Analyzing DNA Microarray Data Using Bioconductor

Sandrine Dudoit and Rafael Irizarry

Short Course on Mathematical Approaches to the Analysis of Complex Phenotypes
The Jackson Laboratory, Bar Harbor, Maine
September 18 - 24, 2002

© Copyright 2002, all rights reserved
Acknowledgements

- **Bioconductor core team**
  - Ben Bolstad, Biostatistics, UC Berkeley
  - Vincent Carey, Biostatistics, Harvard
  - Francois Collin, GeneLogic
  - Leslie Cope, JHU
  - Laurent Gautier, Technical University of Denmark, Denmark
  - Yongchao Ge, Statistics, UC Berkeley
  - Robert Gentleman, Biostatistics, Harvard
  - Jeff Gentry, Dana-Farber Cancer Institute
  - John Ngai Lab, MCB, UC Berkeley
  - Juliet Shaffer, Statistics, UC Berkeley
  - Terry Speed, Statistics, UC Berkeley
  - Yee Hwa (Jean) Yang, Biostatistics, UCSF
  - Jianhua (John) Zhang, Dana-Farber Cancer Institute

- Spike-in and dilution datasets:
  - Gene Brown’s group, Wyeth/Genetics Institute
  - Uwe Scherf’s group, Genomics Research & Development, GeneLogic.

- GeneLogic and Affymetrix for permission to use their data.
References

• **Personal web pages**
  – [http://www.stat.berkeley.edu/~sandrine](http://www.stat.berkeley.edu/~sandrine)
  – [http://www.biostat.jhsph.edu/~ririzarr](http://www.biostat.jhsph.edu/~ririzarr)
    articles and talks on: image analysis; normalization; identification of differentially expressed genes; cluster analysis; classification.

• **Bioconductor** [http://www.bioconductor.org](http://www.bioconductor.org)
  – software and documentation;
  – training materials from short courses;
  – mailing list.

• **R** [http://www.r-project.org](http://www.r-project.org)
  – software; documentation; R Newsletter.
Outline

I. Pre-processing: cDNA microarrays.
II. Pre-processing: Affymetrix GeneChip arrays.
III. Overview of the Bioconductor project.
IV. Object oriented programming: biobase, affy, and marrayXXX packages.
V. Analysis and presentation via web interfaces: genefilter, multtest, and annotate packages.
VI. Bioconductor software demo.
More …

on image analysis, normalization, experimental design, multiple testing, cluster analysis, classification.

Slides from the Bioconductor Summer 2002 short course

[www.bioconductor.org/workshops/Summer02Course/index.html](http://www.bioconductor.org/workshops/Summer02Course/index.html)
Biological question
Differentially expressed genes
Sample class prediction etc.

Experimental design

Microarray experiment

Image analysis

Expression quantification

Normalization

Pre-processing

Estimation

Testing

Clustering

Prediction

Biological verification and interpretation

Role of Statistics
Pre-processing

• cDNA microarrays
  – Image analysis;
  – Normalization.

• Affymetrix oligonucleotide chips
  – Image analysis;
  – Normalization;
  – Expression measures.
Part I. Pre-processing: cDNA microarrays

Sandrine Dudoit and Yee Hwa Yang

© Copyright 2002, all rights reserved
RGB overlay of Cy3 and Cy5 images

- 4 x 4 sectors
- 19 x 21 probes/sector
- 6,384 probes/array
Terminology

- **Target**: DNA hybridized to the array, mobile substrate.
- **Probe**: DNA spotted on the array, aka. spot, immobile substrate.
- **Sector**: collection of spots printed using the same print-tip (or pin), aka. print-tip-group, pin-group, spot matrix, grid.
- The terms *slide* and *array* are often used to refer to the printed microarray.
- **Batch**: collection of microarrays with the same probe layout.
- **Cy3** = Cyanine 3 = green dye.
- **Cy5** = Cyanine 5 = red dye.
Raw data

E.g. Human cDNA arrays

• ~43K spots;
• 16–bit TIFFs: ~ 20Mb per channel;
• ~ 2,000 x 5,500 pixels per image;
• Spot separation: ~ 136um;
• For a “typical” array, the spot area has
  – mean = 43 pixels,
  – med = 32 pixels,
  – SD = 26 pixels.
Image analysis
Image analysis

• The **raw data** from a cDNA microarray experiment consist of pairs of **image files**, 16-bit TIFFs, one for each of the dyes.

• **Image analysis** is required to extract measures of the red and green fluorescence intensities, $R$ and $G$, for each spot on the array.
Image analysis

1. **Addressing.** Estimate location of spot centers.

2. **Segmentation.** Classify pixels as foreground (signal) or background.

3. **Information extraction.** For each spot on the array and each dye
   - foreground intensities;
   - background intensities;
   - quality measures.

→ R and G for each spot on the array.
Segmentation

Adaptive segmentation, SRG  Fixed circle segmentation

Spots usually vary in size and shape.
Seeded region growing

• Adaptive segmentation method.
• Requires the input of seeds, either individual pixels or groups of pixels, which control the formation of the regions into which the image will be segmented. Here, based on fitted foreground and background grids from the addressing step.
• The decision to add a pixel to a region is based on the absolute gray-level difference of that pixel’s intensity and the average of the pixel values in the neighboring region.
• Done on combined red and green images.
• Ref. Adams & Bischof (1994)
Local background

---- GenePix
---- QuantArray
---- ScanAnalyze
Morphological opening

• The image is probed with a structuring element, here, a square with side length about twice the spot-to-spot distance.

• Erosion (Dilation): the eroded (dilated) value at a pixel $x$ is the minimum (maximum) value of the image in the window defined by the structuring element when its origin is at $x$.

• Morphological opening: erosion followed by dilation.

• Done separately for the red and green images.

• Produces an image of the estimated background for the entire slide.
Background matters

Morphological opening

Local background

\[ M = \log_2 R - \log_2 G \quad \text{vs.} \quad A = \frac{\log_2 R + \log_2 G}{2} \]
Quality measures

- **Spot quality**
  - **Brightness**: foreground/background ratio;
  - **Uniformity**: variation in pixel intensities and ratios of intensities within a spot;
  - **Morphology**: area, perimeter, circularity.

- **Slide quality**
  - Percentage of spots with no signal;
  - Range of intensities;
  - Distribution of spot signal area, etc.

- How to use quality measures in subsequent analyses?
Spot image analysis software

• Software package **Spot**, built on the **R** language and environment for statistical computing and graphics.

• Batch automatic addressing.

• Segmentation. **Seeded region growing** (Adams & Bischof 1994): **adaptive** segmentation method, no restriction on the size or shape of the spots.

• Information extraction
  – Foreground. Mean of pixel intensities within a spot.
  – Background. **Morphological opening**: non-linear filter which generates an image of the estimated background intensity for the entire slide.

• Spot quality measures.
Normalization
Normalization

• **Purpose.** Identify and remove the effects of systematic variation in the measured fluorescence intensities, other than differential expression, for example
  – different labeling efficiencies of the dyes;
  – different amounts of Cy3- and Cy5-labeled mRNA;
  – different scanning parameters;
  – print-tip, spatial, or plate effects, etc.
Normalization

• Normalization is needed to ensure that differences in intensities are indeed due to differential expression, and not some printing, hybridization, or scanning artifact.

• Normalization is necessary before any analysis which involves within or between slides comparisons of intensities, e.g., clustering, testing.
Normalization

• The need for normalization can be seen most clearly in self-self hybridizations, where the same mRNA sample is labeled with the Cy3 and Cy5 dyes.

• The imbalance in the red and green intensities is usually **not constant** across the spots within and between arrays, and can vary according to overall spot intensity, location, plate origin, etc.

• These factors should be considered in the normalization.
Single-slide data display

• Usually: \( R \) vs. \( G \)
  \[ \log_2 R \text{ vs. } \log_2 G. \]

• Preferred
  \[ M = \log_2 R - \log_2 G \]
  vs.
  \[ A = (\log_2 R + \log_2 G)/2. \]

• An MA-plot amounts to a 45° counterclockwise rotation of a
  \( \log_2 R \) vs. \( \log_2 G \) plot followed by scaling.
Self-self hybridization

\[ \log_2 R \text{ vs. } \log_2 G \]

\[ M = \log_2 R - \log_2 G, \quad A = \frac{\log_2 R + \log_2 G}{2} \]
Self-self hybridization

Robust local regression within sectors (print-tip-groups) of intensity log-ratio $M$ on average log-intensity $A$.

$$M = \log_2 R - \log_2 G, \quad A = \frac{\log_2 R + \log_2 G}{2}$$
Swirl zebrafish experiment

- **Goal.** Identify genes with altered expression in Swirl mutants compared to wild-type zebrafish.
- 2 sets of dye-swap experiments (n=4).
- Arrays:
  - 8,448 probes (768 controls);
  - 4 x 4 grid matrix;
  - 22 x 24 spot matrices.
- Data available in Bioconductor package `marrayInput`. 
Diagnostic plots

• **Diagnostics plots** of spot statistics
  E.g. red and green log-intensities, intensity log-ratios M, average log-intensities A, spot area.
  – Boxplots;
  – 2D spatial images;
  – Scatter-plots, e.g. MA-plots;
  – Density plots.

• **Stratify** plots according to layout parameters, e.g. print-tip-group, plate.
2D spatial images

Swirl 83 array: Cy3 background

Swirl 83 array: Cy5 background

Cy3 background intensity

Cy5 background intensity
2D spatial images

Swirl 93 array: pre-normalization log-ratio M

Intensity log-ratio, M
Boxplots by print-tip-group

Swirl 93 array: pre-normalization log-ratio M

Intensity log-ratio, M
MA-plot by print-tip-group

\[ M = \log_2 R - \log_2 G, \quad A = \frac{\log_2 R + \log_2 G}{2} \]
Location normalization

\[ \log_2 R/G \leftarrow \log_2 R/G - L(\text{intensity}, \text{sector}, \ldots) \]

- **Constant normalization.** Normalization function \( L \) is **constant** across the spots, e.g. mean or median of the log-ratios \( M \).

- **Adaptive normalization.** Normalization function \( L \) depends on a number of **predictor variables**, such as spot intensity \( A \), sector, plate origin.
Location normalization

• The normalization function can be obtained by robust locally weighted regression of the log-ratios M on predictor variables.
  E.g. regression of M on A within sector.

• Regression method: e.g. lowess or loess (Cleveland, 1979; Cleveland & Devlin, 1988).
Location normalization

- **Intensity-dependent normalization.**
  Regression of M on A (*global loess*).

- **Intensity and sector-dependent normalization.**
  Same as above, for each sector separately (*within-print-tip-group loess*).

- **2D spatial normalization.**
  Regression of M on 2D-coordinates.

- **Other variables:** time of printing, plate, etc.

- **Composite normalization.** Weighted average of several normalization functions.
2D images of L values

Global median normalization

Global loess normalization

Within-print-tip-group loess normalization

2D spatial normalization
2D images of normalized M-L

Global median normalization

Global loess normalization

Within-print-tip-group loess normalization

2D spatial normalization
Boxplots of normalized M-L

Global median normalization

Global loess normalization

Within-print-tip-group loess normalization

2D spatial normalization
MA-plots of normalized M-L

- Global median normalization
- Global loess normalization
- Within-print-tip-group loess normalization
- 2D spatial normalization
Normalization

• Within-slide
  – **Location** normalization - additive on log-scale.
  – **Scale** normalization - multiplicative on log-scale.
  – **Which spots** to use?

• Paired-slides (dye-swap experiments)
  – Self-normalization.

• Between-slides.
Scale normalization

• The log-ratios M from different sectors, plates, or arrays may exhibit different spreads and some scale adjustment may be necessary.

\[
\log_2 \frac{R}{G} \leftarrow \frac{\left(\log_2 \frac{R}{G} - L\right)}{S}
\]

• Can use a robust estimate of scale such as the median absolute deviation (MAD)

\[
\text{MAD} = \text{median} | M - \text{median}(M) |.
\]
Scale normalization

• For print-tip-group scale normalization, assume all print-tip-groups have the same spread in $M$.

• Denote true and observed log-ratio by $\mu_{ij}$ and $M_{ij}$, resp., where $M_{ij} = a_i \mu_{ij}$, and $i$ indexes print-tip-groups and $j$ spots. Robust estimate of $a_i$ is

$$\hat{a}_i = \frac{MAD_i}{\sqrt[\prod_{i=1}^{I}MAD_i}}$$

where $MAD_i$ is MAD of $M_{ij}$ in print-tip-group $i$.

• Similarly for between-slides scale normalization.
Which genes to use?

• **All spots on the array:**
  – Problem when many genes are differentially expressed.

• **Housekeeping genes:** Genes that are thought to be constantly expressed across a wide range of biological samples (e.g. tubulin, GAPDH).
  Problems:
  – sample specific biases (genes are actually regulated),
  – do not cover intensity range.
Which genes to use?

- **Genomic DNA titration series:**
  - fine in yeast,
  - but weak signal for higher organisms with high intron/exon ratio (e.g. mouse, human).

- **Rank invariant set** (Schadt et al., 1999; Tseng et al., 2001): genes with same rank in both channels. Problems: set can be small.
Microarray sample pool

- **Microarray Sample Pool, MSP**: Control sample for normalization, in particular, when it is not safe to assume most genes are equally expressed in both channels.
- MSP: pooled all 18,816 ESTs from RIKEN release 1 cDNA mouse library.
- Six-step dilution series of the MSP.
- MSP samples were spotted in middle of first and last row of each sector.
- Ref. Yang et al. (2002).
Microarray sample pool

MSP control spots

• provide potential probes for every target sequence;
• are constantly expressed across a wide range of biological samples;
• cover the intensity range;
• are similar to genomic DNA, but without intron sequences → better signal than genomic DNA in organisms with high intron/exon ratio;
• can be used in composite normalization.
Microarray sample pool

MSP
Rank invariant
Housekeeping
Tubulin, GAPDH
Dye-swap experiment

• Probes
  – 50 distinct clones thought to be differentially expressed in apo AI knock-out mice compared to inbred C57Bl/6 control mice (largest absolute t-statistics in a previous experiment).
  – 72 other clones.

• Spot each clone 8 times.

• Two hybridizations with dye-swap:
  Slide 1: trt → red,   ctl → green.
  Slide 2: trt → green,  ctl → red.
Dye-swap experiment
Self-normalization

• Slide 1, $M = \log_2 (R/G) - L$
• Slide 2, $M' = \log_2 (R'/G') - L'$

Combine by **subtracting** the normalized log-ratios:

$$M - M' = \frac{[ (\log_2 (R/G) - L) - (\log_2 (R'/G') - L') ]}{2}$$

$$\approx \frac{[ \log_2 (R/G) + \log_2 (G'/R') ]}{2}$$

$$\approx \frac{[ \log_2 (RG'/GR') ]}{2}$$

provided $L = L'$.

**Assumption: the normalization functions are the same for the two slides.**
Checking the assumption

MA-plot for slides 1 and 2
Result of self-normalization

\[(M - M')/2 \text{ vs. } (A + A')/2\]
Summary

Case 1. Only a few genes are expected to change.

Within-slide
   – Location: intensity + sector-dependent normalization.
   – Scale: for each sector, scale by MAD.

Between-slides
   – An extension of within-slide scale normalization.

Case 2. Many genes are expected to change.

   – Paired-slides: Self-normalization.
   – Use of controls or known information, e.g. MSP.
   – Composite normalization.
Pre-processing cDNA microarray data

- **marrayClasses**: class definitions for cDNA microarray data;
  basic methods for manipulating microarray objects: printing, plotting, subsetting, class conversions, etc.

- **marrayInput**: reading in intensity data and textual data describing probes and targets;
  automatic generation of microarray data objects;
  widgets for point & click interface.

- **marrayPlots**: diagnostic plots.

- **marrayNorm**: robust adaptive location and scale normalization procedures.