Feature Level Data

Outline

• Affymetrix GeneChip arrays
• Two color platforms

Affymetrix GeneChip Design
Before Hybridization

More Realistic

Non-specific Hybridization
Affymetrix GeneChip Design

GeneChip Feature Level Data

• MM features used to measure optical noise and non-specific binding directly
• More than 10,000 probesets
• Each probeset represented by 11-20 feature Note 1: Position of features are haphazardly distributed about the array.
  Note 2: There are between 20-100 chip types
• So we have $PM_{gij}$, $MM_{gij}$
  ($g$ is gene, $i$ is array and $j$ is feature)
• A default summary is the avg of the PM-MM

Two color platforms

• Common to have just one feature per gene
• Typically, longer molecules are used so non-specific binding not so much of a worry
• Optical noise still a concern
• After spots are identified, a measure of local background is obtained from area around spot
Local background

GenePix
QuantArray
ScanAnalyze

GenePix does something different these days

Two color feature level data

- Red and Green foreground and and background obtained from each feature
- We have $R_{fgij}$, $G_{fgij}$, $R_{bgij}$, $G_{bgij}$ ($g$ is gene, $i$ is array and $j$ is replicate)
- A default summary statistic is the log-ratio: $(R_{f} - R_{b}) / (G_{f} - G_{b})$

Affymetrix Spike In Experiment
Spike-in Experiment

- Throughout we will be using Data from Affymetrix’s spike-in experiment
- Replicate RNA was hybridized to various arrays
- Some probesets were spiked in at different concentrations across the different arrays
- This gives us a way to assess precision and accuracy
- Done for HGU95 and HGU133 chips
- Available from Bioconductor experimental data package: SpikeIn

Probeset A

<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>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>512</td>
<td>1024</td>
<td>0</td>
</tr>
<tr>
<td>C</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>512</td>
<td>1024</td>
<td>0</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>0</td>
<td>1</td>
<td>2</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>1</td>
<td>2</td>
<td>4</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>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
</tr>
</tbody>
</table>

Probeset B

<table>
<thead>
<tr>
<th></th>
<th>0.25</th>
<th>0.5</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>16</th>
<th>32</th>
<th>64</th>
<th>128</th>
<th>256</th>
<th>512</th>
<th>1024</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</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>512</td>
<td>1024</td>
<td>0</td>
</tr>
<tr>
<td>H</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>512</td>
<td>1024</td>
<td>0</td>
</tr>
<tr>
<td>I</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>512</td>
<td>1024</td>
<td>0</td>
</tr>
<tr>
<td>J</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>512</td>
<td>1024</td>
<td>0</td>
</tr>
</tbody>
</table>

Probeset C

<table>
<thead>
<tr>
<th></th>
<th>0.25</th>
<th>0.5</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>16</th>
<th>32</th>
<th>64</th>
<th>128</th>
<th>256</th>
<th>512</th>
<th>1024</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</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>512</td>
<td>1024</td>
<td>0</td>
</tr>
<tr>
<td>M</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>512</td>
<td>1024</td>
<td>0</td>
</tr>
</tbody>
</table>

Spikein Experiment (HG-U95)

- A similar experiment was repeated for a newer chip
- The 1024 picoMolar concentration was not used. 1/8 was used instead.
- No groups of 12
- Note: More spike-ins to come!
Background Effects
Experiments

Learn about optical effect and NSB

<table>
<thead>
<tr>
<th>label</th>
<th>sample type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Empty</td>
<td>empty</td>
</tr>
<tr>
<td>NoRNA</td>
<td>no RNA</td>
</tr>
<tr>
<td>NoLabel</td>
<td>human</td>
</tr>
<tr>
<td>YeastDNA</td>
<td>yeast genomic DNA</td>
</tr>
<tr>
<td>polyC</td>
<td>poly C</td>
</tr>
<tr>
<td>polyG</td>
<td>poly G</td>
</tr>
</tbody>
</table>

The Background Effects
**Background Effect**

Background Experiment for Affymetrix HGU95 array

- Empty
- No RNA
- Yeast DNA
- polyG

**Why Adjust for Background?**

- This are the no-label and Yeast DNA chips

![Graph showing intensity vs. log intensity with trends and data points]

- Observedaverage intensity vs. Nominal Concentration
Why Adjust for Background?

Notice local slope decrease as the nominal concentration becomes small.

Probe-specific NSB

Why not subtract MM,BG?
Why not subtract MM?

Why not subtract MM?

Solutions
**Direct Measurement Strategy**

The hope is that:

\[ PM = B + S \]
\[ MM = B \]

\[ PM - MM = S \]

But this is not correct!

**Notice**
- We care about ratios
- We usually take log of S

**Stochastic Model**

Better to assume:

\[ PM = B_{PM} + S \]
\[ MM = B_{MM} \]

\[ \text{Var}[\log(PM - MM)] \rightleftharpoons 1/S^2 \]

**Alternative solution:**

\[ E[S | PM] \]

**Simulation**

- We create some feature level data for two replicate arrays
- Then compute \( Y = \log(PM - kMM) \) for each array
- We make an MA using the Ys for each array
- We make a observed concentration versus known concentration plot
- We do this for various values of k. The following “movie” shows k moving from 0 to 1.
RMA Background Adjustment

The Basic Idea:

\[ PM = B + S \]

*Observed: PM*

*Of interest: S*

Pose a statistical model and use it to predict S from the observed PM

---

The Basic Idea

\[ PM = B + S \]

- A mathematically convenient, useful model
  - \( B \sim \text{Normal} (\mu, \sigma) \)
  - \( S \sim \text{Exponential} (\lambda) \)
  \[ \hat{S} = E[S | PM] \]
- No MM
- Borrowing strength across probes

---

MAS 5.0

---
Notice improved precision but worst accuracy

**Problem**

- Global background correction ignores probe-specific NSB
- MM have problems
- Another possibility: Use probe sequence

**Sequence effect**


\[
\text{Affinity} = \sum_{j \in \{1, \ldots, 4\}} \sum_{k \in \{1, \ldots, 25\}} \mu_{j,k} \cdot \alpha_{j,k} \quad \mu_{j,k} = \text{smooth function of } k
\]
**General Model**

\[ PM = \exp(h_i(z_{ij}) + \theta_{ij} + \eta_{ij}) + \exp(f_i(x_{ij}) + \theta_{ij} + \eta_{ij}) \]

\[ MM = \exp(h_i(z_{ij}) + \theta_{ij} + \eta_{ij}) \]

We can calculate: \( E[\theta_{ij} | PM_{ij}, MM_{ij}] \)

**Alternative background adjustment**

- Use this stochastic model
- Minimize the MSE:

\[
E \left[ \left( \log \left( \frac{3}{S} \right) \right)^2 | S > 0, PM, MM \right]
\]

- To do this we need to specify distributions for the different components
- Notice this is probe-specific so we need to borrow strength

*These parametric distributions were chosen to provide a closed form solution*

**Explains Bimodality**

![Bimodality Chart]
C, T in the middle

A, G in the middle