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</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
A Quantitative Overview to Gene Expression Profiling

Affymetrix Chips

A. Reverter – Sept. 2006, UAB, Barcelona, Spain
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. Reverter – Sept. 2006, UAB, Barcelona, Spain
A Quantitative Overview to Gene Expression Profiling

Affymetrix Chips

Procedures for Target Preparation

- cDNA
- Wash & Stain
- Hybridise (16 hours)
- Scan
- IVT (Biotin-UTP Biotin-CTP)
- Fragment (heat, Mg2+)
- Biotin-labeled transcripts

A. Reverter – Sept. 2006, UAB, Barcelona, Spain
**Affymetrix Chips**

RNA fragments with fluorescent tags from sample to be tested

RNA fragment hybridized with DNA on GeneChip® array

**Terminology**

**Probe** \(\rightarrow\) A 25mer oligo complementary to a sequence of interest, attached to a glass surface on the probe array.

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

**Mismatch (MM)** \(\rightarrow\) 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** \(\rightarrow\) A combination of a PM and a MM.

**Probe Set** \(\rightarrow\) A set of 11 – 20 probe pairs.
Affymetrix Chips

Terminology

Gene Sequence: 3’ ———— 5’
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

GeneChip® Expression Array Design

Figure 1-3 Expression tiling strategy

A. Reverter – Sept. 2006, UAB, Barcelona, Spain
**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

---


“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

**Affymetrix Chips**

Ferl et al. (2003)

![Diagram](image)

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

**Converting the signal intensity into numeric values**

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

*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** \( \frac{(PM-MM)}{(PM+MM)} > \tau \), **then probe set is excluded**

*Increasing \( t \) can reduce the number of false positives,*
*but the true present calls might be lost.*

A. Reverter – Sept. 2006, UAB, Barcelona, Spain
A Quantitative Overview to Gene Expression Profiling

**Affymetrix Chips**

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)

Figure 2. In this hypothetical probe set, the Perfect Match (PM) intensity is 80 and the Mismatch (MM) intensity for each probe pair increases from 10 to 100. The probe pairs are numbered from 1 to 10. As the Mismatch (MM) probe call intensity, plotted on the y-axis, increases and becomes equal to or greater than the Perfect Match (PM) intensity, the Discrimination score decreases as plotted on the y-axis. More specifically, as the intensity of the Mismatch (MM) increases, our ability to discriminate between the PM and MM decreases. The dashed line is the user-definable parameter \( \tau \) (default = 0.015).

A. Reverter – Sept. 2006, UAB, Barcelona, Spain

---

A Quantitative Overview to Gene Expression Profiling

**Affymetrix Chips**

Converting the signal intensity into numeric values

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.

\[ P-value \]

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

<table>
<thead>
<tr>
<th>P-value of a probe set</th>
<th>( \alpha_1 )</th>
<th>( \alpha_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>0.04</td>
<td>0.06</td>
</tr>
<tr>
<td>Marginal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Absent</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Default: 0.04 0.06

A. Reverter – Sept. 2006, UAB, Barcelona, Spain
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** → gene is expressed (Present)
  - **M** → gene is Marginally expressed
  - **A** → 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?

Example

A Quantitative Overview to Gene Expression Profiling

A. Reverter – Sept. 2006, UAB, Barcelona, Spain
Affymetrix Chips

Example

Data for a Single Chip

Probes ID  Intensity  Flag  P-Value

<table>
<thead>
<tr>
<th>Probe ID</th>
<th>Intensity</th>
<th>Flag</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>252.2</td>
<td>A</td>
<td>0.000722</td>
</tr>
<tr>
<td>2</td>
<td>350.3</td>
<td>A</td>
<td>0.000052</td>
</tr>
<tr>
<td>3</td>
<td>350.0</td>
<td>A</td>
<td>0.000052</td>
</tr>
<tr>
<td>4</td>
<td>329.9</td>
<td>A</td>
<td>0.000052</td>
</tr>
<tr>
<td>5</td>
<td>226.1</td>
<td>A</td>
<td>0.000052</td>
</tr>
<tr>
<td>6</td>
<td>329.4</td>
<td>M</td>
<td>0.000052</td>
</tr>
<tr>
<td>7</td>
<td>11013.7</td>
<td>M</td>
<td>0.000052</td>
</tr>
<tr>
<td>8</td>
<td>3500.4</td>
<td>M</td>
<td>0.000052</td>
</tr>
<tr>
<td>9</td>
<td>1009.0</td>
<td>M</td>
<td>0.000052</td>
</tr>
<tr>
<td>10</td>
<td>1005.9</td>
<td>M</td>
<td>0.000052</td>
</tr>
<tr>
<td>11</td>
<td>1005.9</td>
<td>M</td>
<td>0.000052</td>
</tr>
<tr>
<td>12</td>
<td>1005.9</td>
<td>M</td>
<td>0.000052</td>
</tr>
<tr>
<td>13</td>
<td>1005.9</td>
<td>M</td>
<td>0.000052</td>
</tr>
<tr>
<td>14</td>
<td>3500.4</td>
<td>M</td>
<td>0.000052</td>
</tr>
<tr>
<td>15</td>
<td>3500.4</td>
<td>M</td>
<td>0.000052</td>
</tr>
</tbody>
</table>

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

Proportions are approx. constant for all chips.

Increasing intensity from A to M to P.

Very good variance stabilisation.
Affymetrix Chips

Example

Use all data and include Flag in the definition Comparison Group

A. Reverter – Sept. 2006, UAB, Barcelona, Spain

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)

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

A Qualitative Overview to Gene Expression Profiling
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