Microarray Data Analysis

Data quality assessment for microarrays
Microarray studies life cycle

- Biological question
- Experimental design
- Microarray experiment
- Image analysis
- Normalization

Here we are

- Analysis
  - Estimation
  - Testing
  - Clustering
  - Discrimination

Biological verification and interpretation
Looking at microarray data
Diagnostic Plots

Was the experiment a success?
Exploring experimental results

• Microarray experiments generate huge quantities of data.
• It is hard to decide if things “seem to be all right” just by looking at the numbers.
• Standard statistical approach: use plots.
  • Show all data together.
  • Highlight structures,
  • May help detect problems (“unusual patterns”)

Diagnostic plots for microarrays

- Microarray data usually considered at two levels
  - Low-level: Data directly coming from the scanner
  - High-level: Processed from low-level data. Expression values, normalized or not.
- Some diagnostic plots may differ between one and two color arrays, specially for looking at low level values.
- Other may be used for any type of arrays or for any level.
- Any classification may be misleading.
# Diagnostic plots

<table>
<thead>
<tr>
<th>Microarray type</th>
<th>One color</th>
<th>Two color</th>
<th>General</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data type</strong></td>
<td><strong>Low level</strong></td>
<td><strong>High-level</strong></td>
<td></td>
</tr>
<tr>
<td>1col: (probe, probe-set)</td>
<td>Layout image</td>
<td>MA-Plots</td>
<td>PCA</td>
</tr>
<tr>
<td>2col: (single channel)</td>
<td>Degradation plots</td>
<td>Scatterplots R,G</td>
<td>Histogram/Density</td>
</tr>
<tr>
<td></td>
<td>Density plots</td>
<td>MA-plot</td>
<td>Boxplot</td>
</tr>
<tr>
<td></td>
<td>Probeset plots</td>
<td>Signal2Noise plots</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Layout image (G, R)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>General</strong></td>
<td></td>
</tr>
<tr>
<td>1 col: Relative expression</td>
<td><strong>High-level</strong></td>
<td><strong>General</strong></td>
<td></td>
</tr>
<tr>
<td>2 col: Absolute expression</td>
<td>MA-Plots</td>
<td>PCA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Model-based plots</td>
<td>Scatterplots R,G</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(NUSE, RLE, Residual)</td>
<td>MA-plot</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Signal2Noise plots</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Layout image (log ratios)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>General</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>PCA</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Histogram/Density Boxplot</td>
<td></td>
</tr>
</tbody>
</table>
Diagnostic plots for affy chips
Image plots for affymetrix chips

read from file: HIVControl4A.CEL.gz

read from file: HIVControl4A.CEL.gz

read from file: HIVControl4B.CEL.gz

read from file: HIVControl4B.CEL.gz
MA-plot for GeneChip arrays (1 color)

\[ M = \log_2(\text{MT}) - \log_2(\text{WT}) \]

\[ A = \frac{\log_2(\text{MT} \times \text{WT})}{2} \quad \text{(signal strength)} \]
Box plots

This plot can be used in both one and two-color arrays.
Density plots
Digestion or (degradation) plots
Distance between arrays
(done on normalized data)

This plot can be used in both one and two-color arrays
Principal components (PCA) plot

This plot can be used in both one and two-color arrays.
Diagnostic plots for two color arrays
Red / Green overlay images

• Start by looking at the slides

Co-registration and overlay offers a quick visualization, providing information on
• colour balance,
• uniformity of hybridization,
• spot uniformity,
• background, and
• artifacts such as
  • dust or
  • scratches

Bad: high bg

Good: low bg
Scatterplots: always log*, always rotate

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

\[ M = \log_2 \frac{R}{G} \text{ vs } A = \log_2 \sqrt{RG} \]

* Other transformations can provide improvement
MA-plot for spotted arrays (2 colors)

Cy3/5-cDNA or aRNA

Mutant (MT)

Wild Type (WT)

Spot

MT and WT intensity for each probe

$M = \log_2(\text{MT/WT})$

$A = \log_2(\text{MT*WT}) / 2$

(signal strength)
Images with high background tend to have lower log_{2}(signal/noise) ratios
Spatial plots for slide backgrounds
Quality between slides
Tools for automatic exploration

• Several packages available in bioconductor allow an easy exploration and quality control:
  • arrayQualityMetrics (state of the art, today)
  • affycoretools
  • simpleaffy
  • Most packages for two colour have been deprecated.