Microarray data analysis

by Binh Nguygen (nguygen.726@osu.edu)

Overview

This section describes a minimum setup requirement and step-by-step procedure to setup an environment for Affymetrix oligo microarray analysis. A sample experiment with input and output files is also described for basic steps in "microarray data analysis". You can modify the procedure to fit your own analysis; however, it may require additional software packages, techniques, and programming to analyze more complex experiment which will not be described herein. We selected Bioconductor freeware and R-statistical package, AffylmGUI (Wettenhall, 2006) for several reasons: (1) Free software that has excellent support from community; (2) Ease of installation and Graphic User Interface to guide non-programmer; (3) Utilization of comprehensive and standardized microarray analysis methods.

Setting up workbench

The minimum requirement for Affymetrix microarray analysis workbench is a Personal Computer (PC) with XP windows environment with Services Pack 2 (SP2) and 512MB RAM with internet connection. For UNIX or MAC R package installation (http://cran.r-project.org) and additional Bioconductor packages (www.bioconductor.org), please refer to website instructions.

Microarray Analysis Freeware installation on Windows:

This section describes step-by-step procedure to download and install software for minimal Affymetrix microarray analysis. The required packages are:

(a) R-statistical packages (http://cran.r-project.org) version 2.2.X.

(b) Tcl/Tk 8.4 or later from Active State ActiveTcl for Windows.

(c) Bioconductor packages including affylmGUI version 1.5.4.

Step-by-step software installation procedure:

1. R-statistics, Tcl/Tk and bioconductor packages installation.
1. Download R-statistics package version 2.2.X from http://cran.r-project.org. After download the binaries version for windows from base directory, double-click on the downloaded file (e.g. R-2.2.1-win32.exe) and follow instructions to install R-statistic base packages.
2. Download Tcl/Tk 8.4 or later from Active State ActiveTcl for Windows.
4. Install additional R-packages and bioconductor from R windows environment.
    . After R package installation, double-click to open the R windows environment.
    . Go to top menu and select "Package → Repositories"; then select by "CRAN, CRAN extras, and bioconductor" by holding down Control Key (CRTL) and use the left mouse button (CTRL+Left button) to select items from the popup menu; then click OK button.
    . Go to top menu and select "Packages → Install package(s)"; then select mirror site (e.g. USA CA1) followed by packages affy, affylmGUI, tkrplot, xtable or affyann.
3. Update latest version of affylmGUI (version 1.5.4)
    a. down load affylmGUI version 1.5.4 zipfile to local disk
    . Go to top menu and select "Packages → Install package(s) from local zip file. (affylmGUI_1.4.5.zip)
    . Select the location of the local zipfile (affylmGUI_1.4.5.zip) and update afflylmGUI.

Affymetrix Microarray Data Analysis procedure using AffylmGUI

The detailed step-by-step Affymetrix Microarray Data Analysis procedure and dataset could be downloaded from the Walter and Elizabeth Institute of Medical Research, Melbourne Australia (WEHI biotech centre) website. This section briefly describes the default step-by-step standard procedure and underlying analysis methods.

Overview

For the default data analysis settings, we recommended Robust Multichip Analysis (RMA) for background adjustment, cross-chip normalization, and robust linear model to summarize probe level expression values. The RMA model logarithmically transforms probe intensities, adjusts background, performs cross-chip normalization, and uses robust linear models to summarize probe level expression values. Briefly, I represents number of probesets (11-20 for ATH1 array), J represents number of probes (~28 K for ATH1 array), and K represents number for arrays. RMA will calculate probe intensities of GeneA as $\log_2(GeneA_{ijk}) = \theta_{ik}+\phi_{jk}+\varepsilon_{ikj}$, $i = 1,...,I$, $j = 1,...,J$, $k = 1,...,K$ for θ represents a quantity proportional to the amount of RNA, ø represents a probe-specific effect, and ε, representing measurement error. The standard RMA algorithm uses median polish to estimate θ (Irizarry, 2003). To identify differential expressed genes, we used a moderated *t*-test (Smyth, 2004) with Benjamini-Hochberg adjustment, and the average log-fold-change to adjust the cutoff values.

Preparation:

1. Sample Experiment: Identify the direct targets of a transcription factor, GL3 in *arabidopsis*.

This is a glucocorticoid-transcription factor fusion (TF-GR) experiment in egl3gl3 double mutant background. The egl3gl3 mutant has defective trichome development (no trichome in seedlings) compared to wild type (Ler). With Dexamethasone treatment (DEX), pGL3::GL3-GR construct drives expression of GL3 targets and rescues egl3gl3 phenotype (trichome formation). In order to identify the direct targets of GL3, cyclohexamide (CHX) is used to stop system-wide translation. Gene expression profile analysis based on DEX treatment in contrast with Mock treatment (EtOH) may identify both direct and indirect targets whereas CHX+DEX treatment in contrast with Mock would give the potential direct targets (refer Analysis section).

2. Input to AffylmGUI:

Before using affylmGUI to analyze the data, we need to obtain all the CEL files for the experiment and create a "DNA Targets" file. The CEL files should be organized in a directory (e.g. c:\GL3Experiment\CELfiles).

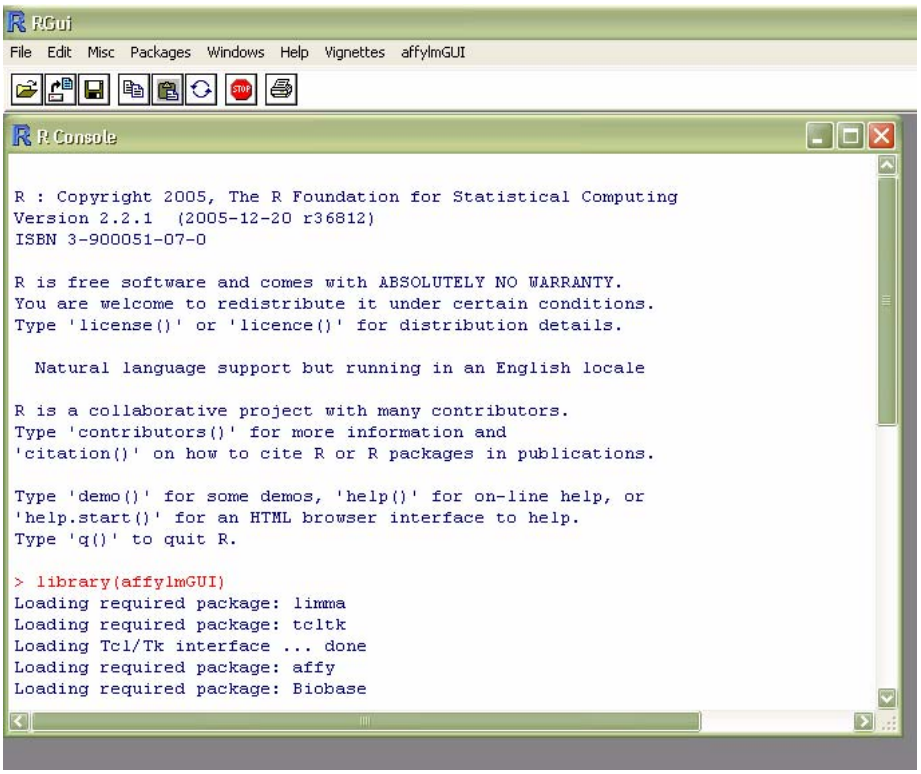
Table 1: DNA Target file contains three columns separated by "tab". The first column, "Name", is a unique identification used by affylmGUI. FileName represents the exact name of CEL file. Target represents the treatment group.

Name	**FileName**	**Target**
GL3M1	GL3-M1-N.CEL	GL3Mock
GL3M2	GL3-M2-N.CEL	GL3Mock
GL3C1	GL3-C1-N.CEL	GL3CHX
GL3C2	GL3-C2-N.CEL	GL3CHX
GL3CD1	GL3-CD1-N.CEL	GL3C+D
GL3CD2	GL3-CD1-N.CEL	GL3C+D
GL3-D1	GL3-D1-N.CEL	GL3DEX
GL3-D2	GL3-D2-N.CEL	GL3DEX

The tab-delimited "DNA target" file should be placed in the same directory as the CEL files.

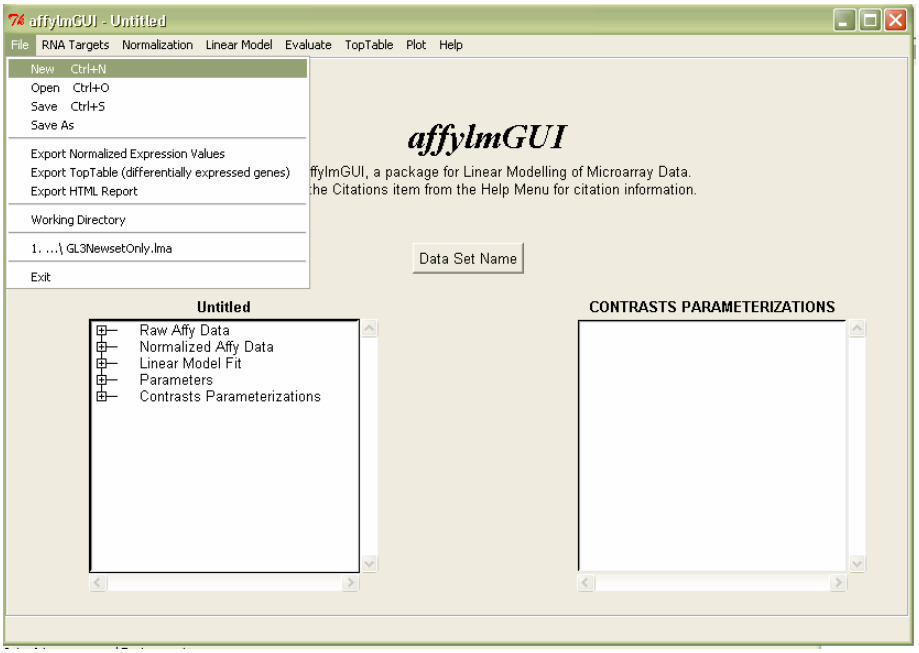
3. Step-by-step procedures:

a. Open R package and load afflyGUI.
b. Double-click on R icon to open R console.
c. At the prompt (>) in R window environment, type "library(affylmGUI)" to open affylmGUI window.

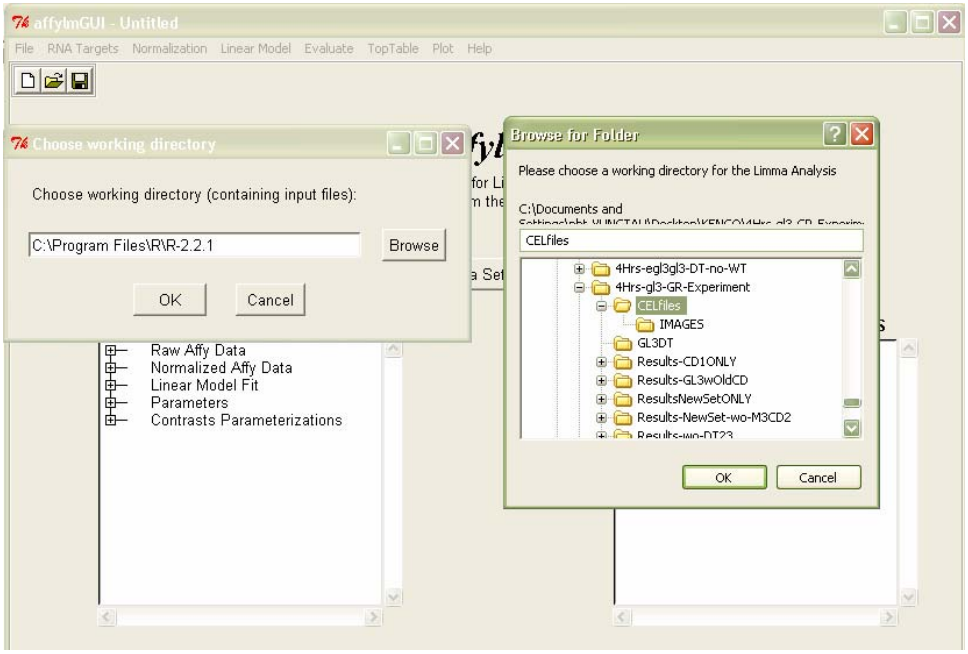


d. Now open afflylmGUI tcl/tk interface (loaded under R environment).

e. From top menu, select File → New, to open location of CEL files and DNA Target file.



f. Open CEL file directory using "Browse", select CEL file directory and click "OK".

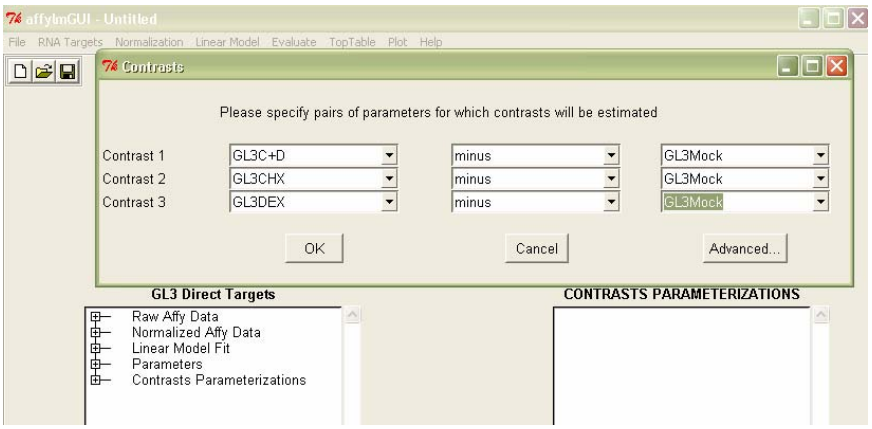


g. AffylmGUI will prompt you to select DNA Targets file; select appropriate DNA target file and click "OK" (Figures not shown).

h. AffylmGUI will automatically load CEL files to "memory" according to specified DNA target files. The system will prompt and ask for dataset name. You can enter your experiment name for the dataset name (e.g. GL3DirectTargets) then click "OK".

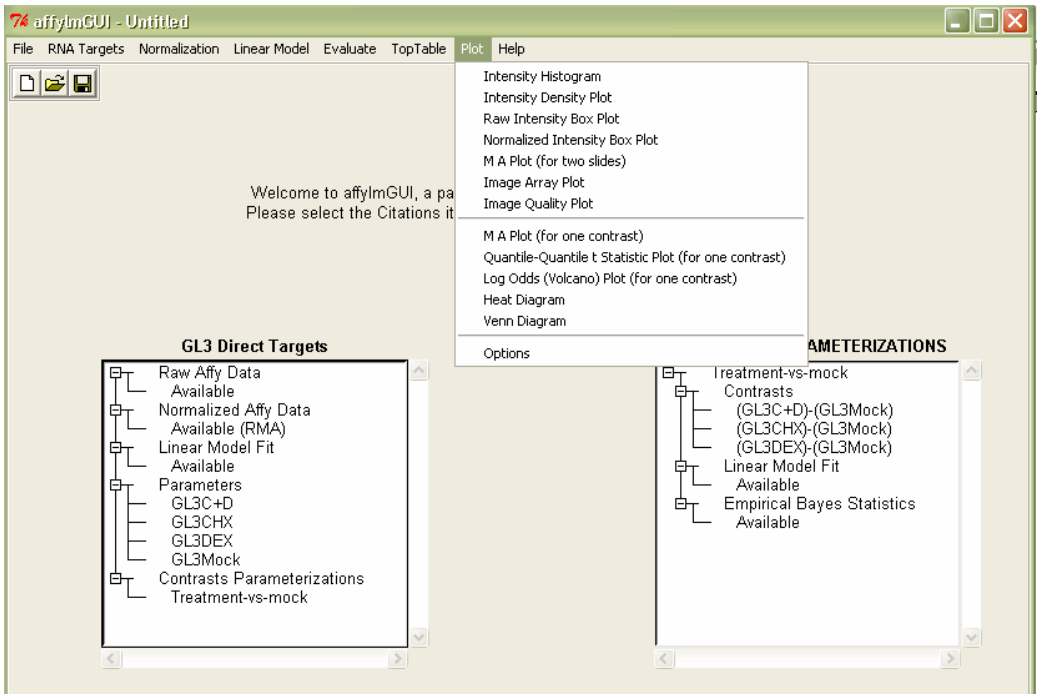
i. Go to top menu and select Normalization→ Normalize. Select RMA from pop-up menu and click OK. At this point, affylmGUI performs background correction, cross-chip normalization, and summarization.

j. Now, you need to make appropriate contrasts between treatments and mock.

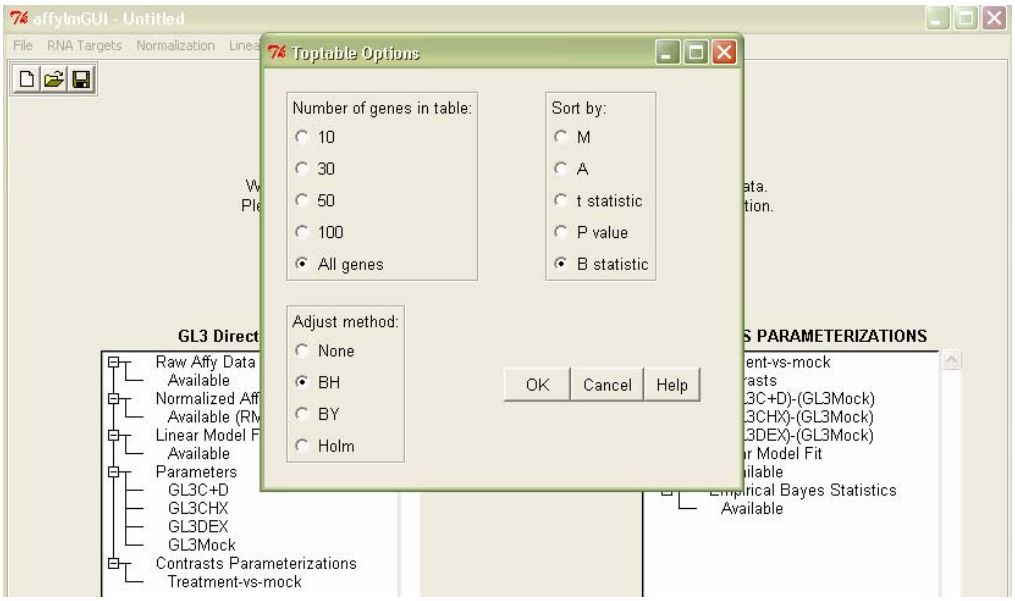


k. Once select of contrasts were made, enter the name for the contrasts (e.g. treatment vs. mock) then click "OK". Note that you can make as many contrasts as possible.

l. There are several plot option for where you can take advantage of these features to have an overview of your data.



m. You now can export the data for further analysis using File → Export HTML report or File → Top table (differential expressed genes) one at a time.



n. Select “All genes”, Sort by “B statistic”, Adjust method “BH” or Benjamini-Hochberg, then click ok and save the contrasts output file (e.g. gl3dex-m).
Sample output: (gl3dex-m) using EXCEL to open the saved tab-delimited file.

	ID	M	A	t	P.Value	B
21852	266752_at	-1.87737	9.972459	-11.3783	0.016479	4.810983
22365	267265_at	1.908659	4.412661	10.6508	0.016479	4.434487
1375	246275_at	1.523087	4.067865	10.59877	0.016479	4.406087
7982	252882_at	1.728133	3.605809	10.55761	0.016479	4.383471
3776	248676_at	-2.19795	7.621175	-10.2282	0.016479	4.19774
18502	263402_at	-2.02942	8.732087	-10.082	0.016479	4.112531

First, column represents the affymetrix probe position, ID is affymetrix probe identification, M represents Fold change for the contrasts, A represents average intensity in log2 transformation, t represents moderated-t statistics, P.value, and B log odds value using Benjamini-Hochberg adjustment for false discovery rate with RMA.

o. What cut off to use depend on number of replicates and quality of arrays. We recommend p.value < 0.05 and/or |moderated-t| > 4 with adjusted Fold change and B value > 2.5. Using the above criteria, the outliner (large variance in the gene probe intensities per gene) will be automatically eliminated (large P.value or small absolute moderated-t statistics).

4. Analysis Consideration:

Affymetrix arrays are typically used to identify differential expressed genes. For one-way comparision, the contrasts and statistical analysis (cutoff) would be sufficient to select significantly differential expressed genes. However, to identify direct targets with the TF-GR fusion experiment, we consider the following criteria:

1. The effect of cyclohexamide should be minimal (CHX/Mock is low).
2. The differential expression in each contrast is significant (i.e. within cut off value P. value < 0.05 and/or |moderated-t| > 4, B.value > 2.5).
3. We prefer to have DEX/Mock and (CHX + DEX)/Mock in the same direction and high.

Note: The side-effect of CHX in system-wide translational inhibition is that cyclohexamide also stabilizes mRNA; Therefore, the combination of CHX and DEX treatment could have a synergistic effect (i.e. the expression of a particular gene (mRNA) induced by DEX and also stabilized by CHX would highly represent whereas its expression induced by DEX alone could be low due to mRNA degradation).

Other Microarray Analysis and Statistical Analysis Software

BASE (BioArray Software Environment). Microarray database server with normalization and some analysis facilities. Lund University, Sweden.

Bioconductor. An R-based development project for open-source genomic software. Includes a number of important packages. Coordinated by Robert Gentleman, Fred Hutchinson Cancer Center.

BioSieve. Produce **ExpressionSieve**, a microarray data analysis, data mining and data visualization software package written in Java.

Broad Institute of MIT and Harvard. Produce the GenePattern package for genomics and microarray analysis.

Eisen Lab. Distribute ScanAlyze, Cluster and/or TreeView free for non-profit researchers.

GeneSifter.Net. A web-based data analysis tool for Affymetrix microarrays created by VizX Labs, Seattle.

Bengtsson, Henrik. Develops **aroma**, an elegant R package for microarray normalization, diagnostics and data analysis using a custom object-orientated system.

National Center for Genome Resources (GeneX).

PAM: Prediction Analysis for Microarrays. Robert Tibshirani, Stanford University.

Pevsner Lab, (Kennedy Kreiger Institute). **SNOMAD** (Standardization and NOrmalization of MicroArray Data).

PermutMatrix. MS Windows software for clustering and seriation analysis of gene expression data. Gilles Caraux, LIRMM, France.

RMAExpress: A Windows program for computing the RMA expression measure

Speed Group (University of California, Berkeley). R package **sma** (Statistical Microarray Analysis), Windows application **RMAExpress** and contributions to the **Bioconductor** project.

TIGR. SpotFinder image analysis, MIDAS data management, MeV differential expression analysis.

Probability of Expression (POE). An approach to the analysis of gene expression microarrays using three-component mixtures. Giovanni Parmigiani, Johns Hopkins.

Townsend Lab, University of Connecticut. Produces **BAGEL** (Bayesian Analysis of Gene Expression Levels) for the statistical analysis of spotted microarray data, **Pathway Processor** to test for overrepresentation of differential expressed genes within known pathways of *S. cerevisiae* or *B. subtilis*, and **SeqPop** for computing population genetics statistics on sequence data.

University of Pittsburgh & UPMC Bioinformatics Web-Tools Collection. Includes expression analysis tool with www interface.

Walter and Eliza Hall Institute of Medical Research. Produce R package **limma** (linear models and differential expression for microarray data) and associated user-interfaces **limmaGUI** and **affylmGUI**.

Wong Lab (Harvard University). **DNA-Chip Analyzer**.

Yang, Jean (University of Sydney). R packages **marray**, **DEDS** and **stepNorm**.

References:

Irizarry, R.A., B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, U. Scherf, and T.P. Speed. (2003). Exploration, normalization, and summaries of high density oligo-nucleotide array probe level data. *Biostatistics 4*, 249-264.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* 3, No. 1, Article 3. (Full Text)

Smyth, G. K. (2005). Limma: linear models for microarray data. In: *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds.), Springer, New York, pages 397-420. (Published 8 August 2005, Publisher web site, PDF)

Wettenhall J., Simpson K. M., Satterley K, and Smyth G. (2006). affylmGUI: a graphical user interface for linear modeling of single channel microarray data. *Bioinformatics Advances Access* (Feb. 2006). (PDF)

http://arabidopsis.med.ohio-state.edu/NSF2010Project/