Step by Step Guide to Importing Expression Data into JMP Genomics

USING JMP GENOMICS 8

Introduction

Data for expression analyses can exist in a variety of formats. Before this data can be analyzed it must imported into one or more properly formatted SAS data sets. JMP Genomics contains numerous processes for importing and formatting data from a variety of sources. In this document, we will explore one of these processes.

Typically the analyses will be performed on tall data sets where mRNA/miRNA/metabolites/proteins are in rows and samples are in columns. Additional sample annotation or phenotype information will be needed as well and will be part of a second file called the Experimental Design File (EDF). A third optional file is the annotation file which is a description of the molecule of interest and various gene/metabolite/protein information along with an ID to link it to the expression data set.

JMP Genomics uses the variable/column/header name conventions of SAS. Typically variables/columns/headers should be no more than 32 characters and start with a letter. The following link provides more information:

http://www.jmp.com/support/downloads/life_sciences_documentation/wwhelp/wwhimpl/js/html/wwhelp.htm#href=getting_started.01.16.html

For more information about the different data sets used by JMP Genomics, go to this link:

http://www.jmp.com/support/downloads/life_sciences_documentation/wwhelp/wwhimpl/js/html/wwhelp.htm#href=getting_started.01.21.html#356832

Objectives

In this document, we will explore the following import processes:

- Creating an Experimental Design File (EDF)
- Importing data from text/Excel format files

Creating the Experimental Design File

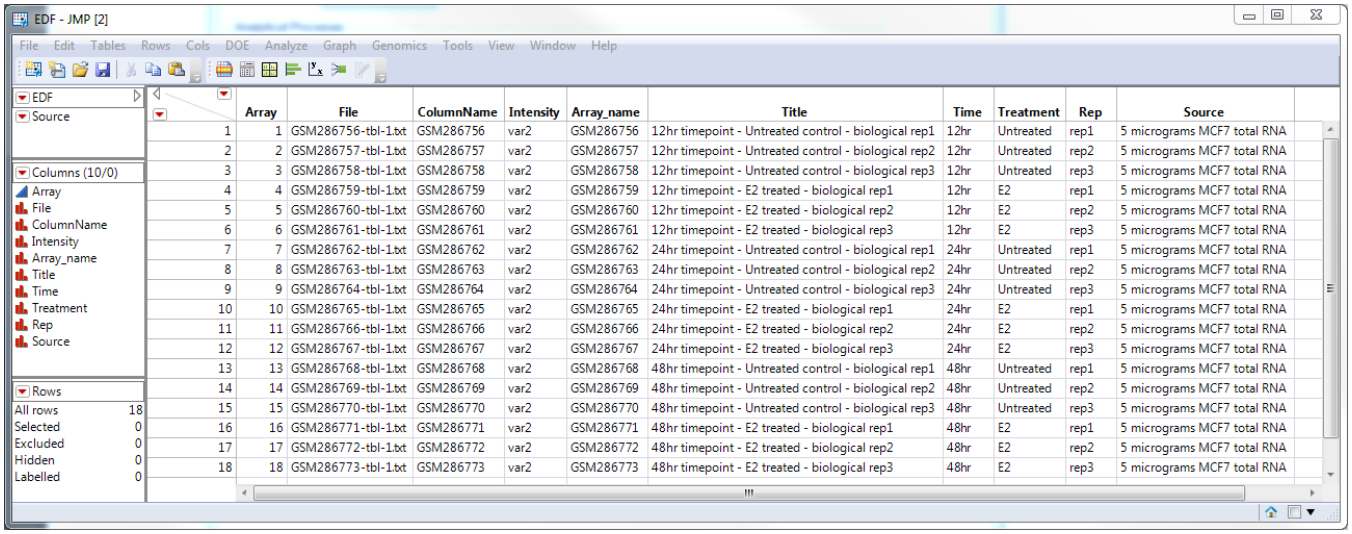
JMP Genomics uses experiment information as well as phenotype and sample information for analyses and graphing purposes. This information is typically found in a separate file where the samples are rows and each column represents the characteristics of that sample as well as the experimental conditions that were applied. In JMP Genomics, we call this the Experimental Design File. It can also be used to import a directory/folder of files that contain the expression data and will be covered in a later section.

Besides information about the samples, we also need a few more columns that have a variety of purposes.

1) The first main column is the **ColumnName**.
    - For expression data, the values of the ColumnName variable are used to name the columns of intensity data corresponding to each array or sample in the imported data set. This variable can, and often should, be different than the file name. Using a more descriptive name can often be helpful when performing analyses or examining sample-level quality control graphics.
    - **ColumnName** must be unique for each array/sample. If you want to use a descriptive variable as your ColumnName that is not unique, adding the **Array** variable to the end of the preferred **ColumnName** will create a unique identifier.
2) There is also an **Array** (or **Chip**) column that must have a unique number for each sample in the case of single channel/one-color/intensity/count based data.
3) A third column is the **File** (or **FileName**) column which is only needed if the EDF is going to be use to import a list of files from a folder. It should contain the file name and extension of the file that will be imported.
4) The last additional column is the **Intensity** column and is required to import a list of files from a folder. It typically will contain the column header/ label of the column that contains the intensity/expression values. If no column header/label is available, then the something must be put in to tell JMP Genomics which column

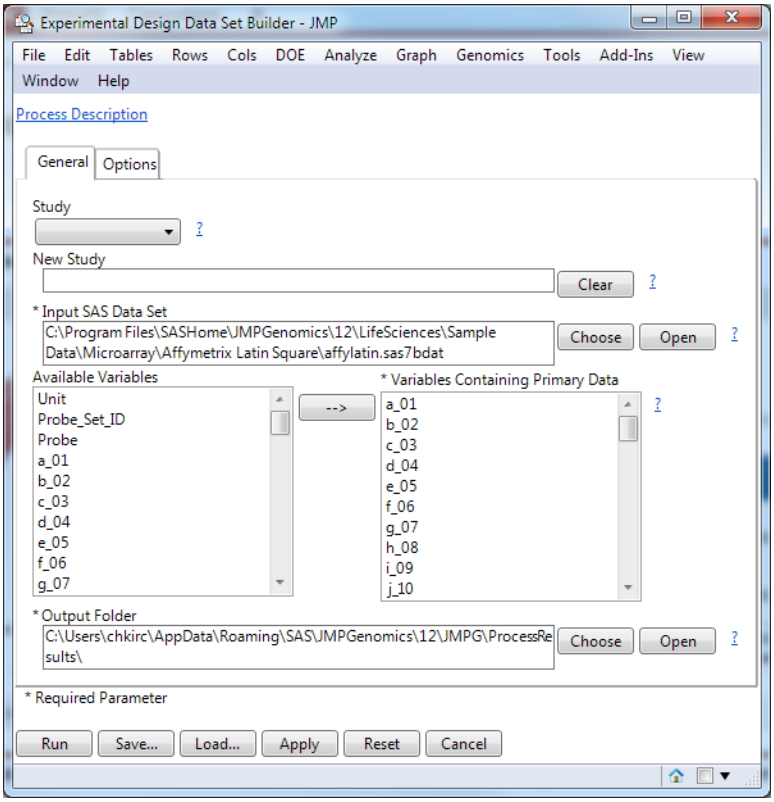
contains the expression data. The syntax is “VarX”, where X is the column number that contains the expression data. This text should be in each row for the intensity column.

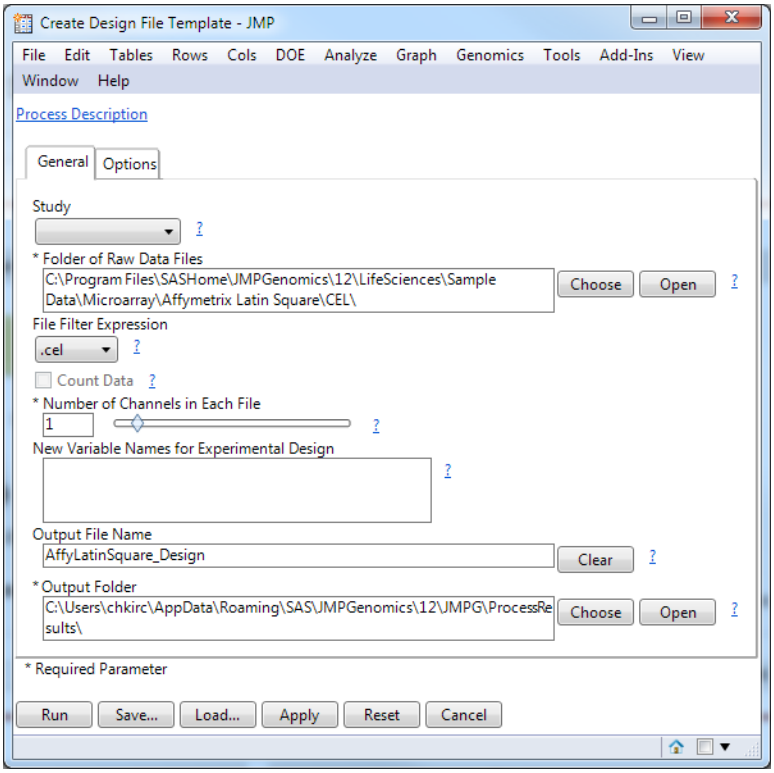
5) An Example of an EDF is shown below:



	Array	File	ColumnName	Intensity	Array_name	Title	Time	Treatment	Rep	Source
1	1	GSM286756-tbl-1.txt	GSM286756	var2	GSM286756	12hr timepoint - Untreated control - biological rep1	12hr	Untreated	rep1	5 micrograms MCF7 total RNA
2	2	GSM286757-tbl-1.txt	GSM286757	var2	GSM286757	12hr timepoint - Untreated control - biological rep2	12hr	Untreated	rep2	5 micrograms MCF7 total RNA
3	3	GSM286758-tbl-1.txt	GSM286758	var2	GSM286758	12hr timepoint - Untreated control - biological rep3	12hr	Untreated	rep3	5 micrograms MCF7 total RNA
4	4	GSM286759-tbl-1.txt	GSM286759	var2	GSM286759	12hr timepoint - E2 treated - biological rep1	12hr	E2	rep1	5 micrograms MCF7 total RNA
5	5	GSM286760-tbl-1.txt	GSM286760	var2	GSM286760	12hr timepoint - E2 treated - biological rep2	12hr	E2	rep2	5 micrograms MCF7 total RNA
6	6	GSM286761-tbl-1.txt	GSM286761	var2	GSM286761	12hr timepoint - E2 treated - biological rep3	12hr	E2	rep3	5 micrograms MCF7 total RNA
7	7	GSM286762-tbl-1.txt	GSM286762	var2	GSM286762	24hr timepoint - Untreated control - biological rep1	24hr	Untreated	rep1	5 micrograms MCF7 total RNA
8	8	GSM286763-tbl-1.txt	GSM286763	var2	GSM286763	24hr timepoint - Untreated control - biological rep2	24hr	Untreated	rep2	5 micrograms MCF7 total RNA
9	9	GSM286764-tbl-1.txt	GSM286764	var2	GSM286764	24hr timepoint - Untreated control - biological rep3	24hr	Untreated	rep3	5 micrograms MCF7 total RNA
10	10	GSM286765-tbl-1.txt	GSM286765	var2	GSM286765	24hr timepoint - E2 treated - biological rep1	24hr	E2	rep1	5 micrograms MCF7 total RNA
11	11	GSM286766-tbl-1.txt	GSM286766	var2	GSM286766	24hr timepoint - E2 treated - biological rep2	24hr	E2	rep2	5 micrograms MCF7 total RNA
12	12	GSM286767-tbl-1.txt	GSM286767	var2	GSM286767	24hr timepoint - E2 treated - biological rep3	24hr	E2	rep3	5 micrograms MCF7 total RNA
13	13	GSM286768-tbl-1.txt	GSM286768	var2	GSM286768	48hr timepoint - Untreated control - biological rep1	48hr	Untreated	rep1	5 micrograms MCF7 total RNA
14	14	GSM286769-tbl-1.txt	GSM286769	var2	GSM286769	48hr timepoint - Untreated control - biological rep2	48hr	Untreated	rep2	5 micrograms MCF7 total RNA
15	15	GSM286770-tbl-1.txt	GSM286770	var2	GSM286770	48hr timepoint - Untreated control - biological rep3	48hr	Untreated	rep3	5 micrograms MCF7 total RNA
16	16	GSM286771-tbl-1.txt	GSM286771	var2	GSM286771	48hr timepoint - E2 treated - biological rep1	48hr	E2	rep1	5 micrograms MCF7 total RNA
17	17	GSM286772-tbl-1.txt	GSM286772	var2	GSM286772	48hr timepoint - E2 treated - biological rep2	48hr	E2	rep2	5 micrograms MCF7 total RNA
18	18	GSM286773-tbl-1.txt	GSM286773	var2	GSM286773	48hr timepoint - E2 treated - biological rep3	48hr	E2	rep3	5 micrograms MCF7 total RNA

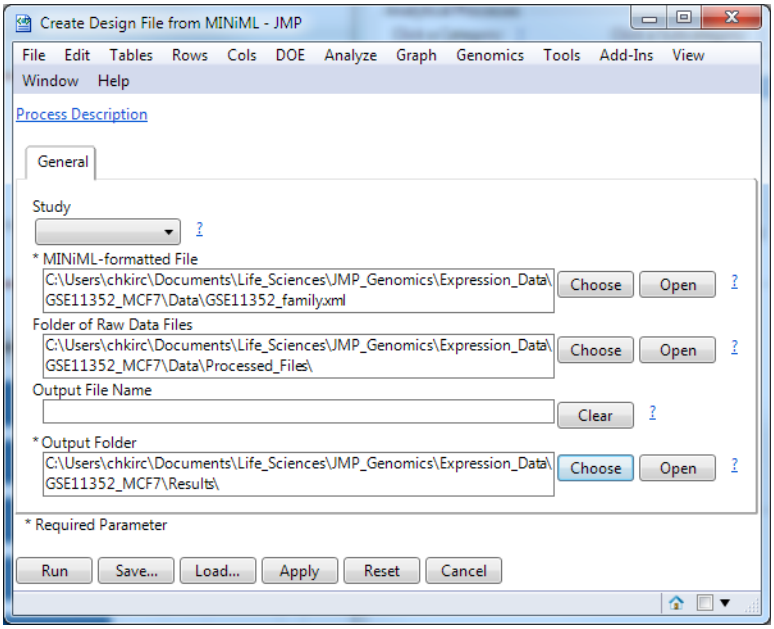
6) To create this file, we have a few options.
   a) If everything is in a single text file, you can open it up in JMP Genomics by using File>Open and add the 4 columns mentioned above. Then use File>Save As... to save as a SAS Data Set (*.sas7bdata) as the file type. This is the main way that many people create their EDFs
      i) You can use JMP or JMP Genomics tools to create the 4 additional columns. Within Import>Experimental Design File section of the Genomics Starter there are two useful tools called Create Array Index and Create ColumnName which help in creating these two columns.
   b) Use **Create Design Data Set from SAS Data Set** or **Create Design File Template**. Both options are available from the Genomics Starter from Import>Experimental Design File> and have examples that can be viewed by clicking on each button then clicking the on the Load button at the bottom of the resulting dialog that appears. See below examples of the dialogs.



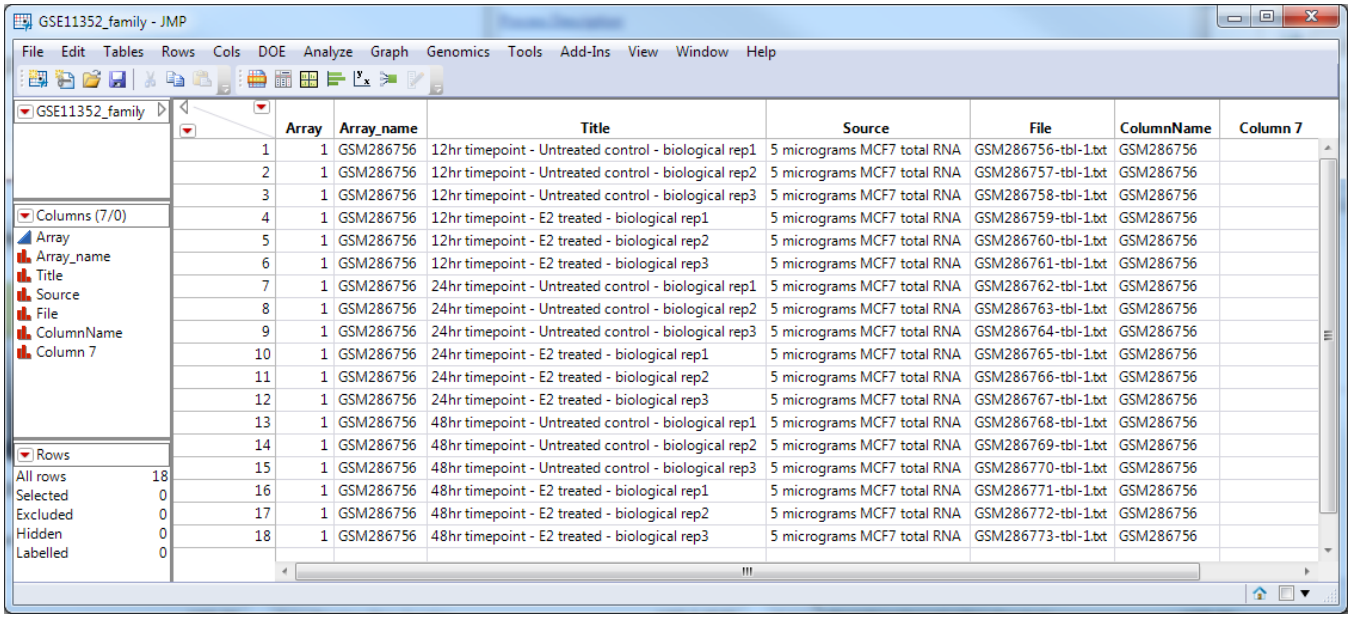


- The result of each will be a file with the following columns created. For **Create Design Data Set from SAS Data Set**, Array column and ColumnName column which will contain the names of the columns from the SAS data set. For **Create Design File Template,** Array, ColumnName, and File columns will be created. File column will contain the file names of the files within the **Folder of Raw Data Files field**.

c) If using a GEO based data set, you can download the MiNiML XML file from GEO and import it to create the EDF for that GEO data set. Be aware that not all MiNiML XML files are properly formatted or contain proper information. See example dialog below.



- The result is an EDF that looks something like this



	Array	Array_name	Title	Source	File	ColumnName	Column 7
1	1	GSM286756	12hr timepoint - Untreated control - biological rep1	5 micrograms MCF7 total RNA	GSM286756-tbl-1.txt	GSM286756	
2	1	GSM286756	12hr timepoint - Untreated control - biological rep2	5 micrograms MCF7 total RNA	GSM286757-tbl-1.txt	GSM286756	
3	1	GSM286756	12hr timepoint - Untreated control - biological rep3	5 micrograms MCF7 total RNA	GSM286758-tbl-1.txt	GSM286756	
4	1	GSM286756	12hr timepoint - E2 treated - biological rep1	5 micrograms MCF7 total RNA	GSM286759-tbl-1.txt	GSM286756	
5	1	GSM286756	12hr timepoint - E2 treated - biological rep2	5 micrograms MCF7 total RNA	GSM286760-tbl-1.txt	GSM286756	
6	1	GSM286756	12hr timepoint - E2 treated - biological rep3	5 micrograms MCF7 total RNA	GSM286761-tbl-1.txt	GSM286756	
7	1	GSM286756	24hr timepoint - Untreated control - biological rep1	5 micrograms MCF7 total RNA	GSM286762-tbl-1.txt	GSM286756	
8	1	GSM286756	24hr timepoint - Untreated control - biological rep2	5 micrograms MCF7 total RNA	GSM286763-tbl-1.txt	GSM286756	
9	1	GSM286756	24hr timepoint - Untreated control - biological rep3	5 micrograms MCF7 total RNA	GSM286764-tbl-1.txt	GSM286756	
10	1	GSM286756	24hr timepoint - E2 treated - biological rep1	5 micrograms MCF7 total RNA	GSM286765-tbl-1.txt	GSM286756	
11	1	GSM286756	24hr timepoint - E2 treated - biological rep2	5 micrograms MCF7 total RNA	GSM286766-tbl-1.txt	GSM286756	
12	1	GSM286756	24hr timepoint - E2 treated - biological rep3	5 micrograms MCF7 total RNA	GSM286767-tbl-1.txt	GSM286756	
13	1	GSM286756	48hr timepoint - Untreated control - biological rep1	5 micrograms MCF7 total RNA	GSM286768-tbl-1.txt	GSM286756	
14	1	GSM286756	48hr timepoint - Untreated control - biological rep2	5 micrograms MCF7 total RNA	GSM286769-tbl-1.txt	GSM286756	
15	1	GSM286756	48hr timepoint - Untreated control - biological rep3	5 micrograms MCF7 total RNA	GSM286770-tbl-1.txt	GSM286756	
16	1	GSM286756	48hr timepoint - E2 treated - biological rep1	5 micrograms MCF7 total RNA	GSM286771-tbl-1.txt	GSM286756	
17	1	GSM286756	48hr timepoint - E2 treated - biological rep2	5 micrograms MCF7 total RNA	GSM286772-tbl-1.txt	GSM286756	
18	1	GSM286756	48hr timepoint - E2 treated - biological rep3	5 micrograms MCF7 total RNA	GSM286773-tbl-1.txt	GSM286756	

- Notice that Array_name and ColumnName contents are in error due to an improperly formatted MiNiML file. These can easly be corrected by using JMP/JMP Genomics tools for parsing out the File column using – at a delimiter and the GSM numbers can be used for the ColumnName (deleting the old column first). Array_name is not a needed column so it can be deleted.

7) More information about creating Experimental Design Files can be found at the link below:

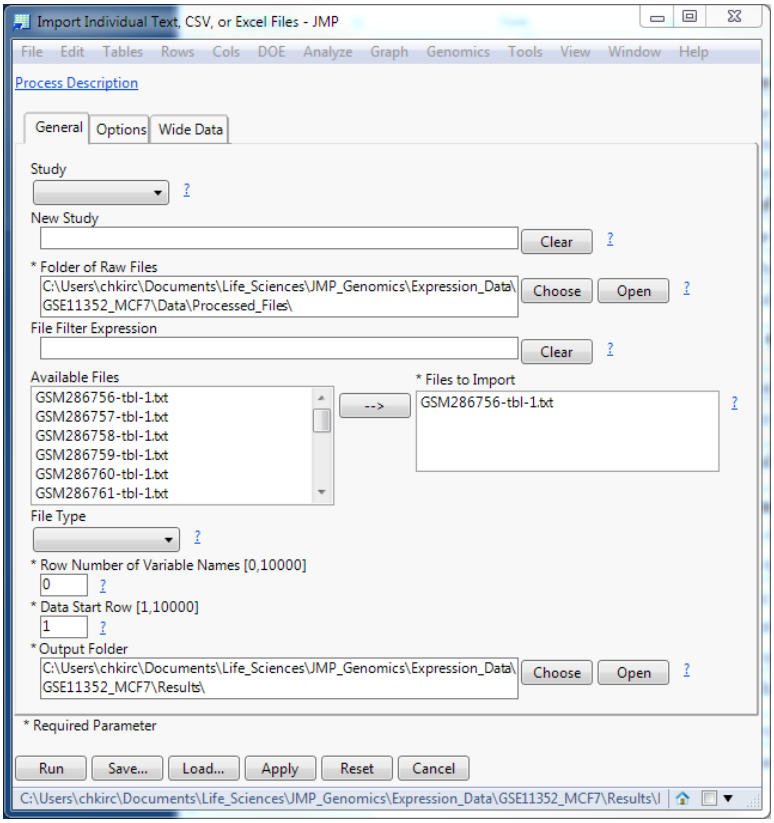
http://www.jmp.com/support/downloads/life_sciences_documentation/wwhelp/wwhimpl/js/html/wwhelp.htm#href=ST_G_IM_0003.html

Importing Text Format Expression Data

One file at a time

The JMP Genomics text file format import engine can be used to import large expression data sets. For small data sets (fewer than 5,000 genes or so), the JMP text file import (**File > Open)** may be the best option, followed by saving the JMP table as a sas7bdat file. In this section, we will cover importing large data sets.

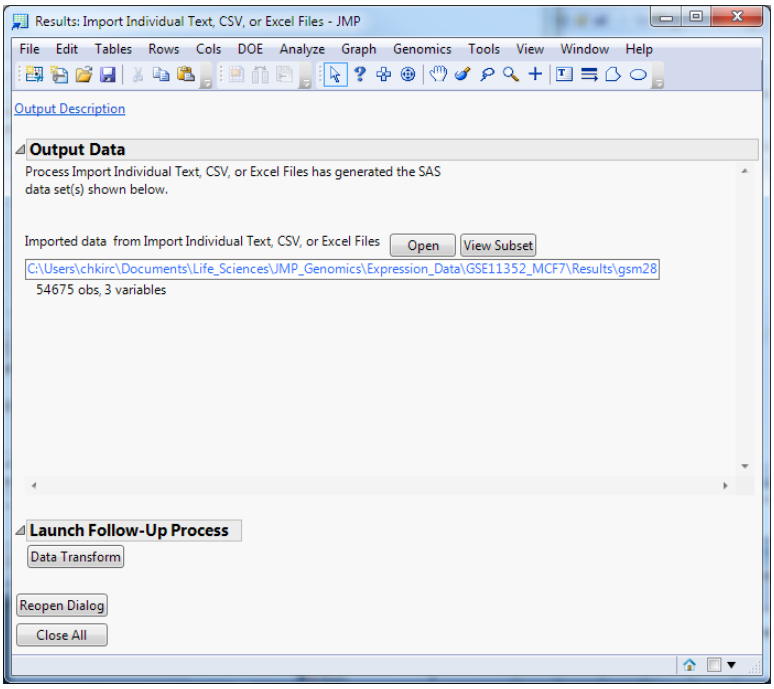
1) Select Import > Text > Import Individual Text, CSV, or Excel Files from the Genomics starter menu.
    - We recommend that you import text or csv formatted files, as Excel formats often have formatting that is incompatible with the import process.
2) Complete the **General** tab as shown below. This file has all samples listed as rows and genotypes as columns. The location of the sample data is listed in the "Folder of Raw Files" field.



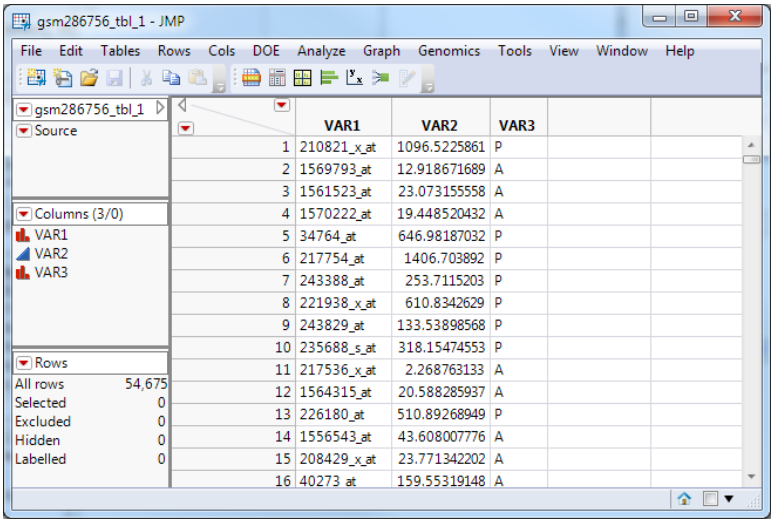
- The **File Filter Expression** box is optional.
- Multiple files can be imported, but all the files must be in the same format.
- **File Type** does not need to be specified if the file suffix is .txt or .csv. If the file has a different suffix but is in one of these formats, specify the type with this pull-down menu.
- The **Row Number of Variable Names** should have the appropriate value entered. If 0 is used, then each column imported will have column names of VAR1, VAR2, etc. If 1 is used, then the first row is used as column headers.
- The **Data Start Row** should have the appropriate value entered.
- **Output Folder** is where you would like to save the resulting SAS data set.
- Each file selected will be imported and saved separately in the Output Folder.

3) Select the **Options** tab.
    - Here you can set the number of rows to scan,
    - Make no changes to the default settings on the **Options** tab.

4) Select the **Wide Data** tab.
   - Complete the **Minimum Number of Columns to Scan** box. Set this value to the column number in the data set past which the type of data (categorical vs. numeric) is constant from that column to the last column of the data table. In this case, you can leave it to its default of 0.
   - When JMP Genomics imports a text or csv file, the software checks each column for the data type. Setting the minimum number of columns to scan will turn off this process for all columns beyond the selected value and significantly improve the import speed.
   - All other fields can be left as the default.
5) Click **Run** to start the process.
6) The result is a window with a link to the created SAS data set:



7) The resulting SAS data set will look something like this (This example shows a data set that does not have column headers in the original file):



	VAR1	VAR2	VAR3
1	210821_x_at	1096.5225861	P
2	1569793_at	12.918671689	A
3	1561523_at	23.073155558	A
4	1570222_at	19.448520432	A
5	34764_at	646.98187032	P
6	217754_at	1406.703892	P
7	243388_at	253.7115203	P
8	221938_x_at	610.8342629	P
9	243829_at	133.53898568	P
10	235688_s_at	318.15474553	P
11	217536_x_at	2.268763133	A
12	1564315_at	20.588285937	A
13	226180_at	510.89268949	P
14	1556543_at	43.608007776	A
15	208429_x_at	23.771342202	A
16	40273_at	159.55319148	A

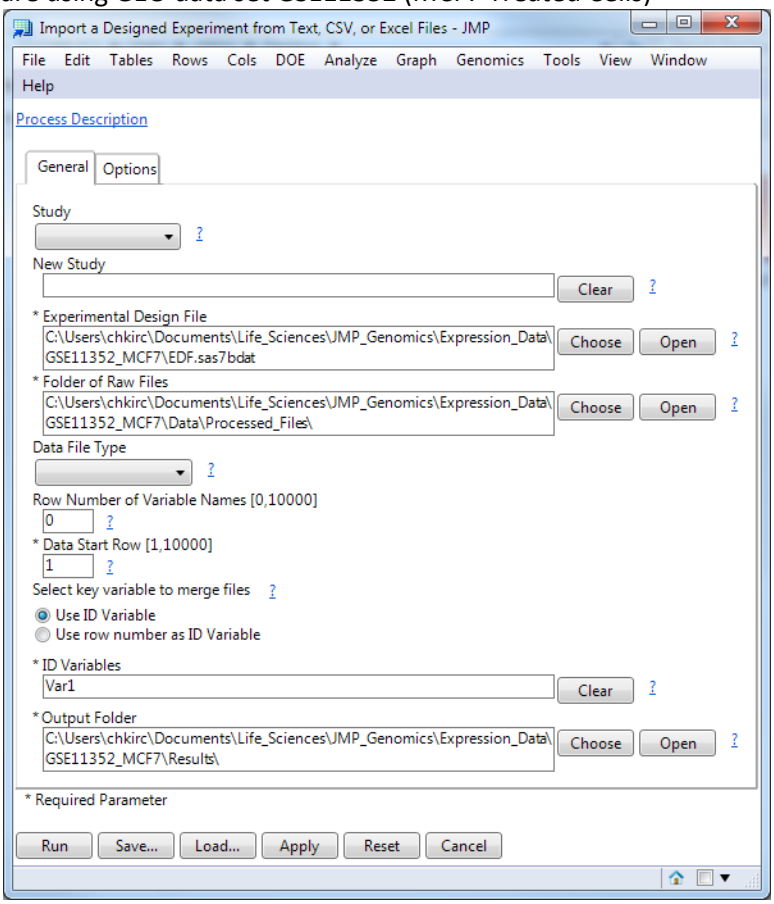
8) Information can be found at this link regarding importing text files:

http://www.jmp.com/support/downloads/life_sciences_documentation/wwhelp/wwhimpl/js/html/wwhelp.htm#href=PR_G_IM_0028.html

Importing multiple files to be combined in one data set

Sometimes the expression profiles of the samples will be in individual files. It will be necessary, then, to import all of them and combine them into a single file. Below are the steps to import multiple expression profiles at the same time and automatically combine them. An EDF file the file names is required for this process.

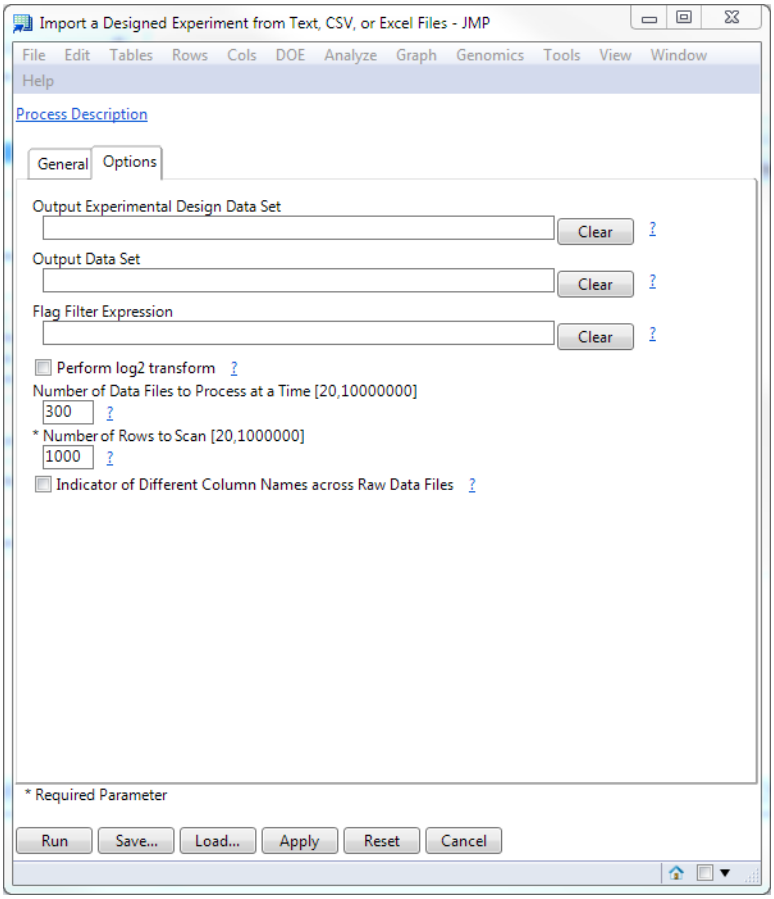
1) Select Import > Text > Import a Designed Experiment from Text, CSV, or Excel Files from the Genomics starter menu.
    - We recommend that you import text or csv formatted files, as Excel formats often have formatting that is incompatible with the import process.
2) Complete the **General** tab as shown below. This file has all samples listed as columns and transcripts/proteins/metabolites as rows. The location of the data is listed in the "Folder of Raw Files" field. In this example, we are using GEO data set GSE11352 (MCF7 Treated Cells)



- **Data File Type** does not need to be specified if the file suffix is .txt or .csv. If the file has a different suffix but is in one of these formats, specify the type with this pull-down menu.
- The **Row Number of Variable Names** should have the appropriate value entered. If 0 is used, then each column imported will be referred to by the column names of Var1, Var2, etc. which refer the order of appearance of the columns (e.g. column 1 is referred to as Var1, column 2 is referred to as Var2 and so on). If 1 is used, then the first row is used as column headers and the column header value must be specified in the **ID Variables section**. In this example a column header does not exist so 0 is used and we will have to specify the first column as containing the IDs (Var1).
- The **Data Start Row** should have the appropriate value entered.

- For the **Select key variable to merge files** option, you can use row number as the ID Variable, but it assumes that all rows in each file are sorted the same, have the same representative transcript/protein/metabolite, and the same number of rows. Typically each file has an ID column and it is the preferred and more robust way to merger data sets into a single file since it can be used as a key for each row that needs to be imported from the different files. In this case, use the use ID Variable option (default) and specify the ID column in the next field. We will keep the default selection for this example.
- **ID Variables** field is where you specify the column header name or the column number (using the syntax VarX, where X is the column number that contains the IDs). In this example, we will enter Var1 to specify that the first column contains the IDs we will use to merge the files together.
- **Output Folder** is where you would like to save the resulting SAS data set.
- Multiple files can be imported, but all the files must be in the same format.

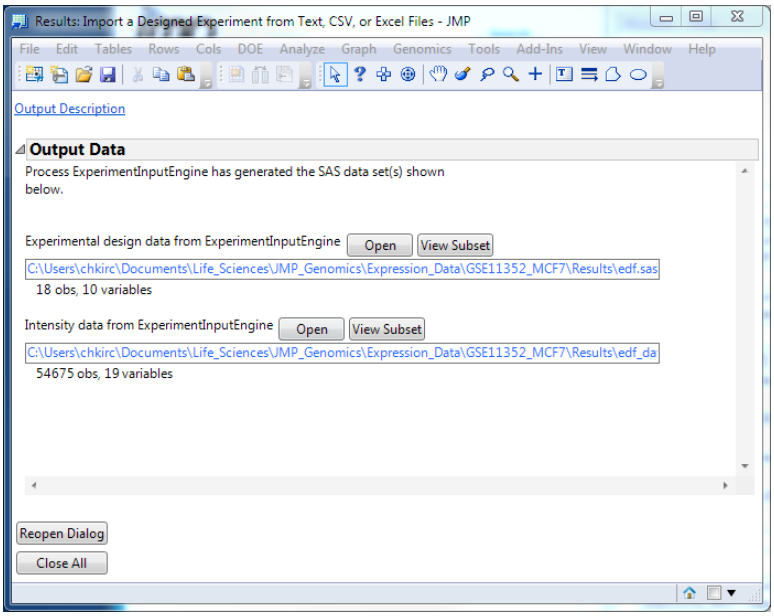
3) Select the **Options** tab as seen below:



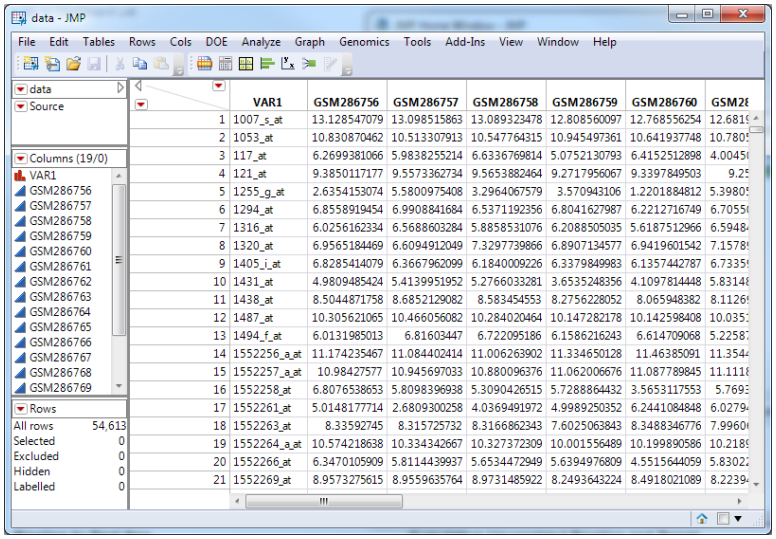
- Here you can set the name of the EDF file that will be generated as well as the name of the new data set that will be created once all of the files have been merged together in the first two field. If nothing is specified, the default names of the SAS data sets will be edf.sas7bdat and edf_data.sas7bdat, respectively.
- There is also a **Flag Filter Expression** option
- One can log2 transform the data at this step if the data has not already been transform and/or normalized already by checking the **Perform log2 transform** option.
- **Number of Data files to Process at a Time** option lets you specify how many files you would like to merge at one time. The larger the number, the better the performance, but more temporary disk space is needed.

- **Number of Rows to Scan** option lets you specify the number of rows from the data file to be scanned to determine the attributes of variables (columns) in the SAS data set
- **Indicator of Different Column Names across Raw Data Files** option. Check here if the column names or their order are different across raw data files.
- Make no changes to the default settings on the **Options** tab.

4) Click on **Run** to begin the import process.
5) The result is a window with a link to the created SAS data set:



6) An example of the final data set is below:



7) Information can be found at this link regarding importing text files based on a designed experiment:

http://www.jmp.com/support/downloads/life_sciences_documentation/wwhelp/wwhimpl/js/html/wwhelp.htm#href=PR_G_IM_0029.html

This concludes the import of expression data. Following import, you should next refer to the **Basic Expression Analysis and Normalization** document which reviews data analysis, sample quality assessment and normalization methods.