Summarizing and comparing GeneChip® data

Terry Speed,
UC Berkeley & WEHI, Melbourne

Affymetrix Users Meeting,
Friday June 7, 2002
Redwood City, CA.
Acknowledgements

• Rafael Irizarry and Leslie Cope, JHU
• Francois Collin, GeneLogic
• Ben Bolstad, UCB
• Bridget Hobbs, WEHI
• Gene Brown’s group at Wyeth/Genetics Institute, and Uwe Scherf’s Genomics Research & Development Group at Gene Logic, for generating the spike-in and dilution data
• Gene Logic and Affymetrix for permission to use their data
Summary of my talk

• Review of GeneChip® technology
• Probe set summaries
• What we do: our 4 steps
• Assessing and using different probe set summaries to measure differential expression
• What is left to do?
A probe set = 11-20 PM,MM pairs

There may be 5,000-20,000 probe sets per chip
Summarizing 11-20 probe intensity pairs to give a measure of expression for a probe set

A key issue. The original Affymetrix solution left something to be desired (just joking).

There are many possible low-level summaries, but they fall into three main classes. In part this talk is about comparing them. In part is about using them to measure differential expression.
Competing measures of expression

• The original GeneChip® software used $AvDiff$

\[
AvDiff = |A|^{-1} \sum_{\{ij \in A\}} (PM_j - MM_j)
\]

where $A$ is a suitable set of pairs chosen by the software. Here 30%-40%- could be $<0$, which was a major irritant.

• $\log PM_j / MM_j$ was also used in the above.
Competing measures of expression, 2

- Li and Wong (dChip) fit the following model to sets of chips

\[ PM_{ij} - MM_{ij} = \square_i \square_j + \square_j \]

where \( \square_i \sim N(0, \square^2) \). They consider \( \square_i \) to be expression in chip \( i \). Their model is also fitted to PM only, or to both PM and MM. Note that by taking logs, assuming the LHS is >0, this is close to an additive model.

- Efron *et al* consider \( \log PM_j - 0.5 \log MM_j \). It is much less frequently <0.

- Another summary is the second largest PM, \( PM_{(2)} \).
Competing measures of expression, 3

- The latest version of GeneChip® uses something else, namely

\[ \text{Log(Signal Intensity)} = \text{TukeyBiweight}\{\log(\text{PM}_j - \text{MM}_j^*)}\] 

with \( \text{MM}_j^* \) a version of \( \text{MM}_j \) that is never bigger than \( \text{PM}_j \). Here *TukeyBiweight* can be regarded as a kind of robust/resistant mean.
What we do: four steps

We use only PM, and ignore MM. Also, we

- Adjust for **background** on the raw intensity scale;
- Carry out **quantile normalization** of PM-*BG with chips in suitable sets, and call the result n(PM-*BG);
- Take $\log_2$ of normalized background adjusted PM;
- Carry out a **robust multi-chip** analysis (RMA) of the quantities $\log_2 n(PM-*BG)$.

We call our approach RMA.
Why we ignore the MM values

The reason is that we haven’t yet found a way to do better using them. They definitely have information, - about both signal and noise - but using it without adding more noise (see below) seems to be a challenge.

We should be able to improve the BG correction using MM, without having the noise level blow up: work in progress.
Why (how) we remove background

White arrows mark the means.
Why (how) we normalize

Density of PM probe intensities for Spike-In chips

After Quartile Normalization
Why (how) we take $\log_2$
Why (how) we write
\[ \log_2(n(\text{PM-}*\text{BG})) = \text{chip effect} + \text{probe effect} \]

Because:

probe effects are additive on the log scale

Spike in data set A: 11 control cRNAs spiked in, all at the same concentration, which varies across 12 chips. The example on the next slide is typical of the set of 11. Every set of experiments should exhibit this parallel behaviour across probes.
Probe level data exhibiting parallel behaviour on the log scale
Why we carry out a Robust Multi-chip Analysis

Why multi-chip?
To put each chip’s values in the context of a set of similar values.
This helps even if we do not do so robustly.

Why robust?
To get even more out of our multi-chip analysis.
In the old human and mouse series, perhaps 10%-15% of probe level values are “outliers”.
Robust summaries really improve over the standard ones, by down weighting outliers and leaving their effects visible in residuals.
How we carry out our Robust Multi-chip Analysis

We base our analysis on the linear model embodying the parallel behaviour noted:

$$\log_2 n(\text{PM}_{ij} - \ast \text{BG}) = m + a_i + b_j + e_{ij}$$

where $i$ labels chips and $j$ labels probes. Our current implementation uses median polish. It is like Tukey’s biweight, but in the 2-way array of chips by probes, and we adjust for probe affects.
RMA in summary

- We *background* correct PM on original scale
- We carry out *quantile* normalization
- We take $\log_2$

Under the *additive* model

$$\log_2 n(\text{PM}_{ij} - \text{BG}) = m + a_i + b_j + \epsilon_j$$

- We estimate chip effects $a_i$ and probe effects $b_j$
  using a *robust/resistant* method.
Nothing is new under the sun

Felix Naef and colleagues at Rockefeller explained a nice way of doing a **background** adjustment, and pioneered **PM only** analyses on the **log** scale.

Wing Wong and colleagues at UCLA, now Harvard pioneered **multi-chip analyses, non-linear normalizations**, and probe effect x chip effect models.

Dan Holder and colleagues at Merck used **additive models** after a linear-**log** hybrid transformation and **fitted robustly**. The software MAS5.0 now uses a robust method too, but only on one or two chips.
Comparisons

A (non-Bayesian) statistician’s view of the world: life is a trade-off, either of

- Bias/variance (accuracy/precision), or
- False positives/true positives.

To place ourselves on the spectrum, we need some truth. Often hard to come by, but we have some special data sets from GeneLogic and Affymetrix.

We begin looking at variability (SD) across replicates.
Average and SD of 5 SN19 Samples stratified by average expression
HG_U133 Chips

Distribution of Average Expression by Stratum

Distribution of SD of Expression by Stratum
Later we consider many different combinations of concentrations.

<table>
<thead>
<tr>
<th>Probe Set</th>
<th>Conc 1</th>
<th>Conc 2</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioB-5</td>
<td>100</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>BioB-3</td>
<td>0.5</td>
<td>25.0</td>
<td>2</td>
</tr>
<tr>
<td>BioC-5</td>
<td>2.0</td>
<td>75.0</td>
<td>4</td>
</tr>
<tr>
<td>BioB-M</td>
<td>1.0</td>
<td>37.5</td>
<td>4</td>
</tr>
<tr>
<td>BioDn-3</td>
<td>1.5</td>
<td>50.0</td>
<td>5</td>
</tr>
<tr>
<td>DapX-3</td>
<td>35.7</td>
<td>3.0</td>
<td>6</td>
</tr>
<tr>
<td>CreX-3</td>
<td>50.0</td>
<td>5.0</td>
<td>7</td>
</tr>
<tr>
<td>CreX-5</td>
<td>12.5</td>
<td>2.0</td>
<td>8</td>
</tr>
<tr>
<td>BioC-3</td>
<td>25.0</td>
<td>100</td>
<td>9</td>
</tr>
<tr>
<td>DapX-5</td>
<td>5.0</td>
<td>1.5</td>
<td>10</td>
</tr>
<tr>
<td>DapX-M</td>
<td>3.0</td>
<td>1.0</td>
<td>11</td>
</tr>
</tbody>
</table>
Displaying differential expression

In the following slides: for each gene we plot the log fold change $M$ across two chips, given by

$$\log (\text{chip 1}/\text{chip 2}) = \log(\text{chip 1}) - \log(\text{chip 2}),$$

vertically, against overall abundance $A$, measured by

$$\log\sqrt{(\text{chip 1})(\text{chip 2})} = [\log(\text{chip 1}) + \log(\text{chip 2})]/2$$

horizontally. This is just a rotated version of the plots everyone else uses.
Quantile-quantile plots

These are used by statisticians to compare distributional shape, and to highlight extremes, relative to a reference distribution for the majority. In a sense they are just cumulative distributions with the axes rescaled.

In the examples which follow, the reference distribution is normal (Gaussian) for the log fold change M. The better the fit of the majority, the straighter the line, and the more the extremes will stand out. Everything is on the log scale.
Differential Expression

Avg.Diff MVA plot

Avg.Diff QQ-plot
Differential expression

MAS 5.0 MVA plot

MAS 5.0 QQ-plot
Differential expression

Li and Wong's θ MVA plot

Li and Wong's θ QQ-plot
Differential expression

Average log(PM–BG) MVA plot

Average log(PM–BG) QQ–plot
Differential expression

RMA MVA plot

RMA QQ-plot
Receiver Operating Characteristic curves: single chip comparisons

ROC curves compare the true and false positive rates at varying cut-off values for a stated criterion such as fold change.

In the next four plots we compare ROC curves from MAS, LW and RMA using either fold change (FC) or the associated p-values, in single chip comparisons using spike-in data.

We have pooled the results from a number of 1 chip vs 1 chip comparisons, to get a smooth ROC curve.
ROC based on fold change for Gene Logic 1

- MAS 5.0
- Li and Wong
- RMA

True Positives vs. False Positives
Conclusions from single chip comparison ROC curves

On the basis of the data just presented, and much more:

With **FC**, RMA is best, LW next. MAS does not do well here.

With **p-values**, RMA is as good as, and usually better than MAS, which is next. MAS does best on the Affymetrix spike-in data sets. LW (dChip) does not do so well here.

All judgements are **comparative**. Everyone does well in absolute terms, but some do better.
Comparing expression summaries and test statistics with replicated data

Here we display ROC curves to compare expression measures and test statistics with replicated spike-in data.

We have used a subset of 24 of the 59 chips from the Affymetrix spike-in study, which is 2 sets of 12 where all probe sets are at the same concentration. This gives us two populations, which for \( N = 2, 3, 4, 6 \) and 12 we divide into \( 12/N \) subgroups and use measures and test statistics to determine true and false positives. One ROC curve for all is calculated.
Key to figures

MAS, LW and RMA: probe set summaries

FC: Decision based on estimated Fold Change (log scale)
t-test: estimated log FC/estimated SE of estimated log FC
Nominal: t-statistic from model (LW & RMA only)
SAM: t-statistic with slightly inflated SE.

FC Comp: Collecting the FC components of MAS, LW & RMA
Comparisons using FC, N = 2, 3, 4, 6 and 12.

a) Comparison of measures (N=2) using Affy Spikeln

b) Comparison of measures (N=3) using Affy Spikeln

c) Comparison of measures (N=4) using Affy Spikeln

d) Comparison of measures (N=6) using Affy Spikeln

e) Comparison of measures (N=12) using Affy Spikeln
MAS: Comparison of FC, t and SAM: N = 2, 3, 4, 6 and 12.
Conclusions from replicate chip ROC curves

Overall **RMA** does better than LW, which in turn does better than MAS 5.0 using FC.

The simple \( t = \frac{\text{est log FC}}{\text{SE(\text{est log FC})}} \) seems best for use with MAS (and RMA, data not shown).

**MAS** looks bad here because we use single chip summaries in our analysis. They need a multi-chip version of their Signal Log Ratio. When done, it will look like the final step in RMA.

With RMA and LW, nominal SEs are not as good as observed ones and p-values are better than (log) fold change: (data not shown).
Observed versus true ratio for all spike in experiments

(a) MAS 5.0

slope = 0.81, R^2 = 0.72
c) RMA

slope = 0.71, R^2 = 0.73
d) Log Fold Change of Non-Differentially-Expressed Genes
Smaller scale comparisons are more revealing
b) Li and Wong

slope = 0.37, R^2 = 0.79
c) RMA

slope = 0.4, R^2 = 0.82
Conclusions from estimated versus true fold change plots.

From many plots of the kind just shown, MAS 5.0 has less bias in comparison with RMA and dChip.

The problem with MAS right now is that it pays a very large price in extra variability for this low bias. Overall, a little bias but greatly reduced variance puts RMA ahead (in my view). There is much more evidence to support this view not shown today.
A few remaining questions

Now that the probe sequences are available, it is a challenge to make use of them to compute better expression summaries.

Use of RMA residuals in quality assessment - of parts or all of chips, of probes - work in progress. Very promising.

To pool or not to pool? How many replicates? What kind replicates? How can we adjust for the host of systematic effects that manifest themselves in GeneChip data?

Low level analysis never ends…as the technology evolves, we need to go with it, answering these questions as we go.
The talk ended here. A web site for more info is

For software, see
www.bioconductor.org
and for the package specifically used here, see
www.biostat.jhsph.edu/ririzarr/affy

The slides that follow give more detail on some aspects of the
talk. They need to be interleaved with the preceding ones. A
couple of papers on this are in preparation.
Dilution experiment

- cRNA hybridized to human chip (HGU95) in range of concentrations. Two different RNA sources were used each time.
- Dilution series begins at 1.25 µg cRNA per GeneChip array, and rises through 2.5, 5.0, 7.5, 10.0, to 20.0 µg per array. 5 replicate chips were used at each dilution
- Normalize just within each set of 5 replicates
- For each of 12,000 probe sets, we compute expression, average and SD over replicates
a) Median Expression vs. Concentration for Liver

- m: MAS 5.0, slope = 0.63
- L: Li and Wong, slope = 0.58
- R: RMA, slope = 0.67

Concentration vs. Median Expression graph with different methods indicated.
c) Log Scale Standard Deviation

- **MAS 5.0**
- **Li and Wong**
- **RMA**

**Axes:**
- Y-axis: Standard Deviation between Replicates
- X-axis: Expression

The graph shows three curves representing the standard deviation between replicates for different expression levels. The curves are color-coded as follows:
- **Red**: MAS 5.0
- **Green**: Li and Wong
- **Blue**: RMA
Comparisons of log fold change estimates:
20μg versus 1.25μg.

a) MAS 5.0
b) Li and Wong

The diagram shows a scatter plot with two axes labeled 'Fold change estimate for 20 Ug' and 'Fold change estimate for 1.25 ug'. The data points are scattered across the plot, with a trend line indicating a linear relationship between the two variables.
Background model: pictorially

\[ \text{Signal} \quad + \quad \text{Noise} \quad = \quad \text{Observed} \]
PM data on log$_2$ scale: raw and fitted model
How we remove background

- Observed PM intensity denoted by S.
- Model S as the sum of a signal X and a background Y, S=X+Y, where we assume X is exponential (a) and Y is Normal (m, s²), X, Y independent random variables.
- Background adjusted values are then E(X|S=s), which is

\[
a + b\left[\frac{f(a/b) - f((s-a)/b)}{F(a/b) - F((s-a)/b) - 1}\right],
\]

where \(a = s - m - s^2\), \(b = \), and \(f\) and \(F\) are the normal density and cumulative density, respectively.

This is our model and formula for background correction. Call the result PM-*BG, the * indicating not quite subtraction.
Observed PM vs PM-*BG

As $s$ increases, the background correction asymptotes to

$$s - m - m^2.$$

In practice, $m >> m^2$, so this is $\sim s - m$. 
How we normalize

• 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.
• The diagram that follows illustrates the transformation.
• Call the result n(PM-*BG).
Quantile normalization: in words

• The two distribution functions are effectively estimated by the sample quantiles.
• Quantile normalization is fast
• After normalization, variability of expression measures across chips reduced
• Looking at post-normalization PM vs pre-normalization PM (natural and log scales), you can see transformation is non linear.
One distribution for all: the black curve
After vs Before: intensity scale
After vs Before: log intensity scale
M v A plots of chip pairs: before normalization
M v A plots of chip pairs: after quantile normalization
Our normalization reduces variability in comparison with nothing and Affy.

Quantile vs Un-normalized
Quantile vs Affymet. normalized

Vertical: log[var q. norm/var other]; Horizontal: Aver. log mean
Note differences in vertical scales
How we normalize

- 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.
- The diagram that follows illustrates the transformation.
- Call the result \( n(\text{PM-}^*\text{BG}) \).
Quantile normalization: in words

- The two distribution functions are effectively estimated by the sample quantiles.
- Quantile normalization is fast.
- After normalization, variability of expression measures across chips reduced.
- Looking at post-normalization PM vs pre-normalization PM (natural and log scales), you can see transformation is non-linear.
One distribution for all: the black curve
After vs Before: intensity scale
After vs Before: log intensity scale
M v A plots of chip pairs: before normalization
M v A plots of chip pairs: after quantile normalization
Our normalization reduces variability in comparison with nothing and Affy.

Quantile vs Un-normalized

Vertical: log[\text{var q. norm}/\text{var other}]; Horizontal: Aver. log mean

Quantile vs Affymet. normalized

Note differences in vertical scales
Average and SD of 5 SN19 Samples stratified by average expression

Distribution of Average Expression by Stratum

Distribution of SD of Expression by Stratum
Average and SD of 5 Liver Samples stratified by average expression

Distribution of Average Expression by Stratum

Distribution of SD of Expression by Stratum
Average and SD of 5 Liver Samples stratified by average expression
HG_U133 Chips

Distribution of Average Expression by Stratum

Distribution of SD of Expression by Stratum
Design summary (5 chips at each point)

Dilution/Mixture study

- cRNA sample A (ug)
- cRNA sample B (ug)

Graph showing dilution and mixture study with data points for cRNA sample A and cRNA sample B.
## Observed ranks

<table>
<thead>
<tr>
<th>Gene</th>
<th>AvDiff</th>
<th>MAS 5.0</th>
<th>Li&amp;Wong</th>
<th>AvLog(PM-BG)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioB-5</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>BioB-3</td>
<td>16</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>BioC-5</td>
<td>74</td>
<td>6</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>BioB-M</td>
<td>30</td>
<td>3</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>BioDn-3</td>
<td>44</td>
<td>5</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>DapX-3</td>
<td>239</td>
<td>24</td>
<td>24</td>
<td>7</td>
</tr>
<tr>
<td>CreX-3</td>
<td>333</td>
<td>73</td>
<td>36</td>
<td>9</td>
</tr>
<tr>
<td>CreX-5</td>
<td>3276</td>
<td>33</td>
<td>3128</td>
<td>8</td>
</tr>
<tr>
<td>BioC-3</td>
<td>2709</td>
<td>8572</td>
<td>681</td>
<td>6431</td>
</tr>
<tr>
<td>DapX-5</td>
<td>2709</td>
<td>102</td>
<td>12203</td>
<td>10</td>
</tr>
<tr>
<td>DapX-M</td>
<td>165</td>
<td>19</td>
<td>13</td>
<td>6</td>
</tr>
<tr>
<td>Top 15</td>
<td>1</td>
<td>5</td>
<td>6</td>
<td>10</td>
</tr>
</tbody>
</table>
Dilution series: before and after quantile normalization in groups of 5

Note systematic effects of scanners 1,…,5 in before boxplots
d) Standard Deviation

- MAS 5.0
- Li and Wong
- RMA

Expression vs. Replicate Standard Deviation graph with different lines representing each method.
Comparison of fold change estimates from mixture and dilution data

Fit the model:

$$\text{Log } E(\hat{a}) = \text{Log } [\hat{a}E(L) + (1-\hat{a})E(CNS)] + \epsilon$$

to estimate $$\log \{E(L)/E(CNS)\}$$.
b) Li and Wong

$r = 0.9$
c) RMA

$r = 0.94$
N = 4

MAS

LW

RMA

FC comp

False Positive

True Positive

False Positive

True Positive

False Positive

True Positive

False Positive

True Positive

False Positive

True Positive

False Positive

True Positive

False Positive

True Positive