

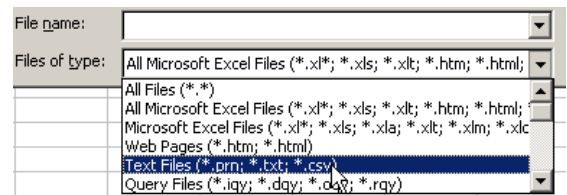
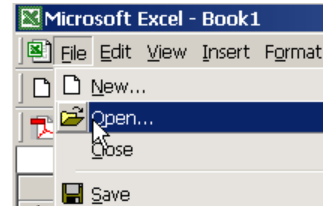
Microarray Excel Hands-on Workshop Handout

Piali Mukherjee (pim2001@med.cornell.edu; <http://icb.med.cornell.edu/>)

Importing Data

Excel allows you to import data in tab, comma or space delimited text formats. Most microarray data is generated in one of these formats.

1. From the menu select **File** and then **Open** (or click on the yellow folder icon in the menubar)
2. Browse to the location of the file using the **Look in** drop down menu and then select **Text Files (*.prn; *.txt; *.csv)** from the **Files of Type** drop down menu.
3. Select the desired text file and click **Open**.
4. A **Text Import Wizard** will then guide you through the steps for the specific file type.
5. For example for **tab-delimited** text files select the **Delimited** radio button and click **Next**
6. Then select the **Tab** radio button and click **Next**
7. You can then select advanced formats for your columns
8. Click **Finish**

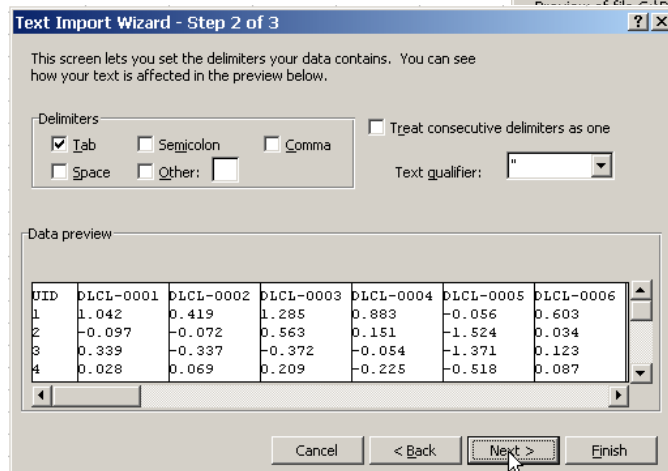
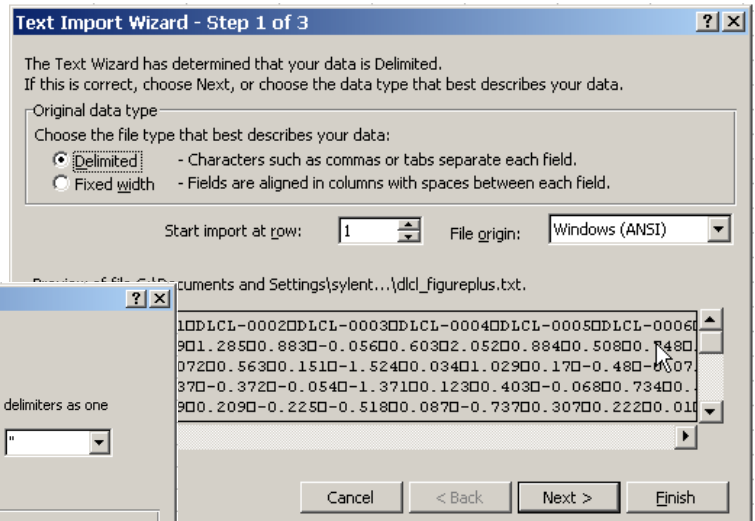


Tip:

If you want to insert a tab-delimited dataset into an already open excel workbook. Select

Data → Get External Data → Import Text File

from the menu bar and then follow *steps 3-8*.



Tip:

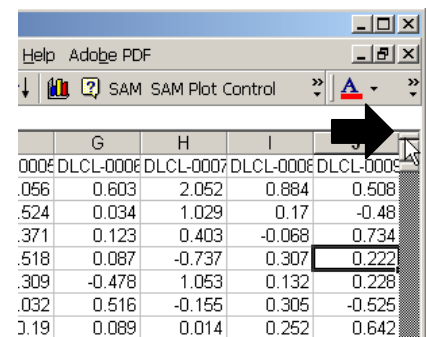
Click on the gray area with a letter above a column to select the entire column or Click on the gray row number to select an entire row.

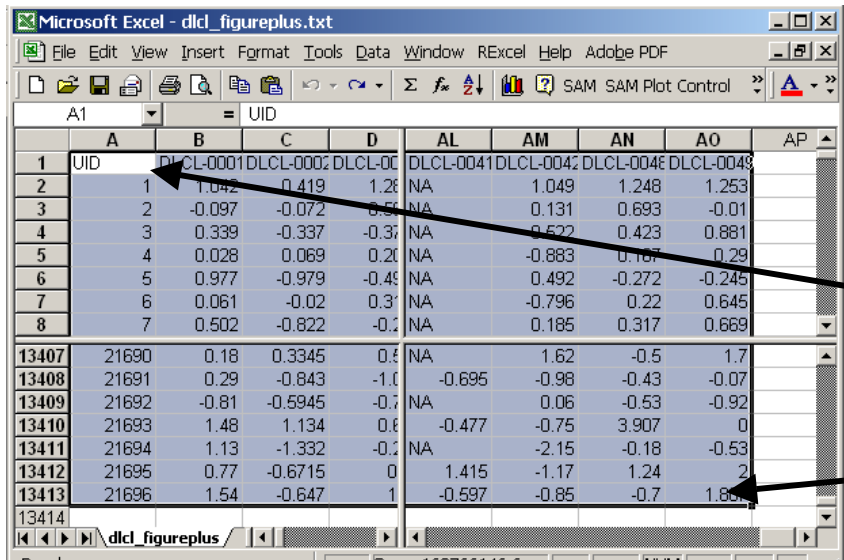
Clicking on the control '**Ctrl**' key while selecting allows you to select discontinuous cells or blocks of cells while clicking on the **Shift** key allows you to select continuous blocks of cells.

Selecting Data

You can select data by either dragging the cursor across the data you want to select or by clicking on one corner of the selection area and then pressing the **Shift** key while clicking on the diagonally opposite corner cell of the selection. To select large datasets it might be easier to use the screen dividers:

1. Click and drag the dividers on the right hand top or bottom corners (see arrow) of the spreadsheet to the desired position in the dataset.
2. Using the **Shift** key you can now select the top left hand corner and diagonally opposite (right bottom) corner of the dataset you want to select (see figure below).



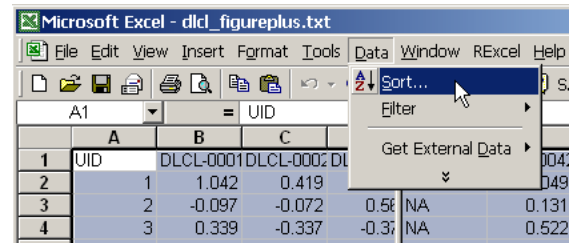


Click here first

Then press the Shift key and click here

Sorting Data

First select the data you would like to sort and then go to **Data** → **Sort** in the menu bar. A Sort menu should pop up allowing you to sort by the desired column.



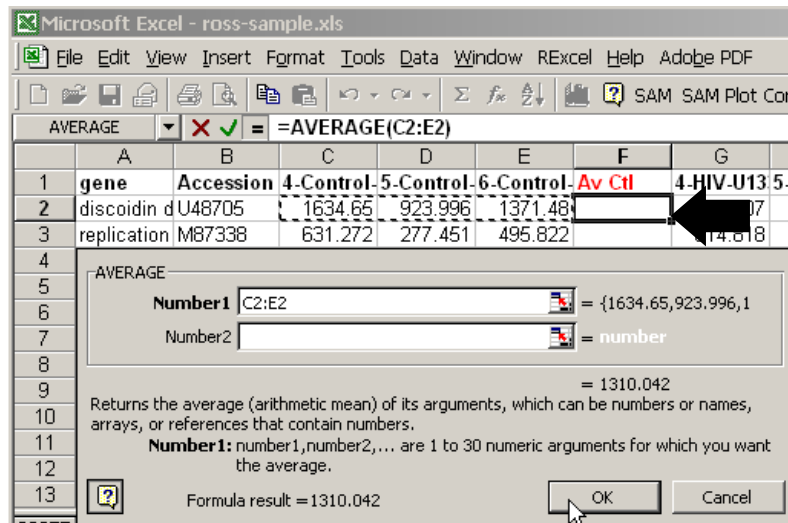
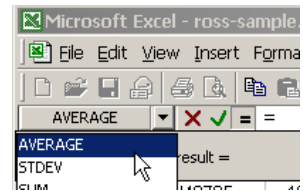
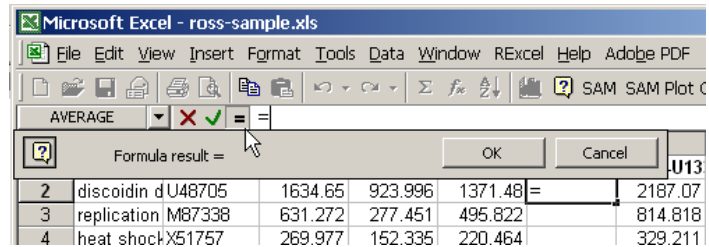
Filtering Data

Similarly to filter data go to **Data** → **Filter** → **AutoFilter** or **Advanced Filter** in the menu bar.

Calculation

For basic calculations (sum, averages, logs, ratios, standard deviations etc.):

1. Click the cell where you want your result to appear and then click on the “=” (equal to sign) in the menu.
2. Next from the function drop down menu pick the function appropriate for your calculation (SUM, AVERAGE, LOG etc.) and then follow the instructions for the specific calculation.
3. For example, to average the numbers in cells C2 – D2 and display the result in cell F2 you can select the cells C2-D2 (by dragging your cursor or by using the Shift key) or you can write **C2:D2** in the space provided and click **OK**.



Tip:

To extend the same calculation to several cells you can click and drag the square from the corner of the formula-containing cell to the adjacent cells where you want the same formula to be applied OR

you can cut a calculation from one cell and paste it to the whole column.

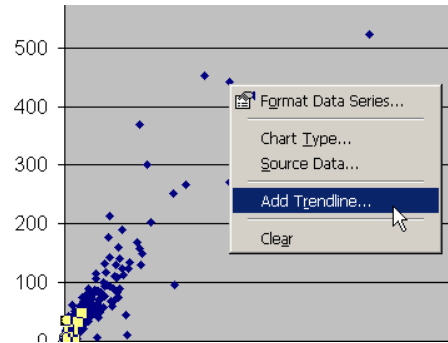
Most statistical analysis like T-Tests, Geometric Means, correlation etc. can be accessed as described above. To access more analysis options (ANOVA, Histogram etc.) you can also go to **Tools** → **Data Analysis**. (This requires installation of an Analysis Pack that is available on your Microsoft Office CD)

Graphs

To draw a basic scatter plot, select the data you want to graph and click on the graph icon in the menubar or go to **Insert → Chart** then pick the kind of graph you want to plot (for example XY Scatter plot) and click **Next**. Then you can select details for your chart such as axes labels, legends etc.

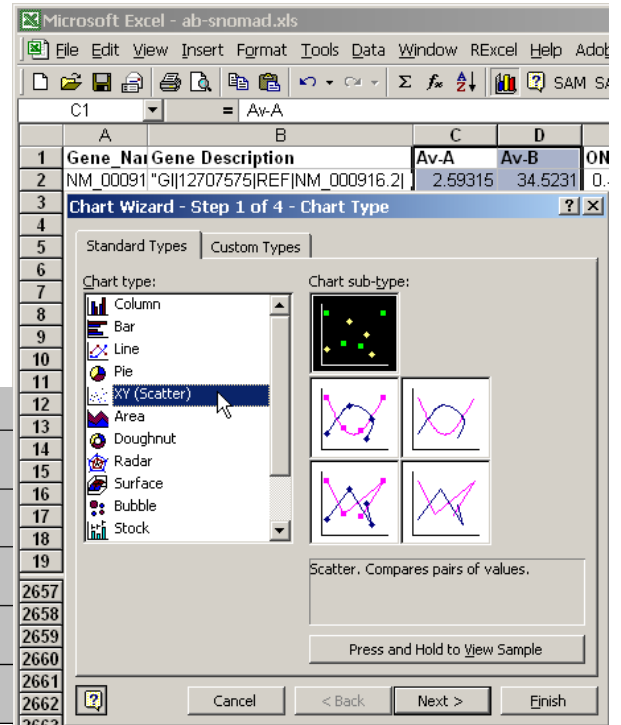
To plot a linear regression line in your scatter plot, right click a data point on the graph and select **Add Trendline** from the resulting pull down menu. To display the R squared

(regression coefficient) value on the chart click the appropriate radio button in the **Options** tab.



Tip

Right-clicking on different areas of the chart allows you to customize the look of your chart.



SAM (Significance Analysis of Microarrays)

Available at: <http://www-stat.stanford.edu/~tibs/SAM/> (free for academic users)

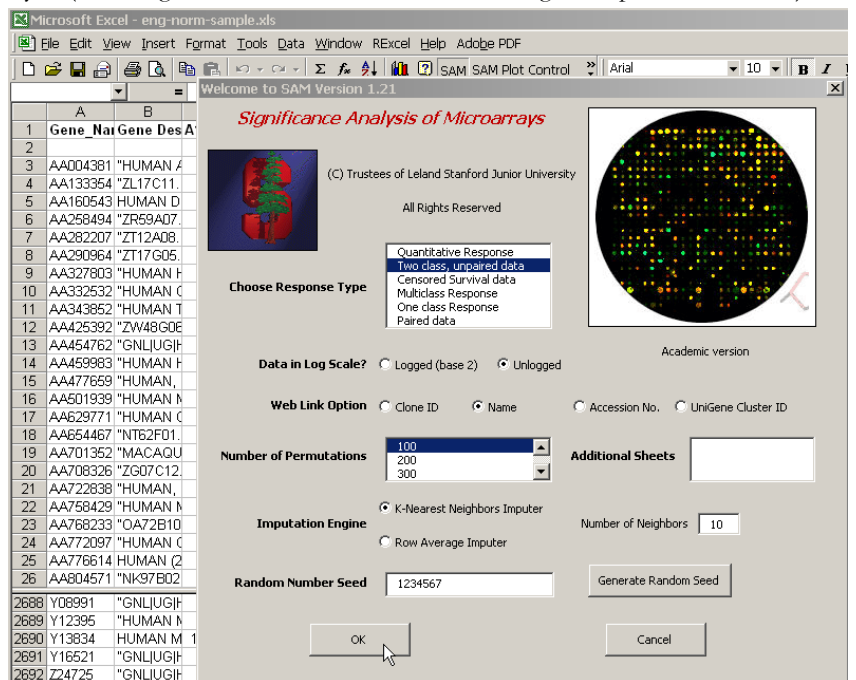
SAM uses repeated permutations of the data to determine if the expression of given genes/clones are significantly related to the response (treated vs. untreated etc.). The cutoff for significance is determined by tuning a parameter δ , chosen by the user based on the false positive rate (also called false significance).

1. Download and install the latest stable release of SAM as per the instructions on the web site.
2. Format normalized dataset for SAM. The SAM installation includes detailed documentation and example datasets for each supported experiment type. [Note: The default location for the SAM installation is **C:\Program Files\SAMVB**. The **doc** folder contains the manual (PDF format) and the **Examples** folder contains the example datasets (excel format)]
3. Select the block of data for analysis (including the identifier columns but excluding the top row with labels).

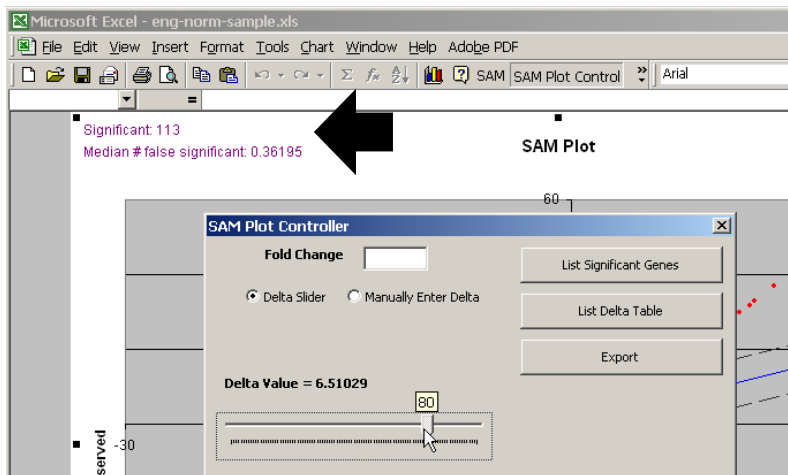
[Note: SAM allows 2 columns of identifier information. The second column will be linked to the SOURCE web site so this should be the accession numbers or gene ids or clone ids etc.]

4. Click the **SAM** button on the Excel menu bar
5. Choose the type of experiment (The example dataset is a **two class unpaired data** experiment) and fill in the options available as desired.

6. If your data is in several worksheets in the same workbook select the desired sheets visible in the **Additional Sheets** text area.



7. Click **OK**
8. A SAM plot will be generated along with a **SAM Plot Control**. Adjust the delta value slider in the SAM Plot Controller. Different delta values give you different numbers of significant and false significant genes (see arrow). Ideally choose a delta value that gives you the smallest false significant value with the largest number of significant genes
9. You can also specify a **Fold Change** at this point although this is not necessary.
10. Finally you can click the **List Significant Genes** button to get a new worksheet with a list of your significantly differentially expressed genes. This list (report) also provides links to the source web site for your specified ID (in the case of our example – Accession numbers)



SAM score (d)

The T-statistic value.

Numerator (r)

The numerator of the T-statistic (a score).

Denominator (s + s₀)

The denominator of the T-statistic (corrected standard deviation).

$$d_i = \frac{r_i}{s_i + s_0}; i = 1, 2, \dots, p$$

q-value

This is the lowest False Discovery Rate at which the gene is called significant. It is like the familiar “p-value”, adapted to the analysis of a large number of genes. The q-value measures how significant the gene is.

As d increases, the corresponding q-value decreases.

The data is put through several (we can specify how many) permutations and the d value is calculated. A delta grid is then calculated from a comparison of the d values from different permutations and a number of false significant genes (the false discovery rate) is calculated for each delta value.

Detailed information about the methods and algorithms used in SAM can be found in the manual located at:

C:\Program Files\SAMVB\doc on your computer;

or in the publication by Tusher, Tibshirani and Chu (2001): **Significance analysis of microarrays applied to the ionizing radiation response** *PNAS* 2001 98: 5116-5121

Samster

Samster can take an Excel spreadsheet or text file and extract the raw data into a text output file, which can be fed directly into Cluster or opened in Treeview. This makes it easy to cluster your significant genes from your SAM analysis. Available at: <http://falkow.stanford.edu/whatwedo/software/software.html>

Tips:

- ❑ Right-clicking on the label or tab of a worksheet allows you to move, copy, delete or rename a worksheet.
- ❑ Right-clicking the gray column letter or row number allows you to insert a new or copied column or row.
- ❑ After you have copied from a cell or column with formulae if you want to paste the value of the cell without the formula go to **Edit** → **Paste Special** and select the **Values** radio button. You can also use **Paste Special** to transpose what you have copied (for example from a row to a column)
- ❑ You can access Help at any time by going to **Help** → **Microsoft Excel Help** or by clicking **F1**

Microsoft Excel Official Site: <http://www.microsoft.com/office/excel/>

Excel Shortcuts: http://www.asap-utilities.com/index.php?page=/p_sh.php

Microarray Analysis Tutorial (Jonathan Pevsner): <http://pevsnerlab.kennedykrieger.org/hinxton.html>

Excel tutorials on the web:

- ❑ <http://www.usd.edu/trio/tut/excel/index.html>
- ❑ <http://www.baycongroup.com/el0.htm>
- ❑ <http://www.georgetown.edu/departments/psychology/researchmethods/computer/excel.htm>

Practice Exercises:

The Sample data is in tab delimited text format (sampledata.txt) containing microarray data for 2 samples A and B where there are 2 replicate experiments for each A and B. The experiments were performed using Amersham Codelink oligonucleotide chips and have been median normalized and flagged based on the negative threshold of each chip.

The columns are labeled as follows:

A1-Raw_Intensity – This is the raw intensity

A1-Normalized_Intensity – This is the median normalized intensity (this is the intensity we will be using most in our exercise)

A1-Above_Threshold – This is the flag indicating whether the normalized intensity is above or below the negative threshold for the chip. [**Note:** This is similar to the Affymetrix Present/Absent call]

1. Open the sample data in Excel.
2. Filter out all genes/clones that have intensities below threshold (flagged NO in our sample dataset).
3. Calculate the average of the filtered replicate normalized intensities for A and for B.
4. Plot a scatter plot of the filtered average normalized intensities for A vs. B and calculate the linear regression on the chart.
5. Do a T-Test on the normalized intensities to calculate the p-values for each gene.
6. Sort the dataset by ascending p-value.
7. Calculate the log to the base 2 of the average normalized intensities.
8. Calculate the mean log intensity between A and B – This is an average of all the logged normalized intensities for a given gene/clone
9. Calculate the Ratio of the logged normalized intensities for A and B.
10. Plot the mean log intensity versus log ratio.
11. Format the dataset of SAM and run SAM