Tutorial - Analysis of Microarray Data

Microarray Core E
Consortium for Functional Glycomics
Funded by the NIGMS
Data Analysis introduction

Warning: Microarray data analysis is a constantly evolving science. The methods and software described here are the current favorites of Core E and the CFG. Please be aware that newer softwares and better methodologies are constantly and swiftly being developed to meet the needs of the microarray community. As newer analysis tools become prevalent, this tutorial will be updated accordingly.

To learn more about Affymetrix array data, “low” level (generating signal intensities) and “high” level (clustering, class comparison, etc.) analysis, click here.
Analysis Tools demonstrated

Tools to learn
- RMAExpress
- BRB ArrayTools
- DAVID website for KEGG or GO

To see a general flow chart describing how to use these softwares, see next slide.
Core E General flow chart for data analysis

3+ Biological Reps each: Class 1 vs Class 2

- Modeled Signal Generation in RMA
- Filter gene list to desired species
  * Optional: Filter for Present on 2 of 3 chips in at least one class (or 3 of 4, etc.)
- Load Data into BRB-Array Tools (Excel)
- Perform Hierarchical Clustering (by Sample)
- Perform Class Comparison Analysis on known groups using:
  * Randomized variance model for univariate tests.
  * Restrict multivariate permutation to 10% False Positive rate.
  * Confidence level (Beta risk) at 80%.
- Annotate Class Comparison List using GLYCOv2 Annotation
- Use DAVID website to generate Gene Ontology breakdown
- Use DAVID website to generate KEGG Pathways of interest
Microarray Data analysis

Using RMA Express
RMA Express:
The easiest way to use RMA is by downloading RMAExpress from the Ben Bostad group at UC Berkeley. You can do that here: 
http://stat-www.berkeley.edu/users/bolstad/RMAExpress/RMAExpress.html
Scroll down to the “How do I download and install it?” section and download the newest version.
NOTE: At this time RMA express is only available in a Windows version!

How do I download and install it?
Click here for the current release Windows version. Use the installer to install the program. The current release version number is 0.2 (released Jan 11, 2004). The Linux version is not currently publicly available. You can download pre-release versions from the following table (the release versions will be more stable, the development versions may have features that are incomplete or will be removed before the next release version).

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1 Release</td>
<td>Jun 11, 2003</td>
</tr>
<tr>
<td>0.2 Release</td>
<td>Jan 11, 2004</td>
</tr>
<tr>
<td>0.3 Release</td>
<td>Dec 16, 2004</td>
</tr>
</tbody>
</table>
Before using RMAExpress

What you need for RMA analysis:

• All the .CEL files in your experiment.
• The .CDF file for your array type (i.e. GLYCOv2).

You can download both these items here

• A newly created folder with only these files in it:
Using RMAExpress

1. Open the RMAExpress application.
2. Select to **File-> Read unprocessed files.** A window will appear that will ask you to select your .CDF file. Select it from its location and click “Open”

3. Another window will immediately open that will ask for all .CEL files. Select all .CEL files in your experiment. (use the shift or control button to select multiple files).
4. RMAExpress will now read in the data.
When it is done reading in datafiles, Select **File-> Compute RMA measure.** In the options box that opens, leave the settings at default (Background Adjust: Yes, Normalization: Quantile, Store residuals: [unclicked]).
5. RMA will now carry out the analysis. When finished, it will display “Done computing RMA expression measure”. Now select **File->Write Results to file**.
6. A “save as” window will appear. Give the file a name and save it where you like.

Next: BRB-ArrayTools
Microarray Data analysis

Using BRB-ArrayTools
BRB ArrayTools is an open-source software integrated package for the visualization and statistical analysis of DNA microarray gene expression data. It is an excel add-in, and is available for download at: [http://linus.nci.nih.gov/BRB-ArrayTools.html](http://linus.nci.nih.gov/BRB-ArrayTools.html)

NOTE: You will need to request a password in order to download this software. This entails filling out a simple registration form asking for name, contact info, and institution. A password is returned relatively quickly, usually within 1-2 days.
Once you have a password, use it to download the latest “Standard version”
This will give you the following options:

• If your computer does not have R installed, you must download and install the “R setup file” (you may need to restart the computer).

• If your computer already has R or you have completed the above installation, Download the “Full installation” and install BRB Array Tools.
Open Excel. To make sure the add-ins are included, go to **Tools -> Add-ins.** Look for both BRB-ArrayTools and BRB-ArrayTools RServer boxes to be checked.
If they do not appear in the list, click “Browse” and look for them in the directory **C: Program Files/ ArrayTools/ Excel**. Select the add-ins and click their boxes once in the add-in list.
Before using BRB ArrayTools

What you need for BRB-ArrayTools analysis:

• The signal intensity values for your experiment. If you used RMAExpress, this is the saved output file. (*If using the GLYCOv2 array, you will need to do some file clean-up first)

• The “Experiment description file”. You can download a template at: http://www.functionalglycomics.org/glycomics/publicdata/microarray.jsp

Also, During the data import wizard, an option to create this file will be provided, so you may begin without it.
GLYCOv2:

If using the GLYCOv2 array, you must first open the signal intensity data file in excel. Select Column A, the column with the Probesets. Select Edit->Find and then click on the “replace tab”. You can also simply press (Control + H).

Replace EXACTLY as follows:
Find what: _Copy1_
Replace with: _

And then select replace all. Repeat this process for “_Copy2_” and then “_Copy3_”. Save this file under a different name, such as RMAdata_NoCopies.xls. This will allow BRB ArrayTools to average the multiple replicate probesets on the GLYCOv2.

Also, you can change the experiment names headers in order to make them easier to read. For example, you could change “MM_021405_BRN_Sample1/GLYCO_v2”.CEL to simply “Sample1”.
Loading Data into BRB-ArrayTools

Using the import wizard
1. In Excel, go to **ArrayTools->Collate Data-> Data import wizard**
2. Select the following:

- In Data type: Under single channel, “Affymetrix probeset-level data”. This will activate the chip type pull down menu. Select your chip type. For GLYCOv1 or v2, select “other”.

- At the bottom- If you are using data from an RMA application, Click “Input data is already logged transformed (base2)”. Otherwise, leave unclicked.

- If working with GLYCOv2 arrays (see above), select “average the duplicate spots within an array.”
3. For “File type”: Select “Arrays are saved in a horizontally aligned file”

4. For “File containing expression data for all arrays”: Click Browse and select the expression data file. If using the GLYCOv2, select the NoCopies version of the file created above. Click next. ArrayTools will warn you it is changing the file into text format. Click OK.
5. This will bring up a window that asks you to identify the rows and columns in the file.

From the pull down menus select:

- the header row
- the first line of data
- which column has the probeset ID, usually col. 1
- which column the data for the first array begins, usually col. 2
- which column the data for the second array begins, usually col. 3
- which column the first array’s signal will appear, usually col. 2
- leave “Detection call” blank.

Excel should show a message window that states the number of arrays you have. If correct, click yes.
6. If you have downloaded a Experiment Descriptors File from the central database, browse to the appropriate file and select. If not, click the box that says “I don’t have an experimental descriptors file, please create a template for me”. This should open a “Save as” box that allows you to save and name the file as you like.

Before proceeding, open this template and add in column B the heading “group” or “class”. In this column distinguish the sample as they are distinguished in your experiment. For example:

<table>
<thead>
<tr>
<th>Experiment Name</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Wild type</td>
</tr>
<tr>
<td>B</td>
<td>Wild type</td>
</tr>
<tr>
<td>C</td>
<td>Wild type</td>
</tr>
<tr>
<td>1</td>
<td>Knockout</td>
</tr>
<tr>
<td>2</td>
<td>Knockout</td>
</tr>
<tr>
<td>3</td>
<td>Knockout</td>
</tr>
</tbody>
</table>

Save file and return to BRB-ArrayTools import wizard.
8. ArrayTools will create a directory for your project. Give your project a name.
9. Give your project excel worksheet a name
10. BRB will import the data. It will give a final tally of the genes in the analysis. You will be asked if you wish to annotate the genes online. Since the GLYCO arrays are both custom designs and ArrayTools will not recognize the probesets, click no.
Clustering and Class comparisons

11. To generate a hierarchical cluster, Select **ArrayTools-> Clustering-> Samples alone**
12. Leave options at default. Center genes, Centered correlation, average linkage. Use all the experiments. Click OK and a cluster will be produced.
13. Class Comparison between groups is used for comparing 2 or more pre-defined classes. To do so, select **ArrayTools-> Class comparison-> Between groups of arrays**
14. Perform Class Comparison Analysis using:

- Experimental design: Groups
- Unpaired Samples
- Randomized variance model for univariate tests.
- Univariate significance test at 0.01
- Restrict multivariate permutation to 10% False Positive rate.
- Maximum proportion of false discoveries 0.1
- Confidence level (Beta risk) at 80%.
15. The class comparison will output a HTML file with the statistical settings and outcome of testing. The following list of significant genes will have a p-value, geometric mean of Group#1, a geometric mean of Group #2, a fold change value, and a Probeset ID. Copy and paste this list into Excel.

Note: The probeset IDs are linked to the Affymetrix database for probesets. Since many of the genes on the GLYCOv1 and GLYCOv2 are custom designed, many of these links will not work.
16. The gene list output from the class comparison contains the best candidate genes for the separation of the 2 classes. You can download an annotation list for the GLYCOv1 or GLYCOv2 from here: http://www.scripps.edu/researchservices/dna_array/glyco_genelist.xls

Use an “advanced filter” (Data->Filter->Advanced Filter) in excel to pull out the annotation for your significant genes list.

Next: DAVID website
Microarray Data analysis

Using DAVID website for KEGG and GO
The DAVID website

DAVID, or Database for Annotation, Visualization and Integrated Discovery, provides integrated solutions for the annotation and analysis of genome-scale datasets derived from high-throughput technologies such as microarray and proteomic platforms.

This tutorial will demonstrate how to use DAVID 1.0. DAVID 2.0 is now available and has many additional options, but both versions operate in similar fashion.

1. On the left hand menu, click “upload new list”. This will link to a page that allows you to either upload a file or cut and paste a list of Affymetrix IDs, locuslink IDs, unigene IDs, or GenBank accession numbers.
2. Cut and paste your list of significant genes IDs into the lower field. Because many of the probes on the GLYCO arrays are custom designed, the AFFYID option is often not useful. It is suggested that GenBank IDs are used. These can be found for GLYCO array probesets here, under the GenBank heading. Click submit text to receive results.
3. This will give a list of options. Two useful ones are GO (gene ontology) classification charts and KEGG pathway charts.

**DAVID**

*Database for Annotation, Visualization and Integrated Discovery*

- **Annotation Tool** - An automated method for the functional annotation of genome-scale datasets.
- **CoCharts** - A visualization tool that graphically displays the distribution of differentially expressed genes among functional categories.
- **KEGGCharts** - A visualization tool that graphically displays the distribution of differentially expressed genes among metabolic pathways.
- **DomainCharts** - A visualization tool that graphically displays the distribution of differentially expressed genes among functional protein domains.
- **rANKonline** - Provides statistical methods for discovering enriched biological themes within gene lists. For detailed information, please see the rANK FAQ.
4. Follow the GoCharts link and you will see a options pages. A good place to start is “Biological Process” classification at level 3 coverage. Click “ChartValues!” and a chart of results will be displayed, as shown below.

*You can mouse over the blue bars to get a list of genes in that category. If you click on the blue bars it will produce an annotated list of gene.

*Click on the category link to see information about that classification.
5. Following the KeggCharts link from the page shown below will provide a simple options menu. Usually, the default settings are good to use.
6. Click “ChartPathways” and the following list will appear, similar in presentation to the GO Charts.

*As with the GO Chart, you can mouse over the blue bars to get a list of genes in that category. If you click on the blue bars it will produce an annotated list of gene.

*Clicking on the category link will display the KEGG chart for that pathway.
7. Below is a screenshot of an example pathway. The genes are represented with boxes and numbers such as “3.2.1.23”. Clicking on the box will pull up an annotation page.

Legend:

**Green boxes** - Gene present in that organism

**White boxes** - gene present in the pathway, but in another organism.

**Red Numbers** - Gene was on your uploaded list
8. From here you can combine the microarray data and the pathways and classification information to arrive at a better understanding of biological processes.

Keep in mind that the KEGG pathways are not complete for glycoproteins. The increase of Glyco pathways is one of the intentions of the Consortium of Functional Glycomics.
For questions or comments concerning this tutorial, contact:

Tim Gilmartin  
CFG Core E  
The Scripps Research Institute  
timgil@scripps.edu

Additional thanks to:
  • Jen Hammond (TSRI DNA Array)  
  • Core B- IT Team (MIT)  
    – Maha Venkataraman  
    – Subu Ramakrishnan  
    – Wei Lang