cDNA chips
Quality control and pre-processing

Ulrich Mansmann
Department of Medical Biometrics and Informatics
University of Heidelberg
- Agilent Bioanalyzer
- Omnimgrid spotter
- Virtek ChipWriter Pro spotter
- 2 GenePix 4A scanners
- 5 thermocycler (4 blocks)
- 2 Packard pipetting robots
- 1 Ventana Discovery hybridization station

DKFZ microarray platform
cDNA - QC - Normalization
Measurements should be unbiased and precise.

- **Biased**
  - Low noise
  - High noise

- **Unbiased**
Quality control: Noise and reliable signal

- Is the signal dominated by noise? Acceptable amount of noise? Quantifying noise?
- Quantifying quality of a signal;
- Guidelines for reasonable thresholds on the quality of a signal;
- Defining strategies for exclusion of probes;
- Probe level: quality of the expression measurement on one particular array
- Gene level: quality of the expression measurement across all arrays
- Array level: quality of the expression measurement on one particular glass slide
This experiment was carried out using zebrafish as a model organism to study early development in vertebrates. Swirl is a point mutant in the BMP2 gene that affects the dorsal/ventral body axis. Ventral fates such as blood are reduced, whereas dorsal structures such as somites and notochord are expanded.

A goal of the Swirl experiment is to identify genes with altered expression in the swirl mutant compared to wild-type zebrafish. Two sets of dye-swap experiments were performed, for a total of four replicate hybridizations. For each of these hybridizations, target cDNA from the swirl mutant was labeled using one of the Cy3 or Cy5 dyes and the target cDNA wild-type mutant was labeled using the other dye.

Target cDNA was hybridized to microarrays containing 8,448 cDNA probes, including 768 controls spots (e.g. negative, positive, and normalization controls spots). Microarrays were printed using 4 times 4 print-tips and are thus partitioned into a 4 times 4 grid matrix. Each grid consists of a 22 times 24 spot matrix that was printed with a single print-tip. Here, spot row and plate coordinates should coincide, as each row of spots corresponds to probe sequences from the same 384 well-plate.
Probe-level quality control

- Individual spots printed on the slide
- Sources: faulty printing, uneven distribution, contamination with debris, size of signal relative to noise, poorly measured spots;
- Visual inspection: hairs, dust, scratches, air bubbles, dark regions, regions with haze
  → set points to NA
  → local normalization procedures which account for regional idiosyncrasies.
Visual inspection

4 x 4 sectors
Sector:
24 rows
22 columns
8448 spots
Mean signal intensity
Visual inspection

**cDNA - QC - Normalization**
Probe-level quality control

- Individual spots printed on the slide
- Sources: faulty printing, uneven distribution, contamination with debris, size of signal relative to noise, poorly measured spots;
- Visual inspection
- 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
**Weak Signal**

\[ \lambda_{\text{low}} = 200 \quad \lambda_{\text{high}} = 500 \]

<table>
<thead>
<tr>
<th>CY5 channel</th>
<th>CY3 channel</th>
<th>Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{FR-BR} &lt; \lambda_{\text{low}} )</td>
<td>( \text{FG-BG} &lt; \lambda_{\text{high}} )</td>
<td>exclude</td>
</tr>
<tr>
<td>( \text{FR-BR} &lt; \lambda_{\text{high}} )</td>
<td>( \text{FG-BG} &lt; \lambda_{\text{low}} )</td>
<td>exclude</td>
</tr>
<tr>
<td>( \text{FR-BR} &lt; \lambda_{\text{low}} )</td>
<td>( \text{FG-BG} &gt; \lambda_{\text{high}} )</td>
<td>( \log_2[\lambda_{\text{low}} / (\text{FG-BG})] )</td>
</tr>
<tr>
<td>( \text{FR-BR} &gt; \lambda_{\text{high}} )</td>
<td>( \text{FG-BG} &lt; \lambda_{\text{low}} )</td>
<td>( \log_2[(\text{FR-BR}) /\lambda_{\text{low}}] )</td>
</tr>
<tr>
<td>( \text{FR-BR} &gt; \lambda_{\text{high}} )</td>
<td>( \text{FG-BG} &gt; \lambda_{\text{high}} )</td>
<td>( \log_2[(\text{FR-BR}) / (\text{FG-BG})] )</td>
</tr>
</tbody>
</table>

\[ \log_2[510/490] = 0.057, \log_2[30/10] = 1.58, \quad \text{FG}=2030, \text{BG}=2000 \]
cDNA - QC - Normalization
Gene-level quality control

- Test of poor hybridization in the reference channel
- Test for exclusion of low variance genes.
Gene-level quality control:
Poor Hybridization and Printing

- Some probes will not hybridize well to the target RNA
- Printing problems such that all spots of a given inventory well have poor quality.
- A well may be of bad quality – contamination
- Genes with a consistently low signal in the reference channel are suspicious: Median of the background adjusted signal $< 200$
Gene-level quality control:
Probe quality control based on duplicated spots

- Printing different probes that target the same gene or printing multiple copies of the same probe.
- Mean squared difference of log₂ ratios between spot r and s:

\[ \text{MSDLR} = \frac{\sum (x_{jr} - x_{js})^2}{J} \text{ sum over arrays } j = 1, \ldots, J \]

recommended threshold to assess disagreement: MSDLR > 1

- Disagreement between copies: printing problems, contamination, mislabelling. Not easy if there are only 2 or 3 slides.

Gene-level quality control:
Low variance genes

- Good for normalization, but uninformative for the analysis
- Threshold for fold change:
  \[ \log_2(\text{max}) - \log_2(\text{min}), \] this difference may depend on sample size
  taking sample size into account: \[ \log_2(q_{0.95}) - \log_2(q_{0.05}) \]
- Variance based criterion
Array-level quality control

- Problems: array fabrication defect
  - problem with RNA extraction
  - failed labelling 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
  - Amount of adjustment needed: signals have to substantially changed to make slides comparable.
### Raw data are not mRNA concentrations

- tissue contamination
- RNA degradation
- amplification efficiency
- reverse transcription efficiency
- hybridization efficiency and specificity
- clone identification and mapping
- PCR yield, contamination
- spotting efficiency
- DNA-support binding
- other array manufacturing-related issues
- image segmentation
- signal quantification
- 'background' correction
Sources of variation

- amount of RNA in the biopsy
- efficiencies of
  - RNA extraction
  - reverse transcription
  - labeling
  - photodetection
- PCR yield
- DNA quality
- spotting efficiency, spot size
- cross-/unspecific hybridization stray signal

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

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

Calibration

Error model

W. Huber
Normalization

- Identify and remove sources of systematic variation, other than differential expression, in the measured fluorescence intensities.
- Normalization is necessary before each analysis, in order to ensure that differences in intensities are indeed due to differential expression and not experimental artefacts.
- Location normalisation: corrects for spatial or dye bias
- Scale normalisation: homogenises the variability across arrays
  \[ \text{MAD} = \text{median}\{ |x_1-m|, ..., |x_n-m| \} \]
- Normalised log-intensity ratios are given by
  \[ M_{\text{norm}} = (M-\text{loc})/s \]
- Normalisation: within arrays (marray) or between arrays (vsn), single channels between arrays, log expression ratios, etc
Swirl Data

cDNA - QC - Normalization
marray: Pre Normalization

Swirl array 93: pre-norm

Swirl arrays: pre-normalization

cDNA - QC - Normalization
marray: Post Normalization

Swirl array 93: post-norm

Swirl arrays: post-normalization

cDNA - QC - Normalization
marray: MA Plots

Swirl array 93: pre-norm MA-Plot

Swirl array 93: post-norm MA-Plot

cDNA - QC - Normalization
marray:
Signals - raw versus normalised

---

cDNA - QC - Normalization
VSN: model and theory

- Huber et al. (2002) Bioinformatics, 18:S96–S104
- Model for measured probe intensity
- Log-transformation is replaced by a transformation (arcsinh) based on theoretical grounds.
- Estimation of transformation parameters (location, scale) based on ML paradigm and numerically solved by a least trimmed sum of squares regression.
- VSN-normalised data behaves close to the normal distribution
- Following slides are borrowed from Wolfgang Huber
measured intensity = offset + gain × true abundance

\[ \gamma_{ik} = a_{ik} + b_{ik} \times x_k \]

\[ a_{ik} = a_i + \varepsilon_{ik} \]

\( a_i \) per-sample offset

\( \varepsilon_{ik} \sim N(0, b_i^2 s_1^2) \)

“additive noise”

\[ b_{ik} = b_i \, b_k \, \exp(\eta_{ik}) \]

\( b_i \) per-sample normalization factor

\( b_k \) sequence-wise probe efficiency

\[ \eta_{ik} \sim N(0, s_2^2) \]

“multiplicative noise”
The two-component model

- "additive" noise
- "multiplicative" noise

B. Durbin, D. Rocke, JCB 2001
Variance Stabilizing Transformations

Let $X_u$ be a family of random variables with $E X_u = u$, $\text{Var} X_u = v(u)$. Define

$$f(x) = \int x d\frac{1}{\sqrt{v(u)}}$$

$\Rightarrow \text{var } f(X_u) \approx \text{independent of } u$

Derivation: linear approximation
v variance stabilizing transformations
variance stabilizing transformations

\[
f(x) = \int x \frac{1}{\sqrt{\nu(u)}} \, du
\]

1.) constant variance ('additive') \( v(u) = s^2 \) \( \Rightarrow \) \( f \propto u \)

2.) constant CV ('multiplicative') \( v(u) \propto u^2 \) \( \Rightarrow \) \( f \propto \log u \)

3.) offset \( v(u) \propto (u + u_0)^2 \) \( \Rightarrow \) \( f \propto \log(u + u_0) \)

4.) additive and multiplicative

\[
v(u) \propto (u + u_0)^2 + s^2 \Rightarrow f \propto \text{arsinh} \frac{u + u_0}{s}
\]
v the "glog" transformation

\[ f(x) = \log(x) \]

\[ h_s(x) = \text{asinh}(x/s) \]

\[ \text{arsinh}(x) = \log \left( x + \sqrt{x^2 + 1} \right) \]

\[ \lim_{x \to \infty} \left( \text{arsinh} x - \log x - \log 2 \right) = 0 \]

P. Munson, 2001

D. Rocke & B. Durbin, ISMB 2002
Parameter estimation

\[
\text{arsinh} \left( \frac{Y_{ki} - a_i}{b_i} \right) = \mu_k + \varepsilon_{ki}, \quad \varepsilon_{ki} \sim N(0, c^2)
\]

- Maximum likelihood estimator: straightforward
  - but sensitive to deviations from normality
- Model holds for genes that are unchanged; differentially transcribed genes act as outliers.
- Robust variant of ML estimator, à la Least Trimmed Sum of Squares regression.
  - Works as long as <50% of genes are differentially transcribed.

Measured intensity = offset + gain * true abundance

\[
Y_{ik} = a_{ik} + b_{ik} x_{ik}
\]

- \(a_{ik} = a_i + L_{ik} + \varepsilon_{ik}\)
  - \(a_i\) per-sample offset
  - \(L_{ik}\) local background provided by image analysis
  - \(\varepsilon_{ik} \sim N(0, b_i^2 s_1^2)\)
  - “additive noise”

- \(b_{ik} = b_i b_k \exp(\eta_{ik})\)
  - \(b_i\) per-sample normalization factor
  - \(b_k\) sequence-wise labeling efficiency
  - \(\eta_{ik} \sim N(0, s_2^2)\)
  - “multiplicative noise”
Least trimmed sum of squares regression

\[ \minimize \sum_{i=1}^{n/2} (y_i - f(x_i))^2 \]

- least sum of squares
- least trimmed sum of squares

P. Rousseeuw, 1980s
evaluation: effects of different data transformations

a) $\Delta y$

b) $\Delta \log(y)$

c) $\Delta h(y)$
Normality: QQ-plot
Swirl - VSN

Swirl array 93: post-norm MA-Plot

Swirl Array 93: VSN MA-plot
Swirl: marray versus VSN
Acknowledgments

Holger Sültmann
Wolfgang Huber
