Statistical Modeling 3

Bias correction and normalization

“Essentially, all models are wrong, but some are useful”

George E.P. Box
RNA-seq data

Log ratio of two Poisson random variables
Biological replicates: Poisson does not fit

SDs across twelve biological reps
Not normally distributed

Model: variances as scaled F-distribution

\[ s^2 \sim s_0^2 F_{d, d_0} \]
A biochemical experiment

Michaelis-Menten equation

\[ V = \frac{V_{\text{max}} \times C'}{K + C'} \]

- \( V \) = initial velocity
- \( C' \) = concentration
- \( V_{\text{max}} \) = maximum velocity
- \( K \) = rate constant
A biochemical experiment

\[ V = \frac{V_{\text{max}} \times C}{K + C} \]

\[ \Rightarrow \frac{1}{V} = \frac{K + C}{V_{\text{max}} \times C} \]

\[ = \frac{K}{V_{\text{max}} \times C} + \frac{1}{V_{\text{max}}} \]

\[ \Rightarrow \frac{1}{V} = \left( \frac{1}{V_{\text{max}}} \right) + \left( \frac{K}{V_{\text{max}}} \right) \times \left( \frac{1}{C} \right) \]
A biochemical experiment

Which is more reasonable?

\[ \frac{1}{V} = \beta_0 + \beta_1 \left( \frac{1}{C} \right) + \text{error} \]

\[ V = \frac{V_{\text{max}} \times C}{K + C} + \text{error} \]

Why so much noise?
Signal is buried in the noise

Non-human RNA on a human microarray
A spike-in experiment (HG-U95)

mRNA reference sequence

Probes:

Gene expression (RNA) microarray

Probes: CAGACATAGTGCTGTTTTTCTCT
Individual probes for each spiked-in gene

Eleven probes from one spiked-in gene
\[ Y_{ij} = \beta_j + \theta_i \phi_j + \varepsilon_{ij} \quad \text{var}(\varepsilon_{ij}) \propto \theta_i \phi_j \]
Why to adjust for background

\[
\frac{Y_{i,j}}{Y_{i-1,j}} \approx 1
\]

\[
\frac{Y_{i,j}}{Y_{i-1,j}} \approx 2
\]

mRNA reference sequence

PM: CAGACATAGTGCTGTGTGTCTCTCTCTCT
MM CAGACATAGTGCTGTGTCTCTCTCTCT
Why not use mismatches (MM)?

\[ PM = \beta + \text{signal} \]
\[ MM = \beta \]
\[ PM - MM = \text{signal} \]

\[ \text{correlation} = 0.83 \]
We can expect issues with the variance

\[ PM = \beta_{PM} + \text{signal and } MM = \beta_{MM} \]

\[ \text{corr}(\beta_{MM}, \beta_{PM}) \approx 0.8 \]

\[ \text{var}\{\log_2(PM - MM)\} \propto \frac{1}{\text{signal}^2} \]
A model based approach

signal \sim \text{exponential}(\lambda)

\beta \sim \mathcal{N}(\mu, \sigma)

E[\text{signal} | PM] = PM - \mu - \lambda \sigma^2 + \sigma \left[ \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{(PM/\sigma)^2}{2\Phi(PM/\sigma)} \right\} \right]
Five technical replicates

log2(pms[, 2])

Frequency

unit = 0.1
More than location and scale changes!

Median shifts do not solve the problem!
And there are non-linear effects!

Three possible approaches

A. Local regression (loess).

B. Quantile normalization.

C. Variance stabilizing normalization.
Local regression (loess)
Local regression (loess)

Before
Quantile normalization

<table>
<thead>
<tr>
<th>Original</th>
<th>Order</th>
<th>Averaged</th>
<th>Re-order</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 4 4 5</td>
<td>2 4 3 5</td>
<td>3.5 3.5 3.5 3.5</td>
<td>3.5 3.5 5.0 5.0</td>
</tr>
<tr>
<td>5 14 4 7</td>
<td>3 8 4 5</td>
<td>5.0 5.0 5.0 5.0</td>
<td>8.5 8.5 5.5 5.5</td>
</tr>
<tr>
<td>4 8 6 9</td>
<td>3 8 4 7</td>
<td>5.5 5.5 5.5 5.5</td>
<td>6.5 5.0 8.5 8.5</td>
</tr>
<tr>
<td>3 8 5 8</td>
<td>4 9 5 8</td>
<td>6.5 6.5 6.5 6.5</td>
<td>5.0 5.5 6.5 6.5</td>
</tr>
<tr>
<td>3 9 3 5</td>
<td>5 14 6 9</td>
<td>8.5 8.5 8.5 8.5</td>
<td>5.5 6.5 3.5 3.5</td>
</tr>
</tbody>
</table>
Densities are forced to be identical

Differential expression can be preserved
Differential expression can be preserved

\[ Y_{ij} = \beta_i + \varepsilon_{ij} + A_i \theta_j \eta_{ij} \]

Variance stabilizing normalization (vsn)

Additive measurement error  Multiplicative error

Array specific background level  Array specific gain
Example

\[ \beta_1 = 24, \beta_2 = 20, A_1 = 1, A_2 = 1.25, \sigma = 1, \eta = 0.05 \]

Background corrected and normalized

\[ \frac{Y_{ij} - \beta_i}{A_i} \]

Raw scale  Log scale
Variance stabilizing transformation

$Y$ with $E(Y) = \mu$ and $\text{var}(Y) = \nu(\mu)$

$$f(y) = \int \frac{1}{\sqrt{\nu(\mu)}} \, d\mu$$

$\text{var}\{f(Y)\}$ does not depend on $\mu$

---

VSN for microarrays

$$f(y_{ij}) = \text{arsinh} \left( \frac{y_{ij} - \beta_i}{A_i} \right)$$

$$\text{arsinh}(y) = \log \left\{ y + \sqrt{y^2 + 1} \right\}$$
Before

After