Low level analysis of Affymetrix microarray expression data (the methods also apply to NimbleGen data)

Statistics 246     Spring 2006
Week 9           Lecture 1
What is low-level analysis?

- Image analysis, including registration, segmentation, summarization of pixel-level data
- Background adjustment, for optical noise and non-specific binding
- Normalization across arrays
- Summarization of multiple probes for the same transcript
- Quality assessment

AIM: To get accurate (low bias) and precise (low variance) gene expression measurements from the raw data
What is not low-level analysis?

- Determining differential expression
- Classification
- Clustering
- All other *high-level* analyses, which

AIM: to do something good with the measurements (discover genes or pathways, detect cancer, etc)
Does it matter?

Three quick slides displaying different aspects of our aim using low-level analyses widely used (at least until recently)
Log$_2$(chip 1/chip 2) vs log$_2$$\sqrt{(chip 1)(chip 2)}$

**Observed:** Black <2-fold, Red > 2-fold change

**Truth:** Green = 2-fold change, Blue = $\infty$-fold change, Black, Red = No change
Comparing estimates of log fold change based on 5 replicates with 20μg or 1.25 μg of RNA resp.

\[ \text{log}_2 \text{FC of CNS vs liver} \]

- **Red**: >4-fold
- **Orange**: >2-fold, <4-fold
- **Yellow**: < 2-fold
Observed vs true log fold change from a large spike-in experiment

Slope: 1.21

Slope: 0.88
A glance at some raw data:
20 probe spike-in set across 14 arrays
16 probe spike-in set: 6 non-responding
16 probe non-spiked set across 59 arrays
16 such sets all from
The Affymetrix human spike-in expt
(14 spiked genes, all but one array in triplicate = 59)
Concentrations in picoMoles

<p>| | | | | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></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>
The need for background adjustment

Red × indicates PM values for spiked-in mRNA
Some background adjustments

• Affymetrix initially subtracted MM, later what I’ll call MM*, a modified MM arranged to be < PM.
• We initially subtracted the mean of the low range normal, with an ad hoc fix.
• Then we fitted a convolution model to be described, currently used in RMA.
• Later, everyone came to realize the value of using the probe sequences to estimate non-specific binding in a more systematic manner. This led to GCRMA, but attempts to do better still continue. See next lecture.
Convoluted background model: pictorially

Signal + Noise = Observed
PM data on $\log_2$ scale: raw and fitted model
How we remove background

Denote the observed PM intensity by $S$. Model $S$ as the sum of a signal $X$ and a background $Y$, $S = X + Y$, where we assume $X$ is exponential($\alpha$), $Y$ is Normal($\mu$, $\sigma^2$), and $X$ and $Y$ are independent random variables.

The background adjusted PM is 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 $a = s - \mu - \sigma^2 \alpha$, $b = \sigma$, and $\phi$ and $\Phi$ are the normal density and cumulative density, respectively.

**Exercise.** Derive this formula.

This is the current RMA model and formula for background correction.
Observed PM vs BG corrected PM

As $s$ increases, the background correction asymptotes to

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

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

The impact of background adjustment can be substantial, especially on observed log fold changes, and on signals in the low range. This is illustrated in the next 3 slides, where we use data from the Affymetrix spike-in experiment, coloring and labelling points by their expected log$_2$ fold change. Black is zero, and the others go from 1 or -1 (red) to infinity or -infinity (purple).
No backround subtraction

91 = 14x13/2 pairwise average MA-plots all pooled
The current RMA bg subtraction
A proposed bg subtraction
The bg adjustment saga will continue

As indicated above, we will return to this important topic. But first, we’ll round out our introduction to low-level analysis by explaining normalization and RMA summarization.
Normalization: why and how

For why, see next slide. In general, probe-level data from different chips look **different**: location, scale, extremes. We can tell from replicate data that this is not due to compositional differences in the mRNA.

Initially Affymetrix simply multiplied each chip’s probe set summaries by a **scale factor**, in order to give every chip’s data a standard mean value (100 or 500).

Later Li & Wong (in dChip) introduced their invariant set normalization, which we won’t describe.

Next, Bolstad *et al* introduced quantile normalization as a way to make the distribution of probe intensities the same for every chip. Though not without problems, this has caught on, and we now describe it.
Density plots of log(PM-*BG) for some of the 59 spike-in chips
The quantile normalized density
Quantile normalization

• Quantile normalization is a way of making the distribution of probe intensities the same for every chip. It extends the familiar idea of matching means and standard deviations (usually to 0 and 1, respectively) to matching all quantiles: medians, upper and lower quartiles, quintiles, deciles, etc, i.e. all quantiles.

• Matched to what? The reference distribution could be anything we wish, but in this case the normalization distribution is chosen by averaging each quantile across the batch of chips currently being analysed. This idea is based on the belief that the data should be adjusted the as little as possible, consistent with achieving our goal.

• The diagram that follows illustrates the transformation from one chip to the normalization distribution.
Quantile normalization: formulae

\[ x_{\text{norm}} = F_{\text{norm}}^{-1}(F_{\text{raw}}(x)) \]
The two distribution functions are effectively estimated by the sample quantiles.

Quantile normalization is fast.

Looking at post-normalization PM vs pre-normalization PM (natural and log scales), we can see that the transformation is non-linear, see next slides.

After normalization, variability of expression measures across chips reduced, at least between replicate chips. Our hope is that this normalization will remove artifactual but not biological variation between non-replicate chips, see slides after next.
After vs Before: log intensity scale
After vs Before: intensity scale
M v A plots of chip pairs: before normalization
M v A plots of chip pairs: after quantile normalization
Quantile normalization reduces variability in comparison with nothing and scaling.

Vertical: log[^var q. norm/var other]; Horizontal: Aver. log mean

Note differences in vertical scales