Affymetrix GeneChips
Oligonucleotide Microarrays

Many commercial microarray platforms are available:

<table>
<thead>
<tr>
<th>Platform</th>
<th>Array Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affymetrix</td>
<td>Oligo arrays</td>
</tr>
<tr>
<td>Qiagen</td>
<td>Oligo arrays</td>
</tr>
<tr>
<td>Amersham Biosciences</td>
<td>Oligo arrays</td>
</tr>
<tr>
<td>MWG Biotech</td>
<td>Oligo arrays</td>
</tr>
<tr>
<td>Rosetta (Merck)</td>
<td>Oligo arrays</td>
</tr>
<tr>
<td>Agilent</td>
<td>Oligo arrays and Oligo arrays</td>
</tr>
<tr>
<td>Clontech, BD Biosciences</td>
<td>cDNA arrays</td>
</tr>
<tr>
<td>UHN MAC (Ontario)</td>
<td>cDNA arrays</td>
</tr>
<tr>
<td>Incyte Gene Album</td>
<td>cDNA arrays</td>
</tr>
<tr>
<td>Genomictree, Inc</td>
<td>cDNA arrays</td>
</tr>
</tbody>
</table>

Plus a variety of custom cDNA arrays
Examples of publicly available

gene expression data repositories

1. **ArrayExpress** - A public repository for microarray based gene expression data maintained by European Bioinformatics Institute.
2. **ChipDB** - A searchable database of gene expression.
3. **Gene Expression Atlas** - A database for gene expression profile from 91 normal human and mouse samples across a diverse array of tissues, organs, and cell lines.
4. **Gene Expression Database (GXD)** - A database of Mouse Genome Informatics at the Jackson laboratory.
5. **Gene Expression Omnibus** - A database in NCBI for supporting the public use and disseminating of gene expression data.
6. **MUSC DNA Microarray Database** - MUSC DNA Microarray Database is a web-accessible archive of DNA microarray data.
7. **NASCArrays** - a repository for Affymetrix data generated by NASC's transcriptomics service.
A Quantitative Overview to Gene Expression Profiling in Animal Genetics

Affymetrix Chips

Procedures for Target Preparation

1. Wash & Stain
2. Scan
3. Hybridise
4. (16 hours)

- IVT (Biotin-UTP, Biotin-CTP)
- Fragment (heat, Mg++)
- Fragmented cRNA
- Biotin-labeled transcripts

RNA → cDNA → B → B → B → B

0.18 cm
1.28 cm

Actual size of GeneChip® array

600,000 locations on each GeneChip® array

Actual strand = 28 base pairs

Armidale Animal Breeding Summer Course, UNE, Feb. 2006
**Probe** → A 25mer oligo complementary to a sequence of interest, attached to a glass surface on the probe array.

**Perfect Match (PM)** → Probes that are complementary to the sequence of interest.

**Mismatch (MM)** → Probes that are complementary to the sequence of interest except for homomeric base change (A-T or G-C) at the 13th position.

**Probe Pair** → A combination of a PM and a MM.

**Probe Set** → A set of 11 – 20 probe pairs.
### Terminology

Gene Sequence:  
Probe Sequences:  


Probe set: 11 to 20 probe pairs (PM & MM) to interrogate each gene  
There may be 5,000-20,000 probe sets per chip  

---

Figure 1-3: Expression tiling strategy
**Pros and Cons of Affymetrix**

**Advantages:**
- Conditions are precisely controlled, chips are identical and can be compared
- Only unique part of sequence is chosen – detection of closely related genes or splice variants is possible

**Disadvantages:**
- The sequences are chosen based on a contemporary UniGene release and might get revised
- Short probes may result in less specific hybridization and reduced sensitivity
  (Agilent prefers 50-100mers)
- Expensive!!! We often have to resort to cDNA arrays

**Bridging Platforms**


"The overall correlations between platforms were in the range 0.7 to 0.8. When concordance was measured for expression ratios significant at P < 0.05, the agreement among the platforms was very high, ranging from 93% to 100%"

Many other references comparing platforms with mixed results:
Pessimistic at the beginning (ie. 2000’s), more optimistic later on (…as the analysis methods to compare were more sophisticated).

A Quantitative Overview to Gene Expression Profiling in Animal Genetics

Affymetrix Chips

Ferl et al. (2003)

27 DE in cDNA
Of which 14 were present in the Affy chip.

R = (PM-MM)/(PM+MM) **Discrimination Score of a Probe Pair.**

Discrimination score R describes the ability of a probe pair to detect its intended target.
If R is close to 1.0 in a majority of pairs in a set, the detection p-value will be lower

Discrimination Score of each probe pair is compared to \( \tau \) - user defined value (default =0.0015)

If \( (PM-MM)/(PM+MM) > \tau \), then probe set is excluded

Increasing \( t \) can reduce the number of false positives, but the true present calls might be lost.
Converting the signal intensity into numeric values

**R = Discrimination Score**

\[ R = \frac{(PM-MM)}{(PM+MM)} \]

Discrimination score of each probe pair is compared to \( t \) (default = 0.0015)

A one-sided Wilcoxon's Signed Rank test is the statistical method used to calculate the **Detection P-value** that reflects the significance of the differences between PM and MM. It assigns each probe pair a rank based on how far the probe pair Discrimination Score is from \( \tau \)

P-value or statistical significance of a result is the probability that the observed change in a sample occurred by pure chance.

\[ \alpha_1 \text{ and } \alpha_2 \text{ are user defined values but have optimized defaults in the software} \]

<table>
<thead>
<tr>
<th>P-value of a probe set</th>
<th>0.01</th>
<th>0.02</th>
<th>0.03</th>
<th>0.04</th>
<th>0.05</th>
<th>0.06</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>Marginal</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
</tr>
<tr>
<td>Absent</td>
<td>0.08</td>
<td>0.08</td>
<td>0.08</td>
<td>0.08</td>
<td>0.08</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Armidale Animal Breeding Summer Course, UNE, Feb. 2006
Converting the signal intensity into numeric values

- Each probe pair in a probe set is considered as having a potential vote in determining the Signal value.
- The real signal is estimated by taking the log of the Perfect Match intensity after subtracting the slide signal estimate (CT: Background correction across the entire array).
- Subsequently, an expression call flag is assigned to each probe set:

  - **P** \(\Rightarrow\) gene is expressed (Present)
  - **M** \(\Rightarrow\) gene is Marginally expressed
  - **A** \(\Rightarrow\) gene is not expressed (Absent)

Conclusions

- Affymetrix arrays can give absolute expression values for a given gene. The software generates a call: **Present, Marginal or Absent** as well as a numeric value for expression level.
- There is a number of "user defined" values used in calculations that we should be aware of while extracting the data.
- Default software values guarantee very stringent cut-offs. The stringency of call generation can be manually changed to include more genes.
Possible Problems

What if

• a small number of the probe pairs hybridize much better than the rest?
• removing the middle base does not make a difference for some probes?
• some MM are PM for some other gene?
• there is need for normalization?
A Quantitative Overview to Gene Expression Profiling in Animal Genetics

Affymetrix Chips

Example

Data for a Single Chip

Probe ID Intens. Flag P-Value

<table>
<thead>
<tr>
<th>Probe ID</th>
<th>Intens.</th>
<th>Flag</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>24,128</td>
<td>252.2</td>
<td>P</td>
<td>0.000072</td>
</tr>
<tr>
<td>551.1</td>
<td>57.8</td>
<td>A</td>
<td>0.040604</td>
</tr>
<tr>
<td>6335.1</td>
<td>591.3</td>
<td>P</td>
<td>0.000593</td>
</tr>
<tr>
<td>7355.1</td>
<td>550.0</td>
<td>P</td>
<td>0.000752</td>
</tr>
<tr>
<td>12377.1</td>
<td>822.0</td>
<td>P</td>
<td>0.000054</td>
</tr>
<tr>
<td>20693.1</td>
<td>229.0</td>
<td>P</td>
<td>0.000004</td>
</tr>
<tr>
<td>8247.4</td>
<td>220.1</td>
<td>P</td>
<td>0.037950</td>
</tr>
<tr>
<td>23024.1</td>
<td>179.4</td>
<td>M</td>
<td>0.000412</td>
</tr>
<tr>
<td>11379.1</td>
<td>11917.7</td>
<td>P</td>
<td>0.000044</td>
</tr>
<tr>
<td>1751.8</td>
<td>1294.0</td>
<td>P</td>
<td>0.000044</td>
</tr>
<tr>
<td>6809.1</td>
<td>1099.0</td>
<td>P</td>
<td>0.000024</td>
</tr>
<tr>
<td>6705.1</td>
<td>1099.0</td>
<td>P</td>
<td>0.000024</td>
</tr>
<tr>
<td>4475.1</td>
<td>339.7</td>
<td>P</td>
<td>0.000025</td>
</tr>
<tr>
<td>2570.1</td>
<td>1669.1</td>
<td>P</td>
<td>0.000025</td>
</tr>
<tr>
<td>5511.1</td>
<td>500.8</td>
<td>P</td>
<td>0.000044</td>
</tr>
<tr>
<td>7002.1</td>
<td>376.0</td>
<td>A</td>
<td>0.037042</td>
</tr>
<tr>
<td>9344.1</td>
<td>593.7</td>
<td>P</td>
<td>0.000053</td>
</tr>
<tr>
<td>1531.1</td>
<td>900.7</td>
<td>P</td>
<td>0.000044</td>
</tr>
<tr>
<td>7235.1</td>
<td>503.2</td>
<td>P</td>
<td>0.000044</td>
</tr>
<tr>
<td>559.1</td>
<td>8.0</td>
<td>A</td>
<td>0.196409</td>
</tr>
<tr>
<td>8092.1</td>
<td>1295.0</td>
<td>P</td>
<td>0.000072</td>
</tr>
<tr>
<td>7397.1</td>
<td>1021.1</td>
<td>P</td>
<td>0.000072</td>
</tr>
<tr>
<td>8044.1</td>
<td>343.2</td>
<td>P</td>
<td>0.000044</td>
</tr>
<tr>
<td>10279.1</td>
<td>170.0</td>
<td>P</td>
<td>0.000121</td>
</tr>
<tr>
<td>20157.1</td>
<td>702.7</td>
<td>P</td>
<td>0.000044</td>
</tr>
<tr>
<td>4161.1</td>
<td>6.3</td>
<td>A</td>
<td>0.460044</td>
</tr>
<tr>
<td>1509.1</td>
<td>991.7</td>
<td>P</td>
<td>0.000027</td>
</tr>
<tr>
<td>22829.1</td>
<td>2656.3</td>
<td>P</td>
<td>0.000150</td>
</tr>
</tbody>
</table>

Each represents the average Mismatch-corrected intensity of 11 – 20 Probe Pairs!

For all 15 Chips

Proportions are approx. constant for all chips.

Increasing intensity from A to M to P.

Very good variance stabilisation.

Armidale Animal Breeding Summer Course, UNE, Feb. 2006
Use all data and include Flag in the definition Comparison Group

Prop. Total Variance:

~ 3% of Genes being DE in a given contrast:
1. Pregnancy – Lactation
2. Pregnancy – Involution
3. Lactation – Involution

CG: Comparison Group
= Expression Intensities from the same chip (15) and flag (3). Hence, 45 Levels.

Gene by Animal (5) for Biological Variability
Gene by Stage (3)
A final list of 4,003 DE genes (16.6%) was generated after exploring three statistical approaches:

- **GS**: GeneSpring (t-stat)
- **MME**: Mixed-Model Equations
- **BCI**: Bootstrap Confidence Intervals