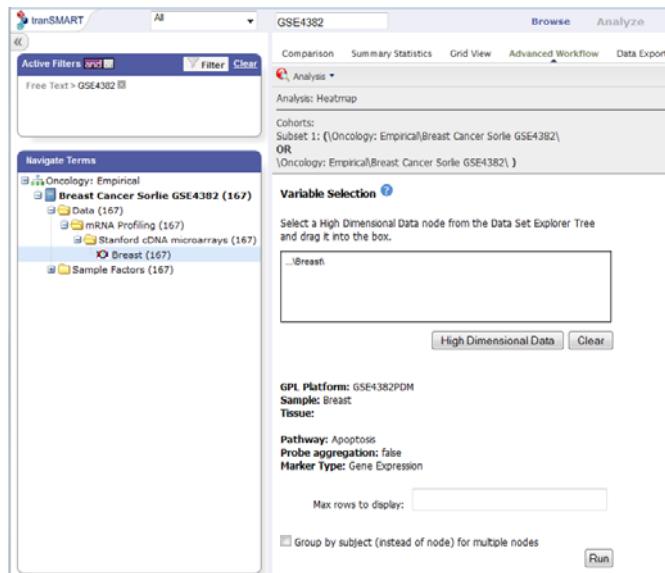
5. Drag and drop the node '\Breast Cancer Sorlie GSE4382\Data\...\\Breast' under Variable Selection.



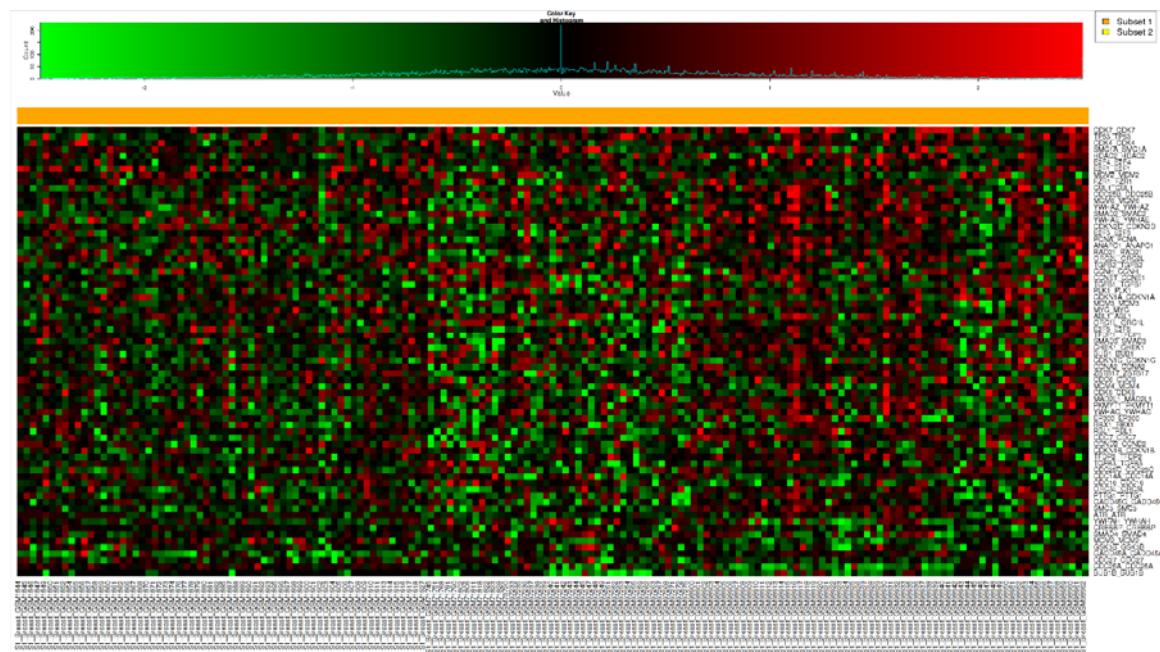
6. Click the **High Dimensional Data** button, add the pathway **Apoptosis** and click **Apply Selections**.



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.



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

Hierarchical Clustering

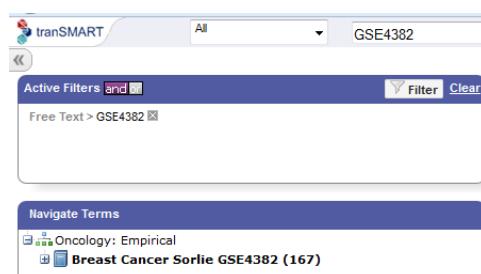
Hierarchical clustering is a type of clustering analysis whose goal is to organize data so that the objects in the same cluster are more similar to each other than to those in other clusters.

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.



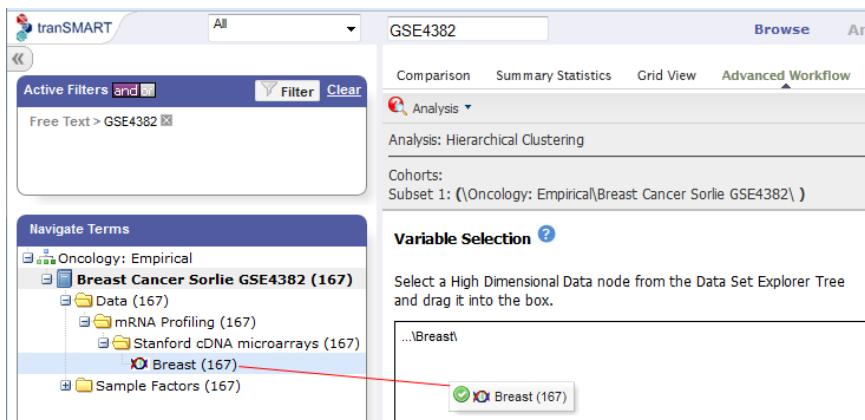
2. Under the Program, open the Study '**Breast Cancer Sorlie GSE4382**'.

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

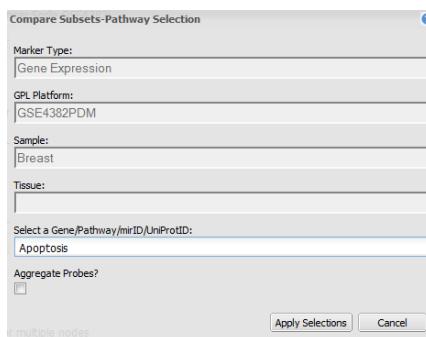


4. Go to the **Advanced Workflow**, then select **Hierarchical Clustering** from the Analysis dropdown menu.

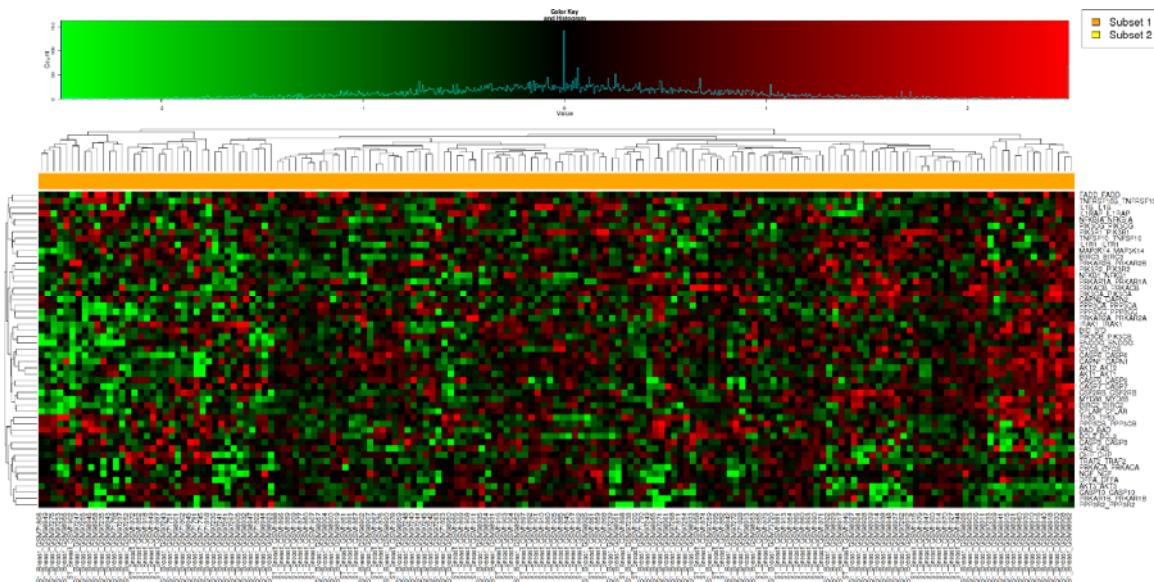
5. Drag and drop the node '\Breast Cancer Sorlie GSE4382\Data\...\\Breast' under Variable Selection.



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

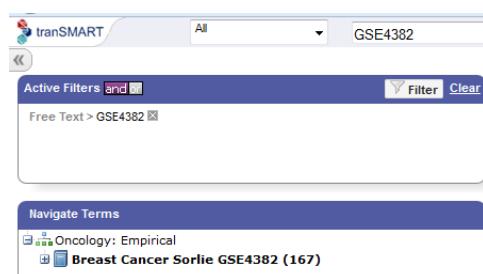
The K-Means clustering heatmap clusters samples into a specified number of clusters. The result is k clusters, each centered around a randomly-selected data point.

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**'.

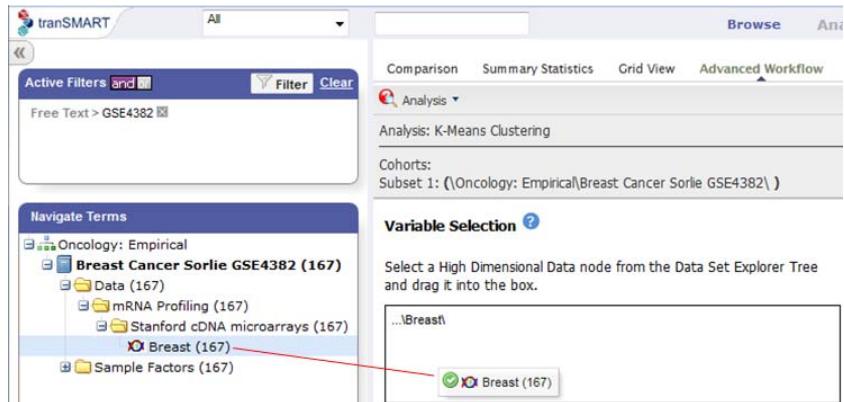


3. In the Comparison tab, drag and drop the Study '**Breast Cancer Sorlie GSE4382**' as Subset 1 cohort.



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**.

Compare Subsets-Pathway Selection

Marker Type: Gene Expression

GPL Platform: GSE4382PDM

Sample: Breast

Tissue:

Select a Gene/Pathway/mirID/UniProtID: Apoptosis

Aggregate Probes?

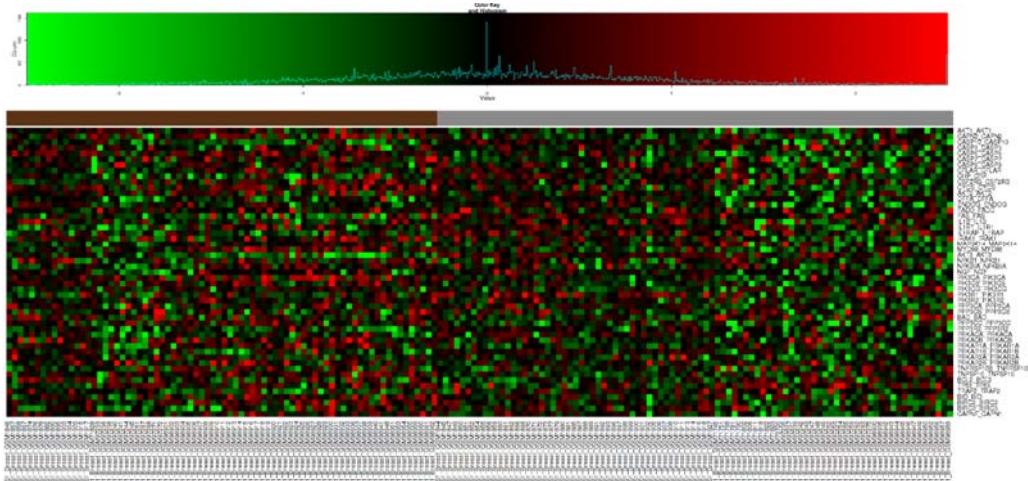
multiple nodes

Apply Selections **Cancel**

The screenshot shows the tranSMART software interface with the following details:

- Top Navigation:** Comparison, Summary Statistics, Grid View, Advanced Workflow, Data E.
- Analysis Section:**
 - Analysis: K-Means Clustering
 - Cohorts: Subset 1: (Oncology: Empirical Breast Cancer Series GSE4382)
- Variable Selection:** A tree view under "Navigate Terms" shows "Oncology: Empirical" expanded, with "Breast Cancer Series GSE4382 (167)" selected. Other nodes include "Data (167)", "mRNA Profiling (167)", "Stanford cDNA microarrays (167)", "Breast (167)", and "Sample Factors (167)". A box labeled "Select a High Dimensional Data node from the Data Set Explorer Tree and drag it into the box." contains the text "Breast". Buttons "High Dimensional Data" and "Clear" are below the box.
- Parameter Settings:**
 - GPL Platform: GSE4382PDM
 - Sample: Breast
 - Tissue:
 - Pathway: Apoptosis
 - Probe aggregation: false
 - Marker Type: Gene Expression
- Run Button:** A "Run" button is located at the bottom right of the parameter section.

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

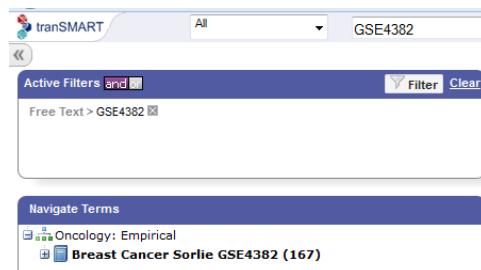
Marker Selection is a display of the top differentially expressed genes between two specified cohorts.

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.



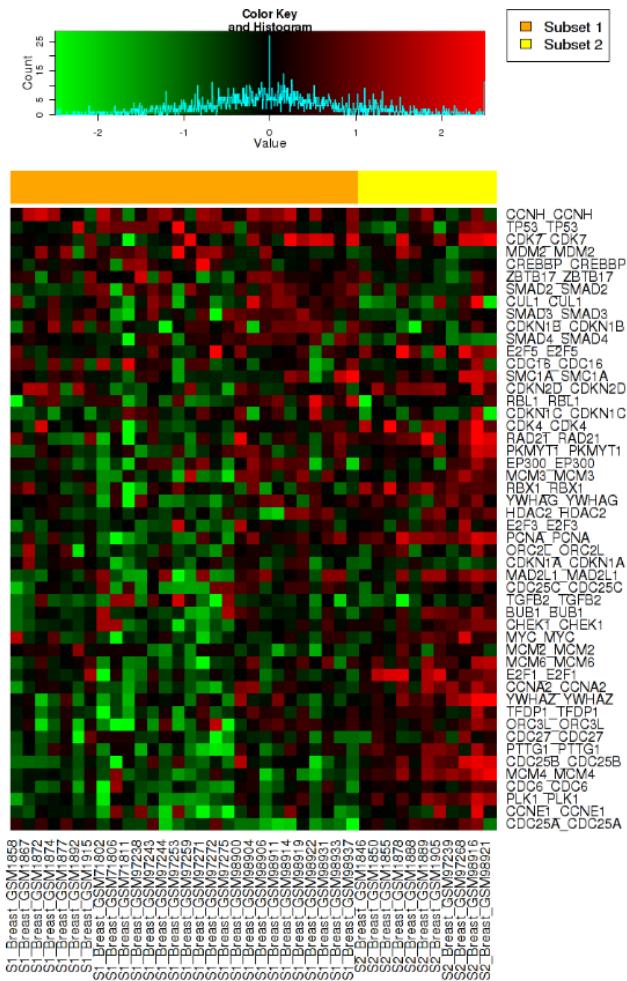
2. Under the Program, open the Study '**Breast Cancer Sorlie GSE4382**'.
3. In the Comparison tab, drag and drop the Study '\Breast Cancer Sorlie GSE4382\Sample Factors\Tumor Subtype\Luminal A' as Subset 1 cohort and '\Breast Cancer Sorlie GSE4382\Sample Factors\Tumor Subtype\Luminal B' as Subset 2 cohort.

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**'.

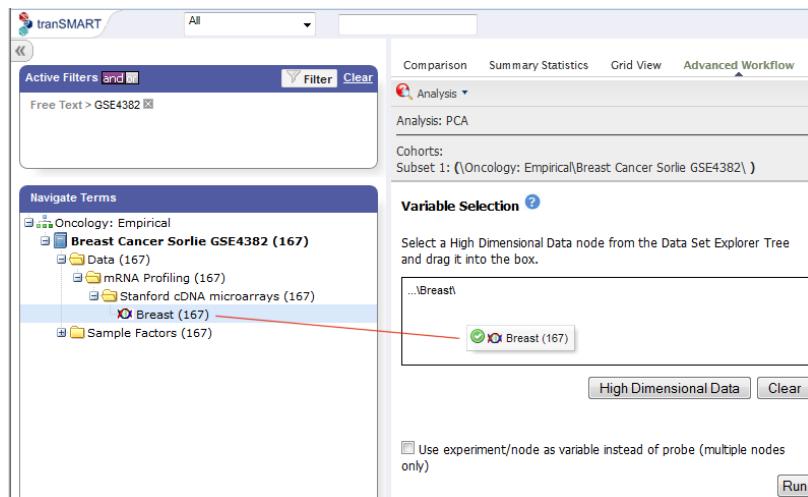


3. In the Comparison tab, drag and drop the Study '**Breast Cancer Sorlie GSE4382**' as Subset 1 cohort.



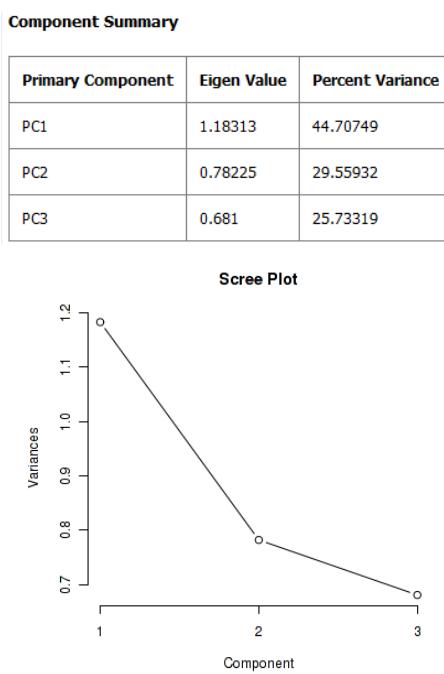
4. Go to the **Advanced Workflow**, then select **PCA** from the Analysis dropdown menu.

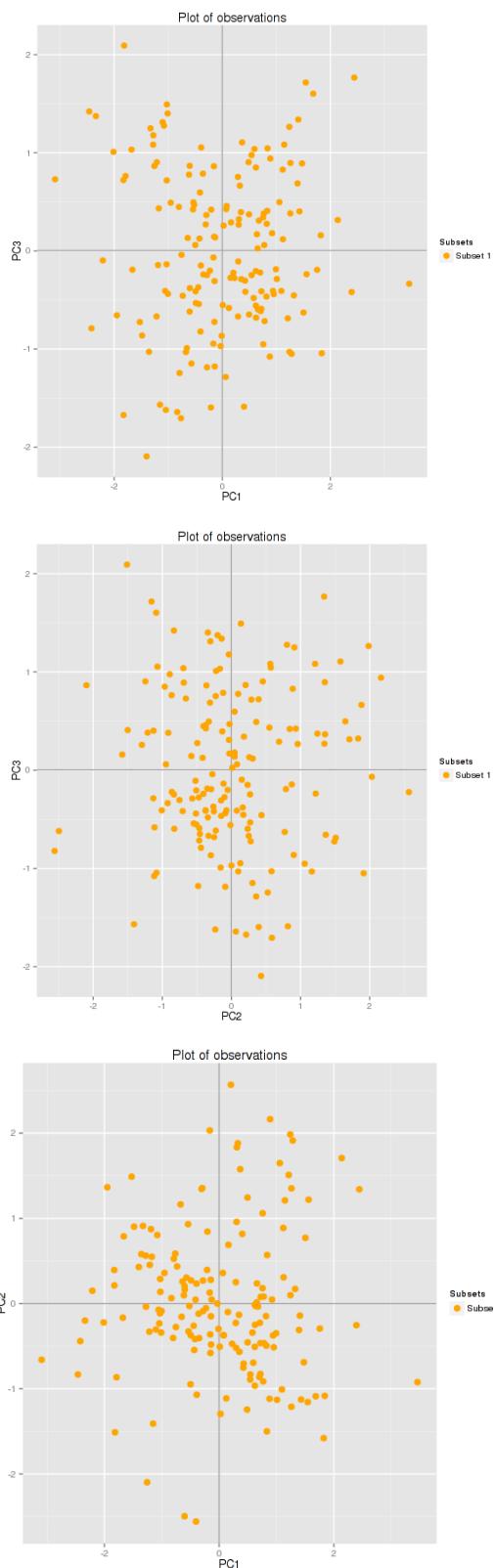
5. Drag and drop the node '\Breast Cancer Sorlie GSE4382\Data\...\Breast' under Variable Selection.

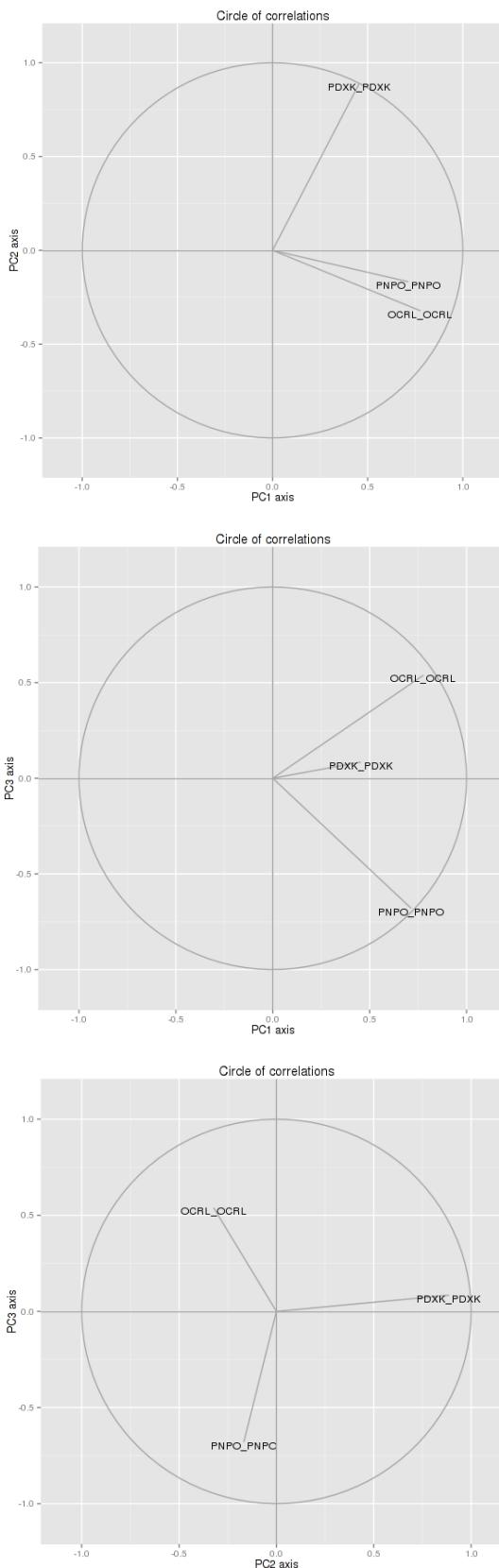


6. Click the **High Dimensional Data** button, add the pathway **Vitamin B6 metabolism** and click **Apply Selections**.

7. Click the **Run** button.







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.



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.

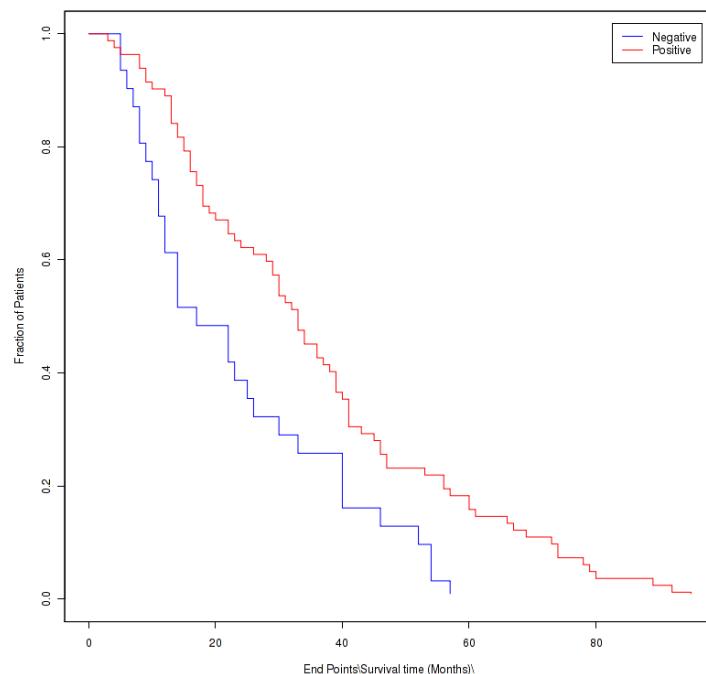
5. Drag and drop the node '\Breast Cancer Sorlie GSE4382\Sample Factors\End Points\Survival time (Months)' under the Time box.

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

Kaplan-Meier estimator



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.
- A blue circular icon containing a white letter 'i'.
- 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.

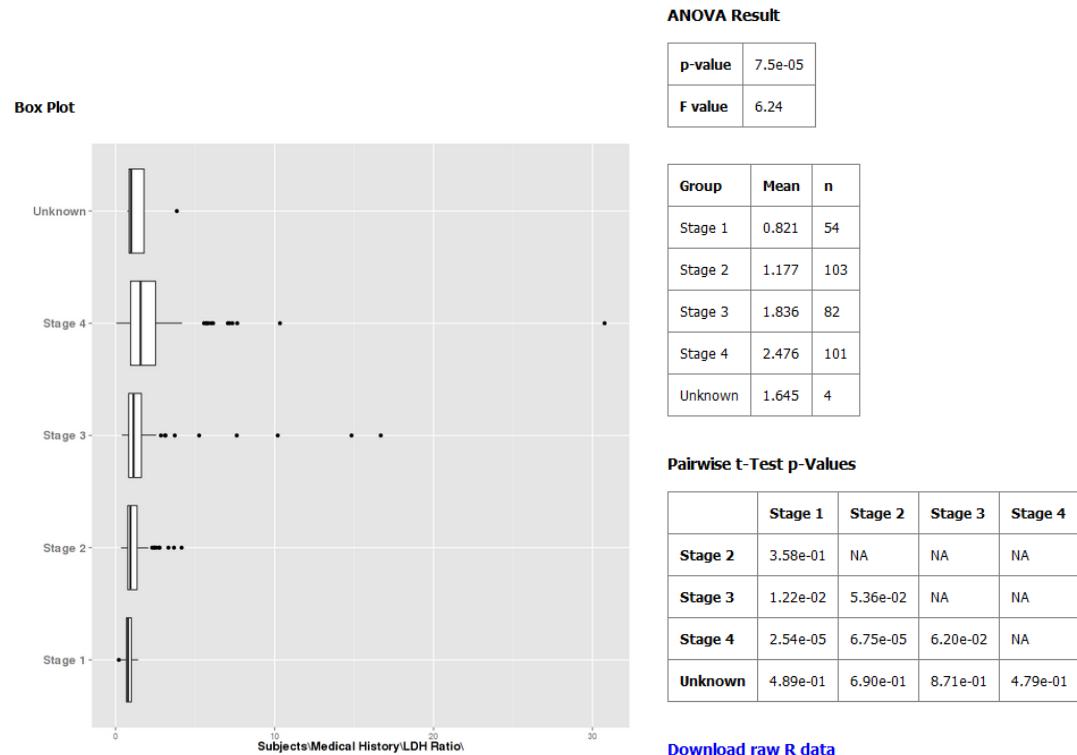
4. Click the **Advanced Workflow** tab, then select **Box Plot with ANOVA** from the Analysis dropdown menu.

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.

The screenshot shows the tranSMART software interface with the 'Analysis' tab selected. In the 'Variable Selection' section, 'LDH Ratio' is selected as the independent variable and 'Cancer Stage (420)' is selected as the dependent variable. A 'Run' button is visible at the bottom right.

7. Click the **Run** button.



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.

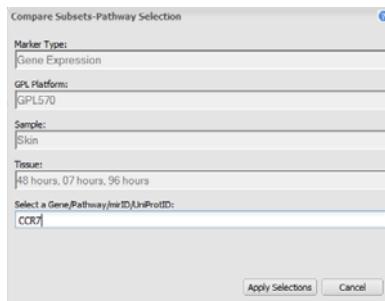
The screenshot shows the 'Comparison' tab in the tranSMART interface. The left side features a 'Navigate Terms' sidebar with a tree view of study terms. The 'Nickel Allergic Contact Dermatitis Pedersen GSE6281 (12)' node is expanded, showing its sub-categories: Biomarker data (12), Skin (12) with time points (00 hour, 07 hours, 48 hours, 96 hours), Subjects (12), and Group (12). The right side shows the 'Subset 1' configuration panel with the study name '...!Nickel Allergic Contact Dermatitis Pedersen GSE6281' and an 'Exclude' button.

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

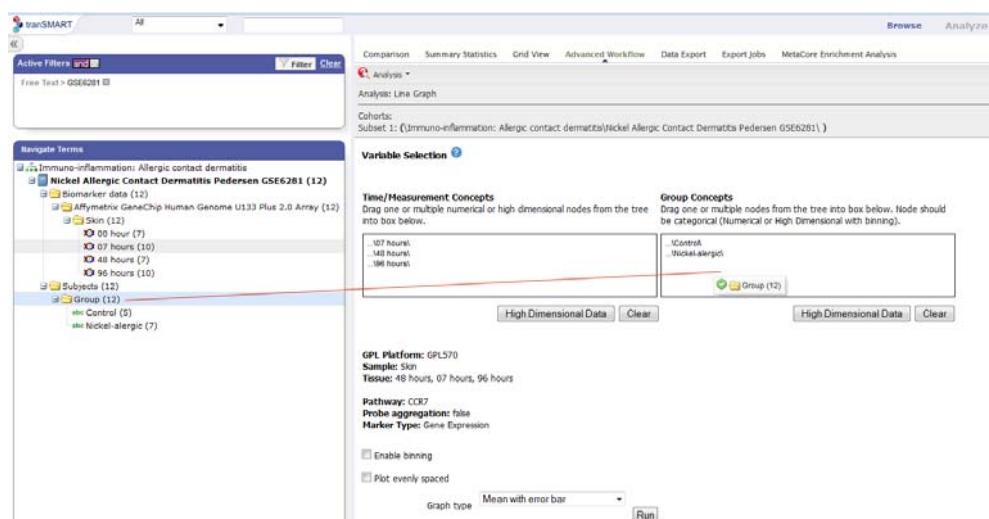
5. Drag and drop the nodes '\Immuno-inflammation: Allergic contact dermatitis\Nickel Allergic Contact Dermatitis Pedersen GSE6281\Biomarker data\Affymetrix GeneChip Human Genome U133 Plus 2.0 Array\Skin\07 hours', '\...\48 hours' and '\...\96 hours' under the Time/Measurement Concepts box.

The screenshot shows the 'Advanced Workflow' tab with 'Analysis: Line Graph'. The 'Variable Selection' section includes a 'Cohorts:' dropdown set to 'Subset 1: (\!Immuno-inflammation: Allergic contact dermatitis\Nickel Allergic Contact Dermatitis Pedersen GSE6281\)'. The 'Time/Measurement Concepts' section contains three boxes: 'Time/Measurement Concepts' (with nodes 07 hours, 48 hours, 96 hours), 'Group Concepts' (empty), and 'High Dimensional Data' (empty). A red arrow points from the '96 hours' node in the first box back to the 'Navigate Terms' sidebar. Below these are checkboxes for 'Enable binning' and 'Plot evenly spaced', and a 'Graph type' dropdown set to 'Mean with error bar' with a 'Run' button.

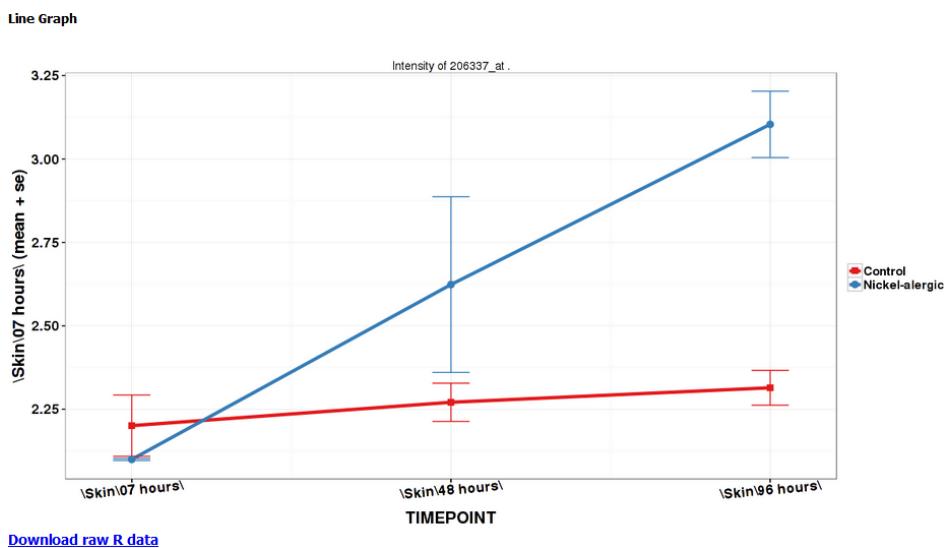
6. Click the **High Dimensional Data** button, add **CCR7** then **Apply Selections**.



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.

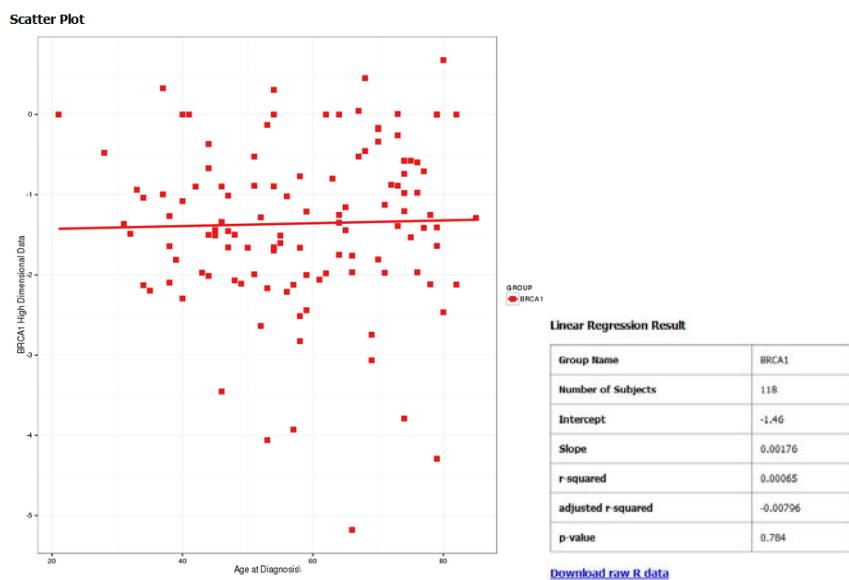
4. Click the **Advanced Workflow** tab, then select **Scatter Plot with Linear Regression** from the Analysis dropdown menu.
5. Drag and drop the node '\Breast Cancer Sorlie GSE4382\Sample Factors\Age at Diagnosis' under the Independent Variable box.
6. Drag and drop the '\Breast Cancer Sorlie GSE4382\Data\mRNA Profiling\Stanford cDNA microarrays\Breast' under the Dependent Variable box.

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

Marker Type: Gene Expression
GPL Platform: GSE4382PDM
Sample: Breast
Tissue: _____
Select a Gene/Pathway/mirID/UniProtID: BRCA1

GPL Platform: GSE4382PDM
Sample: Breast
Tissue: _____
Pathway: BRCA1
Probe aggregation: false
Marker Type: Gene Expression

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
χ^2	159
χ^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.

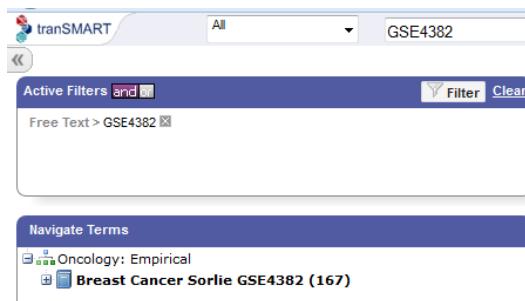
Lesson Goal: Become acquainted with the Correlation Analysis.

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.

The screenshot shows the 'Comparison' tab in the tranSMART interface. On the left, there is a 'Navigate Terms' sidebar with a tree view. A red arrow points from the 'Subset 1' section on the right towards the 'Breast Cancer Sorlie GSE4382 (167)' node in the tree. The 'Subset 1' section contains a box with the study name and an 'Exclude' button.

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.

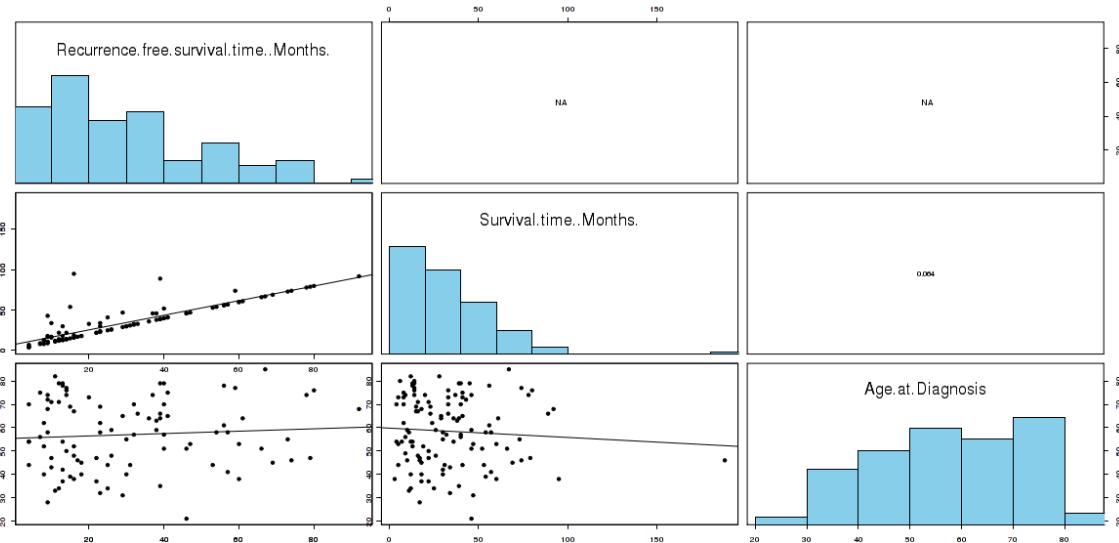
The screenshot shows the 'Advanced Workflow' tab selected. The 'Analysis' dropdown is set to 'Correlation Analysis'. The 'Variable Selection' section contains a box where three numerical concepts have been selected: 'Recurrence-free survival time (Months)', 'Survival time (Months)', and 'Age at Diagnosis'. Below this, there are dropdown menus for 'Run Correlation' (set to 'By variable') and 'Correlation Type' (set to 'Spearman'). At the bottom right is a 'Run' button.

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](#)

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



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

Variable Selection ?

Drag two or more **numerical** concepts from the tree into the box below that you wish to generate correlation statistics on.

...|Recurrence-free survival time (Months)|
...|Survival time (Months)|
...|Age at Diagnosis|

Run Correlation	By variable
Correlation Type	Spearman

Variable Selection ?

Drag two or more **numerical** concepts from the tree into the box below that you wish to generate correlation statistics on.

...|Recurrence-free survival time (Months)|
...|Survival time (Months)|

Run Correlation	By variable
Correlation Type	Spearman

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).

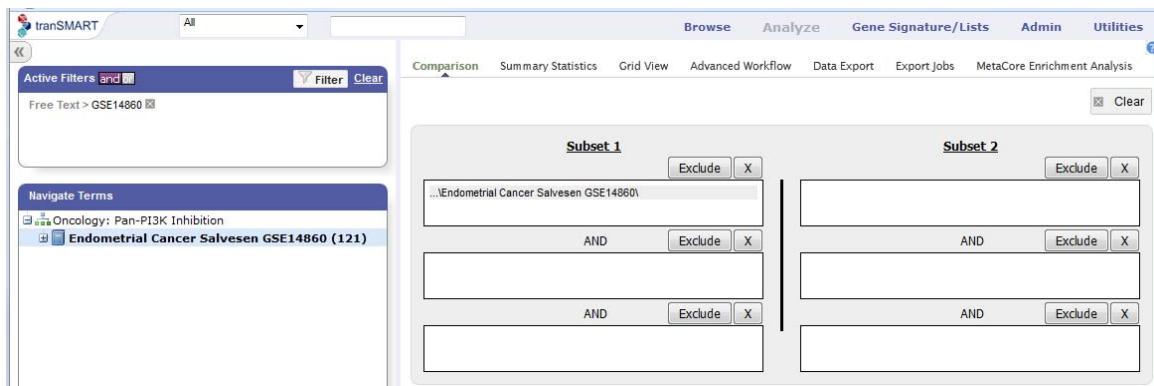
Lesson Goals: (1) Export data for a Study using the check boxes to indicate the data types and file formats that are desired (2) Filter data to export by dragging and dropping some node onto each data type row (3) Use cohorts to export a subset of data (4) Export of SNP data.

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.



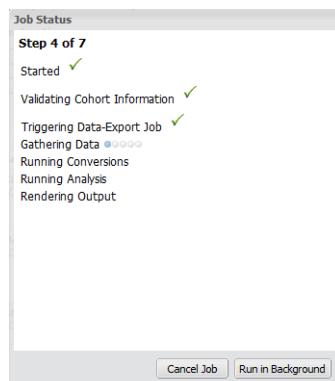
2. In the Comparison tab, drag and drop the Study ‘**Endometrial Cancer Salvesen GSE14860**’ as Subset 1 cohort.



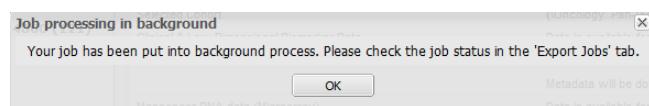
3. Go to the **Data Export** tab.
4. Activate the check boxes of 'Clinical and Low Dimensional Biomarker data' and 'Messenger RNA data (microarray)'.

The screenshot shows the tranSMART interface with the 'Data Export' tab active. On the left, there's a sidebar with 'Active Filters' and a 'Navigate Terms' section. The main area contains three data type sections: 'Selected Cohort', 'Messenger RNA data (Microarray)', and 'SNP data (Microarray)'. Each section provides details about the data source, patient count, and export options. At the bottom right is a large 'Export Data' button.

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.

Name	Query Summary	Status	Started On
-DataExport-13590		Gathering Data	
-DataExport-13589		Cancelled	
-DataExport-13585		Completed	
-DataExport-13584		Completed	
-DataExport-13583		Completed	
-DataExport-13582		Error	
-DataExport-13581		Completed	
-DataExport-13580		Completed	
-DataExport-13579		Completed	
-DataExport-13578		Completed	
-DataExport-13577		Completed	
-DataExport-13576		Completed	
-DataExport-13575		Completed	
-DataExport-13573		Completed	
-DataExport-13572		Completed	
-DataExport-13569		Completed	
-DataExport-13568		Completed	

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.

PATIENT ID	SAMPLE	AUGAL	SAMPLE CODE	TRIALNAME	VALUE	LOG2	ZSCORE	PROB_ID	GENE_ID	GENE_SYMBOL	
2 704	Endometrioid Tumor	GSE14860	3221	GSM177115	GSE14860	0.853	-0.2823522	-0.456498	VAPA	9218	VAPA
3 3005	Endometrioid Tumor	GSE14860	3222	GSM177105	GSE14860	0.792	-0.3746952	-0.20002	VAPA	9219	VAPA
4 4147	Endometrioid Tumor	GSE14860	3223	GSM177118	GSE14860	1.0871	0.148498	0.28511	VAPA	9218	VAPA
5 1417	Endometrioid Tumor	GSE14860	3224	GSM177104	GSE14860	0.0096	-0.3046621	-0.70915	VAPA	9218	VAPA
6 2257	Other	GSE14860	3225	GSM177103	GSE14860	1.2735	0.3476855	0.81626	VAPA	9218	VAPA
7 2797	Other	GSE14860	3226	GSM177102	GSE14860	1.0757	0.9074004	2.12525	VAPA	9210	VAPA
8 1047	Endometrioid Tumor	GSE14860	3227	GSM177087	GSE14860	0.7794	-0.3594158	-0.83718	VAPA	9218	VAPA
9 9567	Other	GSE14860	3228	GSM177139	GSE14860	0.8943	-0.1611534	-0.37358	VAPA	9218	VAPA

clinical_i2b2trans.txt

clinical_i2b2trans.txt: All clinical & LDD data exported.

PATIENT ID	SAMPLE CODES	\Sample Factors\Tumor Status	\Sample Factors\Histology\Histological Type	\Sample Factors\PI3K Regulation	\Sample Factors\
2 1	null NA	NA	NA	NA	?
3 10	null NA	NA	NA	NA	?
4 101N	null NA	NA	NA	NA	?
5 101T	null NA	NA	NA	NA	?
6 104T	GSM377087 FIGO I/II and No Recurrence	Endometrioid Tumor	Up	Female	

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.

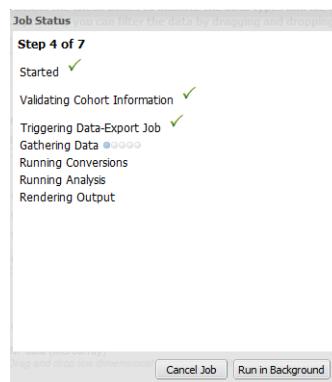
14. Drag and drop the node '\Endometrial Cancer Salvesen GSE14860\Data\mRNA profiling\Agilent Human 1A Microarray G4110\Endometrioid Tumor' in the 'Messenger RNA data (microarray) data' area.



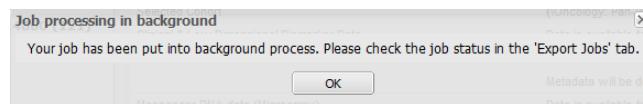
If you have added a node by mistake in the export filter area, right-click the node and remove it.

The screenshots show the 'Selected Cohort' section of the tranSMART interface. On the left, there is a 'Delete' button next to the term 'IPDK Regulation'. On the right, this button has been removed, indicating it has been deleted.

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



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



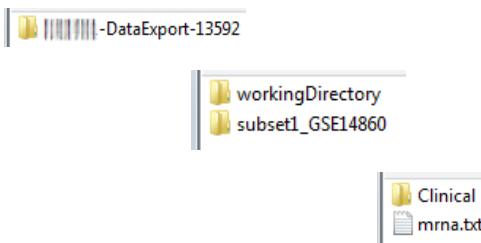
17. Click the **Export Jobs tab.**

The screenshot shows the tranSMART interface with the 'Analyze' tab selected. In the top navigation bar, there are tabs for Comparison, Summary Statistics, Grid View, Advanced Workflow, Data Export, Export Jobs, and MetaCore Enrichment Analysis. Below the tabs is a table titled 'Name' with columns for 'Query Summary', 'Status', and 'Started On'. The table lists numerous entries, all of which are 'Completed'. A 'Refresh' button is located at the bottom right of the table area.

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

19. Click the link when the status is completed.

The screenshot shows the tranSMART interface with the 'Analyze' tab selected. A modal dialog box is open, prompting the user to choose how to handle a file named '-DataExport-13592.zip'. The dialog includes options to 'Ouvrir avec' (Open with) or 'Enregistrer le fichier' (Save the file). There is also a checkbox for 'Toujours effectuer cette action pour ce type de fichier.' (Always perform this action for this type of file). The background table of completed jobs is visible.



Gene expression data exported only for Endometrioid Tumor.

The screenshot shows two tables side-by-side, both titled "clinical_i2b2trans.txt".

Top Table (Gene Expression Data):

PATIENT ID	SAMPLE	ASSAY ID	SAMPLE CODE	TRIALNAME	VALUE	LOG2E	ZSCORE	PROBE ID	GENE ID	GENE SYMBOL
1	78T	Endometrioid Tumor	GSE14860	3221	GSM377135	GSE14860	0.8222	-0.2823522	-0.65698	VAPA
2	306T	Endometrioid Tumor	GSE14860	3222	GSM377105	GSE14860	0.7927	-0.3349735	-0.78002	VAPA
3	41T	Endometrioid Tumor	GSE14860	3223	GSM377118	GSE14860	1.0871	0.1205349	0.28511	VAPA
4	141T	Endometrioid Tumor	GSE14860	3224	GSM377095	GSE14860	0.8096	-0.3046621	-0.70915	VAPA
5									9218	VAPA

Bottom Table (Clinical Factors):

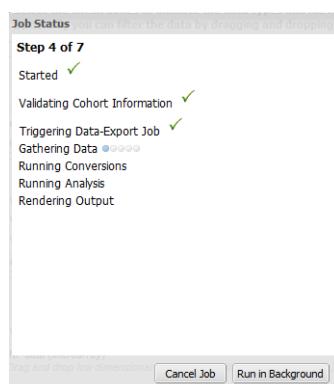
PATIENT ID	SAMPLE CODES	\Sample Factors\Tumor Status	\Sample Factors\Histology\Histological Type	\Sample Factors\PI3K Regulation
1	null	NA	NA	NA
2	1	NA	NA	NA
3	10	NA	NA	NA
4	101N	NA	NA	NA
5	101T	NA	NA	NA
6	104T	GSM377087 FIGO I/II and No Recurrence	Endometrioid Tumor	Up
7	108T	GSM377088 FIGO I/II and No Recurrence	Endometrioid Tumor	Up

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

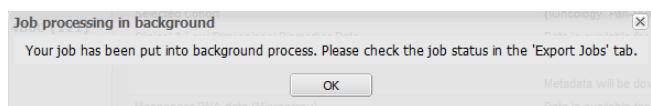
Scenario (3): Use cohorts to export a subset of data for the Study 'Endometrial Cancer Salvesen GSE14860'.

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.

Name	Query Summary	Status	Started On
-DataExport-13595		Gathering Data	
-DataExport-13594		Completed	
-DataExport-13592		Completed	
-DataExport-13590		Completed	
-DataExport-13589		Completed	
-DataExport-13585		Completed	
-DataExport-13584		Completed	
-DataExport-13583		Completed	
-DataExport-13581		Error	
-DataExport-13580		Completed	
-DataExport-13579		Completed	
-DataExport-13578		Completed	
-DataExport-13577		Completed	
-DataExport-13576		Completed	
-DataExport-13575		Completed	
-DataExport-13573		Completed	
-DataExport-13572		Completed	
-DataExport-13569		Completed	

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

30. Click the link when the status is completed.

-DataExport-13595

workingDirectory
subset2_GSE14860
subset1_GSE14860

Clinical
mrna.txt (subset 1)

```

1 PATIENT ID SAMPLE ASSAY ID SAMPLE CODE TRIALNAME VALUE LOG2E ZSCORE PROBE ID GENE ID GENE SYMBOL
2 306T Endometrioid Tumor_GSE14860 3222 GSM377105 GSE14860 0.7927 -0.3349735 -0.78002 VAPA 9218 VAPA
3 436T Endometrioid Tumor_GSE14860 3229 GSM377123 GSE14860 1.5345 0.6178222 1.44793 VAPA 9218 VAPA
4 172T Endometrioid Tumor_GSE14860 3231 GSM377098 GSE14860 0.8830 -0.179378 -0.41619 VAPA 9218 VAPA
5 445T Endometrioid Tumor_GSE14860 3233 GSM377127 GSE14860 1.3190 0.3995302 0.93749 VAPA 9218 VAPA

```

clinical_i2b2trans.txt

Clinical
mrna.txt (subset 2)

```

1 PATIENT ID SAMPLE CODES \Sample Factors\Tumor Status
2 110T GSM377090 FIGO I/II and No Recurrence
3 111T GSM377091 FIGO I/II and No Recurrence
4 113T GSM377092 FIGO I/II and No Recurrence
5 116T GSM377141 FIGO I/II and No Recurrence

```

clinical_i2b2trans.txt

```

1 PATIENT ID SAMPLE CODES \Sample Factors\Tumor Status
2 78T Endometrioid Tumor_GSE14860 3221 GSM377135 GSE14860 0.8222 -0.2823522 -0.65698 VAPA 9218 VAPA
3 41T Endometrioid Tumor_GSE14860 3223 GSM377118 GSE14860 1.0871 0.1205349 0.28511 VAPA 9218 VAPA
4 141T Endometrioid Tumor_GSE14860 3224 GSM377095 GSE14860 0.8096 -0.3046621 -0.70915 VAPA 9218 VAPA
5 104T Endometrioid Tumor_GSE14860 3227 GSM377087 GSE14860 0.7794 -0.3594158 -0.83718 VAPA 9218 VAPA

```

clinical_i2b2trans.txt

```

1 PATIENT ID SAMPLE CODES \Sample Factors\Tumor Status
2 104T GSM377087 FIGO I/II and No Recurrence
3 108T GSM377088 FIGO I/II and No Recurrence
4 10T2 GSM377089 FIGO I/II and No Recurrence
5 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.

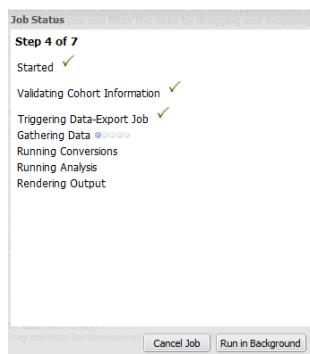
If each SNP node concerns a different cohort, you can use the cohort subset box to export only one SNP node.

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.

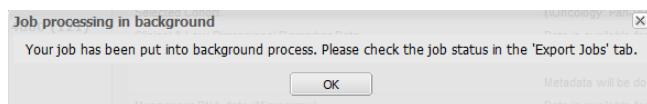
33. Go to the **Data Export** tab.
34. Activate the check boxes for Processed (.PED, .MAP & .CNV files) and Raw data (.CEL files).

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

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



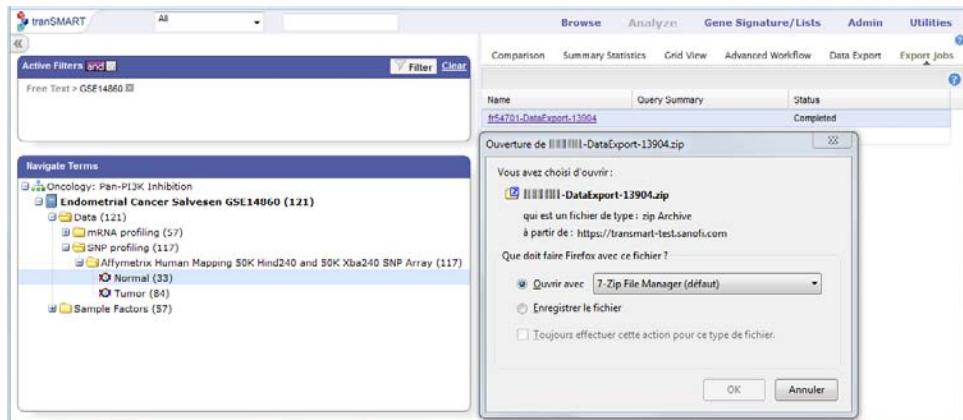
36. Click the button **Run in background** when the pop-up window **Job Status** appears. Then **OK**.



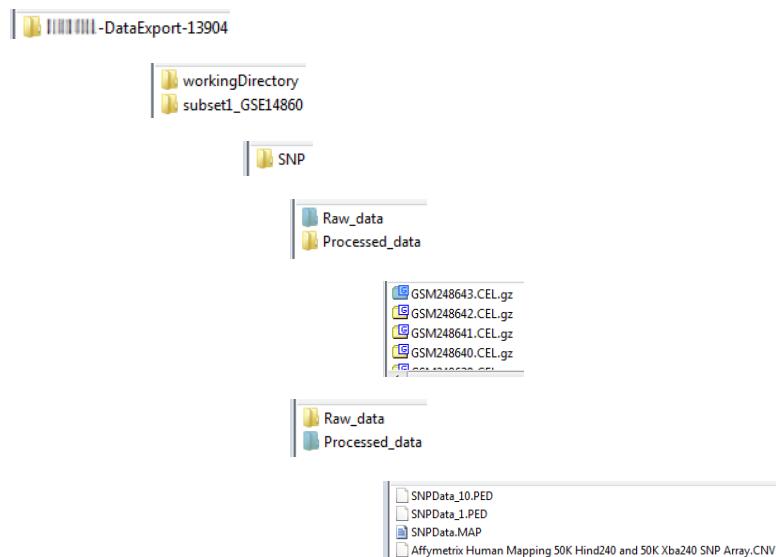
37. Click the **Export Jobs** tab.

Name	Query Summary	Status
DataExport-13904		Gathering Data
DataExport-13717		Completed

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



39. Click the link when the status is completed.



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

tranSMART All

Active Filters

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

Subset 1

...Endometrial Cancer Salvesen GSE14860

AND

Subset 2

AND

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								
					<table border="1"> <thead> <tr> <th>Name</th> <th>Query Summary</th> <th>Status</th> <th>Started On</th> </tr> </thead> <tbody> <tr> <td>-DataExport-13598</td> <td></td> <td>Cancelled</td> <td></td> </tr> </tbody> </table>	Name	Query Summary	Status	Started On	-DataExport-13598		Cancelled		
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.

The screenshot shows the transSMART interface with the 'Comparison' tab selected. On the left, there's a 'Navigate Terms' sidebar with a tree view. A red arrow points from the 'Subset 1' section on the right to the 'Type 2 Diabetes Taneera GSE38642 (63)' node in the tree. The 'Subset 1' section contains a search bar with the same study name and some AND operators.

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

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

The screenshot shows the transSMART interface with the 'MetaCore Enrichment Analysis' tab selected. The 'Variable Selection' section has a 'Data preparation' sub-section. A red arrow points from the 'Pancreatic islets' node in the 'Navigate Terms' tree to the 'High Dimensional Data' button in the 'Variable Selection' section. The 'High Dimensional Data' button is highlighted with a red border.

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

The screenshot shows the 'Compare Subsets-Pathway Selection' dialog box. It has fields for 'Marker Type' (set to 'Gene Expression'), 'GPL Platform' (set to 'GPL6244'), 'Sample' (set to 'Pancreatic islets'), and 'Tissue' (empty). The 'Select a Gene/Pathway/mirID/UniProtID:' field contains 'Metabolic pathways'. At the bottom are 'Apply Selections' and 'Cancel' buttons.

tranSMART Free Text GSE38642

Active Filters and Filter Clear

Free Text > GSE38642

Comparison Summary Statistics Grid View Advanced Workflow Data Export Export Jobs MetaCore Enrichment Analysis

Cohorts (Early Alpha version - only the first cohort will be used)
Subset 1: (Metabolism: Blood glucose regulation by insulin\Type 2 Diabetes Taneera GSE38642\)

Variable Selection

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

... \Pancreatic islets
High Dimensional Data

GPL Platform: GPL6244
Sample: Pancreatic islets
Tissue:

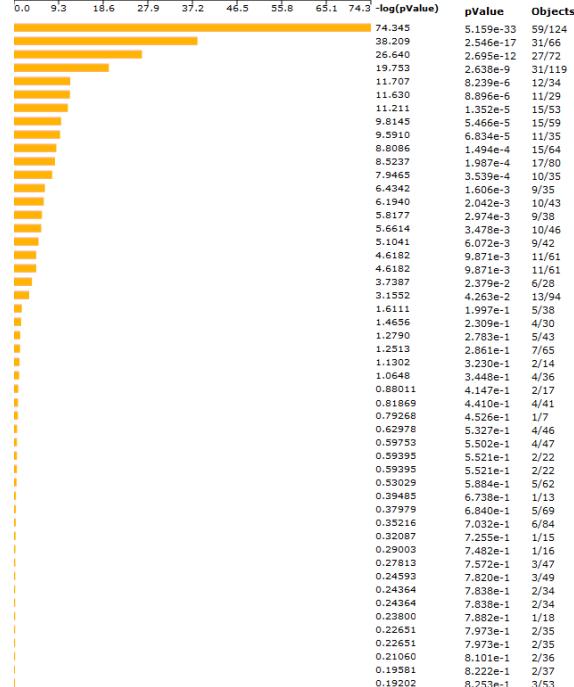
Pathway: Metabolic pathways
Probe aggregation: false
Marker Type: Gene Expression

Specify a Z-Score threshold (optional):
|zscore| >= 0

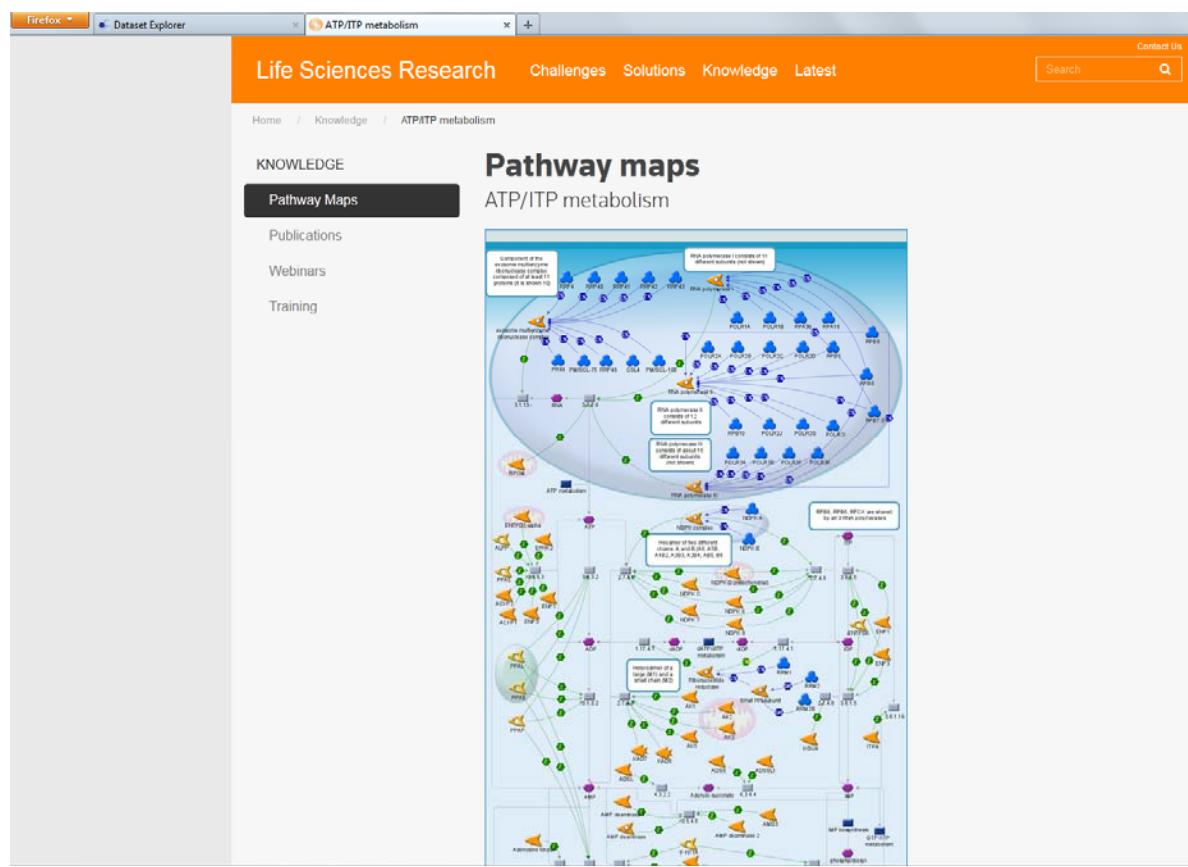
METACORE SETTINGS Run

7. Click the **Run** button.

#	Name
1	ATP/TP metabolism
2	Glycolysis and gluconeogenesis (short map)
3	Retinol metabolism
4	NAD metabolism
5	Methionine metabolism
6	Acetaminophen metabolism
7	Androstenedione and testosterone biosynthesis and metabolism p.i
8	Galactose metabolism
9	Estradiol metabolism
10	Tyrosine metabolism p.i (dopamine)
11	HETE and HPETE biosynthesis and metabolism
12	Estrone metabolism
13	Beta-alanine metabolism
14	Estrogen biosynthesis
15	Glycogen metabolism
16	Arachidonic acid metabolism and metabolism
17	Benzoflavonene metabolism
18	2-Naphthylamine and 2-Nitronaphthalene metabolism
19	Naphthalene metabolism
20	Cell cycle: Transition and termination of DNA replication
21	Prostaglandin 2 biosynthesis and metabolism
22	Vitamin D2 (ergocaliferol) metabolism
23	Transcription Ligand-dependent activation of the ESR1/SP pathway
24	Neurophysiological process: PGE2-induced pain processing
25	Glutathione metabolism
26	CFTTR folding and maturation (normal and CF)
27	G-protein signaling: G-Protein alpha-s signaling cascades
28	Transcription: Transcription factor Tubby signaling pathways
29	Cholesterol and Sphingolipids transport / Transport from Golgi and ER to the apical membrane (normal and CF)
30	G-protein signaling: Rap2B regulation pathway
31	Mechanisms of CFTTR activation by S-nitrosothiophene (normal and CF)
32	NO-dependent CFTTR activation (normal and CF)
33	Immune response: MIF-mediated glucocorticoid regulation
34	Transcription: CREM signaling in testes
35	Development: Alpha-2 adrenergic receptor activation of ERK
36	Cytoskeleton remodeling: Rab5 regulation pathway
37	Immune response: IL16 signaling in NK cells
38	Development: VEGFR signaling: VEGFR2 - generic cascades
39	WcCFTTR and deltaF508 traffic / Late endosome and Lysosome (norm and CF)
40	G-protein signaling: RhoB regulation pathway
41	Transcription: CREB pathway
42	Signal transduction: IP3 signaling
43	G-protein signaling: G-Protein alpha-q signaling cascades
44	G-protein signaling: G-Protein beta/gamma signaling cascades
45	Normal wCFTTR traffic / Sorting endosome formation
46	Inhibitory action of Lipoxin A4 on PDGF, EGF and LTD4 signaling
47	G-protein signaling: RAC1 in cellular process
48	G-protein signaling: Rac2 regulation pathway
49	G-protein signaling: G-Protein alpha-12 signaling pathway
50	Translation: Translation regulation by Alpha-1 adrenergic receptors



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



Lesson 22 – Gene Signature / Gene List

The transSMART 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 transSMART users or is reserved for your private use.

Once you create the gene signature or list, it can be used in transSMART 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.

- O. 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	<i>GeneSymbol</i>	TCN1 IL1RN KIAA1199 GOS2
Gene symbols, actual fold change	<i>GeneSymbol</i> <tab> <i>ActualFC</i>	CXCL5 -19.19385797 IL8RB -18.21493625 FPR1 -17.6056338 FCGR3A -15.69858713
Gene symbols, composite fold change	<i>GeneSymbol</i> <tab> <i>CompositeFC</i>	CXCL5 -1 IL8RB -1 MMP3 0 SOD2 1
Probe set IDs only	<i>ProbesetID</i>	224301_x_at 1398191_at Dr.2473.1.A1_at A_24_P93251
Probe set IDs, actual fold change	<i>ProbesetID</i> <tab> <i>ActualFC</i>	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	<i>ProbesetID</i> <tab> <i>CompositeFC</i>	224301_x_at -1 1398191_at 0 Dr.2473.1.A1_at 1 A_24_P93251 -1

Example

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

GENE_SIGNATURE.txt ...	
Fichier	Édition
Format	Affichage
GLRA1	-2.918
ABCC8	-2.186
CHL1	-2.155
RASGRP1	-1.991
PPP1R1A	-1.973
PFKFB2	-1.972
FAM105A	-1.918
ARG2	-1.879
ENTPD3	-1.836
GLP1R	-1.822



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 transSMART, click the **Gene Signature/Lists** tab.
2. Click the **New Signature** button.

The first page of the gene signature wizard appears:

The screenshot shows the 'Gene Signature Create' wizard on the first page. The top navigation bar includes 'Browse', 'Analyze', 'Gene Signature/Lists', 'Admin', and 'Utilities'. The main section is titled 'Gene Signature Create' with a 'Instructions ▾' button. Below it, 'Page 1: Definition:' is displayed. A red asterisk (*) is present next to the 'Signature/List Name' field, which is currently empty. There is also a 'Description' field with a large text area below it. A note at the bottom states: 'Note, the creator of this signature will be 'Rogerio Martins' at the current system time'. At the bottom of the page are buttons for 'Meta-Data' (highlighted in green) and 'Cancel'.



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:

The screenshot shows the 'Gene Signature Create' wizard on the second page, 'Meta-Data'. The top navigation bar is identical to the previous page. The main section is titled 'Page 2: Meta-Data:' with an 'Instructions ▾' button. It contains several input fields: 'Source of list' (dropdown), 'Owner of data' (dropdown), 'Stimulus' (text area with placeholder 'i.e. LPS, polyIC, etc.'), 'Dose, units, and time:' (text area), 'Treatment' (text area with placeholder 'Drug treatment used in assay:'), 'Dose, units, and time:' (text area), 'OR Enter:' (text area), 'Compound' (dropdown), 'Protocol Number' (text area), 'PMIDs (comma separated)' (text area), 'Species*' (dropdown), 'Technology Platform*' (dropdown), 'Tissue Type' (dropdown), and 'Experiment Type' (dropdown). At the bottom are buttons for 'Definition' (highlighted in green), 'Next', and 'Cancel'.

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:

Name	Author	Date Created	Species	Tech Platform	Tissue Type	Public List	# Genes	# Up-Regulated	# Down-Regulated	Action
Taneera GSE38642 Type 2 Diabetes vs Non Diabetic Regulated 140 genes.txt			Homo sapiens	GPL570	No	No	125	49	76	- Select Action -

Making a New Gene Signature Public

By default, a newly created gene signature is private.

To make a gene signature public:

1. In the **Gene Signature List**, click the **Select Action** dropdown to the right of the gene signature you just created.
2. Click **Make Public** in the dropdown list:

Name	Author	Date Created	Species	Tech Platform	Tissue Type	Public List	# Genes	# Up-Regulated	# Down-Regulated	Action
Taneera GSE38642 Type 2 Diabetes vs Non Diabetic Regulated 140 genes.txt			Homo sapiens	GPL570	No	No	125	49	76	- Select Action -

Other Signatures (14) ▾

- Select Action -
 - Select Action -
 Clone
 Delete
 Edit
 Edit Items
 Excel Download
 Download .GMT file
Make Public

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

Name	Author	Date Created	Species	Tech Platform	Tissue Type	Public List	# Genes	# Up-Regulated	# Down-Regulated	Action
Taneera GSE38642 Type 2 Diabetes vs Non Diabetic Regulated 140 genes.txt			Homo sapiens	GPL570	Yes	No	125	49	76	- Select Action -



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

Performing Actions on Your Gene Signatures

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

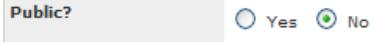
1. In tranSMART, click the **Gene Signature/Lists** tab.
The **Gene Signature List** appears, containing all the genes you have created:



The screenshot shows the 'Gene Signature List' page in tranSMART. At the top, there are tabs for 'Browse', 'Analyze', 'Gene Signature/Lists' (which is selected), 'Admin', and 'Utilities'. A 'New Signature' button is located in the top right. Below the tabs, a table lists 'My Signatures (1)'. The table columns are: Name, Author, Date Created, Species, Tech Platform, Tissue Type, Public List, # Genes, # Up-Regulated, and # Down-Regulated. The single entry is: 'Taneera GSE38642 Type 2 Diabetes vs Non Diabetic Regulated 140 genes.txt'. It was created by 'Homo sapiens' using 'GPL570' platform. The public list status is 'No', and it contains 125 genes, with 49 up-regulated and 76 down-regulated.

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 (<i>except</i> 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 transSMART users have created. The gene signatures you can access and the actions you can perform on them depend on the role assigned to your transSMART 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 transSMART, click the **Gene Signature/Lists** tab.
2. Click **Public Signatures** to open the list of public gene signatures.

The screenshot shows the 'Gene Signature List' interface. At the top, there's a header for 'My Signatures (1)'. Below it is a table with columns: Name, Author, Date Created, Species, Tech Platform, Tissue Type, Public List, Gene #, # Up-Regulated, and # Down-Regulated. A single row is shown: 'Trainee9 Training Account' (Date: 2009-08-08, Species: Human, Tech Platform: GPL8300, Tissue Type: Lung, Public List: No, Gene #: 18, # Up-Regulated: 7, # Down-Regulated: 11). To the right of the table is a dropdown labeled '-- Select Action --'. Below the table, a red oval highlights the 'Public Signatures (11)' link.

i transSMART 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:

Gene Signature List

Click to view the gene signature definition.

My Signatures (1) ▲											
Name	Author	Date Created	Species	Tech Platform	Tissue Type	Public	Gene List	# Genes	# Up-Regulated	# Down-Regulated	-- Select Action --
Trainee9 Gene Signature	Training Account	2009-08-08	Human	GPL8300	Lung	No	No	18	7	11	-- Select Action --

Public Signatures (11) ▲											
Name	Author	Date Created	Species	Tech Platform	Tissue Type	Public	Gene List	# Genes	# Up-Regulated	# Down-Regulated	-- Select Action --
Sys Admin Signature	Sys Admin	2009-08-11	Human	GPL570		Yes	No	113	43	70	-- Select Action --

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

Gene Signature Detail [Trainee9 Gene Signature]

General Infomation ▲

Name: Trainee9 Gene Signature [\[Excel\]](#)

Description: Genes from lung adenocarcinoma experiment with fold change value above absolute 10.

Public Status: Private

Author: Training Account

Create Date: 2009-08-08 10:21:35.924

Modified By: Training Account

Modified Date: 2009-09-08 11:36:19.581

Meta-Data ▼

Analysis Meta-Data ▼

Gene Signature Items ▼

Click to view additional details.

Appendix A – Additional Material

	PCA	Heatmap	Hierarchical Clustering	K-Means Clustering	Marker Selection	Scatter Plot with Linear Regression	Line Graph	Survival Analysis	Table with Fisher Test	Box Plot with ANOVA	Correlation Analysis
1 or 2 cohorts	x	x	x	x	x	x	x	x	x	x	x
2 cohorts											
	High dimensional data					high or low dim variables					
											low dim variables

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.	<ul style="list-style-type: none">■ 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:

Variable Selection ?

Time/Measurement Concepts
Drag one or multiple numerical or high dimensional nodes from the tree into box below.

...\\00 Weeks
...\\12 Weeks
...\\52 Weeks

52 Weeks (8)

Group Concepts
Drag one or multiple nodes from the tree into box below. Node should be categorical (Numerical or High Dimensional with binning).

...\\Asian
...\\Black
...\\Caucasian
...\\Hispanic
...\\Other (specify)

abc Other (specify) (2)

High Dimensional Data **Clear** **High Dimensional Data** **Clear**

GPL Platform: GPL570
Sample: Unknown
Tissue: 00 Weeks, 52 Weeks, 12 Weeks

Pathway: TP53
Probe aggregation: false
Marker Type: Gene Expression

Enable binning
 Plot evenly spaced

Graph type: Mean with error bar **Run**

With Manual Binning:

Variable Selection ?

Time/Measurement Concepts
Drag one or multiple numerical or high dimensional nodes from the tree into box below.

...\\00 Weeks
...\\12 Weeks
...\\52 Weeks

52 Weeks (8)

Group Concepts
Drag one or multiple nodes from the tree into box below. Node should be categorical (Numerical or High Dimensional with binning).

...\\Asian
...\\Black
...\\Caucasian
...\\Hispanic
...\\Other (specify)

abc Other (specify) (2)

High Dimensional Data **Clear** **High Dimensional Data** **Clear**

GPL Platform: GPL570
Sample: Unknown
Tissue: 00 Weeks, 52 Weeks, 12 Weeks

Pathway: TP53
Probe aggregation: false
Marker Type: Gene Expression

Enable binning
 Plot evenly spaced

Graph type: Mean with error bar **Run**

Variable Type: Categorical
Number of Bins: 3
Bin Assignments: Evenly Distribute Population
 Manual Binning

Categories

Drag To Bin

Bin 1
...\\Asian
...\\Black

Bin 2
...\\Caucasian
...\\Hispanic

Bin 3
...\\Other (specify)

Variable Type: Continuous

Without Manual Binning:

Variable Selection ?

Time/Measurement Concepts
Drag one or multiple numerical or high dimensional nodes from the tree into box below.

...100 Weeks
...112 Weeks
...152 Weeks
 52 Weeks (8)

Group Concepts
Drag one or multiple nodes from the tree into box below. Node should be categorical (Numerical or High Dimensional with binning).

...Age At First Infusion (Months)
 123 Age At First Infusion (Months) (18)

High Dimensional Data **Clear** **High Dimensional Data** **Clear**

GPL Platform: GPL570
Sample: Unknown
Tissue: 00 Weeks, 52 Weeks, 12 Weeks

Pathway: TP53
Probe aggregation: false
Marker Type: Gene Expression

Enable binning
 Plot evenly spaced

Graph type: Mean with error bar **Run**

With Manual Binning:

Variable Selection ?

Time/Measurement Concepts
Drag one or multiple numerical or high dimensional nodes from the tree into box below.

...100 Weeks
...112 Weeks
...152 Weeks
 52 Weeks (8)

Group Concepts
Drag one or multiple nodes from the tree into box below. Node should be categorical (Numerical or High Dimensional with binning).

...Age At First Symptoms (Months)
 123 Age At First Symptoms (Months) (18)

High Dimensional Data **Clear** **High Dimensional Data** **Clear**

GPL Platform: GPL570
Sample: Unknown
Tissue: 00 Weeks, 52 Weeks, 12 Weeks

Pathway: TP53
Probe aggregation: false
Marker Type: Gene Expression

Enable binning
 Plot evenly spaced

Graph type: Mean with error bar **Run**

Variable Type: Continuous
Number of Bins: 2
Bin Assignments: Evenly Distribute Population
 Manual Binning

Bin Name	Range
Bin 1	0 - 2
Bin 2	2 - 6

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).

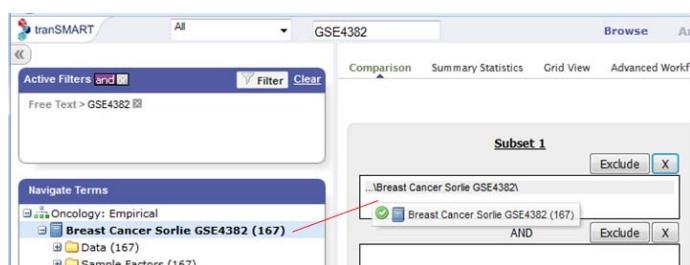
Note: WGCNA was developed by the Department of Human Genetics at UCLA. For more information, see <http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/>.

Example

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

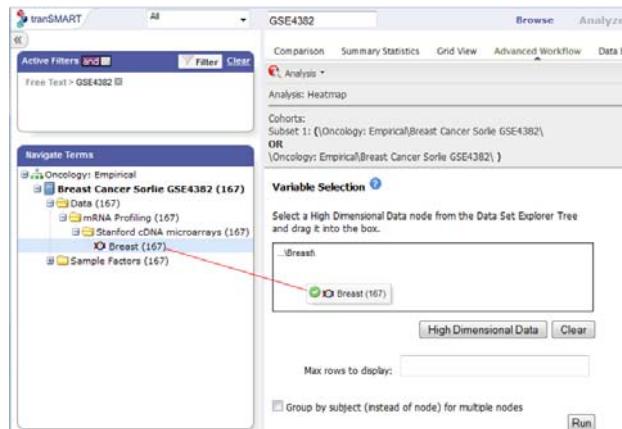


2. In the Navigate Terms area, under the Program, open the Study ‘**Breast Cancer Sorlie GSE4382**’.
3. In the Comparison tab, drag and drop the Study ‘**Breast Cancer Sorlie GSE4382**’ as Subset 1 cohort.

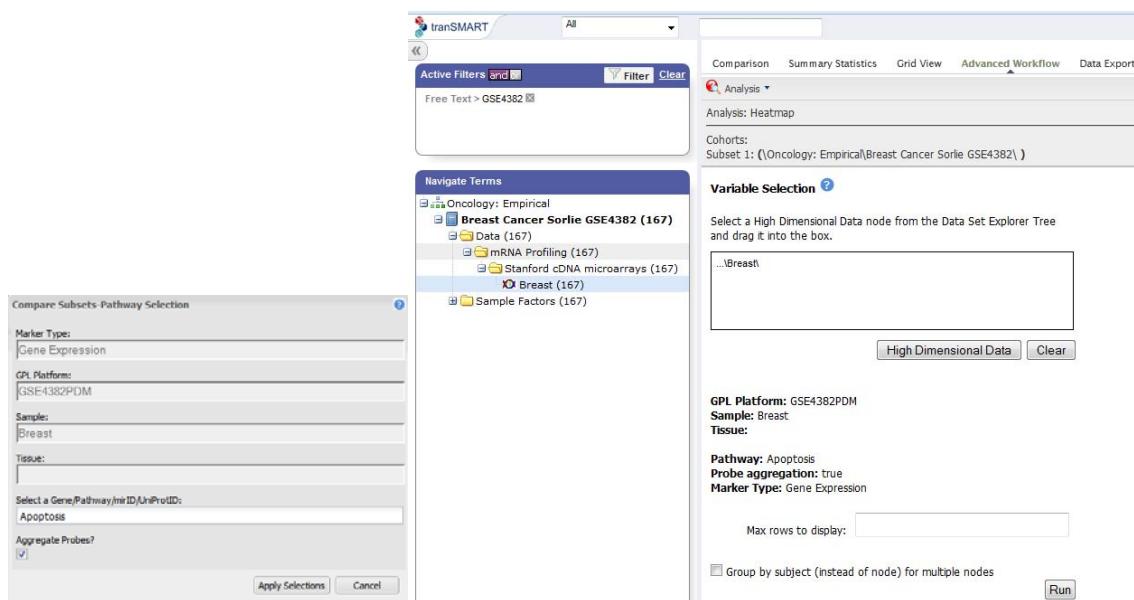


4. Go to the **Advanced Workflow**, then select **Heatmap** from the Analysis dropdown menu.

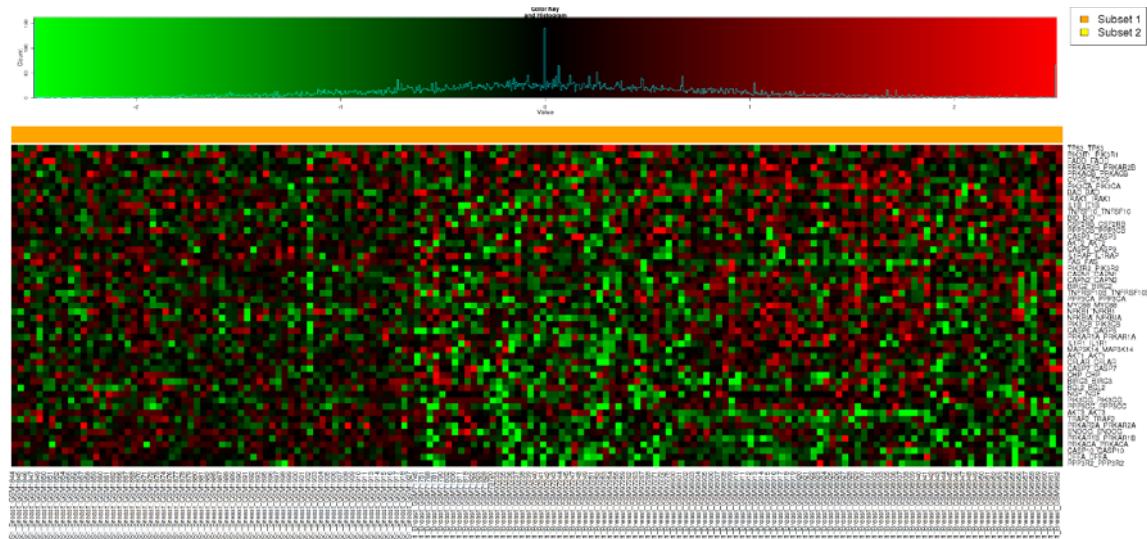
5. Drag and drop the node '\Breast Cancer Sorlie GSE4382\Data\...\Breast' under Variable Selection.



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

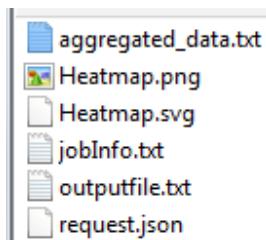


7. Click the **Run** button.



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, \dots, p_k , with observed frequencies m_1, m_2, \dots, m_k . The quantity

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

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

Refers 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 ($\Delta y/\Delta x$).

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.