Assessments Using Spike-In Experiments

Rafael A Irizarry,
Department of Biostatistics JHU
rafa@jhu.edu

http://www.biostat.jhsph.edu/~ririzarr
http://www.bioconductor.org
A probe set = 11-20 PM, MM pairs

GeneChip® Expression Array Design

There may be 5,000-20,000 probe sets per chip
Statistical Problem: Summarize 11-20 probe intensity pairs (PM and MM) to give a measure of expression for a probe set… also background correct and normalize

- GeneChip® software’s default until 2002 was \( \text{Avg. diff} \)

\[
\text{Avg. diff} = \frac{1}{|A|} \sum_{j \in A} (PM_j - MM_j)
\]

with \( A \) a set of “suitable” pairs chosen by software.

Obvious Problems:
- Many negative expression values
- No log transform
Why log?

SD vs. Avg for pm

SD vs. Avg for log2(pm)
Probe Level Analyses

- MAS 4.0 AvgDiff
- Li and Wong’s MBEI (dChip), PNAS 2001
- MAS 5.0 Signal
- RMA, Biostatistics 2003
- Others

How do we assess these?
Spike-In Experiments

• **Gene Logic A:** 11 control cRNAs were spiked in, all at the same concentration, which varied across chips.

• **Gene Logic B:** 11 control cRNAs were spiked in, all at different concentrations, which varied across chips. The concentrations were arranged in 12x12 cyclic Latin square (with 3 replicates)

• **Affymetrix:** 14 human cRNA were spiked in. Latin Square design used as well.
## Gene Logic Set A

### Groups of transcripts

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>B</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>C</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
</tr>
<tr>
<td>D</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>E</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
</tr>
<tr>
<td>F</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>G</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
</tr>
<tr>
<td>H</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
</tr>
<tr>
<td>I</td>
<td>12.5</td>
<td>12.5</td>
<td>12.5</td>
<td>12.5</td>
<td>12.5</td>
<td>12.5</td>
<td>12.5</td>
<td>12.5</td>
<td>12.5</td>
<td>12.5</td>
<td>12.5</td>
</tr>
<tr>
<td>K</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>L</td>
<td>75</td>
<td>75</td>
<td>75</td>
<td>75</td>
<td>75</td>
<td>75</td>
<td>75</td>
<td>75</td>
<td>75</td>
<td>75</td>
<td>75</td>
</tr>
<tr>
<td>M</td>
<td>150</td>
<td>150</td>
<td>150</td>
<td>150</td>
<td>150</td>
<td>150</td>
<td>150</td>
<td>150</td>
<td>150</td>
<td>150</td>
<td>150</td>
</tr>
</tbody>
</table>
Why correct background?

White arrows mark the means.

Concentration of 0

Concentration of 0.5

Concentration of 0.75

Concentration of 1
Why normalize?

Density of PM probe intensities for Spike-In chips

After Quantile Normalization
Why a log scale linear model?
Why ignore MM (for now)?
Comparisons

We study the trade-off 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 spike-in data sets from GeneLogic and Affymetrix.
# Affymetrix Latin Square

## Groups of transcripts

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>0</td>
<td>512</td>
<td>1024</td>
</tr>
<tr>
<td>B</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>0.25</td>
<td>1024</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>0.5</td>
<td>0</td>
<td>0.25</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>1024</td>
<td>1</td>
<td>0.25</td>
<td>0.5</td>
</tr>
<tr>
<td>E</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>1024</td>
<td>0</td>
<td>2</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>G</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>8</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>H</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>16</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>I</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>32</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>J</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>64</td>
<td>16</td>
<td>32</td>
</tr>
<tr>
<td>K</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>128</td>
<td>32</td>
<td>64</td>
</tr>
<tr>
<td>L</td>
<td>256</td>
<td>512</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>256</td>
<td>64</td>
<td>128</td>
</tr>
<tr>
<td>M</td>
<td>512</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>512</td>
<td>128</td>
<td>256</td>
</tr>
<tr>
<td>N</td>
<td>512</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>512</td>
<td>128</td>
<td>256</td>
</tr>
<tr>
<td>O</td>
<td>512</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>512</td>
<td>128</td>
<td>256</td>
</tr>
<tr>
<td>P</td>
<td>512</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>512</td>
<td>128</td>
<td>256</td>
</tr>
<tr>
<td>Q</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>1024</td>
<td>256</td>
<td>512</td>
</tr>
<tr>
<td>R</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>1024</td>
<td>256</td>
<td>512</td>
</tr>
<tr>
<td>S</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>1024</td>
<td>256</td>
<td>512</td>
</tr>
<tr>
<td>T</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>1024</td>
<td>256</td>
<td>512</td>
</tr>
</tbody>
</table>
Later we consider many different combinations of concentrations.
# 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>
We can also use spike-in data for assessing tests
N=3
Acknowledgements

• Terry Speed and Ben Bolstad, UCB
• Leslie Cope, JHU
• Francois Collin, GeneLogic
• 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
Supplemental Slides
Affymetrix GeneChip Arrays

Millions of copies of a specific oligonucleotide probe

Hybridized Probe Cell
- Single stranded, labeled RNA target
- Oligonucleotide probe

Image of Hybridized Probe Array

GeneChip Probe Array

1.28 cm

24 µm

>200,000 different complementary probes

Compliments of D. Gerhold
RMA in summary

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

Under the **additive** model

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

- We estimate chip effects $a_i$ and probe effects $b_j$ using a **robust/resistant** method.
Background model: pictorially

Signal + Noise = 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 (α) and Y is Normal (µ, σ²), X, Y independent random variables.
- Background adjusted values are then E(X|S=s), which is

\[ a + b[\phi(a/b) - \phi((s-a)/b)]/[\Phi(a/b) - \Phi((s-a)/b) - 1], \]

where \( a = s - µ - σ^2 \alpha \), \( b = σ \), and \( \phi \) and \( \Phi \) 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.
As $s$ increases, the background correction asymptotes to

$$s - \mu - \alpha \sigma^2.$$ 

In practice, $\mu >> \alpha \sigma^2$, so this is $\sim s - \mu$. 

Previous Work

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

Cheng Li, 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.
More plots from spike in
Normalization at Probe Level

log(PM) before normalization

A

M

0

1

2

-1

-2

-3

-4

-5

-6

-7

-8

-9

-10

-11

-12
Normalization at Probe Level

$log(PM)$ after normalization
ROC based on p-values for Gene Logic 1

- MAS 5.0
- Li and Wong
- RMA
ROC based on fold change for Gene Logic 1

- MAS 5.0
- LI and Wong
- RMA
How subtracting MM helps

- Affymetrix claims subtracting MMs yields an expression with less bias (accuracy)
- It seems to be true. Especially for lower intensities.
- But they pay a very large price in variance (precision)
Observed versus true ratio for all spikes in experiments
b) Li and Wong

slope = 0.63, R^2 = 0.73
c) RMA

slope = 0.71, R^2 = 0.73
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
Dilution Experiment

- cRNA hybridized to human chip (HGU95) in range of proportions and dilutions
- Dilution series begins at 1.25 $\mu$g cRNA per GeneChip array, and rises through 2.5, 5.0, 7.5, 10.0, to 20.0 $\mu$g per array. 5 replicate chips were used at each dilution
- Normalize just within each set of 5 replicates
- For each probe set compute expression, average and SD over replicates
Design summary (5 chips at each point)
RMA has smaller SD
Especially for low intensities
d) Standard Deviation

The graph shows the relationship between expression levels and replicate standard deviation for different methods:
- Red line: MAS 5.0
- Green line: Li and Wong
- Blue line: RMA

The y-axis represents the replicate standard deviation, while the x-axis represents the expression levels.
Do we sacrifice signal detection (bias)?

**a) Median Expression vs. Concentration for Liver**

- **MAS 5.0, slope = 0.63**
- **Li and Wong, slope = 0.58**
- **RMA, slope = 0.67**
Comparisons of log fold change estimates: 20\(\mu\)g versus 1.25\(\mu\)g.

a) MAS 5.0
b) Li and Wong

Fold change estimate for 1.25 ug vs. Fold change estimate for 20 ug
c) RMA

![Graph showing fold change estimates for 1.25 ug and 20 ug with a diagonal line indicating no change.](image)