New Developments in Affymetrix Probe Level Analysis

Rafael A. Irizarry
Department of Biostatistics, JHU

International Conference on Analysis of Genomic Data
Boston, MA, May 10-11, 2004
# Contact Information

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>e-mail</strong></td>
<td><a href="mailto:rafa@jhu.edu">rafa@jhu.edu</a></td>
</tr>
<tr>
<td><strong>Personal webpage</strong></td>
<td><a href="http://www.biostat.jhsph.edu/~ririzarr">http://www.biostat.jhsph.edu/~ririzarr</a></td>
</tr>
<tr>
<td><strong>Department webpage</strong></td>
<td><a href="http://www.biostat.jhsph.edu/">http://www.biostat.jhsph.edu/</a></td>
</tr>
<tr>
<td><strong>Bioinformatics Program</strong></td>
<td><a href="http://www.biostat.jhsph.edu/bioinfo">http://www.biostat.jhsph.edu/bioinfo</a></td>
</tr>
<tr>
<td><strong>BIOCOCONDUCTOR</strong></td>
<td><a href="http://www.bioconductor.org">http://www.bioconductor.org</a></td>
</tr>
</tbody>
</table>
Acknowledgements

• Zhijin Wu
• Robert Gentleman
• Wolfgang Huber
• James MacDonald
• Laurent Gautier
• Forrest Spencer
• Francisco Martínez-Murillo

• Terry Speed
• Ben Bolstad
• Leslie Cope
• Felix Naef
Outline

• Introduction to Technology
• Expression Measures
• Using Sequence Information
• A Unifying Model
• For more, see Zhijin Wu’s poster 28
Affymetrix GeneChip Design

Reference sequence

...TGTGATGCTGGGGAATGGGTACAGAAGCCCTCCGATGCGCCGATTGAGAAT...

**Perfect match**

**Mismatch**

NSB & SB

NSB
Statistical Problems

• 10-20K genes represented by 11-20 pairs of probe intensities (PM & MM)
• Most applications will want gene summaries, e.g. estimates of a gene’s fold-change across two conditions
• How do we summarize the probe-set pairs?
• **Background adjustment**: How do we adjust for NSB?
• **Normalization**: How do we ensure that measurements from different arrays are comparable?
A popular solution is to give an *expression measure* for each probe-set on each array.

Solutions to more complicated problems are left to others that use expression measures as raw data.

Affymetrix default is MAS 5.0.

There are many alternatives.

Many use **3-steps**: BG adjustment, Normalization, Summarization.

Many are compared by [http://affycomp.biostat.jhsph.edu](http://affycomp.biostat.jhsph.edu)
Previous Work

- **MAS 4.0 (Affymetrix)**
  - Negative expression
  - Very noisy for low expressed genes
  - Averages without taking log
- **dChip (Li and Wong)**
  - Account for the strong probe effect
  - Need for non-linear normalization
  - Multi-chip analysis reveals outliers
- **MAS 5.0 (Affymetrix)**
  - No more negatives
  - Average taking log
- **RMA (Irizarry et al.)**
  - Not subtracting MM improves precision
  - Fit log-scale linear model robustly
- **PerfectMatch (Zhang et al.), GCRMA (Wu et al.)**
  - Use sequence information and physical models
Why so much noise?

Default algorithm seems to be inspired by the following deterministic model for background:

\[ PM = O + N + S \]
\[ MM = O + N \]

\[ PM - MM = S \]

And a multiplicative error model for signal (they take the log before averaging)
Deterministic model is wrong

- Do MM measure non-specific binding?
- Look at Yeast DNA hybridized to Human Chip
- Look at PM, MM log-scale scatter-plot
- $R^2$ is only 0.5
Stochastic Model
(Additive background/multiplicative error)

$PM = O_{PM} + N_{PM} + S,$
$MM = O_{MM} + N_{MM}$

Similar to Durbin et al. 2002

$log (N_{PM}), log (N_{MM}) \sim \text{Bivariate Normal} (\rho \approx 0.7)$

$S = \exp (s + a + \epsilon), \text{ s is the quantity of interest}$

$E[ log(PM - MM) ] \sim s, \text{ but}$
$Var[ log(PM - MM) ] \sim 1/S^2 \text{ (can be very large)}$

Alternative approach: Ignore BG correction, problem is:
$E[ log(PM) ] \sim s+K \text{ (K is bias due to positive BG noise)}$

RMA does global BG correction that strikes a good balance
Accuracy/Precision Assessment from affycomp

Signal is slope of observed versus nominal concentration
Noise is 99.9th percentile of technical replicate log ratios
Accuracy/Precision Assessment

from affycomp

Signal is slope of observed versus nominal concentration
Noise is 99.9th percentile of technical replicate log ratios
Problem is worst at low end
Accuracy/Precision Assessment

Signal is slope of observed versus nominal concentration for low concentrations
Noise is 99.9th percentile of technical replicate log ratios

MAS 5.0
Can we improve accuracy without sacrificing too much precision?

- We need better estimates of means and variances of bivariate normal background noise.
- Use observed MM intensities along with sequence information.
- Solution that do not use MM are desirable.
Problems with MM

• MM detect signal
Problems with MM

- MM detect signal
- The estimate of non-specific binding is sequence-dependent

Plot is courtesy of Felix Naef
Problems with MM

- MM detect signal
- The estimate of non-specific binding is sequence-dependent
Problems with MM

- MM detect signal
- The estimate of non-specific binding is sequence-dependent
Problems with MM

- MM detect signal
- The estimate of non-specific binding is sequence-dependent
- Take up half the array ($250)
Predict NSB with sequence

- Fit simple linear model to yeast on human data to obtain base/position effects (Naef and Magnasco 2003)
Predict NSB with sequence

- Fit simple linear model to yeast on human data to obtain base/position effects (Naef and Magnasco 2003)
- Use these affinities to obtain parameters for background model
- New expression: GCRMA
Does it help?

Signal is slope of observed versus nominal concentration
Noise is 99.9th percentile of technical replicate log ratios

MAS 5.0
Does it help?

Signal is slope of observed versus nominal concentration
Noise is 99.9th percentile of technical replicate log ratios
For specific tasks, such as finding differentially expressed genes, our findings motivate the following model:

\[ Y_{gij} = O_i + \exp(h(\alpha_j) + b_{gj} + \varepsilon_{gij}) + \exp(f(\alpha_j) + S_g + \delta_g X_i + \xi_{gij}) \]

- \( f(\alpha_j) \): accounts for probe effect and normalization
- \( \delta_g \): represents differential expression

- Estimate parameter and standard error
- Standard error accounts for 3-steps
Preliminary Results
Conclusions and Future Work

- When assessing expression measures assess accuracy and precision
- BG adjustment important. Affects both.
- Appropriate use sequence information can be used for BG adjustment
- Unified model permits assessment of uncertainty due to BG adjustment, normalization, and summarization
- Proximity to 3’ end also has an effect
Supplemental Slides
Does it help?

- We can predict empirical results with model
- Accuracy of expression measures improves...
Can this be improved?
RMA

Irizarry et al. (2003) *NAR* 31:e15
• We can predict empirical results with model
• Accuracy of expression measures improves...
• Without adding too much variance
References

• Li and Wong PNAS, 2001
• http://www.biostat.jhsph.edu/~ririzarr
  – Nucleic Acids Research
  – Biostatistics
• http://www.bioconductor.org
• http://affycomp.biostat.jhsph.edu
• Felix Naef’s papers
Why so much noise?

• Default pre-processing uses the following summary:

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

• with \( \text{MM}^* \) a version of MM that is never bigger than PM.

• An ad-hoc optical background procedure and scale normalization are used.
RMA background

- Model observed PM as the sum of a signal intensity $S$ and a background intensity $B$
  \[ PM = B + S, \]
  where it is assumed that $S$ is $\text{Exponential} (\alpha)$, $B$ is $\text{Normal} (\mu, \sigma^2)$, and $S$ and $B$ are independent.
- Background adjusted PM are then $E[S \mid PM]$.
- These parametric distributions were chosen to provide a $\text{closed form}$ solution.
Alternative Solution

• If we can describe the additive background noise with a parametric model, statistical science provides various solutions for adjustment
• e.g. $E[s | PM, MM]$ instead of $PM-MM$
• We need to know means and variances of background noise
Accuracy/Precision Assessment

Signal is slope of observed versus nominal concentration
Accuracy/Precision Assessment

Signal is slope of observed versus nominal concentration
Accuracy/Precision Assessment

This one uses only probes spiked-in at low concentration
Location of compliment on reference sequence has an effect