Single channel normalization for two-channel microarray expression data

Terry Speed
Department of Statistics, UC Berkeley
Walter & Eliza Hall Institute of Medical Research

International Conference on the Analysis of Genomic Data
Harvard Medical School, May 11, 2004
Normalization

**Why?**

To correct for systematic differences between samples on the same slide or chip, or between slides or chips, which do not represent true biological variation between samples.

**How do we know it is necessary?**

By examining replicate and or self-self hybridizations, where no true differential expression is occurring.

**We find** biases which vary with overall probe intensity, location on the array, dye, plate origin, pins, scanner, scanning parameters, ….
How we normalize Affymetrix chips

- Quantile normalization is a method to make the distribution of probe intensities the same for every chip.
- The normalization distribution is chosen by *averaging each quantile* across chips.
Depicting the normalization density
MA-plots of chip pair  PMs: before normalization
MA-plots of chip pair PMs: after quantile normalization
Normalization reduces variance and does not attenuate the signal

It is easy to show that normalization reduces the variance between replicate chips.

A more complex analysis shows that we don’t seem to lose signal, i.e. we don’t introduce any bias. We omit the details here.
Now we turn to two-channel cDNA microarray data

Our plan

Why perform single-channel (SC) analysis?
What is normalization in this context?
Brief review of two-channel normalization.
Some SC normalization methods.
Assessing SC normalization methods.
Why SC analysis?

- Sometimes the question of interest is suited to, or requires single-channel analysis. E.g: what genes are expressed at each point in a time course? … the experiment is not necessarily comparative.
- The complexity of multi-factor experimental designs sometimes means that not all comparisons (contrasts of interest) are estimable from log-ratios within slides. Jin et al (2001) performed a factorial experiment on age, sex and genotype (each at two levels) of Drosophila melanogastor, and age was the only factor compared within slides i.e. age was the only main effect that was estimable based on the log-ratio.
Experiment of Jin et al (2001)

Young Samarkand → Old Samarkand × 6
Young Oregon R → Old Oregon R × 6
Young Samarkand → Old Samarkand × 6
Young Oregon R → Old Oregon R × 6

= 24

- Analysis methods on log-intensities are being used and investigated. Wolfinger et al and Kerr et al are modelling log-intensities rather than log-ratios.
- Our motivation was different. Details omitted.
- Satisfactory single-channel normalization is what is needed for single-channel cDNA microarray analysis to be considered a real option for researchers.
Within and/or between slides?

Since cDNA microarrays are inherently comparative in nature, normalization methods in this context have usually focused on log-ratios within each slide. By contrast, normalizing Affymetrix chip data is on log signal between slides. With SC cDNA normalization, we need both.
Two-channel normalization: Within slides
Global scale, global lowess, pin-group lowess; spatial plot after, smooth histograms of M after
Two-channel normalization: Between slides
Lowess Normalized M

Apo A1 Experiments
The “NCI 60” experiments (no bg)

There are various ways to equalize scale in these examples. We omit the details.
Single channel normalization
Olfactory Epithelium Dataset

We illustrate the problem of single-channel normalization with a time series data set on the olfactory epithelium (OE) of embryonic mice with two replicates of all possible pairwise comparisons of stages E13 till E18.
$E_{16_1} \text{ vs } E_{14_1}$

within slide log-ratios

$E_{16_2} \text{ vs } E_{14_2}$

$E_{16_1} \text{ vs } E_{16_2}$

between slide log-ratios

$E_{14_1} \text{ vs } E_{14_2}$
Log-ratios *within* slides are consistently less variable than log-ratios constructed from single-channels between slides.
Single-channel normalization can also be considered in two stages, **within slide** and **between slide** (i.e. between all single channels) normalization.

Within slide SC normalization is similar to the two-channel normalization: we adapt methods for correcting intensity and spatially dependent dye-imbalances to allow the recovery of single-channel information.

Between slide SC normalization is concerned with making the single channels comparable between slides. Here we use methods that work with **Affymetrix chip data**, in particular the **quantile normalization** method.
p = print-tip-group SC normalization

Let $c_i(A)$ denote the $A$-dependent adjustment to $M$ in print-tip-group lowess normalization for spots in group $i$.

Now adjust the red (Cy 5) signal $R$ and the green (Cy 3) signal $G$ via

$$\log R_p = \log R - 0.5 \, c_i(A)$$
$$\log G_p = \log G + 0.5 \, c_i(A).$$

It follows that $M_p = M - c_i(A)$ and $A_p = A$. 
q = quantile normalization

Here we adapt the idea from normalizing Affymetrix chips. We can do this with either log R and log G for all slides, which we denote by the subscript q ($R_q$, $G_q$), or we can do it to the A-values, in which case we write $A_q$.

If we normalize A-values and M-values separately, we can reconstruct log R and log G via

$$\log R = A + 0.5M, \quad \log G = A - 0.5M.$$ 

In this way we can mix and match, e.g. $M_p A_q$. 25
Density plots of all 60 single-channel log-intensities from the OE dataset
Mₚ Aq; reconstructed R, G(+1)
After MpAq normalization: log-ratios within slides are still consistently less variable than log-ratios constructed from single-channels between slides.
Assessing SC normalization

Reduces variance: relatively easy to show.

Doesn’t introduce bias: harder to show since we need some truth about absolute or relative gene expression.
Here we revisit the Apo AI data, originally 8 KO vs pooled controls and 8 WT vs pooled controls. We summarize the original analysis with 16 2-channel slides, then redo the analysis ignoring the reference samples, i.e. using just 16 single (red) channels. The first analysis is our “truth”.
Which genes have changed?

1. For each gene and each hybridisation (8 ko + 8 ctl), use $M=\log_2(R/G)$.
2. For each gene form the t statistic:

$$\text{average of 8 ko Ms} - \text{average of 8 ctl Ms}$$
$$\sqrt{\frac{1}{8} (\text{SD of 8 ko Ms})^2 + (\text{SD of 8 ctl Ms})^2}$$

3. Form a histogram of 6,000 t values.
4. Do a normal q-q plot; look for values “off the line”.
Histogram & normal q-q plot of t-statistics

ApoA1
SC normalization reduces variance
Tail of normal qq plots of 2-sample t-statistics obtained using only red channels
Another test involves a dataset in which 47 genes were measured by qrt-pcr.

A third test involves many slides on which Scorecard data were available.

Yet a fourth test involves slides with a microarray sample pool in a dilution series.

Conclusion: no evidence of loss of signal following SC normalization
Acknowledgments

WEHI Bioinformatics
Natalie Thorne
Gordon Smyth
Matthew Ritchie
Asa Wirapati

UC San Francisco
Jean Yang

Mt Sinai Medical Center  UC Berkeley
Stuart Sealfon’s lab  Ngai lab
QRT-PCR data

Courtesy of Stuart Sealfon, we have access to QRT-PCR measurements on 47 genes which were on 3 cDNA microarray slides. In each case, determinations were made of mRNA concentrations in 3 experimental samples, E1, E2 and E3, and three control samples C1, C2 and C3, while E1 was compared with C1, on the same slide, etc. Further, each probe was spotted in triplicate.
Intensity distributions for the 3 slides
Reduction of triplicate log ratio variance: within and between slides

Within slides: 1R to 1G, etc

Between slides: 1R to 2G, etc
Reduction of variance in separate channels
No apparent loss of signal (no gain either): 47 genes
As in previous slide: averaged across 3 slides
SC normalization reduces variance: the MSP titration series

In each print-tip-group there are 5 MSP controls put in at known concentrations.

Since we have 32 print-tip-groups, we have 5×32 = 160 controls for which we can measure the variance across the 60 channels before and after different normalisation methods.
SC normalization reduces variance

Different SC normalization methods
SC normalization generally reduces bias: the MSP titration series

Known: relative increase, ratio => slope of the line.
Assessing bias, cont.

Before normalization

Log(Intensity)

Slope = 1

Slope = 0.6

Conc. ng/ul
Assessing bias, cont.

After normalization (we hope!)

Log(Intensity)

Conc. ng/ul

Slope = 1

Slope = 0.9
Measure of slope discrepancy

- For each print-tip-group we calculate \(1 - \text{slope} = \text{discrepancy}\).
- We expect that if normalization is not losing signal, the slope gets closer to 1. Therefore, slope discrepancy should get smaller.
- We measure slope discrepancy for each print-tip-group within each of the 60 single-channels and draw boxplots of these before and after different normalization methods.
Does the variance of the titration series decrease differently for higher or lower concentrations?
Taking scale into account: between slides

Assumption: All slides have the same spread in M

True log ratio is $\mu_{ij}$ where $i$ represents different slides and $j$ represents different spots.

Observed is $M_{ij}$, where

$$M_{ij} = a_i \mu_{ij}$$

Robust estimate of $a_i$ is

$$MAD_i = \text{median}_j \{ |y_{ij} - \text{median}(y_{ij})| \}$$