Statistical Analysis of cDNA microarrays 1

Terry Speed

WEHI Bioinformatics & UCB Statistics
## Spot – output

<table>
<thead>
<tr>
<th>Spot_ID</th>
<th>Gmean</th>
<th>GIQR</th>
<th>Rmean</th>
<th>RIQR</th>
<th>Gvalley</th>
<th>Rvalley</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13177</td>
<td>0.62</td>
<td>10327</td>
<td>0.47</td>
<td>463</td>
<td>1282</td>
</tr>
<tr>
<td>2</td>
<td>8000</td>
<td>0.14</td>
<td>5070</td>
<td>0.13</td>
<td>463</td>
<td>1185</td>
</tr>
<tr>
<td>3</td>
<td>3138</td>
<td>0.27</td>
<td>3211</td>
<td>0.19</td>
<td>481</td>
<td>1213</td>
</tr>
<tr>
<td>4</td>
<td>8433</td>
<td>0.25</td>
<td>8635</td>
<td>0.40</td>
<td>481</td>
<td>1265</td>
</tr>
<tr>
<td>5</td>
<td>4118</td>
<td>0.35</td>
<td>3776</td>
<td>0.22</td>
<td>463</td>
<td>1265</td>
</tr>
<tr>
<td>6</td>
<td>15473</td>
<td>0.90</td>
<td>5603</td>
<td>0.65</td>
<td>456</td>
<td>1231</td>
</tr>
<tr>
<td>7</td>
<td>2399</td>
<td>0.21</td>
<td>2995</td>
<td>0.16</td>
<td>481</td>
<td>1253</td>
</tr>
<tr>
<td>8</td>
<td>1245</td>
<td>0.27</td>
<td>2107</td>
<td>0.11</td>
<td>483</td>
<td>1265</td>
</tr>
<tr>
<td>9</td>
<td>35959</td>
<td>0.81</td>
<td>31807</td>
<td>0.73</td>
<td>483</td>
<td>1247</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gmorphe</th>
<th>Rmorphe</th>
<th>area</th>
<th>circularity</th>
<th>Gsn</th>
<th>Rsn</th>
<th>Iratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>260</td>
<td>1208</td>
<td>31</td>
<td>.88</td>
<td>5.62</td>
<td>2.92</td>
<td>.50</td>
</tr>
<tr>
<td>261</td>
<td>1174</td>
<td>32</td>
<td>.83</td>
<td>4.88</td>
<td>1.73</td>
<td>.99</td>
</tr>
<tr>
<td>261</td>
<td>1146</td>
<td>56</td>
<td>.64</td>
<td>3.46</td>
<td>0.84</td>
<td>.48</td>
</tr>
<tr>
<td>260</td>
<td>1208</td>
<td>11</td>
<td>1.38</td>
<td>4.96</td>
<td>2.62</td>
<td>.13</td>
</tr>
<tr>
<td>262</td>
<td>1163</td>
<td>16</td>
<td>1.03</td>
<td>3.88</td>
<td>1.14</td>
<td>.57</td>
</tr>
<tr>
<td>262</td>
<td>1144</td>
<td>40</td>
<td>.74</td>
<td>5.86</td>
<td>1.95</td>
<td>1.77</td>
</tr>
<tr>
<td>250</td>
<td>1185</td>
<td>30</td>
<td>.94</td>
<td>3.03</td>
<td>0.57</td>
<td>.26</td>
</tr>
<tr>
<td>262</td>
<td>1157</td>
<td>20</td>
<td>1.12</td>
<td>1.91</td>
<td>-0.42</td>
<td>.12</td>
</tr>
<tr>
<td>262</td>
<td>1164</td>
<td>61</td>
<td>.85</td>
<td>7.09</td>
<td>4.68</td>
<td>.22</td>
</tr>
</tbody>
</table>
Statistical problems involving microarray data


Image analysis ones already mentioned.
6. Use of housekeeping genes.
8. Normalization within an experiment:
   * when few genes change.
   * when many genes change.
   * use of red-green and green-red pairs.
9. Normalization between experiments: location and scale effects.
11. Variability.
12. Bias : Use of "truth".
14. Who is up/down?
15. P-values.
16. Planning of experiments:
   1. design.
   2. sample sizes.
17. Analysis of factorial experiments.
18. Discrimination and allocation.
19. Clustering:
   1. of samples.
   2. of genes.
20. Time course experiments.
22. Special problems.
   ✽ Mixture analyses.
   ✽ Pooled cDNA vs amplified DNA.

We start with 9 and go on to 14.
Experiments

Goal. Identify genes with altered expression in the livers of two lines of mice with very low HDL cholesterol levels compared to inbred control mice.

Two experiments: (1) Apo AI knock-out mouse model and (2) SR-BI transgenic mouse model. In each experiment:

- 8 treatment (trt) mice (apo AI ko or SR-BI tg) and 8 control (ctl) mice (C57Bl/6 or FVB).
- 16 hybridizations: mRNA from each of the 16 mice is labeled with Cy5, pooled mRNA from control mice is labeled with Cy3.
- Probes: \( \sim 6,000 \) cDNAs, including 200 related to lipid metabolism.
Single Slide plots

R vs G (intensity scale)
Plotting transformed intensities

Log R vs Log G (base 2)

More informative.
\[ M = \log_2 R/G \text{ vs } A = \frac{1}{2}(\log_2 G + \log_2 R) \]

More informative still.
Within-slide normalization

Normalization balances red and green intensities. Imbalance may be caused by differential incorporation of dyes, different amounts of the two species of RNA, differential scanning, etc. In practice, we usually need to bump up the red intensity a bit to balance the green.

\[
\log \frac{R}{G} \rightarrow \log \frac{R}{G} + c = \log k \frac{R}{G}
\]

A standard choice is to arrange that normalized log ratios have zero mean or median. Our preference is to do this in an A-dependent way: we choose \( c = c(A) \) using lowess.

A proof that this is better than using a constant is currently lacking. It certainly changes things, and we are pretty sure it helps.
Normalization - Median

Assumption: Changes roughly symmetric

First panel: smoothed densities of $\log_2 G$ and $\log_2 R$.
Second panel: $M$ vs $A$ plot with median $M$ put to 0.
Normalization - Lowess

Global lowess. Assumption: changes roughly symmetric at all intensities
Normalization - Print Tip Lowess

Print-tip lowess normalization. Need stronger assumption.
M vs A – after print-tip normalization
Effects of normalization I

![Graph showing density distributions with labels: None, Median, Lowess, Print-tip. N = 6376, Bandwidth = 0.1125.](image)
Effects of normalization II

Before normalization

After normalization
Within print-tip box plots of print-tip normalized M

Print-tip scale effects remain: last four more variable.
Statistical Software

Splus or R (freeware)

SHELL>R --vsize=50M --nsize=2000K

R : Copyright 2000, The R Development Core Team
Version 1.0.1 (April 14, 2000)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type "?license" or "?licence" for distribution details.

R is a collaborative project with many contributors.
Type "?contributors" for a list.

Type "demo()" for some demos, "help()" for on-line help, or
"help.start()" for a HTML browser interface to help.
Type "q()" to quit R.

[Previously saved workspace restored]

> library(Spot)
Spot> library(sma)
Spot>
Which genes have changed expression levels?

Single-slide methods

Existing methods Model dependent rules for deciding whether $(R, G)$ corresponds to a differentially expressed gene. Amounts to drawing two curves in the $(R, G)$–plane and calling a gene differentially expressed if its $(R, G)$ falls outside the region between the two curves. We probably do not know enough about the systematic and random effects within a microarray experiment to justify strong modeling assumptions or theory-based predictions. Conclusion $n = 1$ slide may not be enough.
Existing methods differ in the distributional assumptions they make regarding \((R, G)\).

1. Chen et al. Each \((R, G)\) is assumed to be normally and independently distributed with constant CV. Decision based on \(R/G\) only. (purple)

2. Newton et al. Gamma-Gamma-Bernoulli hierarchical model for each \((R, G)\). (yellow)

3. Roberts et al. Each \((R, G)\) is assumed to be normally and independently distributed with variance depending linearly on the mean.

4. Sapir & Churchill. Each \(\log R/G\) is assumed to be distributed according to a mixture of normal and uniform distributions. Decision based on \(R/G\) only. (turquoise)
Which genes have changed expression levels?
A Bayesian approach for replicated slides

**Motivation** To combine information in $M$ values, taking into account their variability within and between slides. Our Bayesian approach is meant to be a vehicle for doing this, but we do not take the probabilities implicit in it seriously. Mainly, we want to avoid being misled by means involving outliers, while taking care not to be too impressed with unusually small variances.

**Sampling model** Let $M_{ij} = \log (R_{ij}/G_{ij})$ be the log ratio of our green ($G_{ij}$) and red ($R_{ij}$) intensities for a gene, $i = 1 \ldots m$ refers to the slides and $j = 1 \ldots n$ to replicates within slides. We suppose
\[ EM_{ij} = \begin{cases} 
\mu_i & \text{with probability } p \\
0 & \text{with probability } 1 - p 
\end{cases} \]

We also suppose

\[ M_{ij} \sim pN(\mu_i, \sigma^2) + (1 - p)N(0, \sigma^2) \]

but with variances \( \sigma^2_b \) between slides and \( \sigma^2_w \) within slides, \( \sigma^2 = \sigma^2_b + \sigma^2_w \). If \( i = j = 2 \)

\[
Cov(M_{ij}) = \begin{pmatrix}
\sigma^2_b + \sigma^2_w & \sigma^2_w & 0 & 0 \\
\sigma^2_w & \sigma^2_b + \sigma^2_w & 0 & 0 \\
0 & 0 & \sigma^2_w + \sigma^2_b & \sigma^2_w \\
0 & 0 & \sigma^2_w & \sigma^2_b + \sigma^2_w
\end{pmatrix}.
\]
**Priors for $\mu$ and $\tau$**

For an integer $\nu$ and $a > 0$, $c > 0$

\[ a\nu\tau \sim \chi^2_\nu \]

\[ N \left( 0, (c\tau)^{-1} \right)^I \cdot \delta(0)^{1-I} \quad \text{where} \quad \delta(0) = \begin{cases} 1 & \text{if } \mu = 0 \\ 0 & \text{if } \mu \neq 0 \end{cases} \]
Technical point

It will be noted that we have chosen the standard conjugate prior for our normal means and variances. This was with the aim of getting a simple formula to use for the posterior odds ratio, see below. However, there appear to be no closed form expressions (simple or otherwise) when there are two components of variance, even in this balanced case. MCMC methods might work here, but we have 5,000 small samples, and have yet to try it out.

For these technical reasons, we therefore suppose that

$$\sigma_w^2 = k_1 \sigma_b^2, \text{ with } k_1 \text{ known.}$$

In fact we use the parametrization

$$\tau^{-1} = \sigma_w^2 + n\sigma_b^2 \text{ and } \sigma_w^2 = k\tau.$$
The log odds ratio $R$

We are interested in whether or not the gene is differentially expressed, i.e. whether $\mu \neq 0$ or $\mu = 0$, so we calculate the log posterior odds ratio

\[
R = \log \frac{Pr(\mu \neq 0|(M))}{Pr(\mu = 0|(M))} = \frac{p}{1 - p} \left( \frac{c}{c + mn} \right)^{1/2} \left( \frac{va + mnM^2 + SSB + kSSW}{va + mnM^2 + SSB + kSSW - \frac{(mnM^2)}{c + mn}} \right)^{\nu + mn - 1}.
\]
ACKNOWLEDGEMENTS

Yee-Hwa Yang
Ingrid Lonnstedt
Natalie Roberts
Sandrine Dudoit
Suzie Grant