cDNA Microarray
Data Analysis
with BioConductor packages

Nolwenn Le Meur

October MiniCourse
Copyright 2006
Microarrays Experiment

Experimental Design

Image Analysis

Quality Assessment

Pre-processing

Background Correction

Normalization

Summarization

Analysis
Outline

- Data acquisition & Pre-processing (chap. 4)
  - Image analysis
  - Quality assessment
  - Pre-processing

- Lab: case studies (chap 4)
  - marray & arrayQuality (Y.H Yang & A.C. Paquet)
**Terminology**

- **Target**: DNA hybridized to the array, mobile substrate.
- **Probe**: DNA spotted on the array (spot).
- **print-tip-group**: collection of spots printed using the same print-tip (or pin), aka. grid.

- **G, Gb**: Cy 3 signal and background intensities
- **R, Rb**: Cy5 signal and background intensities
- **M** = \( \log_2(R) - \log_2(G) \)
- **A** = \( \frac{1}{2}(\log_2(R) + \log_2(G)) \)
Image Analysis

1. Location

2. Segmentation

3. Quantification

Raw data
Quality Filtering

- Background
- Foreground
Quality Assessment

For at the probe-level:

- **Sources**
  - faulty printing, uneven distribution, contamination with debris, magnitude of signal relative to noise, poorly measured spots

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

- **Action**
  - use weights for measurements to indicate reliability in later analysis.
  - set measurements to NA (missing values)
Quality Assessment

For each array

- **Problems**
  - array fabrication defect
  - problem with RNA extraction
  - failed labeling reaction
  - poor hybridization conditions
  - faulty scanner

- **Quality measures**
  - Percentage of spots with no signal (~30% excluded spots)
  - Range of intensities
  - \((\text{Av. Foreground})/(\text{Av. Background}) > 3\) in both channels
  - Distribution of spot signal area
Quality Assessment

For each array:

- **Visual inspection**
  - hairs, dust, scratches, air bubbles, dark regions, regions with haze

- **Diagnostics plots** of spot statistics
  - e.g. R and G log-intensities, M, A, spot area.
    - 2D spatial images;
    - ECDF plots;
    - Boxplots;
    - Scatter-plots;
    - Density plots.

- **Stratify** plots according to layout parameters, *e.g.* print-tip-group, plate.
Spatial Effects – Image Plots

- R
- Rb
- R-Rb
- color scale by rank
- Print-tip
- Air bubble
- washing
Spatial Effects

1 pin → 1 block
Spotting Pin Quality Decline

after delivery of $5 \times 10^5$ spots

after delivery of $3 \times 10^5$ spots
Print-tip Effects – ECDF plot

\[ F(q) \]

\[ q \text{ (log-ratio)} \]

\[ \log(\text{fg.green}/\text{fg.red}) \]
Print-tip Effects - Boxplots

Q3 75% quantile
median
Q1 25% quantile
Diagnostic plot with arrayQuality
Data Exploration with \textit{limma}
Quality Assessment: Summary

For each spot:
- weight

For each array:
- Diagnostics plots
- Stratify
- Controls

BioC packages:
- arrayQuality
- arrayMagic
- ...
Outline

• Data acquisition & Pre-processing (chap. 4)
  – Image analysis
  – Quality assessment
  – Pre-processing

• Lab : case studies (chap 4)
  – marray & arrayQuality (Y.H Yang & A.C. Paquet)
Variance-Bias trade off

<- Bias

Accuracy ->

Variance ->

Precision <-
Sources of Variation

- RNA extraction
- reverse transcription
- labeling efficiencies
- Scanner settings
- PCR
- DNA concentration
- Printing or pin
- cross-hybridization
- ...

**Systematic**
- similar effect on many measurements
- corrections can be estimated from data

**Stochastic**
- too random to be explicitly accounted for
- “noise”

**Calibration**

**Error Model**
Background Correction

- none
- subtraction, movingmin
  - Minimun, edwards, normexp, …

- More details … limma
  >?backgroundCorrect
Background Correction

![Graphs showing different background correction methods](image)

- None
- Subtraction
- Normexp
Background Correction
Why Normalize?

Theory

Cy5 vs Cy3

10000
1000
100

Cy5

Cy3

Reality

Raw Data - Cy5 vs Cy3

100000
10000
1000
100

Cy5 Raw

Cy3 Raw
Normalization

Identify and remove the effects of systematic variation

- Normalization is closely related to quality assessment. In a ideal experiment, no normalization would be necessary, as the technical variations would have been avoided.

- 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 slide comparisons of intensities, e.g., clustering, testing.
Normalization methods

- median
- loess
- 2D loess
- print-tip loess
- variance stabilisation
- ...

Two channel normalization

- **Location**: centers log-ratios around zero using $A$ and spatial dependent bias
Two channels normalization

slide S  Print-tip lowess  slide S
Two channels normalization

- **Location:** centers log-ratios around zero using A and spatial dependent bias

- **Scale:** adjust for different in scale between multiple arrays
One channel normalization

- As technology improves the spot-to-spot variation is reduced
- Development of normalization techniques that work on the absolute intensities

Ex: quantile normalization (\textit{limma})

variance stabilization (\textit{vsn})
Quantile Normalization

Before

After

Bolstand et al. (2003)
Variance Statibilizing Transformation

- log-transformation is replaced by a arcsinh transformation
  - Meaningful around 0
  - Original intensities may be negatives

- Estimation of transformation parameters (location, scale) based on Maximum Likelihood paradigm
- vsn–normalized data behaves close to the normal distribution

(Huber et al. 2004)
Variance stabilization (vsn)
Preprocessing: Summary

For each array:
- Background correction or not
- Normalization: bias-variance trade-off
- Diagnostic plots

BioC packages:
- marray
- limma
- ...

## BioC Task View: TwoChannel

### 24 packages (18 Bioc1.8)

### Packages in view

<table>
<thead>
<tr>
<th>Package</th>
<th>Maintainer</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>aroma.light</td>
<td>Henrik Bengtsson</td>
<td>Light-weight methods for normalization and visualization of microarray data using only basic R data types</td>
</tr>
<tr>
<td>arrayMagic</td>
<td>Andreas Buesa</td>
<td>two-colour cDNA array quality control and preprocessing</td>
</tr>
<tr>
<td>arrayQuality</td>
<td>A. Paquette</td>
<td>Assessing array quality on spotted arrays</td>
</tr>
<tr>
<td>beadarraySNP</td>
<td>Ian Cossing</td>
<td>Normalization and reporting of Illumina SNP bead arrays</td>
</tr>
<tr>
<td>bridge</td>
<td>Raphael Gottardo</td>
<td>Bayesian Robust Inference for Differential Gene Expression</td>
</tr>
<tr>
<td>convert</td>
<td>Yee Hwa (Jian) Yang</td>
<td>Convert Microarray Data Objects</td>
</tr>
<tr>
<td>cpea</td>
<td>James W. MacDonald</td>
<td>Functions to perform cancer outlier profile analysis</td>
</tr>
<tr>
<td>daMA</td>
<td>Jobst Landgraf</td>
<td>Efficient design and analysis of factorial two-colour microarray data</td>
</tr>
<tr>
<td>genesAssay</td>
<td>IFC Development Team</td>
<td>Microarray Analysis tool</td>
</tr>
<tr>
<td>GEOquery</td>
<td>Sean Davis</td>
<td>Get data from NCBI Gene Expression Omnibus (GEO)</td>
</tr>
<tr>
<td>irma</td>
<td>Gordon Smyth</td>
<td>Linear Models for Microarray Data</td>
</tr>
<tr>
<td>irmaGUI</td>
<td>Keith Batcher</td>
<td>GUI for irma package</td>
</tr>
<tr>
<td>maeDE</td>
<td>Johannes Faller</td>
<td>Microarray database and utility functions for microarray data analysis</td>
</tr>
<tr>
<td>limma</td>
<td>Pierre Nicol</td>
<td>CGH Micro-Array NO normalization</td>
</tr>
<tr>
<td>limma</td>
<td>Yee Hwa (Jian) Yang</td>
<td>Exploratory analysis for two-color spotted microarray data</td>
</tr>
<tr>
<td>mlimma</td>
<td>Luca Laurentini</td>
<td>Spatial and intensity based normalization of cDNA microarray data based on robust neural nets</td>
</tr>
<tr>
<td>microDE</td>
<td>N Dean</td>
<td>Normal Uniform Differential Gene Expression detection</td>
</tr>
<tr>
<td>OILN</td>
<td>Matthias Fuchsik</td>
<td>Optimized local intensity-dependent normalization of two-color microarrays</td>
</tr>
<tr>
<td>OILNrun</td>
<td>Matthias Fuchsik</td>
<td>Graphical user interface for OILN</td>
</tr>
<tr>
<td>rama</td>
<td>Raphael Gottardo</td>
<td>Robust Analysis of Micro Arrays</td>
</tr>
<tr>
<td>ramaCGH</td>
<td>Mike Smith</td>
<td>Segmentation, normalization and processing of aCGH data</td>
</tr>
<tr>
<td>spotSegmentation</td>
<td>Chris Fraley</td>
<td>Microarray Spot Segmentation and Gridding for Blocks of Microarray Spots</td>
</tr>
<tr>
<td>spotNorm</td>
<td>Yuxiang Xiao</td>
<td>Stepwise normalization functions for cDNA microarrays</td>
</tr>
<tr>
<td>tmm</td>
<td>Wolfgang Huber</td>
<td>Variance stabilization and calibration for microarray data</td>
</tr>
</tbody>
</table>