Analysis of cDNA microarrays II

Terry Speed

WEHI Bioinformatics
WHICH GENES HAVE CHANGED EXPRESSION LEVELS?

Using Multiple slides:

1. For each gene and each experiment (mouse), use
   \[ M = \log_2 \frac{R}{G} \]

2. For each gene, form the \( t \) statistic
   \[
   \frac{\text{average of 8 ttt Ms} - \text{average of 8 ctl Ms}}{\sqrt{\frac{1}{8} \left[ (\text{SD of 8 ttt Ms})^2 + (\text{SD of 8 ctl Ms})^2 \right]}}
   \]

3. Form a histogram of the 5,600 \( t \) values

4. Rescale the cumulative distribution function (normal qq plot)
5. Look for values off the line
   Relies on approximate normality of the 5,600 $t$ values
ASSIGNING p-VALUES TO MEASURES OF CHANGE

Controls and replicates essential to this randomization analysis.

1. Choose 8 mice at random from the 16; label them ttt* and the rest ctl*,
2. Calculate statistics (e.g. $t$ values) using ttt* and ctl*,
3. Compute new statistics (e.g. $t_{max}$) from the 5,600 statistics,
4. Go back to 1 and repeat a few thousand times;
5. Summarize the results at 3 (e.g. form the c.d.f of $t_{max}$).
6. Use 5 to assign p-values (e.g. for $t_{max}$). Many variants.
ASPECTS OF GENES WHICH SEEM TO CHANGE

Estimate of change: numerator of $t$ value
Measure of precision this estimate: denominator of $t$
Normalized measure of change: $t$ statistic

Each of these can be related to:

Absolute expression level: given by average
$\mathbf{A} = \log_2 \sqrt{\mathbf{R} \times \mathbf{G}}$

Could they be found in one experiment?
Assess the by plotting the 5,600 $M$ values against their $\mathbf{A}$ values for each mouse.
One M vs A plot
$T$ - plot

- $t$ vs. average $A$
- $t$ denominator vs. average $A$
- $|t$ numerator$|$ vs. average $A$
- $t$ denominator vs. $|t$ numerator$|$
PRECISION OF SINGLE $\log_2(RATIO)$ MEASUREMENTS

Within a single array: very limited analysis
From 3 replicated cDNAs, we find SDs of $M$ values ($\log_2$ scale) ranging from 0.2 to 0.8, corresponding to a 1.2 and 2.2 fold change.

Between arrays (and mice): comprehensive analysis
The 5,600 pooled SDs from each gene’s 16 $M$ values across mice had lower quartile 0.3, median 0.5 and upper quartile 2.5, corresponding to 1.2, 1.4 and 5.6 fold changes.

For averages over 8 replicates:
The median of 5,600 pooled SEs was 0.2, corresponding to 1.1 fold change.
TENTATIVE CONCLUSIONS

• Keeping the $A = \log_2 \sqrt{R \times G}$ values is probably a good idea

• The $M = \log_2 \frac{R}{G}$ numbers are pretty variable, especially those with low $A$ values

• Replication of arrays is desirable, probably essential in this type of experiment

• Replication of cDNAs within arrays is desirable, probably essential in single array experiments

• Detecting changes in $M$ for genes with low $A$ is hard: it is the ultimate measure of the quality of your experiment
Involving a statistician in your microarray work might pay off.
Analysis of 8 ttt mice only

1. For each gene and each experiment (mouse), use
   \[ M = \log_2 \frac{R}{G} \]

2. For each gene, form the \( t \) statistic
   \[
   t = \frac{\text{average of 8 ttt Ms}}{\sqrt{\frac{1}{8} \times (SD \ of \ 8 \ ttt \ Ms)^2}}
   \]

3. Form a histogram of the 5,600 \( t \) values

4. Rescale the cumulative distribution function (normal qq plot)

5. Look for values “off the line” Relies on approximate normality of the 5,600 \( t \) values
Follow-up to Apo AI (and SR-BI) experiment(s)

• Check the clones giving the 20 most extreme changes in each experiment. Some are actually more than one clone. Purify and re-check.

• Spot out eight times each distinct clone from the top 20 clones of both experiments as well as some other clones (\(\sim 50 \times 8\) from Apo AI, \(\sim 50 \times 8\) from SR-BI, some genomic).

• Pool mRNA from 4 Apo AI KO mice (ttt) and similarly from 4 WT C57B6 mice (ctl) (do likewise for the SR-BIs).

• Do two Apo AI KO/WT B6 hybridizations, one with ttt → red, ctl → green, and one with ttt → green, ctl
→ red (repeat for SR-BI).
Analysis of the Apo AI KO data

• Here we analyze that part of the experiment comparing the expression of ~20 genes in the livers of Apo AI KO mice with that from WT C57/B6 mice.

• Genes corresponding to the ~50 clones from the top 20 of the initial SR-BI Tg/WT FEV comparison were not differentially expressed in the Apo AI part of the initial experiment, but are on the slide.

• These genes can be used to normalize the data so that we can (re-)examine differential expression of the ~20 genes from the original Apo AI experiment.

• If we did not have this set of genes in the follow-up experiment, normalization would have been difficult,
as most of the $\sim 20$ genes from the Apo Al experiment are differentially expressed.
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 seem to need to bump up the red intensity a bit.

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

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.
Apo A1: Comparing normalization lines
Self Normalizing plots: Log(R/G) - Log(R'/G')

Log Ratio (base 2)

10 11 12 13 14 15

-1.0 -0.5 0.0 0.5

59, 60: Sterol Desaturase
13, 14: Apo CIII
2: Apo AI
Why taking differences is self-normalizing when...

Suppose that we use a normalizer consisting of adding $c$. This may or may not be gene dependent, i.e. we may use $c = c(A)$.

Then $c$ is \textit{added} to $\log R/G$, and so when R and G dyes are reversed, we would \textit{subtract} $c'$ from $\log G'/R'$.

Now the combined estimate of relative expression

$$\frac{1}{2} [\log R/G + c + \log G'/R' - c'] = \frac{1}{2} \log \frac{RG'}{GR'}$$

provided $c = c'$. 
In other words, subtracting is self normalizing provided the separate normalizations are the same. We may not always be able to check this.
Use of reciprocally labeled paired slides and within-slide replicates

Let $M = \log R/G$, $M' = \log R'/G'$, and $\Delta = \frac{1}{2} (\bar{M} - \bar{M}')$, where ' refers to the slide with R and G reversed.

Then

$$\text{var}(\bar{\Delta}) = \frac{1}{2} [\text{var}(\bar{M}) + \text{var}(\bar{M}') - 2\text{cov}(\bar{M}, \bar{M}')]$$

and the last (covariance) term is usually assumed to be 0.

We specify the model for $M$ as follows:

$$M = \text{true log ratio} + \text{slide effect} + \text{spot effect}$$

We assume that slide effect and spot effect are uncorrelated random variables.
with corresponding variances $\sigma_{slide}^2$ and $\sigma_{spot}^2$

Suppose we have $r$ replicates. Then

$$\text{var}(\bar{M}) = \sigma_{slide}^2 + \frac{\sigma_{spot}^2}{r}.$$ 

Our task now is to estimate $\sigma_{slide}^2$ and $\sigma_{spot}^2$ using ANOVA.
Anova for 2 SLIDES/r SPOTS

Let $M_{1j} = \log \frac{R_{1j}}{G_{1j}}$, $M_{2j} = \log \frac{G_{2j}}{R_{2j}}$, $j = 1 \ldots r$, where the second ratio is from the reverse labelled experiment. Apart from variability, these should estimate the same thing.

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>E(MS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between slides</td>
<td>1</td>
<td>$r \sum_i (M_i - \bar{M})^2$</td>
<td>$r\sigma_{slide}^2 + \sigma_{spot}^2$</td>
<td></td>
</tr>
<tr>
<td>Within slides</td>
<td>$(r - 1)$</td>
<td>$\sum_j (M_{1j} - \bar{M}_1)^2$</td>
<td></td>
<td>$\sigma_{spot}^2$</td>
</tr>
<tr>
<td>Between spots</td>
<td>$(r - 1)$</td>
<td>$+ \sum_j (M_{2j} - \bar{M}_2)^2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$2r - 1$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The variance components are estimated by $\hat{\sigma}_{slide}^2 = \frac{(MS_{between} - MS_{within})}{r}$
and $\hat{\sigma}^2_{\text{spot}} = MS_{\text{within}}$.

Note that $\hat{\sigma}^2_{\text{slide}}$ is estimated on 1 df only: the estimate is very noisy and very non-robust.
ACKNOWLEDGEMENTS

To the biologists who have drawn me into this business: Lynn Corcoran, WEHI; Chuang Fong Kong and David Bowtell, PMCI; Tito Serafini, UCB; and Matt Callow, LBNL.

To the numbers people who have helped me so much: Mary Pat Reeve and Natalie Roberts, WEHI; Sandrine Dudoit, MSRI; and Jean Yee-Hwa Yang, UCB.

And to: Alyssa Barry, WEHI.