Module on Microarray Statistics for Biochemistry: Metabolomics & Regulation

Part 2: Normalization of Microarray Data

By Johanna Hardin and Laura Hoopes

Instructions and worksheet to be handed in

NAME_____

Lecture/Discussion

Why normalize? (scanning at high and low intensities)

MAD/median/scaling/location

MA plots (Minus, Add) Quotation from *Functional Genomics*, 2003, Smyth, G, Yang, Y, and Speed, T, Methods in Molecular Biology (electronic books) vol 224, Humana Press. "The log differential expression ratio is M = $\log_2$ R/G for each spot. Finally, the log intensity of the spot is A = 1/2 $\log_2$ RG, a measure of the overall brightness of the spot. Note that the letter M is a mnemonic for minus, as M = log R - log G; while A is a mnemonic for add, as A = (log R + log G)/2. It is convenient to use base 2 logarithms for M and A so that M is in units of twofold change and A is in units of twofold increase in brightness. On this scale, M = 0 represents equal expression, M = 1 represents a twofold change among the RNA samples, M = 2 represents a fourfold change, and so on." LH note: this plot detects whether or not you have excess variation at low sample intensities due to poor signal/noise ratios.

In our experiment we will use comparative versions of A and M:

A: Average log value - (Log(Red)+Log(Green))/2

M: Normalized log ratio - Normalized Log(Red/Green)

Loess plots. (Lowess plots) This type of normalization will 'fix' variable intensity that seems related to the position on the slide.

Within-slide/print tip/global normalization

Combining replicates and eliminating outliers (Draghici, section 12.2.2)

Background correction or not/ elimination of flagged spots, imputing missing data, normalizing overall array intensity, normalizing for color distortion, calculating ratio, log transform, combining replicates. (Draghici, Section 12.6)

Reading assignments:

Draghici, S (2003) *Data Analysis Tools for DNA Microarrays* Chap12, 13.

Schuchhardt, J Beule, D, Malik, A, Wolski, E, Eickhoff, H Lebrach, H and Herzel, H (2000) Normalization strategies for cDNA microarrays. *Nucleic Acids Research* 28 (10): E47, i-v.

Yang, Y, Dudoit, S, Lin, D, Peng, V, Ngai, J ,Speed, T (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* 30 (4):

Dry laboratory work:

Download BTB ArrayTools: http://linus.nci.nih.gov/~brb/download.htm

This is a government (NIH) sponsored free microarray statistics package that acts as an add-on to Excel. The program accepts gpr (gene pix results) files.

1. Load into this program the set of files of your class data that will be provided on a CD. In order to import the files, you will need to create an

Experiment Description file. Do not put this file into the same folder as your gpr data files or it will confuse the import system. Here is an example of an Experiment Description file:

Experiment Names	organism	Cy3strain	Cy3growth	Cy3 flask	C5strain	C5growth	Cy5 flask
73Bottom_0635	yeast	WT	stat	3	snf1	stat	1
73Top1_0635	yeast	Wt	stat	3	pde2	stat	2
74Top_0635	yeast	WT	stat	3	WT	stat	4
74_bottom_results	yeast	WT	stat	3	WT	log	4
75TOPRESULTS	yeast	WT	stat	3	asr1	stat	5
75_bottom_results	yeast	WT	stat	3 1/3	WT	log	3

For this file, you must have the left column entitled as is, and must list the gpr names exactly (except without the gpr extensions). But the other columns are up to you to create, the program does not care what they are called or what they contain. If your data contain dye flips, you may need to make the data in one column reflect that the sample is a dye flip so that corrections can be made automatically. But in 2006 Biochemistry, your data do not contain dye flips.

Save this file to another location that you can recall easily, because you will need to point the browser to it during the import process.

Do not ask the program to subtract the background during importation. (Refer to the section called 'procedure help' at the bottom of this assignment for more methods descriptions by Ryan Murphy).

During the import process, do not apply any filters. You can add filters later to remove spots with flags, spots with low intensities, etc. Sometimes you can import successfully with filters, but just in case, leave them out for now.

2. Click on scatterplot under the tools menus. Select 'experiment versus experiment" and then choose Normalized log(Red/Green) on the left and ((Log Green)-(Log Red))/2 on the right side. Choose one of the class files (the same one) for experiment on both sides. For the optional label, use M for the left side and A for the right side.
2. For the file selected, set the ratio to 2.
3. Click on Set gene subsets. Click on the lower level box on the left and enter Unique ID in the gene identifier box and Empty in the 'contains the following string' box. Click OK. An MA plot will be produced that is NOT NORMALIZED (unless you have leftover normalizing settings from a previous run). You will probably see a line fitted to the plot that is nowhere near a straight line at 1.0 ratio.
4. Print the MA plot, label it with the file to which it applies, and attach to this worksheet.
5. Prepare an MA plot on a second file using the same settings. This time, select one that is different from your first one in intensities. That is, try to select one that has a lot of low intensities if your first one had high intensities, and vice versa. Print the second MA plot and attach.
6. Describe the 'quality' of the MA plots you prepared, i.e. is there any indication that intensity level was affecting variability in either array?

7. Apply a loess normalization to one of the data sets you chose above. To do that, look under Tools for Filter and Subset the Data and choose that. At this stage, skip the first screen (Spot filters). Highlight the "normalization" bar near the top of the screen. Click normalize and loess. Truncating is not recommended here. Now click on the third bar at the top, 'Gene Filters'. If you want to only consider genes that underwent a particular degree of change in the experiment, then choose the top checked box and fill in a test cutoff point (the default is 20% of the arrays with at least a 1.5 fold change). The final box is the gene subsets box, which you have already filled in. So now you can check the OK box. It will tell you how many genes have passed your filters. If it is zero, don't try to make an MA plot, instead go back and lift or modify one of the filters. The Gene Filters are the most likely ones to cause this result.
   Once you have a reasonable number of genes that pass (more than 100), then go back to the Scatterplot, experiment versus experiment choices under Tools and choose them. They will be applied to the current data selected, with the filters you have given. Print the Loess smoothed MA plot, label with data set and loess, and attach to this worksheet.
8. Describe the difference in the MA plots as a result of the loess normalization.

9. Flagged spots were called abnormal in shape/position by either the GenePix software or the person who gridded the spots on the array. Try filtering out the flagged spots in one of your selected arrays and then redo the MA plot without the loess correction (attach to the worksheet). Click on Tools and Filter and Subset the Data. On the first screen (Spot filters), select the right side and choose 0 and 0 for the two fill in openings. That will remove flagged spots. Decide whether or not you want to take out spots at low intensities. If you do, try the default settings on the left by checking the left box. Run the scatterplot function again. Print out, label, and attach the MA plot. Comment on the effect of the removal of 'bad' data.

10. Log base 2 transformation is useful because it makes down-regulation and up-regulation have equivalent effects on the data. Using Excel, copy out 10 rows of data from one of your gpr files, including the column headers. From the 10 lines, copy and paste (on a second sheet of your workbook) the column headed ID, the column headed 'Ratio of Medians (635/532)' and the column headed 'Log ratio(635/532)'. From the 10 genes, select 5 for which you will plot these two values, choosing ones that with ratios of medians greater than 2 and less than 0.5. Copy the 5 chosen gene IDs to make a new table. Make (under the Insert Chart command) a column graph using the gene ID and the

Ratio of Medians and another column graph using gene ID and Log Ratios. Attach plots. Comment on the comparative visibility of large changes (for example two fold changes) up or down in gene regulation using the two plots.

11. Dye inversion or dye flipping is a method that should minimize the effects of the different dyes used in the probes on the outcome. Since the dyes are of different size and chemistry, it isn't possible to completely avoid dye effects. However, if the same sample pair is tested with red dye on RNA1 and green dye on RNA2, and then with green dye on RNA1 and red dye on RNA2, the dye effects should be cancelled out. Explain how you could make red/green ratios from such a pair of arrays comparable.

12. Background subtraction can be hard to use in microarray data. For comparison, you will need to re-import one of the files you have been working with, asking for background subtraction during the import process. Apply the loess filter and make a new MA plot. Also scan down the tabulated data to examine the results. Comment on what might cause problems in the data analysis:

13. Combining replicates is a complex process for microarray analysis; now that you have looked at normalization methods, describe some of the issues in averaging values from 3 microarrays that purport to be identical.

You now will have data files that are imported, flagged spots removed, and loess smoothed that you can use for your comparison of your data for mutants with your and the summer researchers' data for wild type. To do those comparisons, you can use the additional directions given about comparisons below. You might want to group the two most different mutants together and compare them with all the other arrays, for example, to see if they are statistically significant.

PROCEDURE HELP PREPARED BY RYAN MURPHY and annotated by LH:

Importing Data

Open Excel.
Go to ArrayTools → Collate data → Data import wizard
Set data type: dual-channel intensities
Check "Average the duplicated spots within an array"

Hit Next

File type: Arrays are saved in separate files stored in one folder

Hit Next

Folder containing expression data files: Browse and select the folder containing expression data files

If data are not saved as tab-delimited text files, ArrayTools will save them as such and prompt you before doing so. (Make sure none of the files are opened when you are trying to select them or ArrayTools will not be able to change the file extension).

Make sure that the number of expression data files matches what you would expect. Additional files in the folder can often be confused for expression data files. If the number is correct, hit yes and continue. Hint: most of the expression files are collected in Gene Pix and have the Gene Pix Results (gpr) extension, for example: Diauxie23_1.gpr.

Follow the instructions in the Data Import Wizard, selecting the header line, first data line, etc. We generally prefer to use the median intensities because any contamination or non-uniform loss will not affect the median very much; the other choice would be mean intensities that could be thrown off by a dust bunny intensity. In most cases, these two numbers are very similar, though.

Unique ID – Col 5: ID
Red Intensity – Col 9: F635 Median
Green Intensity – Col 21: F532 Median
Spot Size – leave blank
Spot Flag - Col 54: Flags
Print-tip groups - Col 1: Block

(Note: You cannot average duplicated spots if the dataset contains print-tip groups! Array Tools will prompt you to decide between averaging duplicate spots and using print-tip group data. For this lab, selecting averaging duplicates is fine.)

If wanted, check the box for apply background adjustment (in the exercise above, you don't do this at first, but you will come back and re-import with this checked later)
Red background - Col 14: B635 Median (if wanted)

Green background - Col 26: B532 Median (if wanted)

Hit Next

Gene identifiers: The gene identifiers are placed alongside the expression data.

Unique ID – should already read Col 5: ID
Gene Name - Col 4: Name (in some of our current gene lists, 'name' is actually annotation telling something about what the protein does; in others, it's the common name of the gene)

Hit Next

For the experiment descriptors file, simply locate the file using the browse menu. If you are creating an experiment descriptors file, be sure to save it **outside** of the folder containing expression files. BRBArray tools will confuse it for an expression file.

Hit Next

1. Spot Filters (skip for importing and apply later)

   Intensity Filter:
   Uncheck (for fine data analysis, this would be used)

   Spot Flag Filter:
   EXCLUDE if Spot Flag contains values outside the range 0 to 0.

   Spot Size Filter:
   Uncheck

2. Normalization (skip for importing and apply later)

   Normalize each array:
   Use Lowess smoother (regardless of whether or not you chose to use print-tip group data)

   (Note: when importing data for the first time, you must use lowess or loess smoother or Array Tools will crash. Once the data are imported, you can then perform lowess within print-tip group if you want to; we will not do this).

3. Gene Filters (skip for importing and apply later)

   EXCLUDE a gene using minimum fold change where:
   Less than XX% of the expression data values have at least a XX-fold change in either direction from the gene's median value

This criterion can be used to choose only those genes that changed in expression during the experiment, so you can look at 100 or 1000 and not 6000 data points. A loose criterion would be looking for a 2x change; a tighter one would seek at 3x, 4x, or 5x change.

Percent Missing:

Percent of data missing out or filtered out exceeds XX%

This filter will remove a gene if it is usually flagged or too low or otherwise filtered out, even from those arrays where it is present. The idea is not to conclude something based on unreliable data (guilt by association: this gene is usually not read well).

(Note: Ideal values vary between data sets. If there are too few points, the filter is likely too stringent and adjustments should be made to one or both of the minimum-fold change or percent missing values)

Hit OK

When asked if the user would like to annotate the data, select no. It would be nice if BRB array tools would annotate yeast genes, but it can't. If we were using human or mouse, it could.

Save the file when prompted.

Once the data has been imported the user can reset the filters and change the type of normalization by clicking in the menu ArrayTools → Filter and subset the data.
(Note: Now that the data has been imported, the user can now perform lowess within print-tip group normalization if desired. We won't in this lab.)

MA Plots

With the datafile you just created opened, generate a scatterplot by going to ArrayTools → Scatterplot → Experiment vs. Experiment

X-values:

Variable: Average log value - (Log(Red)+Log(Green))/2
Experiment: 52ISBHWk2
Name: A

Y-values:

Normalized log ratio - Normalized Log(Red/Green)
Experiment: 52ISBHWk2
Name: M

To exclude control and empty spots, click “Select gene subsets”

Click specify which gene to exclude
Where the following gene identifier: unique id
Contains the following string: CONTROL (or empty or null)

If desired, try different normalization methods and observe effect on scatterplot (e.g. lowess smoother vs. lowess within print-tip group).

Class Comparison

Go to ArrayTools → Class Comparison → Between groups of arrays

Column defining classes: grouping variable
Select unpaired samples
Block by
Average over replicates of
Paired samples
Select use random variance model for univariate tests

Find gene lists determined by:

Univariate significance tests:
Significance threshold of univariate tests = 0.001

Restriction on multivariate permutation probability of false discoveries:
Max number of false discoveries = 10
Max proportion of false discoveries = 0.1
Confidence level (between 0 and 100%) = 90

Class Prediction

Column defining classes: Grouping variable

Prediction methods (check the following):

Compound covariate predictor
Diagonal linear discriminant analysis
K-nearest neighbors
Nearest centroid
Support vector machines

Use random variance model for univariate tests

Individual genes:

Significant univariately at alpha level = 0.001