Quality control and normalization of microarray data

Lara Lusa

Istituto Nazionale per lo Studio e la Cura dei Tumori
Fondazione Istituto FIRC di Oncologia Molecolare (IFOM)
Milano

Lara.Lusa@ifom-ieo-campus.it
• After ....
  - Study design (*StatCourse*, day 3)
  - Performance of the experiment (*BioCourse*, day 3)
    • Sample preparation
    • Hybridization
    • Image analysis *(not covered)*

**Quality control and normalization of microarray data**

• Before...
  - Class comparison (*StatCourse*, day 1)
  - Class prediciton  (*StatCourse*, day 2 and day 3)
  - Class discovery   (*StatCourse*, day 2)
Outline

• Why quality control and normalization are needed in microarray experiments?

• Affymetrix GeneChip and spotted arrays
  - Why do they need distinct QC and normalization methods?

• (Quantification of gene expression)
• Quality control
• Normalization
Why QC?

• Signal intensities can be very noisy and some data might not be reliable
  - a probe on one array: **Probe level**
  - a probe on all arrays: **Gene level**
  - many probes on one array: **Array level**

• Using data with many outliers can lead to biased and unstable results...
Why QC?

- What can you do once you decide that some data do not satisfy your quality control criteria?
  - Exclude them
  - Impute values to excluded observations
  - Use model-based approaches

- How much noise is acceptable and how can you quantify it?
  - A lot of useful info can be derived from image analysis software...
  - ...but to a great extent it is a subjective matter and requires careful inspection of the data
  - ...anyhow some guidelines can be given
Why normalization?

Experimental artifacts

Systematic biases and random variation unrelated to the biology of the samples

Biased results

Need to remove or adjust for systematic biases

Imbalance between RNA amounts
RNA amplification
RNA degradation
RT efficiency
Settings of PMT voltage
Dye incorporation
Dye fading...

Why different QC and normalization procedures for Affy and spotted arrays?

• Gene expression is measured in different ways

• The sources of systematic biases are different

• The image analysis is performed in a different way, with different algorithms and different outputs
Affymetrix vs spotted arrays

- **Probe = “Probe set”**
  - 1 probe = 25-mer oligo
  - 1 probe set =
    - 11-20 Perfect Match probes (PM)
    - 11-20 Mismatch probes (MM)

- 1 gene = 1 or more probe sets = 11-20 probe pairs
  - Due to complex chip design, many genes are represented by more than 1 probe set (with different sequences)
    - **ALWAYS IGNORED IN THE ANALYSIS!**

- **Probe = clone or Probe=long oligo**
  - clone = physical copy of a gene, fragment of cDNA amplified by PCR
  - clone = hundred base pairs long
  - long oligo = as Affy probes, but longer (60-80 bp)

- 1 gene = 1 clone = 1 spot
  - Sometimes, replicates of the same clone for quality control purposes
Affymetrix vs spotted arrays

• 1 chip = 1 sample = 1 fluorochrome = 1 image

• 1 array = 2 samples = 2 fluorescent dyes = 2 images
  - Competitive hybridization

• Different print tips are used to deposit the probe on the array
Some background: how do you quantify the gene expression?

**Spotted arrays**

- FG = foreground intensity (by channel)
- BG = background intensity (by channel)
- Signal =
  - FG
  - FG-BG (local)
  - FG-BG (regional)

**Probe-level summary**

- Ratio: $\log_2(\text{Signal}_{ch1}/\text{Signal}_{ch2})$
- Model based: $\log_2(\text{Signal}_{ch1})$ and $\log_2(\text{Signal}_{ch2})$
  - ANOVA models
Some background: how do you quantify the gene expression?

**Affymetrix MAS5-GCOS**

- For all probes (PM and MM)
  - Grid alignment
  - 10x10 pixels -> remove 36 outer pixels -> 8x8
  - FG=75th percentile of the 8x8 pixel values
  - BG= Average of the lowest 2% probe cell values
  - Signal= FG-BG

- Summarize over probes to get the probe set expression value
  - Anti-log of Robust average (Tuckey biweight) of log(PM-CT), CT=MM if PM>MM, adjusted to be <PM if PM<MM

http://www.affymetrix.com
Some background: how do you quantify the gene expression?

**Affymetrix: alternatives to MAS5-GCOS**

- **MAS5** improved **MAS4** (negative intensities)
- But many alternatives are available, the most used are the model-based
  - **MBEI** (dChip, Li & Wong 2001) (models the sensitivity of single probes)
  - **RMA** (Irizarry et al. 2003) (no MM!, different BG and quantile normalization)
  - **GC RMA** (Wu et al, 2004) (BG depends on G-C content of the probe)
QC: what can you do?

1. Visual inspection of images for identification of
   - Poor quality arrays
   - Poor quality regions on the arrays

2. Use outputs from image analysis to
   - Eliminate probes that are flagged as poor quality
   - Identify other probes that might be problematic
   - STATISTICIAN’S IDEAL: use QC information in a model based approach, for instance to downweight less reliable info (weighted regression)
     • More complicated and for certain kind of problems not really useful

3. Use data exploration tools to identify sources of systematic biases
QC1 Visual inspection of image

DAT file

TIFF file
QC1 Visual inspection of image
QC1 Visual inspection of image
QC2 From image analysis software

- Many of the problems are identified at image analysis, manually and automatically
- Omit
  - Flagged spots
  - “Masked” probes
- Still, more QC can be performed before starting normalization
QC2 From image analysis software
Spotted arrays: problematic spots

- Spots with small size
  - Min # of pixels used to detect the spot
- Spots with large relative background
  - Signal and BG intensities are estimated by image analysis (locally)
  - Signal=FG-BG (mean or median values)
  - But the effect of BG on Signal might not be additive!
  - Spots with BG~FG can be problematic
  - Exclude spots with overlapping distributions between FG and BG
    - High % of FG pixels: FG < mean(BG)+1SD(BG)
    - Detailed information, provided in gpr files (GenePix)
Spotted arrays: problematic spots

- Spots with weak signal
  - Signals are usually log-transformed
    - No information is lost with transformation
    - More symmetric distributions
    - Natural scale for describing fold changes
      - \( \log_2(2000/500)=2 \)  \( 2000/500=4 \)
      - \( \log_2 (500/2000)=-2 \)  \( 500/2000=1/4 \)
  - Additive noise increased at weak signals
    - \( \log_2 [(1000+30)/(1000-30)]=0.09 \)
    - \( \log_2 [(50+30)/(50-30)]=2 \)
      - Same amount of noise (30), very different FC!
  - It is the most common “flagging reason”
Table 5.1. Recommended Filtering for Low Intensity Spots on Dual-label Arrays

<table>
<thead>
<tr>
<th>Red Signal</th>
<th>Green Signal</th>
<th>Description</th>
<th>Recommended log ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>$FR - BR &lt; \lambda_{Low}$</td>
<td>$FG - BG &lt; \lambda_{High}$</td>
<td>Red signal too low, Green signal not high enough to be conclusive</td>
<td>EXCLUDE</td>
</tr>
<tr>
<td>$FR - BR &lt; \lambda_{High}$</td>
<td>$FG - BG &lt; \lambda_{Low}$</td>
<td>Green signal too low, Red signal not high enough to be conclusive</td>
<td>EXCLUDE</td>
</tr>
<tr>
<td>$FR - BR &lt; \lambda_{Low}$</td>
<td>$FG - BG &gt; \lambda_{High}$</td>
<td>Red signal too low, but Green signal conclusively higher than green</td>
<td>$\log_2 \left( \frac{\lambda_{Low}}{FG - BG} \right)$</td>
</tr>
<tr>
<td>$FR - BR &gt; \lambda_{High}$</td>
<td>$FG - BG &lt; \lambda_{Low}$</td>
<td>Green signal too low, but Red signal conclusively higher than green</td>
<td>$\log_2 \left( \frac{FR - BR}{\lambda_{Low}} \right)$</td>
</tr>
<tr>
<td>$FR - BR &gt; \lambda_{Low}$</td>
<td>$FG - BG &gt; \lambda_{Low}$</td>
<td>Both Red and Green are high enough to be used directly</td>
<td>$\log_2 \left( \frac{FR - BR}{FG - BG} \right)$</td>
</tr>
</tbody>
</table>

QC2 From image analysis software
Spotted arrays: problematic spots

• Saturated spots
  - Intensities can reach values above limits of representable values and be truncated (max=$2^{16}$)
  - These values are informative only when just one of the channels is saturated
  - Techniques similar to those for censored survival times can be used to deal with it
    • Wit & McClure, 2003, Bioinformatics
## QC3 Overall array quality

**Spotted arrays: a report**

<table>
<thead>
<tr>
<th>Array</th>
<th>SmallDiam</th>
<th>Fpix</th>
<th>R. Sat</th>
<th>G. Sat</th>
<th>R. SatG notSat</th>
<th>G. SatR notSat</th>
<th>R. Low</th>
<th>G. Low</th>
<th>R. LowG High</th>
<th>G. LowR High</th>
<th>R. B. Go overlap</th>
<th>G. B. Gov overlap</th>
<th>GP. flags 50</th>
<th>GP. flags 75</th>
<th>GP. flags 100</th>
<th>Total Flagged</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.35</td>
<td>0</td>
<td>0.6</td>
<td>0.4</td>
<td>0.21</td>
<td>0.01</td>
<td>13.4</td>
<td>20.9</td>
<td>0.27</td>
<td>2.08</td>
<td>24.3</td>
<td>33.51</td>
<td>3.78</td>
<td>5.52</td>
<td>0.06</td>
<td>39.01</td>
</tr>
<tr>
<td>2</td>
<td>1.78</td>
<td>0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.03</td>
<td>0</td>
<td>41.9</td>
<td>50.6</td>
<td>0.54</td>
<td>2.41</td>
<td>52.52</td>
<td>34.62</td>
<td>8.01</td>
<td>5.49</td>
<td>0.52</td>
<td>59.2</td>
</tr>
<tr>
<td>3</td>
<td>0.64</td>
<td>0</td>
<td>0.3</td>
<td>0.2</td>
<td>0.16</td>
<td>0.02</td>
<td>17.3</td>
<td>23.8</td>
<td>0.34</td>
<td>2.4</td>
<td>30.59</td>
<td>42.15</td>
<td>7.86</td>
<td>5.52</td>
<td>0.04</td>
<td>46.86</td>
</tr>
<tr>
<td>4</td>
<td>1.35</td>
<td>0</td>
<td>0.1</td>
<td>0.1</td>
<td>0.02</td>
<td>0.03</td>
<td>34.2</td>
<td>42.2</td>
<td>0.6</td>
<td>2.73</td>
<td>40.8</td>
<td>35.95</td>
<td>7.28</td>
<td>5.49</td>
<td>0.29</td>
<td>51.37</td>
</tr>
<tr>
<td>5</td>
<td>1.01</td>
<td>0</td>
<td>0.3</td>
<td>0.6</td>
<td>0.06</td>
<td>0.35</td>
<td>21.1</td>
<td>24.3</td>
<td>0.56</td>
<td>1.91</td>
<td>29.21</td>
<td>29.52</td>
<td>5.24</td>
<td>5.46</td>
<td>1.15</td>
<td>38.01</td>
</tr>
<tr>
<td>6</td>
<td>0.88</td>
<td>0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.01</td>
<td>0.01</td>
<td>23.2</td>
<td>37.9</td>
<td>0.24</td>
<td>3.86</td>
<td>31.73</td>
<td>51.8</td>
<td>11.14</td>
<td>5.52</td>
<td>0</td>
<td>54.09</td>
</tr>
<tr>
<td>7</td>
<td>1.12</td>
<td>0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.04</td>
<td>0</td>
<td>50.5</td>
<td>49.5</td>
<td>1.72</td>
<td>0.69</td>
<td>59.13</td>
<td>50.95</td>
<td>5.2</td>
<td>5.42</td>
<td>2</td>
<td>68.27</td>
</tr>
<tr>
<td>8</td>
<td>0.37</td>
<td>0</td>
<td>0.9</td>
<td>0.1</td>
<td>0.79</td>
<td>0</td>
<td>14.5</td>
<td>19.8</td>
<td>0.49</td>
<td>1.01</td>
<td>24</td>
<td>34.25</td>
<td>4.27</td>
<td>5.5</td>
<td>0.25</td>
<td>39.77</td>
</tr>
<tr>
<td>9</td>
<td>0.49</td>
<td>0</td>
<td>0.9</td>
<td>0.1</td>
<td>0.82</td>
<td>0</td>
<td>15.8</td>
<td>18.6</td>
<td>0.62</td>
<td>0.76</td>
<td>25.56</td>
<td>37.05</td>
<td>5.89</td>
<td>5.52</td>
<td>0.06</td>
<td>42.78</td>
</tr>
<tr>
<td>10</td>
<td>1.35</td>
<td>0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.13</td>
<td>0</td>
<td>46.6</td>
<td>47.1</td>
<td>1.48</td>
<td>0.64</td>
<td>46.02</td>
<td>44.69</td>
<td>7.42</td>
<td>5.49</td>
<td>0.31</td>
<td>59.41</td>
</tr>
<tr>
<td>11</td>
<td>0.99</td>
<td>0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.06</td>
<td>0</td>
<td>27.3</td>
<td>32.4</td>
<td>0.91</td>
<td>1.3</td>
<td>38.47</td>
<td>56.06</td>
<td>14.59</td>
<td>5.51</td>
<td>0.08</td>
<td>59.32</td>
</tr>
<tr>
<td>12</td>
<td>0.7</td>
<td>0</td>
<td>0.9</td>
<td>0.4</td>
<td>0.53</td>
<td>0.04</td>
<td>19.2</td>
<td>20.2</td>
<td>0.96</td>
<td>0.7</td>
<td>28.64</td>
<td>34.79</td>
<td>2.99</td>
<td>5.45</td>
<td>0.98</td>
<td>42.86</td>
</tr>
</tbody>
</table>
QC3 Overall array quality
Spotted arrays

Yang et al. 2002

\[ M = \log_2(\text{GREEN signal}) - \log_2(\text{RED signal}) \]
\[ A = (\log_2(\text{GREEN signal}) + \log_2(\text{RED signal}))/2 \]

Can be plotted also for Affymetrix data using a reference array, a “median” array or plotting all the possible combinations
MA plot Before and After QC

![MA plot images]

- MA plot before QC: A scatter plot showing the distribution of data points before quality control (QC) with a trend line indicating the behavior of the data.
- MA plot after QC: A similar scatter plot after applying QC, showing an improved distribution pattern compared to before QC.
QC3 Overall array quality
Spotted arrays: from Bioconductor
Box and whiskers plot

\[ \text{IQR} = 3 \text{rd Quantile} - 1 \text{st quantile} \]
(heigth of the box)
QC3 Overall array quality
Spotted arrays: from Bioconductor
QC3 Overall array quality
Spotted arrays: from Bioconductor
QC3 Overall array quality
Spotted arrays: from Bioconductor

Batch effect

StatCourse Day2
QC2 From image analysis software Affymetrix

- **CDF file** = Chip description
- **DAT file** = Image
- **CEL file** = Raw data
- **CHP file** =
  - Intensity values according to Affymetrix statistical algorithm (MAS5, GCOS)
  - Present/Absent calls
- **RPT file** = quality report file
QC2 From image analysis software Affymetrix

- **MAS5** provides a detection algorithm that gives Present/Absent/Marginal calls
  - Absent = gene is not expressed
  - Based on Wilcoxon signed ranked test between PM and MM intensities
  - Assumes independence of the probe pairs
  - p-values not strictly valid
  - But useful for screening genes (always A)

- **dChip** includes a method for automatic outlier and artifact detection (Li & Wong, 2003)
  - Compares multiple probes for the same genes across arrays to detect outlier probes or arrays
  - Outliers = replicates that do not follow pattern
  - Automatically replaced by imputed values

- But...needs at least 10 or 20 chips.
QC3 Overall array quality
Affymetrix RPT file

Report Type: Expression Report
Date: 02:15PM 02/23/2005

Filename: a_U118_sc2.CHP
Probe Array Type: HG-U133_Plus_2
Algorithm: Statistical
Probe Pair Thr: 8
Controls: Antisense

Alpha1: 0.05
Alpha2: 0.065
Tau: 0.015
Noise (RawQ): 2.670
Scale Factor (SF): 7.121
TGT Value: 500
Norm Factor (NF): 1.000

Background:
- Avg: 78.90
  81.00
  Std: 1.32
  Min: 74.80
  Max:
Noise:
- Avg: 4.02
  Std: 0.14
  Min: 3.60
  Max: 4.50
Corner+
- Avg: 135
  Count: 32
Corner-
- Avg: 19476
  Count: 32
Central-
- Avg: 21615
  Count: 9

Total Probe Sets: 54675
Number Present: 17938 32.8%
Number Absent: 35907 65.7%
Number Marginal: 830 1.5%

Average Signal (P): 2312.3
Average Signal (A): 134.4
Average Signal (M): 426.3
Average Signal (All): 853.3
**QC3 Overall array quality**

**Affymetrix RPT file**

<table>
<thead>
<tr>
<th>Housekeeping Controls:</th>
<th>Probe Set</th>
<th>Sig(5')</th>
<th>Det(5')</th>
<th>Sig(M')</th>
<th>Det(M')</th>
<th>Sig(3')</th>
<th>Det(3')</th>
<th>Sig(all)</th>
<th>Sig(3'/5')</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFFX-HUMISGF3A/M97935</td>
<td>504.7</td>
<td>P</td>
<td>1176.5</td>
<td>P</td>
<td>2331.4</td>
<td>P</td>
<td>1337.54</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4.62</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFFX-HUMRGE/M10098</td>
<td>14667.0</td>
<td>P</td>
<td>5520.4</td>
<td>P</td>
<td>14576.0</td>
<td>P</td>
<td>11587.78</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.99</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFFX-HUMGAPDH/M33197</td>
<td>72937.9</td>
<td>P</td>
<td>67489.1</td>
<td>P</td>
<td>75214.9</td>
<td>P</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFFX-HSAC07/X00351</td>
<td>58057.9</td>
<td>P</td>
<td>63609.6</td>
<td>P</td>
<td>66569.8</td>
<td>P</td>
<td>62745.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFFX-M27830</td>
<td>9741.7</td>
<td>P</td>
<td>10870.7</td>
<td>P</td>
<td>72.8</td>
<td>A</td>
<td>6895.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Spike Controls:</th>
<th>Probe Set</th>
<th>Sig(3'/5')</th>
<th>Sig(5')</th>
<th>Det(5')</th>
<th>Sig(M')</th>
<th>Det(M')</th>
<th>Sig(3')</th>
<th>Det(3')</th>
<th>Sig(all)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFFX-BioB</td>
<td>1727.8</td>
<td>P</td>
<td>2726.2</td>
<td>P</td>
<td>1734.5</td>
<td>P</td>
<td>2062.84</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFFX-BioC</td>
<td>5172.9</td>
<td>P</td>
<td>6943.6</td>
<td>P</td>
<td></td>
<td></td>
<td>6058.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.34</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFFX-BioD</td>
<td>12128.3</td>
<td>P</td>
<td>28000.5</td>
<td>P</td>
<td></td>
<td></td>
<td>20064.38</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFFX-Cre</td>
<td>48125.3</td>
<td>P</td>
<td>53003.5</td>
<td>P</td>
<td></td>
<td></td>
<td>50564.38</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*BioB<BioC<BioD<Cre, and Present*

*Sig(3'/5')<3: good labelling reaction*
QC: summary

• Before starting analyzing data for answering to the biological questions, it is wise to perform QC
• Many information are readily available from image analysis
• Others can be derived looking at the data
QC: summary

- We have seen when to exclude single probes
- Genes which are not reliable on most arrays might be excluded from the analysis
- Arrays that present low overall quality should be excluded
  - Too many probes flagged
  - High BG and/or uniformly low intensities
  - Limited range of expression values
  - Too many saturated probes
  - Too different from the other arrays... but might it be a biologically interesting sample???(Tutorial1)
Back to normalization...

Experimental artifacts

- Systematic biases and random variation unrelated to the biology of the samples
- Biased results
  - Need to remove or adjust for systematic biases

- Imbalance between RNA amounts
- RNA amplification
- RNA degradation
- RT efficiency
- Settings of PMT voltage
- Dye incorporation
- Dye fading
...
Normalization

- Aim: remove artifacts from data before performing the analysis
- Statistician’s Ideal: **MODEL BASED METHODS**
  - Adjust for artifacts in data while performing the analysis
  - Carry over uncertainty of normalization into uncertainty of effects of interest
  - Some examples using ANOVA models (Kerr & Churchill 2001)
  - Computationally more challenging
  - AND nonlinear effects exist

- Common practice: normalize and analyze separately
- Common practice: normalize each array separately, BUT use the same method for all arrays!
Norm: which genes to use?

- **Housekeeping genes?**
  - Perfect candidates... but difficult to identify
  - Tissue specific?
- **Spiked Controls?**
  - Technically challenging
  - Not to be used in Affymetrix
- **All genes?**
  - Simplest
  - Relies on the hypothesis that
    - % over expressed ~ % under expressed
    - Very often true
  - But cannot be used with some custom chips!
Norm: spotted arrays

- **Global normalization**
  - Using all genes or housekeeping genes
  - For each array separately
    - Evaluate the median log ratio
    - Subtract it from log ratios of all the probes on the array
      - \( \text{Log}(\text{ch1}/\text{ch2})_{\text{norm}} = \text{Log}(\text{ch1}/\text{ch2}) - \text{median}(\text{Log}(\text{ch1}/\text{ch2})) \)
  - Can be used even if small number of normalizing genes are available
  - Does not overfit data
Norm: spotted arrays

• Need for Global normalization vs need for Intensity dependent normalization

Norm: spotted arrays

- Intensity-based normalization
  - Using all genes or housekeeping genes
  - Adjustment needs to be intensity-dependent
  - Fit a curve to the MA plot: \( f(A) \)
    - Locally weighted regression curves (loess)
    - Splines

- \( \log(ch1/ch2)_{\text{norm}} = \log(ch1/ch2) - f(A) \)

- Different adjustments for different genes on the same array
- Same adjustments for different genes on the same array with the same average intensity in the 2 channels
Norm: spotted arrays

- Intensity based normalization
  - Assumption: at each intensity level equal number of over and under-expressed genes

- Overfitting data?

- If housekeeping genes are used, they need to cover the intensity range
Norm: spotted arrays

- Location based normalization
  - Intensities depend on the position of the probe on the array
Norm: spotted arrays

- Location based normalization
  - Print tips (grids) can be used as blocks within which separate normalizations are carried out
    - Global normalization
    - Intensity based normalization
    - Or combine
  - Needs a large number of normalizing spots within each grid
  - Again, danger of overfitting data
Norm: Affymetrix

- Included in the probe summary procedures
  - MAS5
    - Global normalization ("scaling factor")
      - Target value (500) or target array
      - Housekeeping or all genes
    - OK to apply other normalization procedures AFTER MAS5
  - dCHIP
    - Intensity based as in spotted arrays, all against a reference median array
Norm: Affymetrix

• Included in the probe summary procedures
  - RMA
    • quantile normalization
      - Imposes an identical distribution to all arrays (all quantiles are the same)
      - Nonparametric and nonlinear normalization across arrays
      - Assumption: distribution of expression values does not change between arrays
**Norm: Affymetrix**

- “Extra” intensity based normalization can be performed using signal values
- Define a baseline array
  - Scaling factor close to median sf
- Use MA plots to determine if it is necessary to perform extra-normalization
  - X=array, Y=baseline array
Norm: Summary

• Normalization method used must depend on the data at hand
• A careful inspection of data and study design characteristics should guide the choice of normalization method
• Use the simplest methods when data characteristics allow it!
  - Less assumptions
  - Less overfitting
• After ....
  - Study design \textit{(StatCourse, day 3)}
  - Performance of the experiment \textit{(BioCourse, day 3)}
    • Sample preparation
    • Hybridization
    • Image analysis \textit{(not covered)}

\begin{boxed_text}
Quality control and normalization of microarray data
\end{boxed_text}

• Before...
  - Class comparison \textit{(StatCourse, day 1)}
  - Class prediction \textit{(StatCourse, day 2 and day 3)}
  - Class discovery \textit{(StatCourse, day 2)}
How can YOU do it?

• Always get data which is as row as possible!
• In the tutorial...
• Use R and Bioconductor libraries
  – [http://www.r-project.org](http://www.r-project.org)
  – [http://www.bioconductor.org](http://www.bioconductor.org)
    • marray
    • affy
    • Metadata
  – Your own (and our) scripts