

ArrayStar

Tutorial 1: Affymetrix Microarrays v1.0 - Public datasets

Jean-Yves Sgro

© 2014

Jean-Yves Sgro

Biochemistry Computational Research Facility

Note: ArrayStar is a registered trademark part of the DNASTAR Lasergene Suite

Table of Contents

Foreword	3
Microarrays	5
Wikipedia definition	5
ArrayStar installation	7
Before you begin	9
1. ArrayStar help	9
2. Video Tutorials.....	9
3. Affymetrix Arrays.....	9
10 steps for using ArrayStar	11
1. ArrayStar overview.....	12
Public Dataset	13
1. Gene Expression Omnibus Data.....	13
1.1. Dataset web page.....	13
1.2. Dataset files.....	14
1.3. Download the data.....	15
2. Rename files and/or attributes file.....	16
2.1. Preparation for Windows.....	16
2.2. Preparation on Unix Linux Mac.....	21
Getting started	25
1. Locating & launching ArrayStar	25
2. Creating a project	26
2.1. Import CEL files.....	27
2.2. Preprocessing	27
2.3. Attributes and replicates.....	28
2.4. Import Annotations	29
3. Organizing generic data files	32
4. Using Attributes files	34
Identifying genes of interest	37
1. Scatter plot.....	38
2. Gene Table	40
3. Filters.....	40
3.1. Down-regulated genes	40
4. Clustering	42
4.1. Hierarchical clustering.....	42
4.2. Line clustering	44



4.3. K-means.....	44
5. Gene sets – up-regulated genes.....	44
5.1. Filter up-regulated genes.....	45
5.2. Create a new gene set.....	46
5.3. Hierarchical clustering of up-regulated gene set.....	46
6. Gene Table: annotation & export.....	47
6.1. Add experiment values.....	48
6.2. Add fold-change columns.....	48
6.3. Add statistics columns.....	49
6.4. Add annotations.....	50
6.5. Export gene table to text file.....	51
GO: Gene Ontology.....	53
1. GO Enrichment Analysis.....	54
1.1. Retrieve the 138 gene set:.....	54
1.2. Display the GO results.....	54
1.3. ArrayStar Help.....	55
2. Biological process.....	56
Comparing three groups or more.....	59
1. Identifying genes of interest.....	59
2. Clustering:.....	60
2.1. Heat map.....	60
2.2. Line Graphs.....	61
2.3. K-Means.....	62
3. Export Gene List.....	62
Appendix.....	63
1. Online Tutorials.....	63

Foreword

This is a tutorial to start using ArrayStar, part of the Lasergene suite from DNASTAR, as licensed at the Biochemistry Department.

As a general guide, some marks are placed along most of the tutorials to indicate action to be taken by the reader: typically, **bold** commands are to be typed and often `typewritten styled text` illustrates a software output. Various check marks along the text are suggesting reading or taking action. For example, the following mark requests that the users take action and implement the specific commands or mouse clicks indicated in the text:

 **TASK**

Other markers suggest that the information should be read ( **READ**) to better understand the method(s) or command(s) used, or placed there as general information ( **INFO**) but not critical for the actual performance of the tutorial.

Microarrays

Wikipedia definition

A **DNA microarray** (also commonly known as **DNA chip** or **bio-chip**) is a collection of microscopic DNA spots attached to a solid surface.

Scientists use DNA microarrays to measure the **expression** levels of large numbers of genes simultaneously or to **genotype** multiple regions of a genome.

Each DNA spot contains **picomoles** (10^{-12} **moles**) of a specific DNA sequence, known as **probes** (or *reporters* or *oligos*).

These can be a short section of a **gene** or other DNA element that are used to **hybridize** a **cDNA** or cRNA (also called anti-sense RNA) sample (called *target*) under high-stringency conditions.

Probe-target hybridization is usually detected and quantified by detection of **fluorophore**-, silver-, or **chemiluminescence**-labeled targets to determine relative abundance of nucleic acid sequences in the target.

ArrayStar installation

ArrayStar is a Windows-only software, part of the Lasergene Suite from DNASTAR.

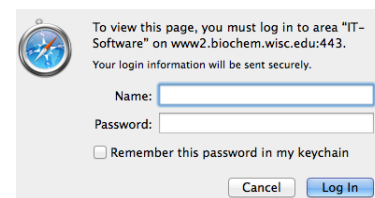
Before you start the tutorial you should install the Lasergene Suite onto a suitable computer, either running Windows 7 or up, or on a Macintosh computer with a Bootcamp partition or a virtual Windows computer under a virtualization program such as VMWare, Parallels or VirtualBox.

The installer for the Lasergene Suite can be found within the Biochemistry Intranet:

- 1) on www.biochem.wisc.edu click on “**Biochemistry Intranet**” or go directly to intranet.biochem.wisc.edu
- 2) Login with your Biochem ID
- 3) On the top bar click the Tab labelled “IT & Medialab”
- 4) On the left-hand side click on “**Information Technology**”
- 5) On the left-hand list click on “**Available Software**”

- 6) On the next page list locate and click on “**Lasergene (Biochem username/password required)**”

- 7) Enter your Biochem username and password



- 8) The next page has a barebone interface, choose the Windows installer, the only one that contains ArrayStar. The file is named:

[Win-DNASTARLasergene..>](#)

- 9) Download the installer file onto your computer and follow the specific installation procedures contained within the file “[Lasergene Installati..>](#)” also located on that same basic web page. **It is important to follow these instructions of the installation will fail.**

Here are the relevant portions of the installation instruction for PC from the installation instructions:

Lasergene Installation for PC

1. Download and save the Win-DNASTARLasergene[version].zip onto your Desktop or Downloads folder. Do not OPEN the file when downloading. Just save it onto your computer.
2. Right---click on the Win-DNASTARLasergene[version].zip and click on Extract All...
3. On the pop---up window, make sure “Show extracted files when complete” is checked and click Extract
4. Run the installer and select all defaults. You will be prompted a number of times to continue or run different parts of the install.

You should now have ArrayStar available. During the installation process you can choose to install the complete Lasergene suite or select only ArrayStar. If you already have Lasergene installed for the Macintosh you do not indeed need to duplicate it.

ArrayStar can now be launched from the “**Start**” button. If you are not familiar with the Windows operating system you should first learn some Windows basic before attempting to use ArrayStar.

VPN connection:

If connecting from outside the Biochemistry department it will be necessary to connect via a Virtual Private Network (VPN) to mimic local presence.

Please refer to the following resources to install and activate VPN:

General University description:

<https://www.doit.wisc.edu/network/vpn/>

Biochemistry Department Intranet instructions:

<https://intranet.biochem.wisc.edu/supportgroups/information-technology/remote-access>

Before you begin

1. ArrayStar help

ArrayStar has a help menu embedded within it. However, the most recent help can be found online with all the other Lasergene software under the “Training and Support” category:

- 1) Lasergene help: www.dnastar.com/t-support-help.aspx
- 2) ArrayStar help: www.dnastar.com/t-help-arraystar.aspx

Software support:

We're here to help! If you have any difficulties with or questions about this application, please contact a DNASTAR support representative:

- E-mail: support@dnastar.com
- Phone (Madison, WI, USA): 608-258-7420

2. Video Tutorials


For many steps short video tutorials are available on the DNASTAR web site. Each video is 1 to 4 min. long and shows what is possible to accomplish with the software. Videos are available on this link:

www.dnastar.com/t-support-videos.aspx

3. Affymetrix Arrays

ArrayStar can automatically read Affymetrix arrays but one or more files describing the array are necessary to the annotation of genes. ArrayStar can download the necessary files automatically but this will require a “login” to the Affymetrix web site.

Registration is free and your username and password can later be entered from an ArrayStar prompt.

 **TASK** If you intend to analyze Affymetrix data you should register on the Affymetrix web site:

- 1) go to www.affymetrix.com
- 2) At the top of the page click “Register”
- 3) Accept the online terms of use and click continue
- 4) Enter your information. Your ID will be your email

Once you are registered you can browse within the web site but also access the necessary files for any Affymetrix analysis from within ArrayStar.

Note: Currently “Exon Level” arrays are not supported by ArrayStar.

10 steps for using Array-Star

From the online training web page:

- 1) Begin a project by clicking one of the hyperlinks on the left of the [ArrayStar Welcome Screen](#).
- 2) Normalize your data (if necessary), import annotations and attributes, and organize your experiments into [replicate sets](#) using the [Project Setup Wizard](#).
- 3) View your imported data, and further organize it, if desired, in the [Experiment List](#) view.
- 4) Select **Graphs > Scatter Plot** to view a pairwise comparison in the [Scatter Plot](#) view. Select statistical tests from the Info Pane, and then click on the resulting links to select genes of interest.
- 5) Select **Clustering > Hierarchical** to view a [Hierarchical clustering](#) of your data set in the [Heat Map](#) view. Select genes of interest by clicking on a node in the [Gene Tree](#), shown on the vertical axis of the Heat Map.
- 6) Select **Data > Show Gene Ontology** to view significant ontology/classification terms for the selected genes in the [Gene Ontology](#) view.
- 7) Select **Clustering > k-Means** to cluster the selected genes with the [k-Means](#) method, and display your results

in the [Line Graph Thumbnails](#) view. Double-click on a graph in the Line Graph Thumbnails view to launch the [Line Graph](#) view for the cluster you selected.

- 8) Select **Edit > Copy Image of Graph** to copy the active graphical view to your clipboard so that you can paste it into another application such as Microsoft PowerPoint or Adobe Illustrator. (When pasting, use **Edit > Paste**, or **Edit > Paste Special** and then select “Enhanced Metafile”).
- 9) Select **Data > Show [n] Table** to explore a [variety of tables](#) showing details for the genes, isoforms, exons, SNPs, peaks or fragments in your project. Add columns to tables by using the associated toolbar tools. Click on hot-links for Gene Ontology terms and other annotations to view more information online.
- 10) Select **File > Save Project** to save your work as an ArrayStar project (*.astar).

1. ArrayStar overview / Help

The Help online features a complete list of ArrayStar capabilities.

Since the URL is long and complex, a short version for the HELP entry is:
bit.ly/1klgtvP

Public Dataset

At this point you should have the following necessary steps accomplished:

- 1) ArrayStar installed and functional.
- 2) Created an Affymetrix username

For this tutorial we will use data available on the public repository Gene Expression Omnibus (GEO).

The detailed explanation below can be skipped if you already have a dataset ready or if the files are provided to you in a class setting.

1. Gene Expression Omnibus Data

This web site is a repository of datasets of microarray and sequence data made public with or without an accompanying published paper.


Note that there exists a European version with cross-referencing to the GEO datasets called ArrayExpress¹.

1.1. Dataset web page

The dataset uses the Affymetrix Human Genome U133 Plus 2.0 microarray to study *Human airway epithelial responses to rhinovirus infection and cigarette smoke extract alone and in combination*.

Each dataset has a unique accession code, for this study it is GSE27973 and the direct link to the data and relevant information is coded as a single line.

¹ www.ebi.ac.uk/arrayexpress/


 **TASK** Go to the GEO web page

www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE27973

or as a shortened URL version: 1.usa.gov/WiUkmp

The top of the page contains the following information about the dataset:

Status	Public on Mar 15, 2011
Title	Human airway epithelial responses to rhinovirus infection and cigarette smoke extract alone and in combination
Organism	Homo sapiens
Experiment type	Expression profiling by array
Summary	This study was performed to test the hypothesis that cigarette smoke extract would alter the responses of primary cultures of human bronchial epithelial cells to infection with purified human rhinovirus 16. The data show marked alterations in rhinovirus-induced expression profiles of a number of genes in the presence of cigarette smoke extract (CSE).
Overall design	Cultured epithelial cells from each of 4 donors were exposed to medium alone, rhinovirus 16 (RV16) alone, CSE alone, or RV16 in the presence of CSE. After a 24 h incubation gene expression was assessed.
Contributor(s)	Proud D , Leigh R
Citation(s)	Proud D, Hudy MH, Wiehler S, Zaheer RS <i>et al.</i> Cigarette smoke modulates expression of human rhinovirus-induced airway epithelial host defense genes. <i>PLoS One</i> 2012;7(7):e40762. PMID: 22808255

 **TASK** Please take a moment to read the “Summary” and the “Overall design” entries on this page (here or on the web) before continuing.

1.2. Dataset files

Further down the page is where we can find and download the data files that we can analyse.

The “Platform” entry is simply the name and type of microarray that was used.

The samples are listed in table form on the web page. Click “More...” below



the entry “Samples (16) to see the complete list:

GSM692115	Donor 1 - medium
GSM692116	Donor 1 - RV16
GSM692117	Donor 1 - CSE
GSM692118	Donor 1 - RV16+CSE
GSM692119	Donor 2 - medium
GSM692120	Donor 2 - RV16
GSM692121	Donor 2 - CSE
GSM692122	Donor 2 - RV16+CSE
GSM692123	Donor 3 - medium
GSM692124	Donor 3 - RV16
GSM692125	Donor 3 - CSE
GSM692126	Donor 3 - RV16+CSE
GSM692127	Donor 4 - medium
GSM692128	Donor 4 - RV16
GSM692129	Donor 4 - CSE
GSM692130	Donor 4 - RV16+CSE

Note that the sample names are GSM-and-a-number and that is not very useful later in the software to organize the data by group. We will have to rely on this list to know “who is who” or “what is what” in the sample list.

For each sample the standard CEL and CHP files created by the Affymetrix software at the source will be available. We will use the CEL files since they are the closest to the “raw” intensity data.

1.3. Download the data

Locate the “Supplementary file” entry at the bottom of the page.



TASK click on the “(http)” link to download all the data to your local hard drive

Supplementary file	Size	Download	File type/resource
GSE27973_RAW.tar	76.5 Mb	(http)(custom)	TAR (of CEL, CHP)

This will download a file called **GSE27973_RAW.tar** on your drive. Its location will depend on the web browser and operating system you use. Typically it will be within a folder called “Downloads” and some browsers may ask you where you want to save the file.

The file **GSE27973_RAW.tar** is in the Unix “*Tape ARchive*” format and if you are downloading the file on a Linux or a Macintosh it may either un-archived automatically or it is easy to do so (e.g. double-click on the Mac.)

However, on a Windows system you will need an external software to un-archive this non-Windows format.

One free option is the software called **7zip**: www.7-zip.org which offers a 32 and a 64 bit version.

Another free option is **IZarc** from www.izarc.org

Both also support .gz (gzip) file format that is also needed later.

Note: The files might be provided in class if you are using this document within at directed tutorial.

If you are using this document on your own, complete the un-archiving of all the CEL files that first appear as *.CEL.gz when first extracted from the .tar file.

Note: If you are on a Mac you can either double-click on each file (tedious!) or you can open a Terminal (within Applications/Utilities) and issue the following command after moving within the directory containing the files:

```
gunzip *.gz
```

2. Rename files and/or attributes file

The files are named GSM692115 through GSM692130 and this notation is not at all useful to group samples by category of treatment within the software (ArrayStar or any other) and therefore we'll make some efforts to either change the name of each file to contain treatment attributes (*e.g.* treated with rhinovirus) and/or create an attributes file list.

This is easier on a Unix/Linux/Mac environment, but since ArrayStar runs only under Windows we'll cover all options. There are only 16 samples in this set, but the methods presented below would be even more useful for larger datasets with dozens or hundreds of files.

2.1. Preparation for Windows

This is only one method and other methods could be devised.

Goal: convert each file into a better name, for example rename based on the info provided on the GEO web page:

GSM692115.CEL into **Medium_Donor-1.CEL** and continue with
GSM692116.CEL into **RV16_Donor-1.CEL** etc.



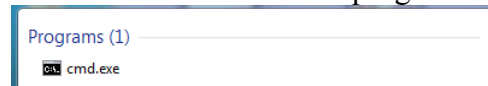
The DOS command to rename a file is simply `RENAME`. We will therefore create a command file to rename all the files.

The first step is to obtain a simple plain text file containing the list of all CEL files, one per line.

This is how it can be easily done in Windows using the DOS command line (as far as this author knows, there is no GUI way to accomplish this task!)

✓ TASK

- 1) Locate the DOS command line interface: in the START menu type “**cmd**” and the “`cmd.exe`” program will be offered. Click on it.



- 2) When “`cmd.exe`” is started it is looking within the default directory for the current user, *e.g.* `C:\Users\jsgro` in my case. On your computer it will reflect your current user name.
- 3) Navigate to the area where the CEL files are located and let’s call it `DATA` for example. If it is difficult to find it, you can move the directory with the mouse *e.g.* to the Desktop to more easily locate it. Assuming that is the case type the following commands:

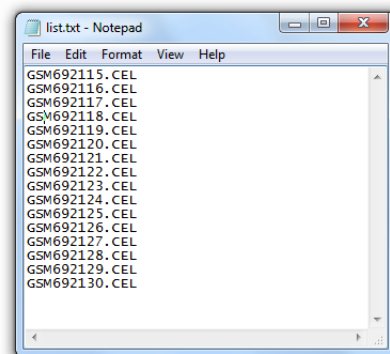
```
cd Desktop
```

```
cd DATA
```

```
DIR /b *.CEL > list.txt
```

- 4) This will create a file called “`list.txt`” containing the name of all CEL files.

If you double click it, the file will open under the default software to open plain text files, most likely Notepad.



This will be the start of a procedure to rename the files.

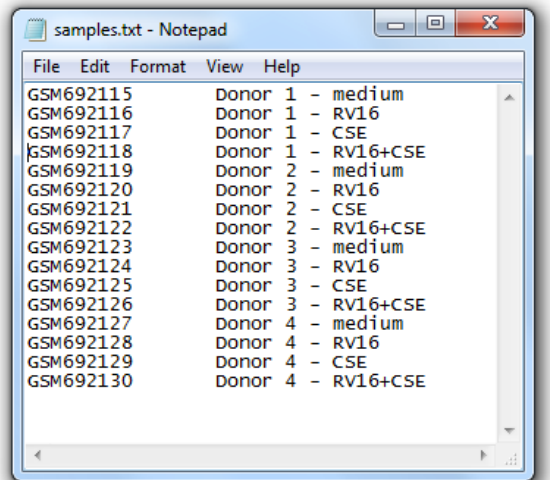
Note: The same file could also be used to create a tabulated entry of attributes as is required by other software (*e.g.* R/Bioconductor.) However, we

will need DOS program later in any case.

We now need to find a way to add the attributes listed within the Samples list on the GEO page. The easiest way is to copy/paste this data from the web page itself into a new plain text document.

✓ **TASK** Copy/paste the sample information within a new file, let's call it samples.txt:

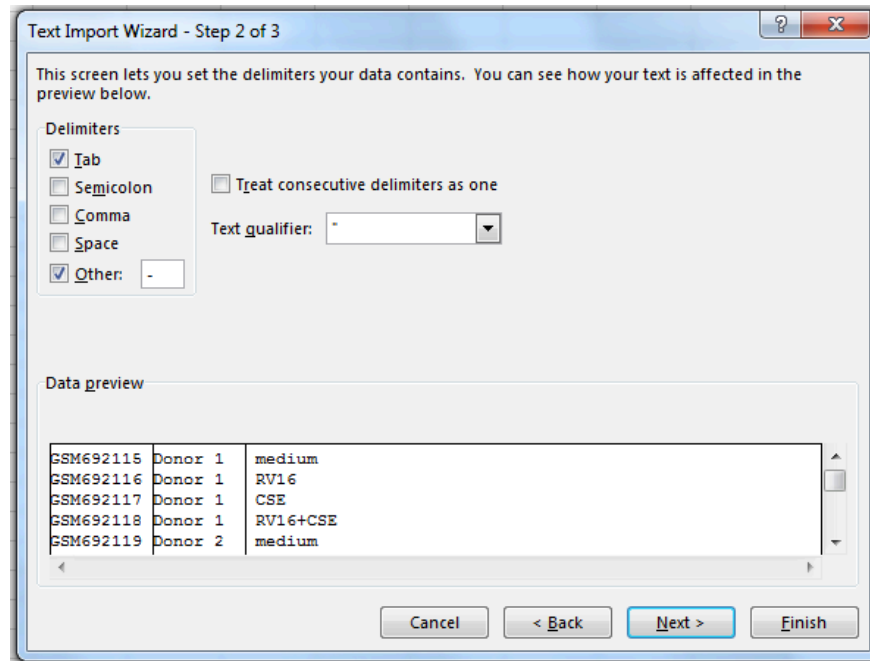
- 1) Open Notepad or use **File> new** if it is already opened
- 2) Copy the Samples(16) data from the GEO page
- 3) Paste the text within a blank file and save it as samples.txt



```
File Edit Format View Help
GSM692115 Donor 1 - medium
GSM692116 Donor 1 - RV16
GSM692117 Donor 1 - CSE
GSM692118 Donor 1 - RV16+CSE
GSM692119 Donor 2 - medium
GSM692120 Donor 2 - RV16
GSM692121 Donor 2 - CSE
GSM692122 Donor 2 - RV16+CSE
GSM692123 Donor 3 - medium
GSM692124 Donor 3 - RV16
GSM692125 Donor 3 - CSE
GSM692126 Donor 3 - RV16+CSE
GSM692127 Donor 4 - medium
GSM692128 Donor 4 - RV16
GSM692129 Donor 4 - CSE
GSM692130 Donor 4 - RV16+CSE
```

✓ **TASK** Next step: open both files in with Excel

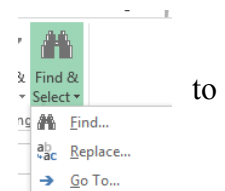
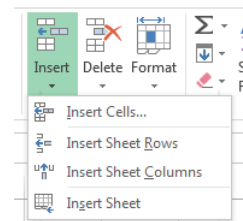
Important: when opening sample.txt add the dash “-“ as one of the delimiters within the Wizard:



Note: the file list.txt could be made optional with the extra effort to “glue” .CEL after each GSM entry within the samples.txt file.

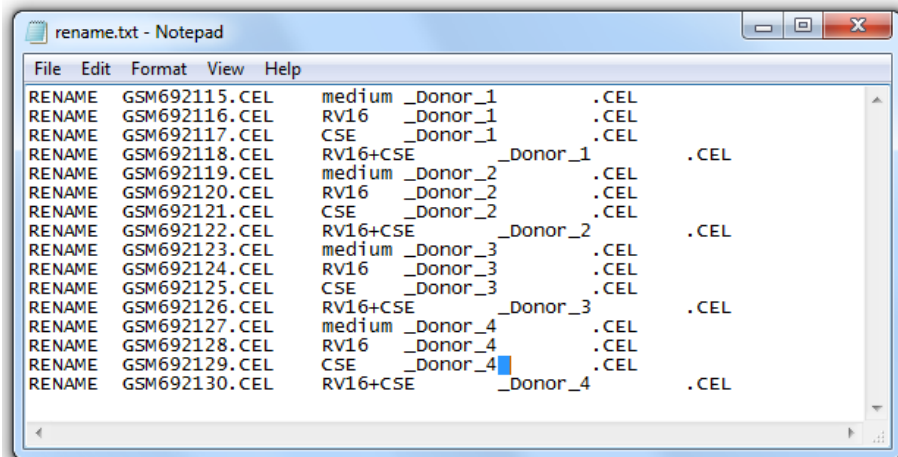
Now copy/paste the following columns between files:

- 1) Copy column C of samples.txt into column A of list.txt
- 2) Copy column B of samples.txt into column C of list.txt
- 3) add .CEL to each cell within column D
- 4) Add a new column before the GSM “A” column, and all columns will be shifted to the right by one letter.
- 5) Fill this new column “A” with the word “RENAME”
- 6) Optional: substitute **RENAME** with **COPY** if you prefer duplicate the data rather than rename it.
- 7) Within now column D replace “Donor ” (with a blank space) with “_Donor_” with an underscore before and after the word.
- 8) At this point your Excel sheet should look like this:



	A	B	C	D	E
1	RENAME	GSM692115.CEL	medium	_Donor_1	.CEL
2	RENAME	GSM692116.CEL	RV16	_Donor_1	.CEL
3	RENAME	GSM692117.CEL	CSE	_Donor_1	.CEL
4	RENAME	GSM692118.CEL	RV16+CSE	_Donor_1	.CEL
5	RENAME	GSM692119.CEL	medium	_Donor_2	.CEL
6	RENAME	GSM692120.CEL	RV16	_Donor_2	.CEL
7	RENAME	GSM692121.CEL	CSE	_Donor_2	.CEL
8	RENAME	GSM692122.CEL	RV16+CSE	_Donor_2	.CEL
9	RENAME	GSM692123.CEL	medium	_Donor_3	.CEL
10	RENAME	GSM692124.CEL	RV16	_Donor_3	.CEL
11	RENAME	GSM692125.CEL	CSE	_Donor_3	.CEL
12	RENAME	GSM692126.CEL	RV16+CSE	_Donor_3	.CEL
13	RENAME	GSM692127.CEL	medium	_Donor_4	.CEL
14	RENAME	GSM692128.CEL	RV16	_Donor_4	.CEL
15	RENAME	GSM692129.CEL	CSE	_Donor_4	.CEL
16	RENAME	GSM692130.CEL	RV16+CSE	_Donor_4	.CEL

- 9) Now save the file as a plain text file using the Save As... menu. Name the file **rename.txt** and dismiss the warning that format will be lost (we don't want any format at this point.)
- 10) Now double-click on the file just saved and it will open within Notepad.



```

rename.txt - Notepad
File Edit Format View Help
RENAME GSM692115.CEL medium _Donor_1 .CEL
RENAME GSM692116.CEL RV16 _Donor_1 .CEL
RENAME GSM692117.CEL CSE _Donor_1 .CEL
RENAME GSM692118.CEL RV16+CSE _Donor_1 .CEL
RENAME GSM692119.CEL medium _Donor_2 .CEL
RENAME GSM692120.CEL RV16 _Donor_2 .CEL
RENAME GSM692121.CEL CSE _Donor_2 .CEL
RENAME GSM692122.CEL RV16+CSE _Donor_2 .CEL
RENAME GSM692123.CEL medium _Donor_3 .CEL
RENAME GSM692124.CEL RV16 _Donor_3 .CEL
RENAME GSM692125.CEL CSE _Donor_3 .CEL
RENAME GSM692126.CEL RV16+CSE _Donor_3 .CEL
RENAME GSM692127.CEL medium _Donor_4 .CEL
RENAME GSM692128.CEL RV16 _Donor_4 .CEL
RENAME GSM692129.CEL CSE _Donor_4 .CEL
RENAME GSM692130.CEL RV16+CSE _Donor_4 .CEL


```

- 11) The remaining spaces and Tabs have to be changed to create the final commands. The necessary steps are to Find/Replace the following:
- Remove the TAB before _Donor
 - Remove Space + TAB before .CEL
 - Replace TAB with a blank space
- 12) The final file should look like this:



```
rename.txt - Notepad
File Edit Format View Help
RENAME GSM692115.CEL medium_Donor_1.CEL
RENAME GSM692116.CEL RV16_Donor_1.CEL
RENAME GSM692117.CEL CSE_Donor_1.CEL
RENAME GSM692118.CEL RV16+CSE_Donor_1.CEL
RENAME GSM692119.CEL medium_Donor_2.CEL
RENAME GSM692120.CEL RV16_Donor_2.CEL
RENAME GSM692121.CEL CSE_Donor_2.CEL
RENAME GSM692122.CEL RV16+CSE_Donor_2.CEL
RENAME GSM692123.CEL medium_Donor_3.CEL
RENAME GSM692124.CEL RV16_Donor_3.CEL
RENAME GSM692125.CEL CSE_Donor_3.CEL
RENAME GSM692126.CEL RV16+CSE_Donor_3.CEL
RENAME GSM692127.CEL medium_Donor_4.CEL
RENAME GSM692128.CEL RV16_Donor_4.CEL
RENAME GSM692129.CEL CSE_Donor_4.CEL
RENAME GSM692130.CEL RV16+CSE_Donor_4.CEL
```

Note: you could use `COPY` instead of `RENAME` if you wanted to keep the originals (but these could be downloaded again if need be.)

 **TASK** Use the created commands to rename (or copy) new file names: one easy method is to simply copy/paste the commands within the DOS cmd.exe window. However, doing so pastes “rename.txt” and therefore this simple copy/paste method may not work.

The easiest work around is simple: change the filename extension and double click it as explained below:

1) within the cmd.exe DOS window type the following command:

```
rename rename.txt rename.cmd
```

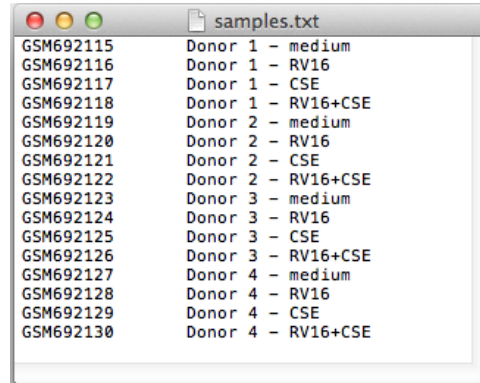
- 2) Look into the DATA directory and simply double-click rename.cmd (the newly renamed file:) the renaming of all GSM--- files will occur!
- 3) You can now delete the `rename.cmd` file or rename it back to `rename.txt` for safekeeping.

Note: DO NOT rename the `rename.txt` file with the mouse as the GUI is “lying” to you as to what the real file name is! You might end-up with a file named `rename.cmd.txt` even though it may *appear* as `rename.cmd` in your folder. That is how modern computers try to “make life easier” but mess things up!

2.2. Preparation on Unix|Linux|Mac

The first step will be the same: copy the information from the web page, but then we’ll use clever command-line text processing to reorganize the data.

✓ **TASK** Go to the GEO web page link (see above on page 13) and Copy/Paste the sample table information into a simple word processor (vi, nano or TextEdit on Mac) and save the file as `samples.txt` somewhere on e.g. the Desktop. (If using TextEdit make sure you save as plain text – see **Format** menu.)



GSM ID	Donor	Sample Type
GSM692115	Donor 1	medium
GSM692116	Donor 1	RV16
GSM692117	Donor 1	CSE
GSM692118	Donor 1	RV16+CSE
GSM692119	Donor 2	medium
GSM692120	Donor 2	RV16
GSM692121	Donor 2	CSE
GSM692122	Donor 2	RV16+CSE
GSM692123	Donor 3	medium
GSM692124	Donor 3	RV16
GSM692125	Donor 3	CSE
GSM692126	Donor 3	RV16+CSE
GSM692127	Donor 4	medium
GSM692128	Donor 4	RV16
GSM692129	Donor 4	CSE
GSM692130	Donor 4	RV16+CSE

What we need to do now is simple word processing:

- 1) Add `.CEL` to the GSM names to mimic the actual file name.
- 2) invert the second and last columns so that the Donor will be last
- 3) “glue” the donor number (third column) to the Donor word with an underscore
- 4) add a rename command to change the name of the file as we did in the Windows version.

To accomplish these tasks there are multiple Unix style solutions, here is one using the line editors `sed` and `awk` (on some system it is called `nawk`.) The commands are “piped” together with the symbol `|` so that results of one program are passed to the next.

The first 2 columns are separated by a TAB character, and there are differences between systems as to deal with that (`\t` on a Unix system and `Ctrl-V+Tab` on a Mac) Therefore we will use a method that does not require using the TAB. The command is written with the “line-continuation” (`\`) to facilitate reading but could be typed as a single command.

(The `$` is the command-line prompt and is not typed:)

```
$ cat samples.txt \
| sed -e 's/Donor/.CEL Donor/g' -e 's/--//g' \
| awk '{print "mv "$1$2,$5"_"$3"_"$4".CEL"}' \
> rename.txt
```

The above command creates “mv” (move or rename) commands that look like this:

```
mv GSM692115.CEL medium_Donor_1.CEL
mv GSM692116.CEL RV16_Donor_1.CEL
```




```
mv GSM692117.CEL CSE_Donor_1.CEL
mv GSM692118.CEL RV16+CSE_Donor_1.CEL
etc.
```

Everything is sent (redirected with `>`) to file `rename.txt` that lists the rename commands. The file can be activated with a single command line. Unlike in Windows the filename extension does not matter and we don't need to add `.cmd` at the end. Simply type the following command after the `$` prompt:

```
$ sh rename.txt
```

The files will be renamed with the naming scheme we devised, with attributes separated by an underscore.

Getting started

At this point you should have the following necessary steps accomplished:

- 1) ArrayStar installed and functional.
- 2) Created an Affymetrix username.
- 3) Have prepared the public dataset CEL files as detailed in the previous chapter or received them in class.

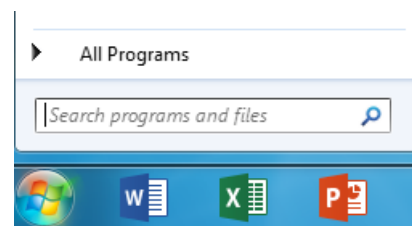
1. Locating & launching ArrayStar

The exact location of the software may depend on the version of Windows you are using. However, you should be able to locate it within the Windows “Start” button:

TASK

- 1) Click on the “**Start**” button
- 2) ArrayStar may appear on the list of programs
- 3) If it is not on the list click “All Programs” and locate it in the updated list.

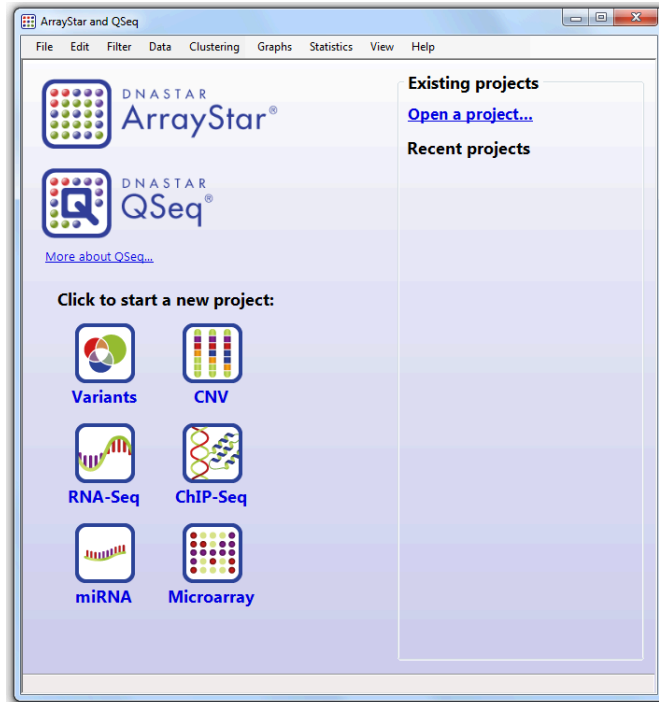
Note: it may be possible that a shortcut exist on the desktop depending on the local installation.



When you have located the software select it or double-click it to start it.

Once launched a main screen will appear with various options as shown below.

Depending on licensing various options might be present such as the QSeq option.

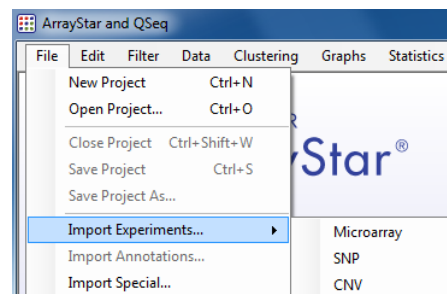


2. Creating a project

A new “project” is automatically started when new data is entered into the software. The “project” is then saved as a special file that contains, in compressed form, all of the data and ancillary file information. ArrayStar project files end with the filename extension: `.astar`.

 **TASK:** Create a new project.

A new microarray project can be created by either clicking on the Microarray logo located under “Click to start a new project:” on the main page or by using the menu cascade **File > Import Experiments > Microarray**



In all cases the “Microarray Project Set-up” wizard window will appear and we can navigate to the location of the `.CEL` files we want to import.

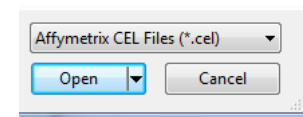


2.1. Import CEL files

✓ **TASK:** Click on the “Add files...” button.

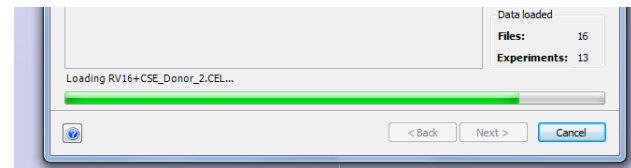
Then **navigate** to the location of your CEL files. In the previous chapter we organized files in a directory called DATA placed on the Desktop. Navigate to the directory/folder where yours are located.

Select the CEL files you want to read. In case your folder contains many other types of files a name filter is provided at the bottom right to display only CEL files. Other filters are also available.



Then **click “Open”** to load the files into ArrayStar.

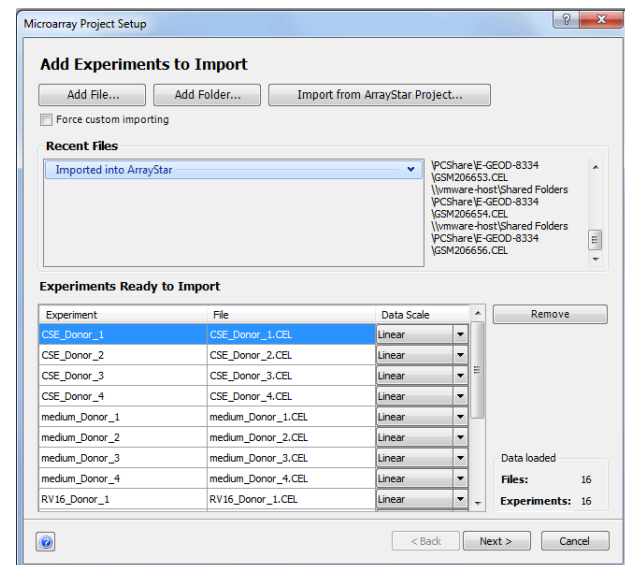
Files are imported one by one.



The files are read-in and displayed at the bottom of the window. The “Data Scale” is set to “Linear” as this is the kind of data that CEL files contain. Data from other sources could be detected to be in log form for example and confirmation might be needed.

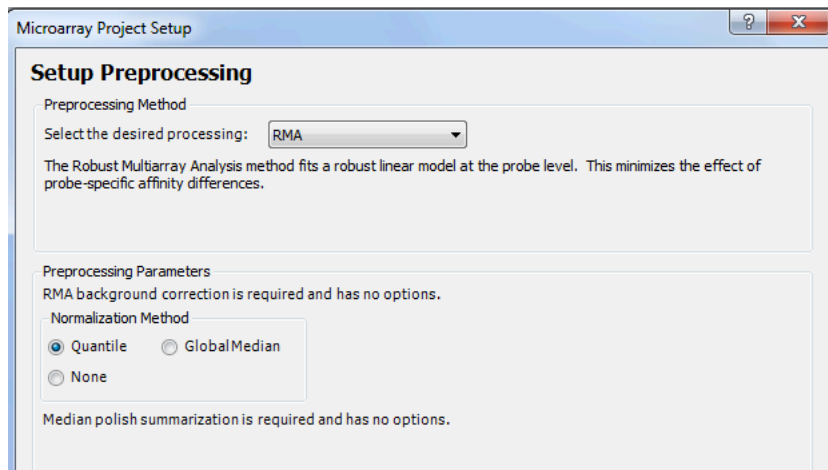
Click Next>

to continue the import process.



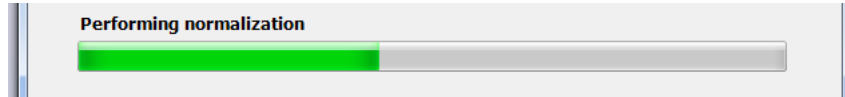
2.2. Preprocessing

The next window will be about “pre-processing” with RMA by default. RMA (Robust Multiarray Analysis) was developed on Affymetrix arrays and is well suited for all typical cases.



Unless you have unusual data simply **press Next>** to accept and continue.

Files will be processed:



2.3. Attributes and replicates

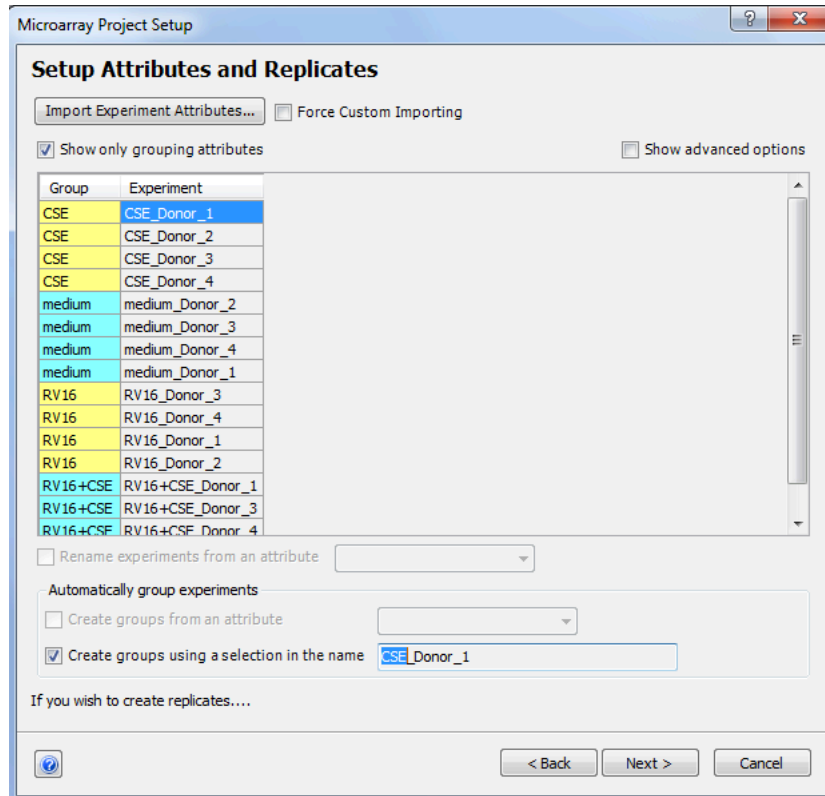
This section is the reason we worked so hard in the previous chapter to rename the files!

✓ TASK Click "Create groups using a selection in the name" at the bottom of the window.

Then **partially select** CSE from the name of the first experiment, in this case CSE_Donor1, only until the underscore. This will automatically create group categories by attributes:

- *CSE*
- *medium*
- *RV16*
- *RV16+CSE*

as illustrated below:



Our hard work renaming files has paid off!

Click **Next >** to continue the import process.

2.4. Import Annotations


Each array contains millions of probes, each a little square on the surface of the array. Each probe corresponds to a portion of a gene, synthesized based on sequence on that specific location on the array. Each probe has a unique, but cryptic name.

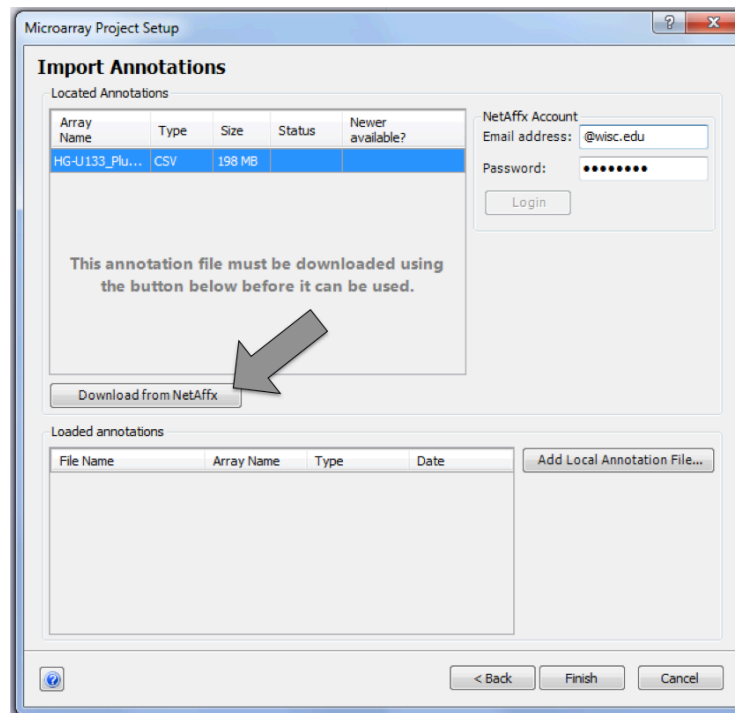
But which gene is where?

This is the purpose of the annotation file that provides a correspondence between the probe name and the gene name as well as further annotations such as the complete gene sequence accession code for the RNA, the protein, from various databases. Other data is available depending on the type of array or the manufacturer.

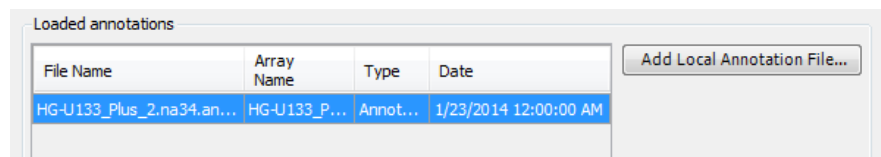
You should be registered on the Affymetrix web site (also called NetAffx) to be able to continue with the next step. If you need to register follow the step detailed in a previous section (Affymetrix Arrays on page 9.)

The annotation file will be downloaded in a reserved folder on your Windows computer.

 **TASK:** Click on “Download from NetAffx” on the Import Annotations window (shown with arrow below:)

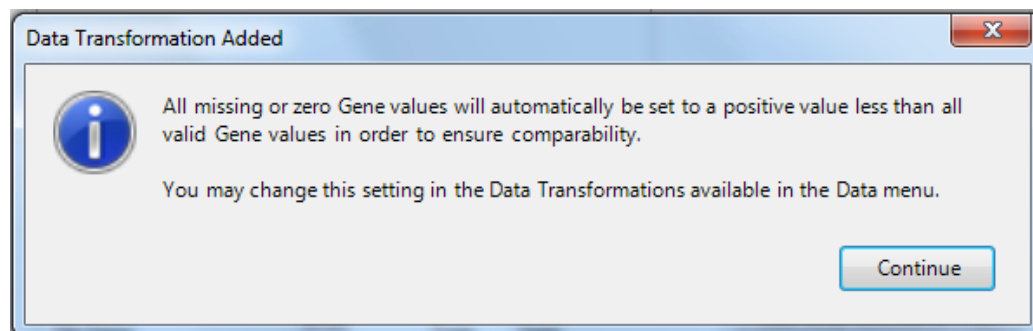


Once the file is downloaded locally it can be re-used for the exact same type of microarrays in future analyses and will appear within the bottom window

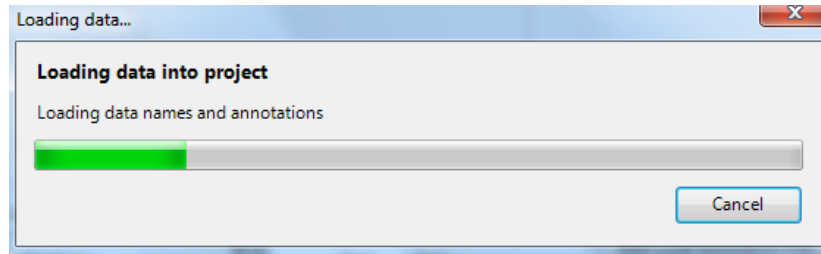


Click the Finish button

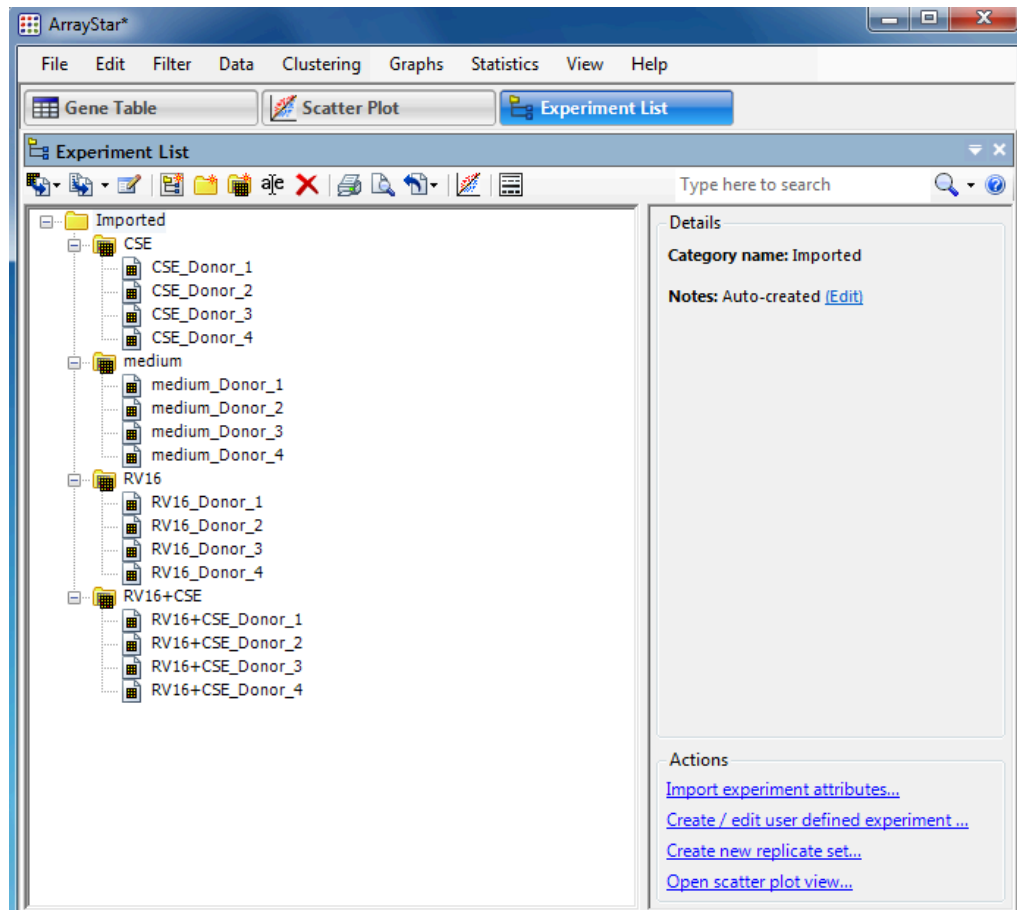
A data transformation warning might appear, just click Continue.



The data will now be loaded into the project.



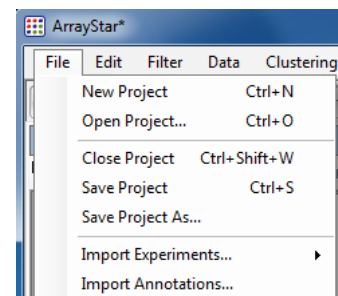
And the ArrayStar page will update to the Experiment List view.



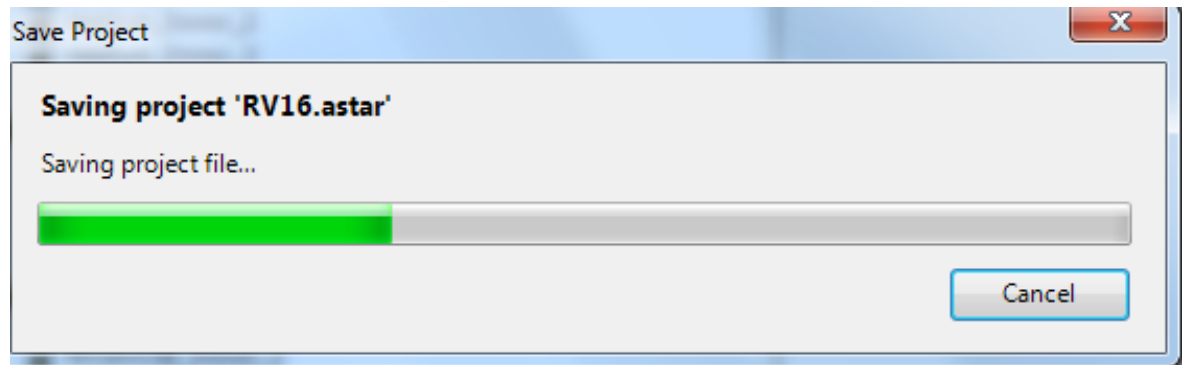
✓ TASK

Save the project into a file with the menu **File> Save Project As...**

Call the file *e.g. RV16.astar* that will be saved by default in the current directory.



Note: You can save the file elsewhere as the project file is stand-alone and no longer need the CEL or annotation files.



The project files are extremely compressed and take much less disk space than the sum of the original data files and annotations.

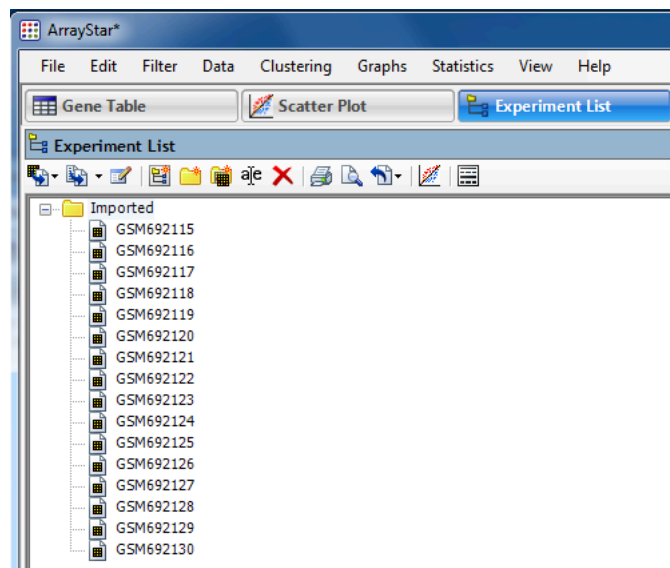
3. Organizing generic data files

What if I don't have files with such clever names? What do I do?

We had a whole chapter to prepare the files with somewhat “generic” GSM names from dataset GSE27973_RAW.


Files named GSM692115 through GSM692130, together with additional information about the samples, were changed into filenames that reflect the samples attributes.

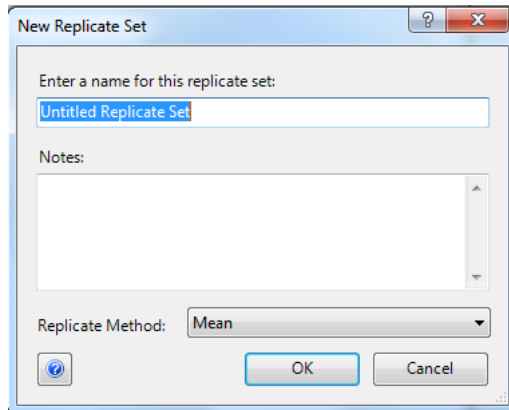
However, if we had used the original file names we would have obtained the following Experiment List in ArrayStar:



All the work has to now be done “manually” to create the groups of interest.

Here is a quick summary on how to do that:

- 1) Click on the **Imported** folder to select it.
- 2) Click the **Replicates** button  on the toolbar above.
- 3) This will open a New Replicate Set window



- 4) Enter the name of the first group, for example **RV16**. Leave the Replicate Methods as **Mean** and Click OK.
- 5) We now need to select and drag within the new replicate directory the corresponding GSM files. We can identify them visually looking at the samples information on the original web page. The RV16 files are:

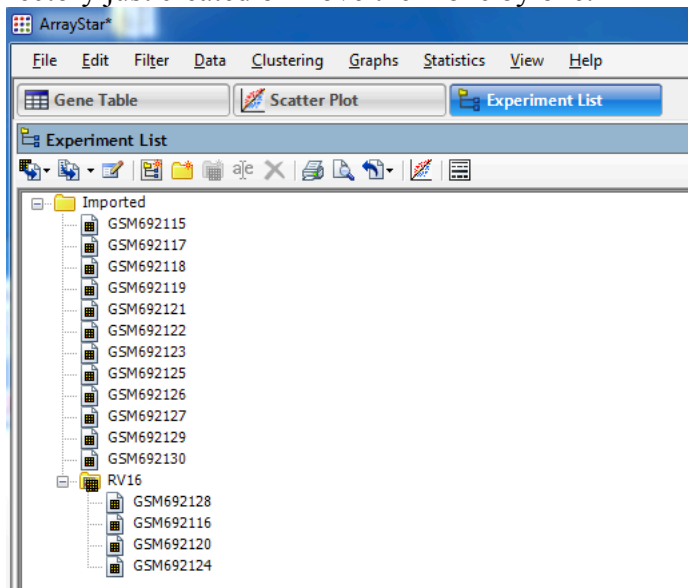
GSM692116

GSM692120

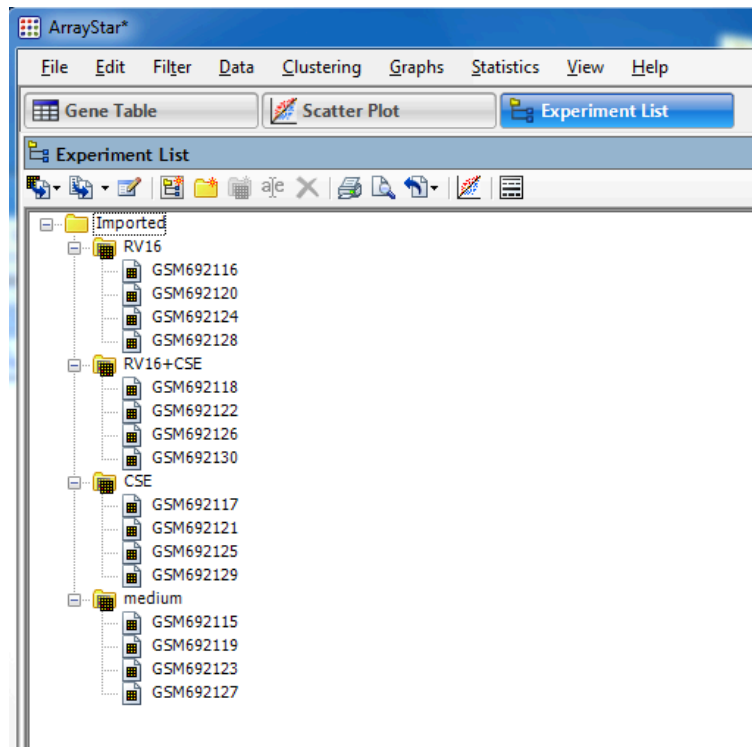
GSM692124

GSM692128

- 6) Use **Ctrl-Click** to select these files and drag them into the RV16 Replicate set directory just created or move them one by one.



- 7) **Repeat the process** for the other 3 groups: CSE, medium, RV16+CSE



Clearly this is much more cumbersome but also prone to errors, and the filename change should be preferred.

Note that “Donor” and its associated number does not appear as an accessible attribute unless the name of the files are edited manually (right-click.)

If you are creating your own data, it is now clear how the naming of files can be chosen wisely and clearly.

4. Using Attributes files

There is another method to import attributes related to a set of data files. However, this requires to either obtain original attributes files such as Affymetrix ARR files but these are not always available.

ArrayStar will accept a plain text file with tab-delimited columns of attributes.

Using the same GSE27973_RAW data as before we can construct an attributes file easily in Unix/Linux/Mac with the following 2 commands if the samples attributes have been copied and pasted into a file called samples.txt (see Preparation on Unix|Linux|Mac on page 21.)

Note: The file needs to have a header for each column or one sample will not be assigned attributes.

```
$ echo -e "ExperimentID\tCase\tDonor" >
samplesAttributes.txt
```



```
$ cat samples.txt | awk '{print $1"\t"$5"\t"$2"_"$3}' >>
samplesAttributes.txt
```


The final file will look like this:

ExperimentID	Case	Donor
GSM692115	medium	Donor_1
GSM692116	RV16	Donor_1
GSM692117	CSE	Donor_1
GSM692118	RV16+CSE	Donor_1
GSM692119	medium	Donor_2
GSM692120	RV16	Donor_2
GSM692121	CSE	Donor_2
GSM692122	RV16+CSE	Donor_2
GSM692123	medium	Donor_3
GSM692124	RV16	Donor_3
GSM692125	CSE	Donor_3
GSM692126	RV16+CSE	Donor_3
GSM692127	medium	Donor_4
GSM692128	RV16	Donor_4
GSM692129	CSE	Donor_4
GSM692130	RV16+CSE	Donor_4


The reader is invited to try the methods shown above (Preparation for Windows on page 16) to create a similar text file in Windows e.g. with Excel or come up with an alternative editing method.


Once the attributes file (`samplesAttributes.txt`) is at hand, it can be incorporated within the ArrayStar information. However, this will not change the file names displayed within the Experiment List.

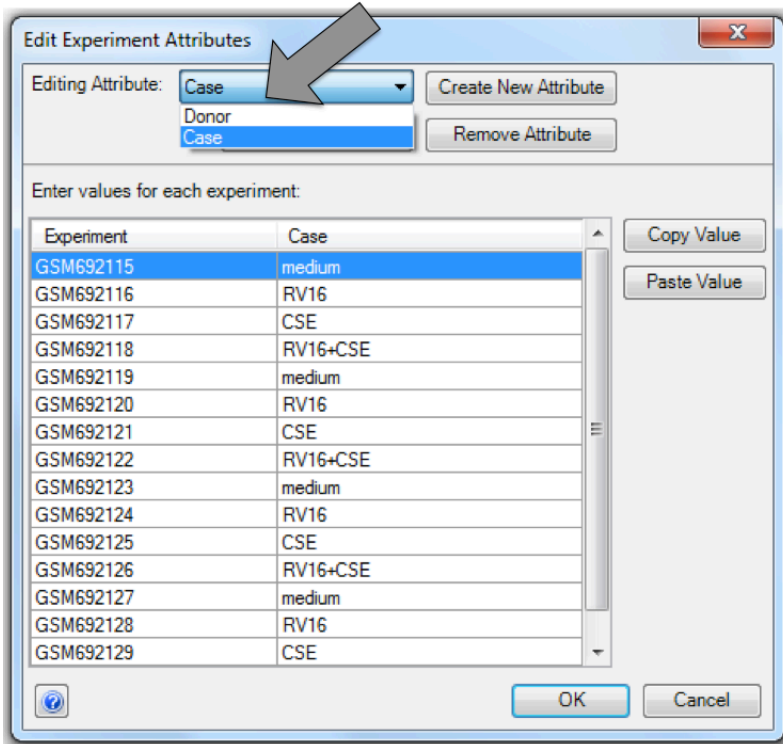
Here are summary step to incorporate the above attributes:

- 1) **Click** on the “Import Experiments Attributes...” button  in the toolbar shown for Experiment List view.
- 2) Navigate to the location of the `samplesAttributes.txt` file or equivalent file (with attributes and header)
- 3) Click **Open**
- 4) Nothing clearly visible will happen but on the right hand side a report will be made on how many user defined attributes have been assigned:

```
Details
20 experiments in 1 group
2 user defined attributes total for 16 experiments.
```

- 5) Attributes can be manually created and reviewed with the toolbar button  (Create / Edit User Defined Experiment Attributes)

- 6) **Click** this attributes  button to reveal the list of attributes. Attributes are shown one column at a time, and a pull-down menu at the top (see Arrow) is used to select which attribute is displayed (the name of which is that of the header columns in the imported attributes file)



- 7) Attributes could have been created manually here, one by one, and can also be edited. ***However, this is a potential risk for Human error!***

Identifying genes of interest

✓ **READ** Let's first remember that a microarray contains millions of probes to account for thousands of genes and/or exons. The experiment performed on the array is “*looking at*” **all** of those probes and genes at the same time. This requires the help of statistical methods to sort out what can be different between sample groups.

Statistical methods include “hypothesis testing” with the underlying assumption that there are relatively few (less than 10%) changes between samples between groups.

Another important assumption is that the data behaves mostly as a “normal” (Gaussian) distribution even though even in the best cases it cannot be exactly so and in some other cases it is completely not so!

In addition, due to the fact that we are looking at so many probes or genes “at the same time” implies using the “multiple testing correction” to try at best to limit false positive and false negative results (false discovery rate or FDR.) This implies “confidence levels” such as 5% or 95% depending on perspective, in that case “allowing” up to 5% of the results to be “false.”

There are two main schools to select differentially expressed genes: p-values and fold change. Depending on the “normality” of the data and the ranking of p-values one or the other might be more appropriate for your data.

If there are enough biological replicates, ArrayStar will automatically pre-compute t-test at confidence intervals of 99%, 95% and 90% and fold change sets of genes at 2-fold, 4-fold and 8-fold change by default. Other values can then be computed with help of filters.

Groups: if there are more than 2 groups, at some point pair-wise comparisons (“contrasts” in statistical jargon) can be defined to obtain ratio between 2 measurements as a way to compare samples.

With that in mind, ArrayStar has a very easy graphical interface and provides “Views” that are pre-computed or easily computed on the spot.

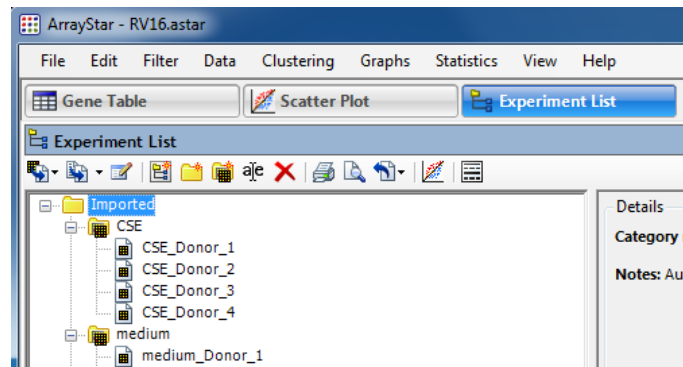
1. Scatter plot

The **Experiment View** is the default view we have used in the previous chapter to arrange how samples are sorted in groups.

We now want to be able to compare groups and this can be done first in a graphical manner.

✓ TASK:

Click on the “**Scatter Plot**” tab in the main window.

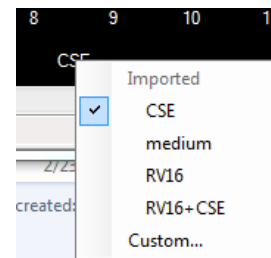


By default the first 2 groups in the list (here CSE and medium) will be compared to each other. However, we want to compare the medium and RV16 groups so we will need to change this accordingly.

✓ TASK: change the groups that are plotted in the scatter plot:

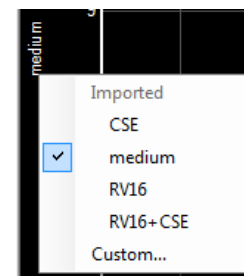
At the bottom of the plot (X axis) click on **CSE** and select **medium** instead.

Since the Y axis is currently also selected as **medium**, the plot will temporarily change to a straight line.



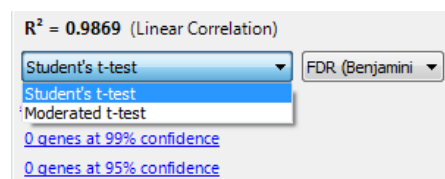
In the same way, on the left (Y axis) click on **medium** and select **RV16** instead.

The plot will now change from a straight line to a scatter plot.



At the top right, change from **Student's t-test** to **moderated t-test**.

The pre-computed genes for confidence levels (quick links) will be updated accordingly.





Note that under Student's t-test all confidence quick links are at 0.

Moderated t-test (from Help:) The primary difference between the two methods is in the calculation of variance. The Student's t-Test calculates variance from the data that is available for each gene. The Moderated t-Test uses information from all of the selected genes to calculate variance.

The pre-computed genes for confidence levels (quick links) will be updated accordingly.

Click on the link “**144 genes at 90% confidence**” to select those genes. They will be colored white on the scatter plot and have 90% confidence of being differentially expressed between the 2 groups.

$R^2 = 0.9869$ (Linear Correlation)

Moderated t-test

FDR (Benjamini)

[Advanced Filtering...](#)

[23 genes at 99% confidence](#)

[96 genes at 95% confidence](#)

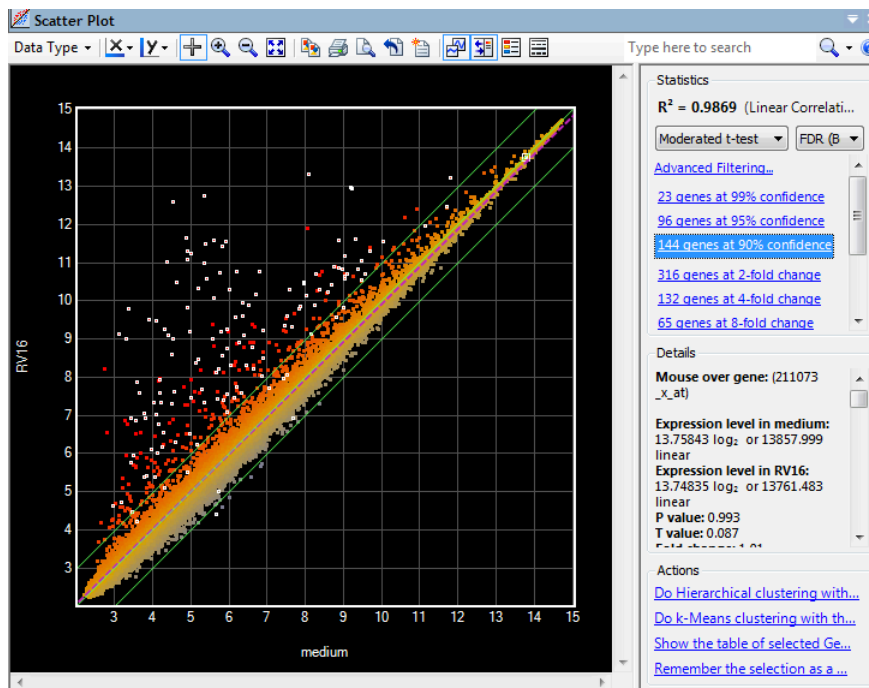
[144 genes at 90% confidence](#)

[316 genes at 2-fold change](#)

[132 genes at 4-fold change](#)

[65 genes at 8-fold change](#)

Note that some genes are below the 2-fold level represented by the 45° angle green lines on both sides of the scatter points, representing the 2-fold up (upper line) and 2-fold down-regulated genes (bottom line.) The dashed line represent the absolute “no change” line.



It is interesting to note that most these selected genes are on the top section (up-regulated) and just a handful of genes are in the down-regulated group (bottom.)

Note: One feature of ArrayStar is that genes or probes selected in one view are automatically selected in all view and tables. This is a very useful feature to be consistent and know exactly what is currently selected regardless of where you are looking.

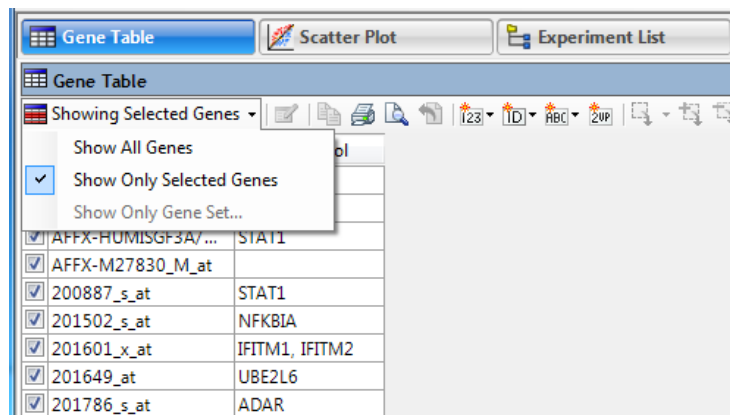
2. Gene Table

✓ **TASK**: Click on the “**Gene Table**” tab in the main window. All genes selected in the scatter plot are selected but not yet clearly visible.

✓ **TASK**: change the option to **show only selected genes** within the gene table as shown.

All other genes will not be listed.

Currently there are only 2 columns.



We will add more column of information later when we have selected a final set of genes.

3. Filters

Filters can be used for many filtering options, including fold change levels and statistics, but also for selecting genes that belong to a specific group into a list.

3.1. Down-regulated genes

Here we'll select based on fold change to list all the few down-regulated genes spotted earlier.

✓ **TASK**: Click on the “**Scatter Plot**” tab in the main window to return to that view and at the top right **click on Advanced Filtering...**

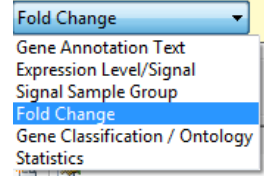
A filter window will open.

The first filter is that of the moderated t-test. It is set at 90% significance as this is what we selected before.



Do not change this first filter.

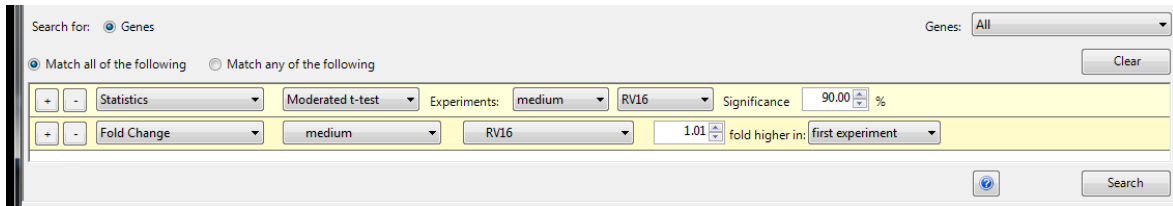
Change the second filter line from the default value (gene Annotation Text) **to Fold Change**



Change the rest of the filter to reflect a fold change of 1.01 fold higher in the first experiment. First experiment is listed as **medium**.

Click Search

The final filter should look like this:



The result will be a handful of genes (6 total) that are below the median line on the scatter plot as we spotted them earlier. The list will contain many more columns that we'll review later when we make a final list:

Probe Set ID	Gene Symbol	SEQ_URL	GeneChip Array
210282_at	ZMYM2	https://www.affymetrix.com/Li...	Human Genome U133 Plus 2.0...
224410_s_at	LMBR1	https://www.affymetrix.com/Li...	Human Genome U133 Plus 2.0...
228433_at	NFYA	https://www.affymetrix.com/Li...	Human Genome U133 Plus 2.0...
228495_at	GPATCH11	https://www.affymetrix.com/Li...	Human Genome U133 Plus 2.0...
229983_at	TIGD2	https://www.affymetrix.com/Li...	Human Genome U133 Plus 2.0...
240066_at	AC000123.3, OTTHUMG0000066283	https://www.affymetrix.com/Li...	Human Genome U133 Plus 2.0...

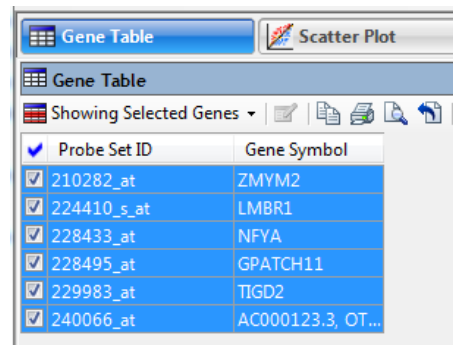
Gene found: 6

Sequence Source	Transcript ID(Array Design)	Target Description	Representative Public ID	Archival UniGene Cluster	UniGene ID
GenBank	g12052767	g>AL136621.1 /DB_XREF=gi:1...	AL136621	Hs.109526	Hs.507433 , Hs.609078 , Hs.613...
GenBank	g13591769	g>AF348513.1 /DB_XREF=gi:1...	AF348513	Hs.900623	Hs.209989 , Hs.594986 , Hs.615...
GenBank	Hs.10441.0	g>AU157605 /DB_XREF=gi:11...	AU157605	Hs.10441	Hs.10441
GenBank	Hs.194293.0	g>AI880633 /DB_XREF=gi:555...	AI880633	Hs.194293	Hs.595250
GenBank	Hs.58924.0	g>AI610112 /DB_XREF=gi:461...	AI610112	Hs.58924	Hs.58924
GenBank	Hs.105623.0	g>AI038071 /DB_XREF=gi:327...	AI038071	Hs.105623	Hs.663241

Gene Ontology Biological Process	Gene Ontology Cellular Component	Gene Ontology Molecular Function	Pathway	InterPro	Trans Memb
transcription, DNA-templated...	nucleus , cytosol , PML body	zinc ion binding , ubiquitin conj...		IPR010507 , IPR010507 , IPR01...	
embryonic digit morphogenesis	membrane , integral componen...			IPR006876 , IPR006876 , IPR00...	
transcription, DNA-templated...	nucleus , nucleoplasm , CCAAT-...	DNA binding , sequence-specifi...			
		nucleic acid binding		IPR000467	
	nucleus	nucleic acid binding , DNA bin...		IPR004875 , IPR006600 , IPR00...	

✓ **TASK:** Click on the button  on the middle task bar in the filter window to “**Select and Show Results in Gene Table**”

The gene table will be updated to this gene list



Probe Set ID	Gene Symbol
<input checked="" type="checkbox"/> 210282_at	ZMYM2
<input checked="" type="checkbox"/> 224410_s_at	LMBR1
<input checked="" type="checkbox"/> 228433_at	NFYA
<input checked="" type="checkbox"/> 228495_at	GPATCH11
<input checked="" type="checkbox"/> 229983_at	TIGD2
<input checked="" type="checkbox"/> 240066_at	AC000123.3, OT...

We could follow a similar procedure to select the other, up-regulated genes. However, it may be best to keep all genes that are differentially expressed in a single table.

4. Clustering

Since we have only 6 genes selected at the moment, we can test clustering options.

4.1. Hierarchical clustering

By default all genes from all experiments will be plotted. However, it may sometimes be better to plot only the genes that are within specific experiments, as in our case we are more interested in the 2 groups medium and RV16.

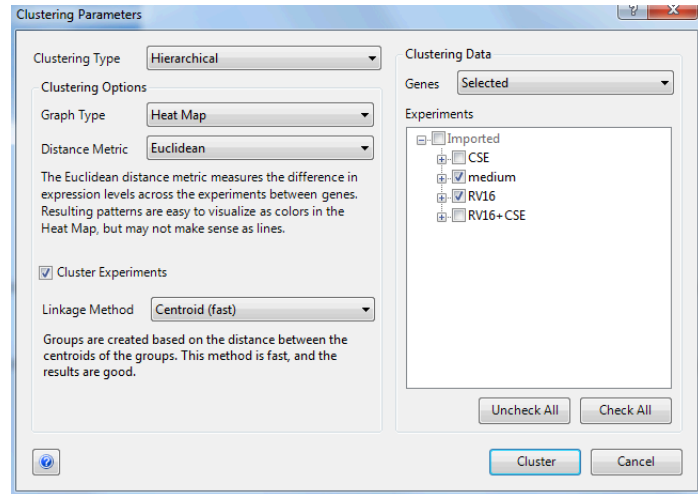
To accomplish this, rather than use the default buttons or quick links offered at various places, we'll use the menu **Clustering > Advanced Clustering...**

✓ **TASK:** use the menu **Clustering > Advanced Clustering...**

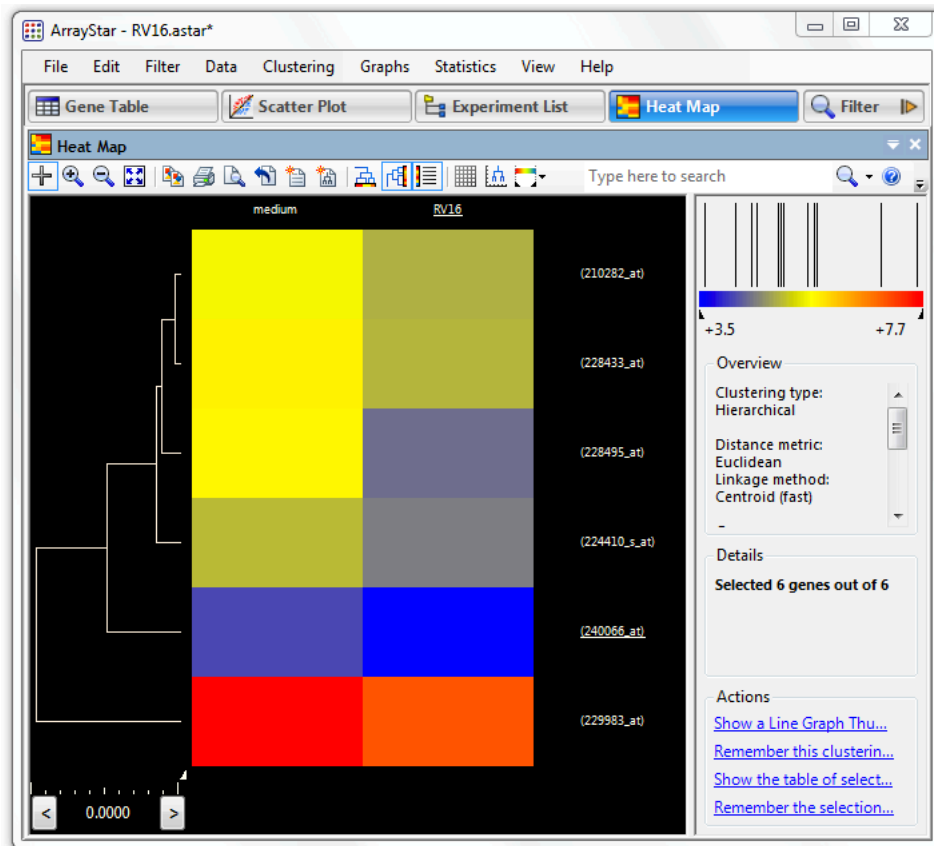


This will open a window within which we can select the experiments we want to plot.

Keeping all other selections as is, **uncheck** CSE and RV16+CSE and **click Cluster**



This will cluster the 6 genes by the listed methods and plot the values for the 2 selected experiments:



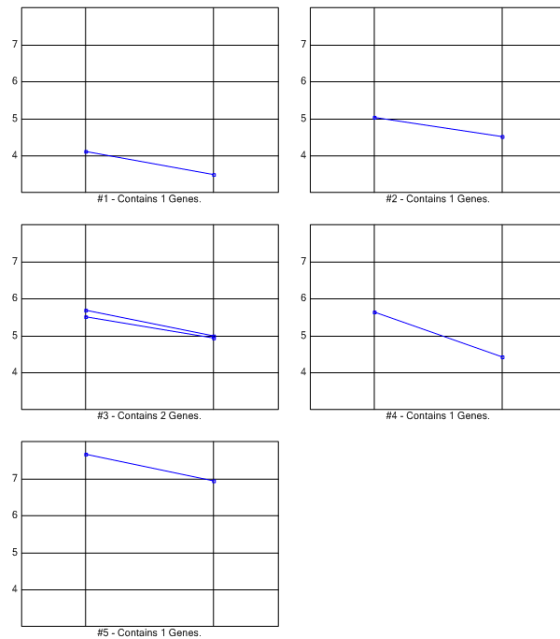
The name of the genes will appear within the right hand side of the window when hovering the mouse over.

4.2. Line clustering

Within the cluster results that we just accomplished above, at the bottom right under “Actions” :

✓ TASK: click on “Show a Line Graph Thumbnails view of this clustering result.”

This will present the data in groups with the log-intensity value on the Y axis and the group on the X axis. The image shown here is that obtained from printing as a white background offers more clarity than the default black background color.



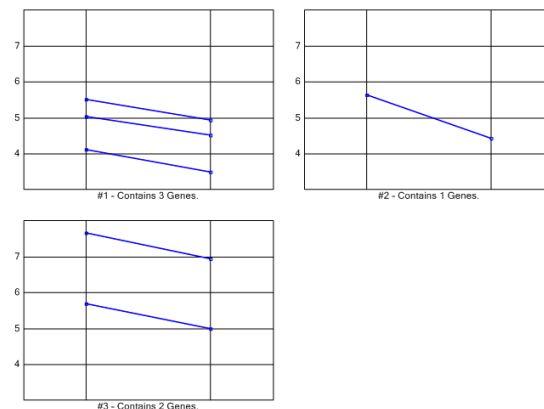
4.3. K-means

A similar plot can be obtained for the K-means clustering.

Making sure that those 6 genes are still selected:

✓ TASK: use the menu cascade **Clustering > Advanced Clustering...** and select the K-means as the method.

This image is show as the Print preview for clarity.



5. Gene sets – up-regulated genes

Let’s review the various steps above and apply them to up-regulated genes, and at the same time save the list into a “set” that can be retrieved later.



5.1. Filter up-regulated genes

✓ **TASK:** Using the same method as previously let's retrieve the up-regulated genes:

- 1) On the **Scatter Plot** view, under **Moderated t-test**, click on **144 genes** at 90% confidence
- 2) Click **Advanced Filtering...**
- 3) Change the method to **Fold Change**
- 4) Verify that the first experiment is **medium** and the second is **RV16**
- 5) **Adjust to fold 1.01 higher** in the **second experiment** (pull-down menu)
- 6) **Click Search**

The result will display 138 genes (= 144 – 6 down-regulated we found earlier.)

Filter: Moderated T test: medium vs. RV16, P value \geq 90.00% AND Fold Change: medium/RV16 1.01 fold higher in second experiment

Search Criteria:

Search for: Genes Genes: All

Match all of the following Match any of the following Clear


+	-	Statistics	Moderated t-test	Experiments: medium	RV16	Significance	90.00 %
+	-	Fold Change	medium	RV16	1.01 fold higher in:	second experiment	

Search Results

Notes	Probe Set ID	Gene Symbol	SEQ_URL	GeneChip Array	Species Scientific Name	Annotation Date	Sequence Type
	AFFX-HUMISGF3A/M97935_M...	STAT1	https://www.affymetrix.com/Li...	Human Genome U133 Plus 2.0...	Homo sapiens	Oct 23, 2013	Control sequence
	AFFX-HUMISGF3A/M97935_M...	STAT1	https://www.affymetrix.com/Li...	Human Genome U133 Plus 2.0...	Homo sapiens	Oct 23, 2013	Control sequence
	AFFX-HUMISGF3A/M97935_3_...	STAT1	https://www.affymetrix.com/Li...	Human Genome U133 Plus 2.0...	Homo sapiens	Oct 23, 2013	Control sequence
	AFFX-M27830_M_at		https://www.affymetrix.com/Li...	Human Genome U133 Plus 2.0...	Homo sapiens	Oct 23, 2013	Control sequence
	200887_s_at	STAT1	https://www.affymetrix.com/Li...	Human Genome U133 Plus 2.0...	Homo sapiens	Oct 23, 2013	Exemplar sequence
	201502_s_at	NFKBIA	https://www.affymetrix.com/Li...	Human Genome U133 Plus 2.0...	Homo sapiens	Oct 23, 2013	Consensus sequenc
	201601_x_at	IFITM1, IFITM2	https://www.affymetrix.com/Li...	Human Genome U133 Plus 2.0...	Homo sapiens	Oct 23, 2013	Exemplar sequence
	201649_at	UBE2L6	https://www.affymetrix.com/Li...	Human Genome U133 Plus 2.0...	Homo sapiens	Oct 23, 2013	Exemplar sequence
	201786_s_at	ADAR	https://www.affymetrix.com/Li...	Human Genome U133 Plus 2.0...	Homo sapiens	Oct 23, 2013	Exemplar sequence
	202086_at	MX1	https://www.affymetrix.com/Li...	Human Genome U133 Plus 2.0...	Homo sapiens	Oct 23, 2013	Exemplar sequence
	202307_s_at	TAP1	https://www.affymetrix.com/Li...	Human Genome U133 Plus 2.0...	Homo sapiens	Oct 23, 2013	Exemplar sequence

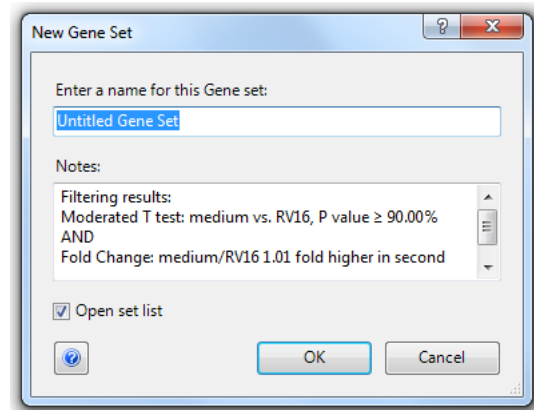
Gene found: 138

5.2. Create a new gene set

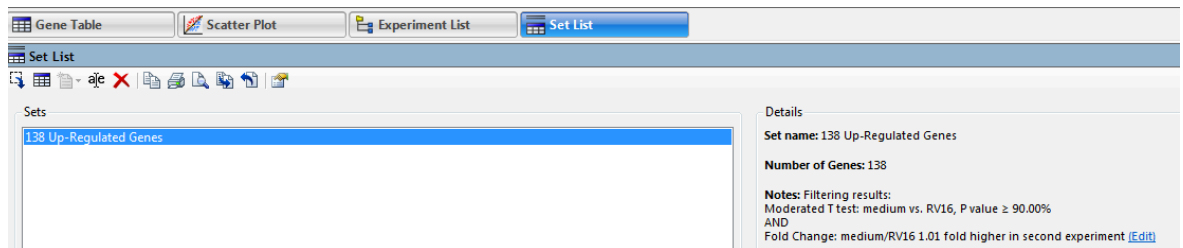
✓ **TASK:** Click on the button  on Search Results task bar in the middle of the window to “**Remember Results as a Gene Set**”

Change the name from “Untitled Gene Set” to e.g. “**138 Up-Regulated Genes**”

Click **OK**




Note: a new Tab will appear next to the “Experiment List” Tab named “Set List” containing one line: our newly created set. On the right hand side is presented a summary of the set: number of genes and method used to select them.



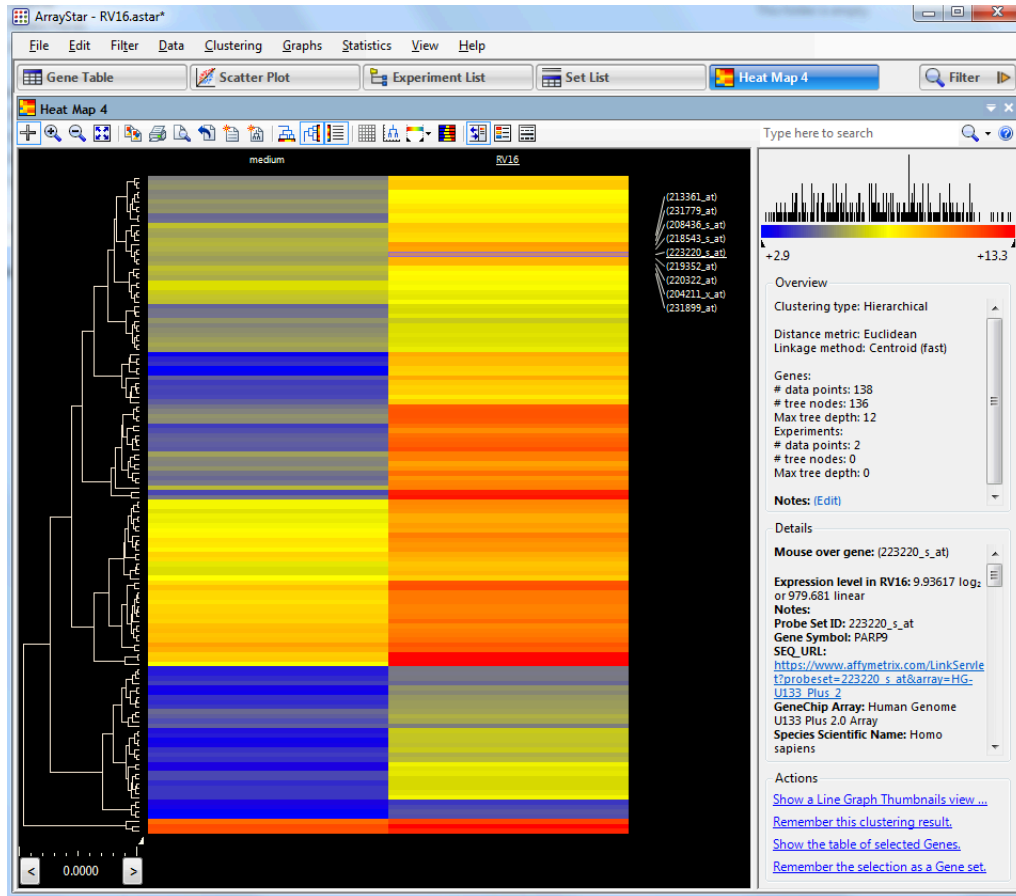
5.3. Hierarchical clustering of up-regulated gene set

✓ **TASK:**

- 1) Click on the button  on the middle task bar in the filter window to “**Select and Show Results in Gene Table**”
- 2) Use the menu cascade **Clustering > Advanced Clustering...**
- 3) Keep only **medium** and **RV16** selected in the **Experiments** window
- 4) **Click Cluster**

138 genes should be used for the cluster.

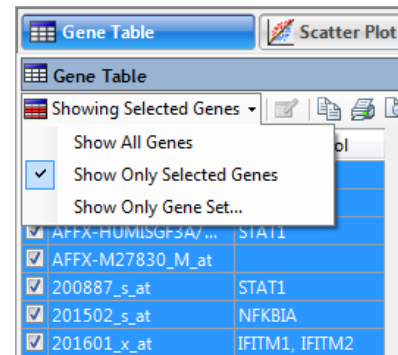
With mouse hovering over the heat map, the probe name is underlined within a local list and information about the probe is shown on the right hand panel.



6. Gene Table: annotation & export

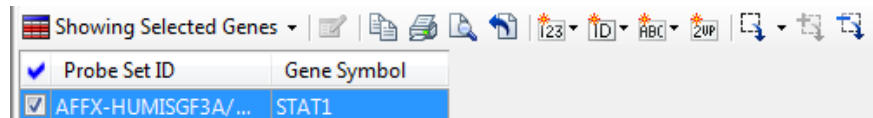
✓ **TASK:** First let's (re)create a gene table for all 144 genes at 90% confidence:

- 1) On the **Scatter Plot** view, under **Moderated t-test**, click on **144 genes** at 90% confidence
- 2) At the bottom right, under **Actions**, click on “**Show the table of selected Genes**”
- 3) Show only selected genes



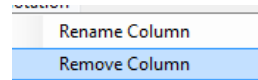
The gene table currently has 2 columns: “ProbeSet ID” and “Gene Symbol”

Now make note of the tool bar presented in this view:



We will use the “123”, “ABC” and “2UP” buttons to populate the gene table with new columns of information in the steps below.

Note: If you add columns that you don’t want: simply right-click and choose “remove column”



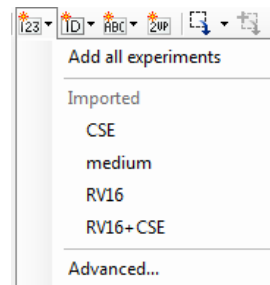
6.1. Add experiment values

✓ TASK: We will now add the intensity (log values) for the experiment of choice as new columns within the gene table:

- 1) **Click** on the “123” button and select **medium**: a new column is added
- 2) **Repeat** and select **RV16**: a new column is added

The gene table has now 4 columns with the 2 we just added.

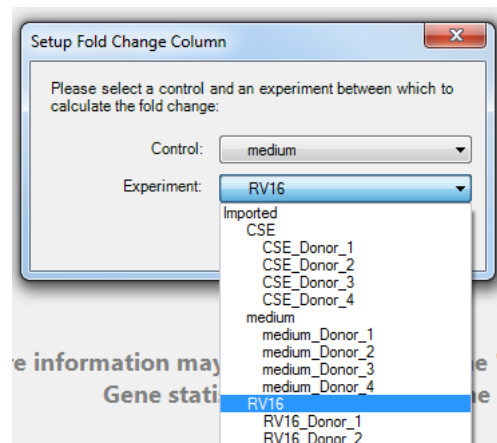
The values are log₂ values of the intensity detected.



6.2. Add fold-change columns

✓ TASK: We will now add the fold change for the 2 experiments listed.

- 1) **Click** on the “2UP” button
- 2) Select **medium** for the Control value
- 3) Select **RV16** for the Experiment value
- 4) **Click OK**






After the ratio the word “up” or “down” is shown to indicate the regulation level.

We now have a total of 5 columns in our gene list.

Probe Set ID	Gene Symbol	medium	RV16	Fold change
<input checked="" type="checkbox"/> AFFX-HUMISGF3A/...	STAT1	5.54747	9.01416	11.055 up
<input checked="" type="checkbox"/> AFFX-HUMISGF3A/...	STAT1	4.98356	7.45248	5.536 up
<input checked="" type="checkbox"/> AFFX-HUMISGF3A/...	STAT1	7.92150	10.49684	5.960 up
<input checked="" type="checkbox"/> AFFX-M27830_M_at		8.97251	9.81341	1.791 up
<input checked="" type="checkbox"/> 200887_s_at	STAT1	8.83401	11.57062	6.665 up
<input checked="" type="checkbox"/> 201502_s_at	NFKBIA	8.92968	10.83499	3.745 up
<input checked="" type="checkbox"/> 201601_x_at	IFITM1, IFITM2	9.15149	12.96854	14.094 up

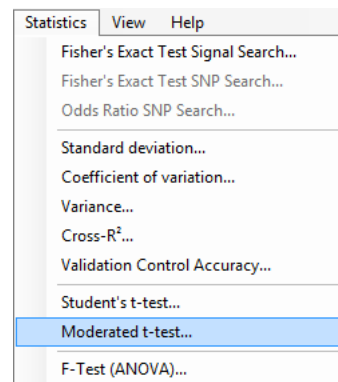
6.3. Add statistics columns

Note: Statistical calculation columns are added via a menu selection rather than clicking a button.

 **TASK:** We will add a column for the moderated t-test.

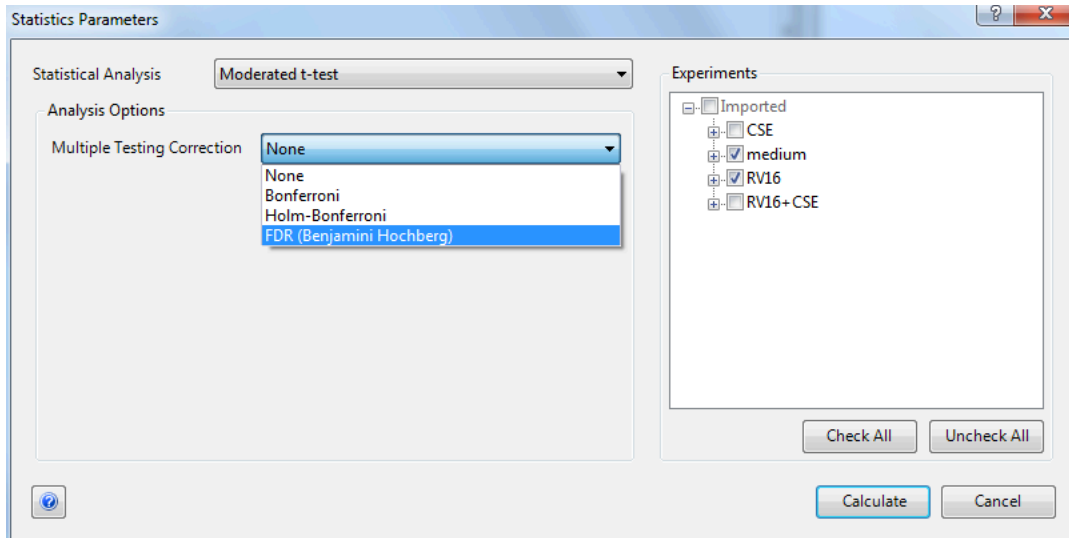
Follow the menu cascade:

Statistics > Moderated t-test...



In the Statistics Parameter window verify or change the following:

- select (check) the **medium** and **RV16** experiments at right
- choose FDR for Multiple Testing Correction in the pull-down menu



Click Calculate

Two new columns will be added to the table:

Probe Set ID	Gene Symbol	medium	RV16	Fold change	P value	T value
AFFX-HUMISGF3A/...	STAT1	5.54747	9.01416	11.055 up	0.0385	-7.306
AFFX-HUMISGF3A/...	STAT1	4.98356	7.45248	5.536 up	0.0251	-8.096
AFFX-HUMISGF3A/...	STAT1	7.92150	10.49684	5.960 up	0.0245	-8.265

P value is the FDR result and T value is the t statistic calculation.

If more columns are added they will be named P2 and T2, P3 and T3 etc.

6.4. Add annotations

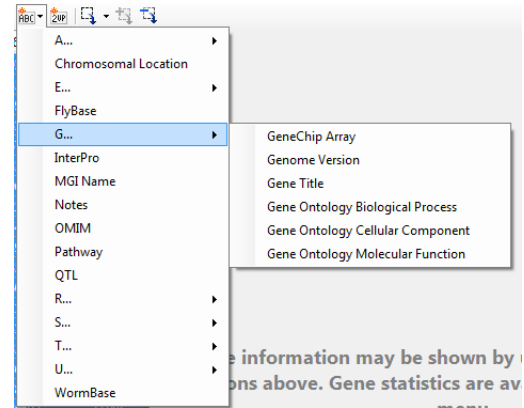
The annotations available here are dependent on the array type and manufacturer and are collected as part of the annotation imported when the project is set-up. See section Import Annotations on page 29.



Annotations are added by clicking on the “ABC” button

Some annotation headings are grouped together into sub-menus for compactness.

Highlighted here is the content of the G... submenu.



TASK: using the “ABC” button add the Gene Ontology options for “Biological Process” and for “Molecular Function.”

We now have 9 columns in our gene table:

Probe Set ID	Gene Symbol	medium	RV16	Fold change	P value	T value	Gene Ontology Biological Process	Gene Ontology Molecular Function
<input checked="" type="checkbox"/> AFFX-HUMISGF3A/...	STAT1	5.54747	9.01416	11.055 up	0.0385	-7.306	negative regulation of transcription from B...	RNA polymerase II core promoter proximal...
<input checked="" type="checkbox"/> AFFX-HUMISGF3A/...	STAT1	4.98356	7.45248	5.536 up	0.0251	-8.096	negative regulation of transcription from B...	RNA polymerase II core promoter proximal...
<input checked="" type="checkbox"/> AFFX-HUMISGF3A/...	STAT1	7.92150	10.49684	5.960 up	0.0245	-8.265	negative regulation of transcription from B...	RNA polymerase II core promoter proximal...
<input checked="" type="checkbox"/> AFFX-M27830_M_at		8.97251	9.81341	1.791 up	0.0747	-6.144		
<input checked="" type="checkbox"/> 200887_s_at	STAT1	8.83401	11.57062	6.665 up	0.00857	-10.457	negative regulation of transcription from B...	RNA polymerase II core promoter proximal...
<input checked="" type="checkbox"/> 201502_s_at	NFKBIA	8.92968	10.83499	3.745 up	0.00807	-11.946	protein import into nucleus, translocation, t...	protein binding, transcription factor binding...

6.5. Export gene table to text file

For this purpose we will use the button “Export Selected Table Row...”

TASK:

Click the export button and give a name to the file, e.g. 144-Modt-90.txt

Note: This file name is chosen to reflect the number of genes (144), the method used to select them (Modt) and the confidence level (90) while txt indicates that it is by default a tab-delimited file.

Save the file within the same directory or choose the desktop for easy find.

The file can then be opened within a spread-sheet software such as Microsoft Excel, Apple Numbers or Open Office Calc.

Hint: Within ArrayStar the fold-change (FC) columns shows “**up**” or “**down**” next to the fold change value. However, these are the result of a ratio calculation and in the exported file the “down” version represents 1/FC. For example a fold change “down” of 2 is shown as 0.5. To make sorting easier, simply create a new column labelled 1/FC to contain 1/FC from the FC column. A simple formula can be used e.g. =1/E2 – Then double click the little square at the bottom right of the cell and the formula will be spread all the way down the column till the last row.

E	F
Fold change	1/FC
0.431613886	=1/E2
0.606505885	
0.617536929	

E	F
Fold change	1/FC
0.431613886	2.316885607
0.606505885	1.648788617
0.617536929	1.619336355

Adding a new column in spreadsheet to better reflect down regulated fold change

The rest is just “spreadsheet expertise.”

GO: Gene Ontology

In the previous section we created a gene table that contained information about gene ontology. Here are a few words about the nature of GO:

Blake JA (2013) **Ten Quick Tips for Using the Gene Ontology**. *PLoS Comput Biol* **9(11)**: e1003343. doi:10.1371/journal.pcbi.1003343

The GO annotation stream focuses on the capture of the knowledge about the functional activities of specific proteins, the larger biological role or process (such as photosynthesis) as part of which these specific functions collectively act, and the cellular locality where all this occurs. The terms defined for these components are represented by the “**Molecular Function**,” “**Biological Process**,” and “**Cellular Component**” domains of the overall Gene Ontology. Protein annotation associations with these terms capture our best knowledge about protein actions, and annotations capture different granularity of knowledge based on the experimental data available.

A few caveats (based on the GO documentation.²)

The following areas are outside the scope of GO, and terms in these domains will not appear in the ontologies:

- Gene products: e.g. cytochrome c is not in the ontologies, but attributes of cytochrome c, such as oxidoreductase activity, are.
- Processes, functions or components that are unique to mutants or diseases: e.g. **oncogenesis** is not a valid GO term, as "causing cancer" is the result of reprogrammed, not normal cells and thus it is not the normal function of a gene.

² geneontology.org/page/documentation

1. GO Enrichment Analysis

From the GO documentation³:

One of the main uses of the GO is to perform enrichment analysis on gene sets. For example, given a set of genes that are up-regulated under certain conditions, an enrichment analysis will find which GO terms are over-represented (or under-represented) using annotations for that gene set.

ArrayStar has a GO analysis function that we'll explore these functions with the 138 genes set saved in the previous section. However, if you are following directly from the previous section you may also use the 144 genes that are currently selected. The example images below will be for the 138 genes set.

1.1. Retrieve the 138 gene set:

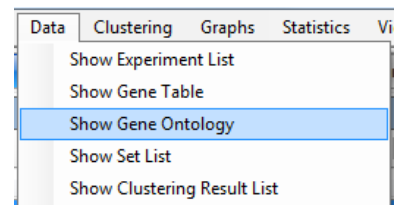
✓ TASK:

- Click on the “Set List” tab
- Click on the previously saved list “138 Up-Regulated Genes”
- Within the window, under **Actions** click “Select and show the table of this set’s Genes.”
- The view will **automatically** be **changed** to that of the “Gene Table” tab, displaying all annotation options selected previously.

1.2. Display the GO results

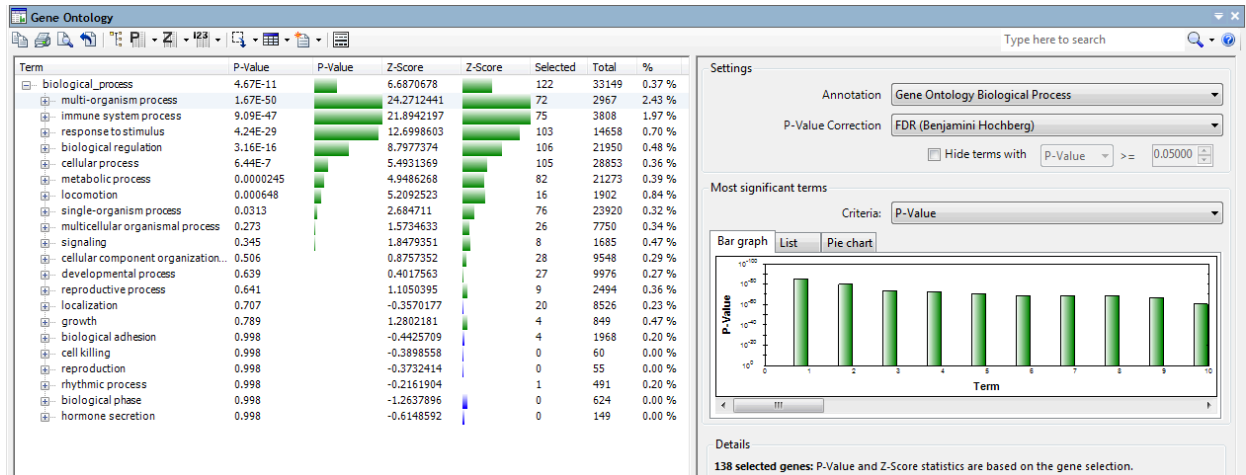
✓ TASK: While all genes are selected in the table use the menu cascade:

Data > Show Gene Ontology



A new “Gene Ontology” tab will be created to display the results for “biological process” data.

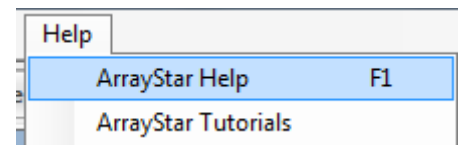
³ geneontology.org/page/go-enrichment-analysis



1.3. ArrayStar Help

Many details can be found within the Help menu of the software. Use the menu cascade:

Help > ArrayStar Help



If you are currently viewing the Gene Ontology tab, Help will open directly with that subject.

Some info glanced from the Help: Column names:

The columns in the Gene Ontology view are defined as follows:

- **Term** – The Gene Ontology or classification term.
- **P-value** – A P-value, representing the probability that the number of genes currently selected for the term occurred by chance. In general, the lower the P-value, the greater the likelihood that the term is significant.
- **P-value graph** - A graphical representation of the P-value. The number graphed is a logarithmic calculation $[-10 \times \log_{10}(P)]$ of how closely the P-value approximates “zero.” A P-value of 1.0 is graphed as “zero,” a P-value of 0.1 is graphed as “one,” etc. If all P-values are equal to 1.0, the graph will therefore appear blank.
- **Z-score** – The number of standard deviations the term is away from the expected number of genes selected. This is useful in determining if the term is over- or under-represented in the current gene selection.

Note: A positive Z-score indicates a positive correlation, while a negative number denotes a negative correlation. Since both types of correlation are equally important, any filtering or sorting uses the absolute value of the Z-score and disregards whether values are positive or negative.

- **Z-score graph** – A graphical representation of the Z-score.
- **Selected** – The number of genes in the current selection for the term.
- **Total** – The number of genes contained within the entire project for the term.
- **%** - The percentage of genes for each term that are currently selected.

2. Biological process

The list can be expanded by clicking on the [+] signs by each category to expand it further.

✓ **TASK:** expand categories and subcategories within the “multi-organism process” and further to reflect the image below:

Term	P-Value	P-Value	Z-Score	Z-Score	Selected	Total
biological_process	4.67E-11		6.6870678		122	33149
multi-organism process	1.67E-50		24.2712441		72	2967
response to other organism	9.71E-68		39.2702996		61	903
response to virus	1.33E-80		50.9150254		61	550
defense response to virus	6.25E-86		63.0978066		56	306
detection of virus	0.0000263		22.4664483		3	7
cellular response to virus	0.0000395		14.5363399		4	29
cytoplasmic pattern reco...	3.21E-6		20.3912296		4	15
cellular process regulatin...	0.998		-0.188239		0	14
defense response to other organ...	4.77E-71		46.2520605		56	559
response to bacterium	0.000618		6.8520396		7	325
response to host	0.00134		11.0356991		3	28
response to defenses of other or...	0.00134		11.0356991		3	28
detection of other organism	0.998		-0.3019147		0	36

Note the high Z-score under the “defense response to virus” and “detection of virus” and their associated p-values with 56 and 3 genes respectively (see in the column named “Selected.”)

Since the data is from a rhinovirus A16 infection it is clear the GO information reflects the experiment well in this section.

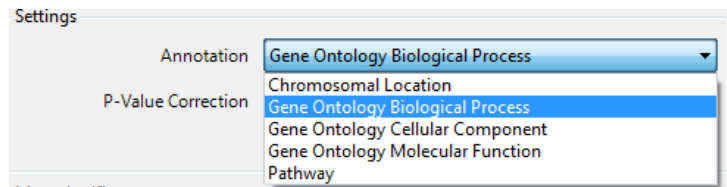
Expand a section further down the list named “**immune system process**” that also reflect virus response:

immune system process	9.09E-47	21.8942197	75	3808
immune effector process	6.24E-61	35.8474984	58	971
defense response to virus	6.25E-86	63.0978066	56	306
cell activation involved in immun...	0.594	2.192756	2	191

✓ **TASK:** on your own explore the other GO annotations



By default the data will be for the “biological process” data, but this can be changed on the right-hand side with the “Annotation” pull-down menu.



Comparing three or more groups

We have so far used only 2 of the 4 groups in the data to simplify the process.

While it may be more “precise” to compare relevant groups in a pairwise fashion, it may also be desirable to compare three or more groups.

Within ArrayStar this can be accomplished with the F-test ANOVA statistical calculation.

Keeping with the 4 groups in the experimental data explored so far the following steps detail how to obtain a list of genes for all 4 groups. Most of the steps after that selection would be identical, or similar to those seen in the previous section: scatter plot, clustering etc.

1. Identifying genes of interest



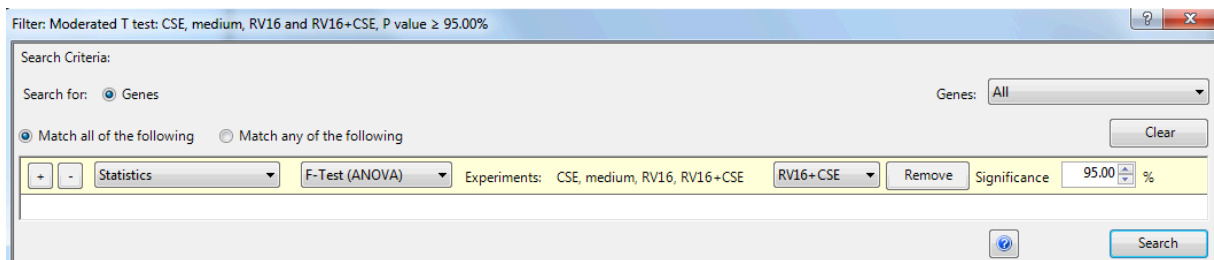
TASK: Go to the “Experiment List” tab.

- Follow the menu cascades:
 - **Edit > Select All**
 - **Filter > Filter All**


In the filter window: organize a single filter with the following options within the pull-down menus

- **Statistics: F-Test (ANOVA)**
- Experiments: **all 4** experiments should be listed: CSE, medium, RV16, RV16+CSE. [Note: any unwanted experiment could be removed with the “Remove” button.]
- Significance: **95%**

Click Search



This will find 91 genes

Click on the button  on the middle task bar in the filter window to “**Select and Show Results in Gene Table**”

The genes are shown within the “Gene Table” tab accompanied with annotation options previously selected for columns.

Note: data for all experiments could be inserted by clicking the “123” button.

Note: genes are automatically selected within all other views (e.g. Scatter plot, previously calculated heat maps etc.)

The next steps are quasi identical to the 2-groups versions.

2. Clustering:

2.1. Heat map

While all 91 genes are selected, follow the menu cascade:

Clustering > Advanced Clustering

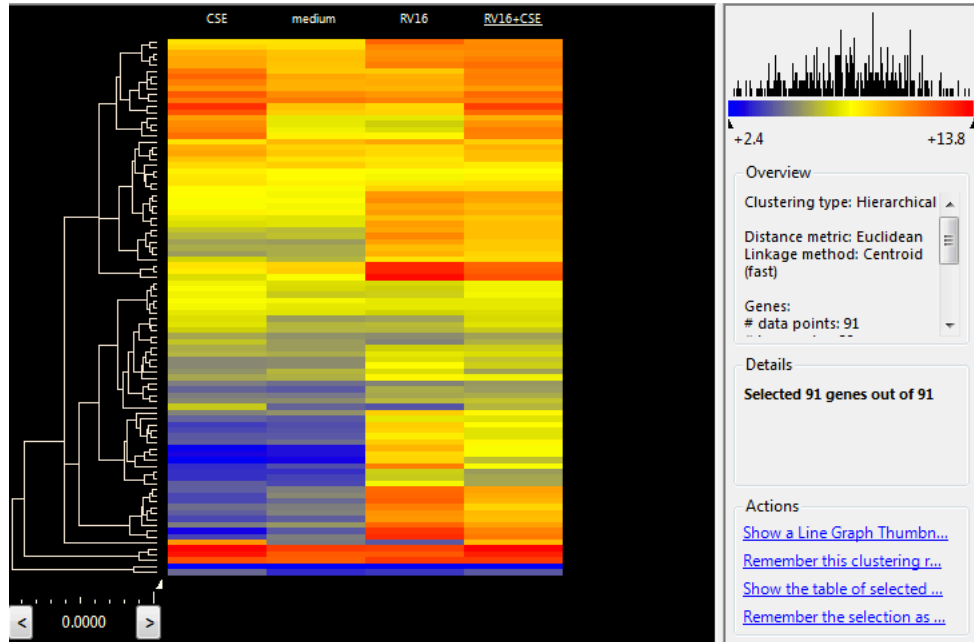
And choose the following options:

- Clustering Type: **Hierarchical** (*default*)
- Graph Type: **Heat Map** (*default*)
- Distance Metric: **Euclidian** (*default*)
- Cluster Experiments: **Centroid** (fast) (*default*)

On the right hand side: Clustering Data

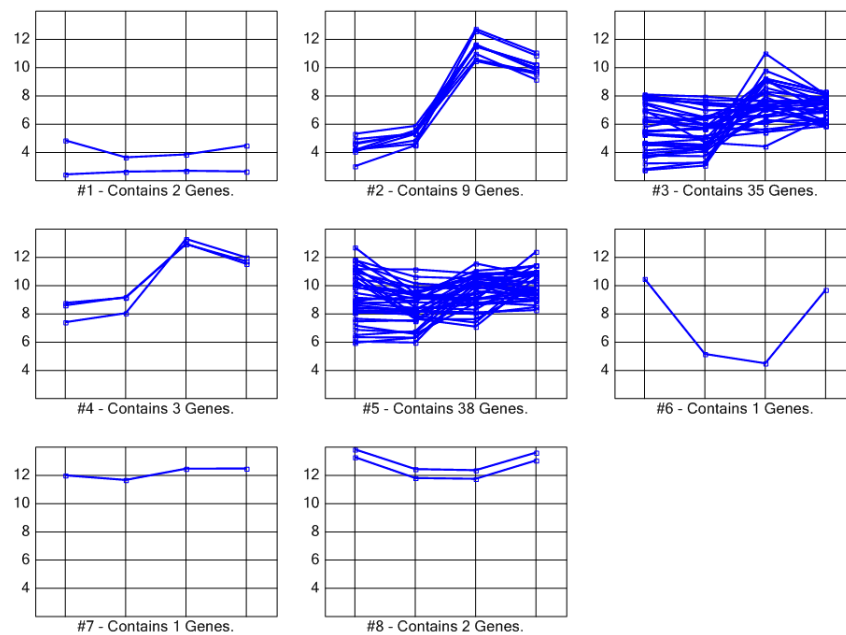
- Genes: **Selected** (*default*)
- Experiments: **Click on all 4**

Click Cluster



2.2. Line Graphs

Using the same menu cascade **Clustering > Advanced Clustering** simply change “Heat Map” for “Line Graph Thumbnails” and **Click Cluster**

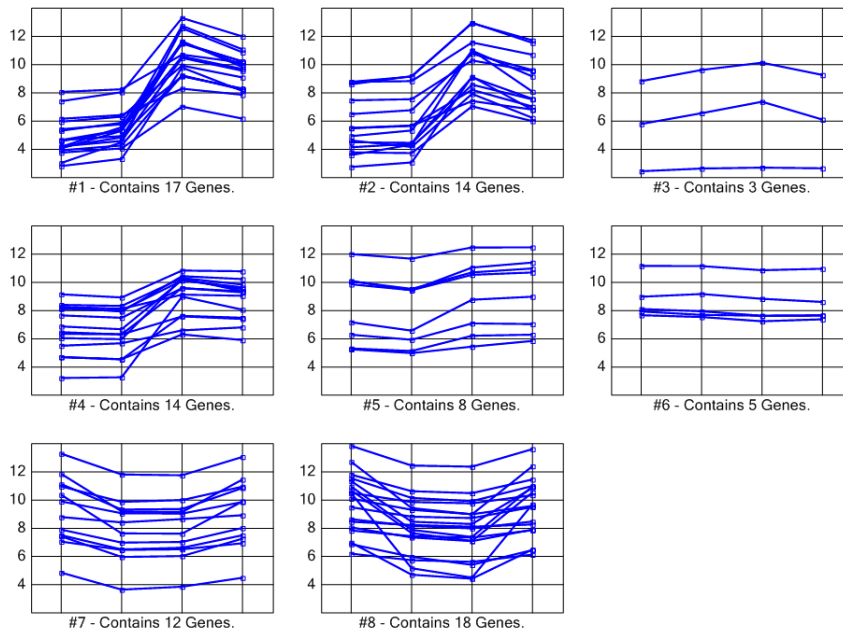


2.3. K-Means

Using the same menu cascade **Clustering > Advanced Clustering** change the following:

- Clustering Type: **k-Means**
- Graph Type: **Line Graph Thumbnails**
- Distance Metric: **Standard Pearson (default)**

Click Cluster



3. Export Gene List

 **TASK:** 

Click the export button and give a name to the file, e.g. 91-Anova-95.txt

Note: any additional annotation column should be added before export.

Appendix

1. Online Tutorials

www.dnastar.com/t-support-videos.aspx

In Fall 2014 the following videos for “**Gene Expression Analysis**” are available:

- DNASTAR - Clustering Your Data
- DNASTAR - Gene Ontology Workflow
- DNASTAR - Analyzing Gene Expression Patterns
- DNASTAR - Identifying Co-Regulated Genes
- DNASTAR - Clustering Gene Expression Data
- DNASTAR - Creating Gene Sets in ArrayStar
- DNASTAR - Using the Gene Table in ArrayStar
- DNASTAR - Using the Gene Ontology View
- DNASTAR - Using the Scatter Plot
- DNASTAR - Using Venn Diagrams in ArrayStar
- DNASTAR - Importing Microarray Data and Creating Replicate Sets
- DNASTAR - Importing Annotations in ArrayStar
- DNASTAR - Importing Custom Annotation and Ontology Data
- DNASTAR - Combining RNA-Seq and Microarray Data