Data QA of a Simple GeneChip Experiment

Carter, DE[1]; Andrews, Joseph D; Khan, HMR; Hegele, RA

London Regional Genomics Centre, Robarts Research Institute, Room 4.01, 100 Perth Drive, London, Ontario, N6A 5K8

Introduction

Many steps are required to produce statistically and biologically significant microarray data. These include proper experimental design, extraction of high quality RNA, labeling and hybridization controls, data normalization and various mining techniques, followed-up by some biological verification such as real time PCR or Northern blotting.

The London Regional Genomics Centre uses numerous analysis packages to assess data quality, and continually explores new software and methodologies. Data quality will be assessed using a simple treatment vs. control design with three technical replicates.

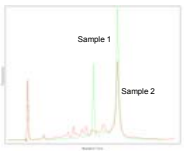
Experimental Design

Experimental design is a critical planning step to ensure the question you wish to explore will yield the information you desire and enough statistical power is built into your experiment. In this example, we used a human cell line, with three technical replicates for each control and treatment. The allows us to statistically explore technical variation of one biological sample.

RNA Quality

The **Agilent 2100 Bioanalyzer** is used to assess the quality and quantity of the starting RNA sample to ensure the highest quality data possible. Degraded RNA will result in an incomplete expression profile due to lost message.

A scan of high quality total RNA will show 1 marker peak, 2 ribosomal peaks and a flat baseline.



Sample	Alert	%Dgn/18S	Conc	T18S	T28S	18S	28S	28S/18S
Sample 1		4.14	196	41.45	48	25.77	31.48	1.22
Sample 2	RED	73.27	98	41.75	48	3.03	12.52	4.12

Figure 1: Overlaid electropherogram outputs from the Agilent 2100 Bioanalyzer using the Data Review Software (Agilent Technologies) [3]

Using the **Degradometer**[2] (Ohio State University) the amount of degradation of each sample can be quantified. No Alert (Sample 1) is of highest quality, whereas Red Alert is lowest. The percent of degradation is calculated as a ratio to the 18S peak.

QC Data Analysis

The QC data analysis pipeline starts with visual inspection of the image file in MAS for obvious defects (scratches, bubbles, etc.). Probe signal intensities are qualified statistically for outliers in MAS and dChip. Quality indices are evaluated in Probe Profiler and Bioconductor. When the data is assessed to be technically acceptable, data mining can begin. Typical data mining steps include transformation (log2) , normalization (MAS, MBEI, or RMA), filtering (ANOVA, t-test, or P/M/A flags), ratio of fold changes, clustering and pathway mapping. For QA purposes, the LRGC is interested in the global behavior of arrays within an experiment, and the performance of QC probe sets on the GeneChips.

GCOS is used to view the scanned image for array surface defects and to flag array within outliers. Probe level data, written to the .CEL file, and QC reports are generated for use in downstream analysis programs.

TIGR MeV provides fourteen different clustering methods, can perform t-tests, ANOVAs and contains SAM functionality.

Volcano plots graph log ratios vs. statistical significance, thus displaying values that are statistically increased or decreased at fold change levels.

MA plots display variance across the range of signal intensities.

Bullfrog filters comparison files from MAS and is one way to estimate the FDR.

GeneSpring is an excellent tool for data visualization, normalization as well as filtering, clustering, ANOVA and PCA.

GeneChip QC Metrics

Poly-A RNA Labeling Controls

Known amounts of four exogenous **RNA positive controls** spiked into each sample allow assessment of the overall success of the cRNA probe generation. The control concentrations are staggered, and must be called Present.

Hybridization Controls

Hybridization is monitored by four **spike-in controls** in the hybridization cocktail. These controls must be in ascending order and called Present.

Test3 Array

This less expensive GeneChip contains probe sets for various species. For a given experiment, the **3'/5' ratio** for the housekeeping genes **β-Actin** and **GAPDH** must be <3.0, the **Percentage of genes called present** must be within 10%, the **Background** under 100 and a **Scale factor** less than 3 fold.

Probe Profiler

Probe Profiler v2.0.0 (Corimbia Inc., Berkeley, CA) is used to look at the variance of individual GeneChip indices. This can reveal technically aberrant arrays undesirable for further analysis. Fold change graphs highlighting statistically significant changes between two conditions can also be generated.

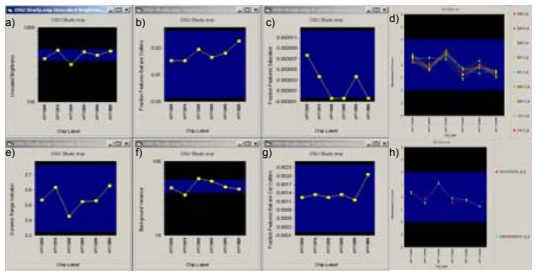


Figure 2: Using Affymetrix .cel files as input, Probe Profiler will display the variation of a) chip Brightness b) Outlier features c) Saturated features d) Hybridization control probes e) Dynamic range f) Background g) MAS array outliers and h) GAPDH and β-Actin genes, in an experiment. Blue bands delineate 2 Standard Deviations.

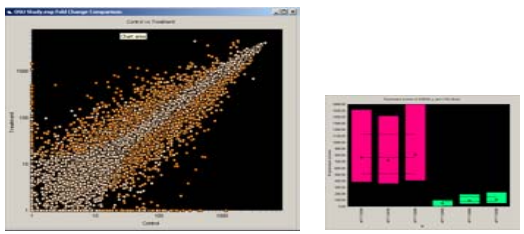


Figure 3: Probe Profiler can also generate a Fold Change scatter plot with statistically significant genes highlighted. Clicking a data point will display the corresponding mean and confidence intervals.

dChip

dChip[3] (Cheng Li and Wing Hung Wong, Harvard) is academic freeware that uses a model based expression algorithm for probe pair normalization. dChip's outlier detection algorithm can warn of probe pair and probe set outliers within and between arrays. dChip can also be used for hierarchical clustering, LDA, PCA, ANOVA, FDR and GoSurfer to visualize GO ontology relationships.

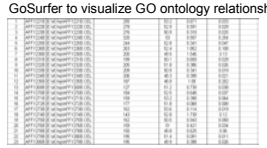


Figure 4: Experiment analysis using PM/MM difference model. Less than 5% outliers is considered acceptable, where an array with > 5% outliers will produce a warning flag.

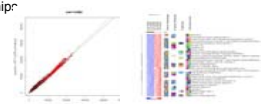


Figure 5: Using the R-com software, the R programming environment can be interfaced to produce graphical visualizations. Pearson correlation coefficients close to 1.0 are desirable for replicates.

Iobion GeneTraffic Uno

GeneTraffic UNO (Iobion Informatics, La Jolla, CA) is a web based client-server analysis program for Affymetrix data. Users can select MAS, MBEI or RMA normalization algorithms, and visualize their data with scatter plots and hierarchical clustering.

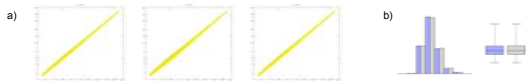


Figure 6: a) Scatter plots of RMA normalized data from three replicates versus mean intensities b) Histograms and box plots of intensity from all arrays in an experiment should have similar mean and ranges.

Bioconductor And R

Bioconductor[4] is an open source and open development software project to provide tools for the analysis and comprehension of genomic data (bioinformatics). There are numerous packages written in R[5] (a language and environment for statistical computing and graphics, similar to S) for use with Affymetrix data.

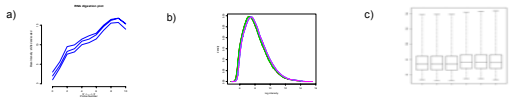


Figure 7: Data quality can be visualized using functions for plotting summarized probe level values. a) RNA degradation plots can display any 5' to 3' trends b) histograms and c) box plots provide graphical summaries of probe level intensities.

References

[1] Corresponding author, email microarray@robarts.ca

[2] Chipping away at the chip bias: RNA degradation in microarray analysis. Auer H, Lyianarachi S, Newsom D, Klisovic MI, Marcucci G, Kornacker (2003) Nature Genetics 35:292-293.

[3] Cheng Li and Wing Hung Wong (2001b) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application, *Genome Biology* 2(8): research0032.1-0032.11

[4, 5] Ross Ihaka and Robert Gentleman. R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics. 1996, 5, 3, 299-314.