Normalization of microarray data

Anja von Heydebreck

Dept. Computational Molecular Biology, MPI for Molecular Genetics, Berlin
Systematic differences between arrays

The boxplots show distributions of log-ratios from 4 red-green 8448-clone cDNA arrays hybridised with zebrafish samples. Some are not centered at 0, and they are different from each other.
Experimental variation

amount of RNA in the sample
- efficiencies of
  - RNA extraction
  - reverse transcription
  - labeling
  - photodetection

Normalization:
Correction of systematic effects arising from variations in the experimental process

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

Normalization
Ad-hoc normalization procedures

• 2-color cDNA-arrays: multiply all intensities of one channel with a constant such that the median of log-ratios is 0 (equivalent: shift log-ratios). Underlying assumption: equally many up- and downregulated genes.

• One-color arrays (Affy, radioactive): multiply intensities from each array $k$ with a constant $c_k$, such that some measure of location of the intensity distributions is the same for all arrays (e.g. the trimmed mean (Affy *global scaling*)).
log-log plot of intensities from the two channels of a microarray

comparison of kidney cancer with normal kidney tissue, cDNA microarray with 8704 spots

• red line: median normalization
• blue lines: two-fold change
Assumptions for normalization

• When we normalize based on the observed data, we assume that the majority of genes are unchanged, or that there is symmetry between up- and downregulation.

• In some cases, this may not be true. Alternative: use (spiked) controls and base normalization on them.
1. Loess normalization

- **M-A plot** (minus vs. add):
  \[ \log(R) - \log(G) = \log(R/G) \]
  vs.
  \[ \log(R) + \log(G) = \log(RG) \]

- With 2-color-cDNA arrays, often “banana-shaped” scatterplots on the log-scale are observed.
Loess normalization

• Intensity-dependent trends are modeled by a regression curve, $M = f(A) + \varepsilon$.

• The normalized log-ratios are computed as the residuals $\varepsilon$ of the loess regression.
Loess regression

• Locally weighted regression.
• For each value $x_i$ of $X$, a linear or polynomial regression function $f_i$ for $Y$ is fitted based on the data points close to $x_i$. They are weighted according to their distance to $x_i$.
• Local model: $Y = f_i(X) + \varepsilon$.
• Fit: Minimize the weighted sum of squares
  $\sum w_j(x_j)(y_j - f_i(x_j))^2$
• Then, compute the overall regression as:
  $Y = f(X) + \varepsilon$, where $f(x_i) = f_i(x_i)$. 
Loess regression

regression lines for each data point

The user-defined width $c$ of the weight function determines the degree of smoothing.

$$w(x) = \left[1 - \left(\frac{|x - x_0|}{c}\right)^3\right]^3, |x - x_0| \leq c$$
Print-tip normalization

- With spotted arrays, distributions of intensities or log-ratios may be different for spots spotted with different pins, or from different PCR plates.
- Normalize the data from each (e.g. print-tip) group separately.
Print-tips correspond to localization of spots

Slide: 25x75 mm
Spot-to-spot: ca. 150-350 µm

4x4 or 8x4 sectors
17...38 rows and columns per sector
ca. 4600...46000 probes/array

sector: corresponds to one print-tip
Print-tip loess normalization
2. Error models, variance stabilization and robust normalization
Sources of variation

amount of RNA in the sample
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”

Normalization

Error model
A model for measurement error


\[ Y_k = \alpha + \beta_k \epsilon^n + \nu \]

- \( Y_k \): measured intensity of gene \( k \)
- \( \beta_k \): true expression level of gene \( k \)
- \( \alpha \): offset
- \( \eta, \nu \): multiplicative/additive error terms, independent normal

For large expression level \( \beta_k \), the multiplicative error is dominant.
For \( \beta_k \) near zero, the additive error is dominant.
A parametric form for the variance-mean dependence

The model of Durbin and Rocke yields:

\[ u_k = \mathbb{E}(Y_k) = \alpha_i + m_\eta \beta_k \]
\[ v_k = \text{Var}(Y_k) = s_\eta^2 \beta_k^2 + s_\nu^2 , \]

\( m_\eta, s_\eta^2 \) : mean/variance of \( e^\eta \),

\( s_\nu^2 \) : variance of \( \nu \)

Thus we obtain a quadratic dependence

\[ v_k = \nu(u_k) = (c_1 u_k + c_2)^2 + c_3^2 . \]
Quadratic variance-vs-mean dependence

For each spot $k$, the variance $(R_k - G_k)^2$ is plotted against the mean $(R_k + G_k)/2$.

$$v(u) = (c_1 u + c_2)^2 + c_3^2.$$
The two-component model
The two-component model

"additive" noise

"multiplicative" noise

raw scale

log scale
Variance stabilizing transformations

Let $X_u$ be a family of random variables with $EX_u = u$, $Var X_u = v(u)$. Define a transformation

$$h(x) = \int \frac{1}{\sqrt{v(u)}} \, du$$

$$\Rightarrow \text{Var } h(X_u) \approx \text{independent of } u$$
Derivation of the variance-stabilizing transformation

Let $X_u$ be a family of random variables with $E X_u = u$, $\text{Var} X_u = \nu(u)$, and $h$ a transformation applied to $X_u$. Then, by linear approximation of $h$,

$$h(X_u) \sim h(u) + h'(u)(X_u - u)$$

$$\text{Var}(h(X_u)) \sim h'(u)^2 \text{Var}(X_u - u) = h'(u)^2 \text{Var}(X_u).$$

Thus, if $h'(u)^2 = \nu(u)^{-1}$, $\text{Var}(h(X_u))$ is approx. independent of $u.$
Variance stabilizing transformations

$f(x)$

$x$

0 20000 40000 60000

8.0 8.5 9.0 9.5 10.0 11.0

raw scale
transformed scale
Variance stabilizing transformations

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

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

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

3.) additive and multiplicative

\[ v(u) \propto (u + u_0)^2 + s^2 \Rightarrow f \propto \text{arsinh} \frac{u + u_0}{s} \]
The “generalized log” transformation

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

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

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

W. Huber et al., ISMB 2002
D. Rocke & B. Durbin, ISMB 2002
A model for measurement error

Now we consider data from different arrays or color channels $i$. We assume they are related through an affine-linear transformation on the raw scale:

$$ Y_{ki} = \alpha_i + \gamma_i \beta_{ki} e^{\eta} + \nu $$

- $Y_{ki}$: measured intensity of gene $k$ in array/color channel $i$
- $\beta_{ki}$: true expression level of gene $k$
- $\alpha_i, \gamma_i$: additive/multiplicative effects of array/color channel $i$
- $\eta, \nu$: multiplicative/additive error terms, independent normal with mean 0
A statistical model

\[
\begin{align*}
\text{arsinh}\left(\frac{Y_{ki}}{b_i} - a_i\right) &= \mu_k + \varepsilon_{ki}, \\
\varepsilon_{ki} &\sim N(0, \sigma^2)
\end{align*}
\]

- Assume an affine-linear transformation for normalization between arrays, and, after that, common parameters for the variance stabilizing transformation. The composite transformation for array/color channel \(i\) is given by \(a_i\) and \(b_i\).
- The model is assumed to hold for genes that are unchanged; differentially expressed genes act as outliers.
Robust parameter estimation

\[
\text{arsinh} \left( \frac{y_{ki} - a_i}{b_i} \right) = \mu_k + \varepsilon_{ki}, \quad \varepsilon_{ki} : \mathcal{N}(0, \sigma^2)
\]

• Assume that the majority of genes is not differentially expressed.
• Use robust variant of maximum likelihood estimation:
• Alternate between maximum likelihood estimation (= least squares fit) for a fixed set \( K \) of genes and selection of \( K \) as the subset of (e.g. 50%) genes with smallest residuals.
Robust normalization

assumption:
majority of genes
unchanged

location estimators:
• mean
• median
• least trimmed sum of squares

(generalized) log-ratio
Normalized & transformed data

![log scale](image1)

![generalized log scale](image2)
Validation: standard deviation versus rank-mean plots

a) $\Delta y$

b) $\Delta \log(y)$

c) $\Delta h(y)$
Which normalization method should one use?

• How can one assess the performance of different methods?
• Diagnostic plots (e.g. scatterplots)
• **Performance measures:**
  • The variance between replicate measurements should be low.
  • Low bias: Changes in expression should be accurately measured. How to assess this (in most cases, the truth is unknown)?
Evaluation: sensitivity / specificity in quantifying differential expression

- **Data**: paired tumor/normal tissue from 19 kidney cancers, hybridized in duplicate on 38 cDNA slides à 4000 genes.

- Apply 6 different strategies for **normalization** and quantification of differential expression

- Apply **permutation test** to each gene

- Compare numbers of genes detected as differentially expressed, at a certain significance level, between the different normalization methods
Comparison of methods

a) test for up-regulation (kidney data)

b) test for down-regulation (kidney data)

Number of significant genes vs. significance level of permutation test
Parametric vs. non-parametric normalization

- **Loess** is **non-parametric**: it makes no assumptions which sort of transformation is appropriate. Disadvantage: Degree of smoothing is chosen in an arbitrary way.

- **vsn** uses a **parametric** model: affine-linear normalization. Disadvantage: the model assumptions may not always hold. Advantage: If the model assumptions do hold (at least approximately), the method should perform better.
vsn may also correct “banana shape”

M-A plot of vsn-normalized zebrafish data, loess fit

Different additive offsets may lead to non-linear scatter plots on the log scale.
References

• Software: R package modreg (loess), Bioconductor packages marrayNorm (loess normalization), vsn (variance stabilization)
