Why do you need normalisation?
From: lymphoma dataset
vsn package
Alizadeh et al., Nature 2000
Scatterplot, colored by PCR-plate

Two RZPD Unigene II filters (cDNA nylon membranes)
PCR plates
PCR plates: boxplots

Intensities

PCR plates: normal

Intensities

PCR plates: tumor
print-tip effects

\[ F(q) \]

\[ q (\text{log-ratio}) \]
spotting pin quality decline

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

H. Sueltmann DKFZ/MGA
spatial effects

spotted cDNA arrays, Stanford-type
Batches: array to array differences \( d_{ij} = \text{mad}_k(h_{ik} - h_{jk}) \)

arrays \( i=1...63 \); roughly sorted by time
A complex measurement process lies between mRNA concentrations and intensities.

- other array manufacturing-related issues
- hybridization efficiency and specificity
- amplification efficiency
- signal quantification
- PCR yield, contamination
- RNA degradation
- image segmentation
- clone identification and mapping
- tissue contamination
- reverse transcription efficiency
- other array manufacturing-related issues
- hybridization efficiency and specificity
- amplification efficiency
- signal quantification
- PCR yield, contamination
- RNA degradation
- image segmentation
- clone identification and mapping
- tissue contamination
- reverse transcription efficiency

The problem is less that these steps are ‘not perfect’; it is that they vary from array to array, experiment to experiment.
Why do you need statistics?
Which genes are differentially transcribed?

**same-same**

**tumor-normal**
Statistics 101:

\[ \text{bias} \quad \text{accuracy} \]

\[ \text{variance} \quad \text{precision} \]
Basic dogma of data analysis

Can always increase sensitivity on the cost of specificity, or vice versa, the art is to
- optimize both
- then find the best trade-off.
Fold changes are useful to describe continuous changes in expression.

But what if the gene is "off" (below detection limit) in one condition?
ratios and fold changes

The idea of the log-ratio (base 2)
0: no change
  +1: up by factor of $2^1 = 2$
  +2: up by factor of $2^2 = 4$
-1: down by factor of $2^{-1} = 1/2$
-2: down by factor of $2^{-2} = \frac{1}{4}$

A unit for measuring changes in expression: assumes that a change from 1000 to 2000 units has a similar biological meaning to one from 5000 to 10000.

What about a change from 0 to 500?
  - conceptually
  - noise, measurement precision
How to compare microarray intensities with each other?

How to address measurement uncertainty ("variance")?

How to calibrate ("normalize") for biases between samples?
Sources of variation

amount of RNA in the biopsy
- efficiencies of 
  - RNA extraction
  - reverse transcription
  - labeling
  - fluorescent detection

probe purity and length distribution
- 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
- remain as “noise”

Calibration

Error model
Error models describe the possible outcomes of a set of measurements.

**Outcomes depend on:**
- **true value of the measured quantity**
  (abundances of specific molecules in biological sample)
- **measurement apparatus**
  (cascade of biochemical reactions, optical detection system with laser scanner or CCD camera)
Purpose:

1. **Data compression**: summary statistic instead of full empirical distribution

2. **Quality control**

3. **Statistical inference**: appropriate parametric methods have better power than non-parametric (this has practical, financial, and ethical aspects)
The two component model

measured intensity = offset + gain × true abundance

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

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

\( a_i \) per-sample offset

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

“additive noise”

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

\( b_i \) per-sample normalization factor

\( b_k \) sequence-wise probe efficiency

\( \eta_{ik} \sim N(0, s_z^2) \)

“multiplicativc noise”
The two-component model

- "multiplicative" noise
- "additive" noise

B. Durbin, D. Rocke, JCB 2001
Parameterization

\[ y = a + \epsilon + b \cdot x \cdot (1 + \eta) \]

\[ y = a + \epsilon + b \cdot x \cdot e^\eta \]

two practically equivalent forms \((\eta \ll 1)\)

<table>
<thead>
<tr>
<th>a systematic background</th>
<th>same for all probes (per array x color)</th>
<th>per array x color x print-tip group</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\epsilon) random background</td>
<td>iid in whole experiment</td>
<td>iid per array</td>
</tr>
<tr>
<td>b systematic gain factor</td>
<td>per array x color</td>
<td>per array x color x print-tip group</td>
</tr>
<tr>
<td>(\eta) random gain fluctuations</td>
<td>iid in whole experiment</td>
<td>iid per array</td>
</tr>
</tbody>
</table>
Important issues for model fitting

Parameterization

- variance vs bias

"Heteroskedasticity" (unequal variances)

⇒ weighted regression or variance stabilizing transformation

Outliers

⇒ use a robust method

Algorithm

If likelihood is not quadratic, need non-linear optimization. Local minima / concavity of likelihood?
Models are never correct, but some are useful

**True relationship:**

\[ y = x - \frac{1}{2} x^2 + \epsilon \quad \epsilon \sim N(0, 0.15^2) \]
variance stabilizing transformations

\(X_u\) a family of random variables with \(E X_u = u, \ Var X_u = v(u)\). Define

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

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

derivation: linear approximation
variance stabilizing transformations

\[
f(x)
\]

\[
x
\]
Variance stabilizing transformations

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

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

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

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

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

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

P. Munson, 2001

D. Rocke & B. Durbin, ISMB 2002

W. Huber et al., ISMB 2002
**glog**

**raw scale**

**log**

**glog**

**variance:**

- **constant part**
- **proportional part**
Parameter estimation

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

- Maximum likelihood estimator:
  - 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 well as long as \(< 50\%\) of genes are differentially transcribed (and may still work otherwise).

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

- \(a_{ik} = a_i + L_{ik} + \varepsilon_{ik}\)
  - \(a_i\) 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

\[
\text{minimize} \quad \sum_{i=1}^{n/2} (y_{(i)} - f(x_{(i)}))^2
\]

- least sum of squares
- least trimmed sum of squares

P. Rousseeuw, 1980s
“usual” log-ratio: \[ \log \frac{x_1}{x_2} \]

‘glog’ (generalized log-ratio): \[ \log \left( \frac{x_1 + \sqrt{x_1^2 + c_1^2}}{x_2 + \sqrt{x_2^2 + c_2^2}} \right) \]

\(c_1, c_2\) are experiment specific parameters (~level of background noise)
Variance Bias Trade-Off

Estimated log-fold-change vs Signal intensity

- q
- h
- glog
Variance-bias trade-off and shrinkage estimators

Shrinkage estimators:
pay a small price in bias for a large decrease of variance, so overall the mean-squared-error (MSE) is reduced.
Particularly useful if you have few replicates.

Generalized log-ratio:
= a shrinkage estimator for fold change

There are many possible choices, we chose “variance-stabilization”:
+ interpretable even in cases where genes are off in some conditions
+ can subsequently use standard statistical methods (hypothesis testing, ANOVA, clustering, classification…) with less worries about heteroskedasticity than with many alternative methods
evaluation: effects of different data transformations

a) $\Delta y$

b) $\Delta \log(y)$

c) $\Delta h(y)$
Normality: QQ-plot
"Single color normalization"

n red-green arrays \((R_1, G_1, R_2, G_2, \ldots, R_n, G_n)\)

within/between slides

for \((i=1:n)\)

- calculate \(M_i = \log(R_i/G_i)\), \(A_i = \frac{1}{2} \log(R_i*G_i)\)
- normalize \(M_i\) vs \(A_i\)
- normalize \(M_1\)…\(M_n\)

all at once

- normalize the matrix of \((R, G)\)
- then calculate log-ratios or any other contrast you like
What about non-linear effects

- Microarrays can be operated in a linear regime, where fluorescence intensity increases proportionally to target abundance (see e.g. Affymetrix dilution series)

Two reasons for non-linearity:

- At the high intensity end: saturation/quenching. This can and should be avoided experimentally - loss of data!

- At the low intensity end: background offsets, instead of $y = k \cdot x$ we have $y = k \cdot x + x_0$, and in the log-log plot this can look curvilinear. But this is an affine-linear effect and can be correct by affine normalization. Non-parametric methods (e.g. loess) risk overfitting and loss of power.
Non-linear or affine linear?
Definitions

linear

affine linear

genuinely non-linear
References


Acknowledgements

Anja von Heydebreck (Darmstadt)
Robert Gentleman (Seattle)
Günther Sawitzki (Heidelberg)
Martin Vingron (Berlin)
Annemarie Poustka, Holger Sültmann, Andreas Buness, Markus Ruschhaupt (Heidelberg)
Rafael Irizarry (Baltimore)
Judith Boer (Leiden)
Anke Schroth (Heidelberg)
Friederike Wilmer (Hilden)