Measuring Gene Expression
Part 2

David Wishart
Bioinformatics 301
david.wishart@ualberta.ca

Measuring Gene Expression

• Differential Display
• Serial Analysis of Gene Expression (SAGE)
• RT-PCR (real-time PCR)
• Northern/Southern Blotting
• DNA Microarrays or Gene Chips
Microarrays

- Principle is to analyze gene (mRNA) or protein expression through large scale non-radioactive Northern (RNA) or Southern (DNA) hybridization analysis
- Essentially high throughput Northern Blotting method that uses Cy3 and Cy5 fluorescence for detection
- Allows expressional analysis of up to 20,000 genes simultaneously

DNA Microarrays
Four Types of Microarrays

- Photolithographically prepared short oligo (20-25 bp) arrays (1 colour)
- Spotted glass slide cDNA (500-1000 bp) arrays (2 colour)
- Spotted nylon cDNA (500-1000 bp) arrays (1 colour/radioactive)
- Spotted glass slide oligo (30-70 bp) arrays (1 or 2 colour)

Principles of 2 Colour Microarrays
Microarray Definition of Probe and Target

- There are two acceptable and completely opposite definitions. We will use:
  - **Target** = the DNA that is spotted on the array
  - **Probe** = the DNA that is labeled with the fluorescent probe
2-Colour Microarray Principles

Scan and detect with confocal laser system
overlay images and normalize
Image process and analyze

Typical 2-Colour Data
Microarrays & Spot Colour

Principles of 1 Colour Microarrays

Prepare cDNA Probe

"Normal"

\[ \text{RT-PCR} \]

Prepare Microarray

Hybridize probe to microarray

SCANT
Two Colour vs. One Colour

- Two-colour hybridization eliminates artifacts due to variation in:
  - quantity of DNA spotted
  - stringency of hybridization
  - local concentration of label

- However,
  - both samples *must* label with equivalent efficiency
  - Information is lost for genes not expressed in the reference or control sample
Two Colour vs. One Colour

- One-colour hybridization may have artifacts due to variation in:
  - quantity of DNA spotted
  - stringency of hybridization
  - local concentration of label
- However good quality control (QC) means,
  - fewer artifacts
  - less manipulation, lower cost
  - reduced loss of information (due to reference sample transcript content)

Specific Arrays of Interest

- Home-made Spotted Oligo Arrays
  - Made using glass slides, Operon oligos and robotic spotting equipment
- Amersham CodeLink Microarrays
  - Made using specially treated slides, QC’d oligos and robotic spotting equipment
- Affymetrix Gene Chips
  - Made using photolithographically produced systems with multi-copy oligos
Array Images

Oligo Microarray
2 colour

Amersham
1 colour

Array Images

Oligo Microarray
2 colour

Affymetrix Gene Chip
1 colour
Home-made Spotted Arrays

- Probes are >100µm and are usually deposited on glass
- Probes can be:
  - oligos (usually >40mers)
  - PCR fragments from cDNA/EST or genomic templates
- Not reused; 2-colour hybridizations
Standard Spotted Array

Home-made Microarrays

Prepare cDNA Probe

"Normal" → Tumor → RT / PCR → Label with Fluorescent Dyes → Combine Equal Amounts → Hybridize probe to microarray → SCAN

Prepare Microarray

Microarray Technology
Common Home-made Microarray Errors

- Irregular Spot
- Comet Tail
- Streaking

- Hi Background
- Low Intensity
- A Good Array

Testing Reproducibility

- Breast tumor tissue biopsy
- mRNA prepared using standard methods
- Control sample made from pooled mRNA from several cell types
- 3 RNA samples prepared from 1 tissue source – arrayed onto two sets of home-made chips from different suppliers
- Conducted pairwise comparison of intensity correlations & no. of spots
Home-made Arrays

Oligo Microarray 1

1) $R=0.7$ 95%CI=(0.68-0.72) N=2027
2) $R=0.65$ 95%CI=(0.62-0.67) N=2818
3) $R=0.61$ 95%CI=(0.59-0.64) N=2001

Oligo Microarray 2

1) $R=0.66$ 95%CI=(0.62-0.69) N=1028
2) $R=0.86$ 95%CI=(0.85-0.87) N=1925
3) $R=0.64$ 95%CI=(0.61-0.68) N=1040
Advantages to Home-made Systems

• Cheapest method to produce arrays ($100 to $300/slide)
• Allows lab full control over design and printing of arrays (customizable)
• Allows quick adaptation to new technologies, new probe sets
• Allows more control over analysis

Disadvantages to Home-made Systems

• Quality and quality-control of oligo probe set is highly variable
• Quality of spotting and spot geometry is highly variable
• Technology is very advanced, difficult and expensive to maintain (robotics)
• Reproducibility is poor
Amersham CodeLink Arrays

• Amersham synthesizes its 30-nucleotide oligos offline, tests them by mass spectrometry, deposits them on specially coated array, and then assays them for quality control.

• Uses a special Flex Chamber™—a disposable hybridization chamber already attached to the slide to improve hybridization consistency.
Amersham CodeLink Oligo Chip

DNA
Hydrophilic polymer
Glass

CodeLink Special Coating

- Most glass substrates are quite hydrophobic
- This hydrophobicity affects the local binding and surface chemistry of most glass-slide chips making most of the attached DNA oligo inaccessible
- Coating the slide with a hydrophilic polymer allows the cDNA to pair up with the substrate oligos much better
Morphology Does Not Affect Dynamic Range
CodeLink Bioarrays Can Achieve Linearity Across 3 Logs*

- The red line indicates the signal level for non-spiked target.
- Error bars represent one standard deviation for each mean (n=18) signal

*Data obtained from cRNA dilution series.
Testing Reproducibility

- Breast tumor tissue biopsy
- mRNA prepared using standard methods
- 3 RNA samples prepared from 1 tissue source – arrayed onto 3 different sets of CodeLink chips
- Conducted pairwise comparison of intensity correlations, intensity ratio correlations & number of “passed” spots

Intensity, Pairwise Comparisons

Amersham Slides

1) $R=1$ 95%CI=$(1-1)$ N=8258
2) $R=0.99$ 95%CI=$(0.99-1)$ N=8332
3) $R=0.99$ 95%CI=$(0.99-0.99)$ N=8290
Ratio, Pairwise Comparisons

Amersham Slides
1) R=0.98  95%CI=(0.98-0.98)  N=7694
2) R=0.97  95%CI=(0.97-0.98)  N=7873
3) R=0.97  95%CI=(0.97-0.97)  N=7694

General Comparison
Amersham Intensity
1) R=1  95%CI=(1-1)  N=8258
2) R=0.99  95%CI=(0.99-1)  N=8332
3) R=0.99  95%CI=(0.99-0.99)  N=8290

Amersham Ratio
1) R=0.98  95%CI=(0.98-0.98)  N=7694
2) R=0.97  95%CI=(0.97-0.98)  N=7873
3) R=0.97  95%CI=(0.97-0.97)  N=7694

Vancouver
1) R=0.7  95%CI=(0.68-0.72)  N=2027
2) R=0.65  95%CI=(0.62-0.67)  N=2818
3) R=0.61  95%CI=(0.59-0.64)  N=2001

Calgary I
1) R=0.66  95%CI=(0.62-0.69)  N=1028
2) R=0.86  95%CI=(0.85-0.87)  N=1925
3) R=0.64  95%CI=(0.61-0.68)  N=1040

Calgary II
1) R=0.49  95%CI=(0.44-0.54)  N=942
2) R=0.81  95%CI=(0.8-0.83)  N=1700
3) R=0.57  95%CI=(0.52-0.61)  N=973
## Comparative Accuracy

<table>
<thead>
<tr>
<th>GENES</th>
<th>RT-PCR</th>
<th>Spotted Array</th>
<th>CodeLink</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Expression Pattern</td>
<td>Expression Pattern</td>
<td>Expression Pattern</td>
</tr>
<tr>
<td></td>
<td>TaqMan</td>
<td>Operon</td>
<td>Amersham</td>
</tr>
<tr>
<td>hENT1</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>hENT2</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>hCNT1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>hCNT2</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>dck</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>ER</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>

## CodeLink Advantages

- Exceptional reproducibility because of:
  - careful probe design
  - QC of oligo preparations and spotting
  - high proportion of oligo binding to cDNA substrate due to hydrophilic coating
  - well controlled/uniform hybridization
- Allows users to continue using same scanners/software as in spotted arrays
CodeLink
Disadvantages

- Lack of flexibility or customizability (users depend on Amersham to provide & design chips)
- Dependent on proprietary kits and reagents
- More expensive than spotted arrays ($700/chip)

Cost per Sample in Triplicate

- Amersham Slides (single channel)
  - $2000

- Vancouver Spotted Arrays (two colour)
  - $800

- Calgary Spotted Arrays (two colour)
  - $1100
Affymetrix Gene Chips

- Chips are 1.7 cm²
- 400,000 oligo probe pairs
- Probe “spots” are $20\mu \times 20\mu$
- Each probe is 25 bases long
- 11-20 “match” probes and 11-20 “mismatch” probes per gene
Affy Chip

- 11-20 probes for each gene/EST
- Each probe is 25 bases long
- 1 has exact match, the other is mismatched in the middle base
- Match (M) and mismatch (MM) pairs are placed next to each other
- Expression levels calculated using intensity difference between M & MM for all probe pairs
Affymetrix Hybridization

Affy Chips
Affy Chips

Human Genome U133A GeneChip® Array

(1) Probe Array

(2) Probe Set

Each Probe Set contains 11 Probe Pairs (PM MM) of different probes.

(3) Probe Pair

Each Perfect Match (PM) and Mismatch (MM) Probe Cells are associated by pairs.

The Human Genome U133 A GeneChip® array represents more than 21,000 full-length genes and EST clusters.

(4) Probe Cell

Each Probe Cell contains \( \approx 4 \times 10^6 \) copies of a specific probe complementary to genetic information of interest. probe - single stranded, sense; fluorescently labeled oligonucleotide (25 mers).

Match

Mismatches

control

match

mismatch

schizophrenic

match

mismatch
Comparison of Affymetrix and Spotted cDNA Arrays

161 620 matched pairs of measurements from 56 cell lines

Affymetrix GeneChip Advantages

- High precision because of:
  - careful probe design
  - up to 20 probes per gene
  - up to 20 mismatch probes
- Very precise measurements
- Very high density (500,000 elements/array)
Affymetrix GeneChips
Disadvantages

• Inflexible: each array requires custom photolithographic masks
• More expensive than spotted arrays ($1000-$1200 per chip)
• Proprietary technology
  – not all algorithms, information public
  – only one manufacturer of readers, etc.

General Comments

• Spotted arrays are still wildly popular and widely used – a great learning tool for expression analysis
• Spotted arrays are generally unreliable and provide only gross indications of RNA expression
• Commercial systems (CodeLink and Affy) offer much greater reliability but are expensive & inflexible
Microarray Production

- Probe design and selection
- Printing
- RNA extraction
- Labeling
- Hybridization and washing
- Scanning
- Data analysis

Probe Design & Selection

- Synthetic oligos 25-70 bases in length
- Choose sequences complementary to mRNA of interest
- Random base distribution and average GC content for organism
- Avoid long A+T or G+C rich regions
- Minimize internal secondary structure (hairpins or other loops)
- 1 M salt + 65 °C thermostability
Probe Design & Selection

- Design and select oligo sequences that are less than 75% identical to existing genes elsewhere in the genome (i.e. do a BLAST search)
- Sequences with >75% sequence identity to other sequences will cross-hybridize – leading to confounding results

Cross-hybridization

Analysis of a cross-hybridization within the CYP450 superfamily

Xu et al. (2001) Gene
Microarray Printing

- Probes are deposited by robots using:
  - piezo-electric jets
  - microcapillaries
  - split or solid pins

- Coated glass is the most common substrate
  - aminosilane, poly-lysine, etc. give non-covalent linkages
  - covalent linkage is possible with modified oligos + aldehyde (etc.) coatings
RNA Extraction

• RNA is extremely unstable
• Probably the most problematic step in all microarray analysis
• RNA is extracted as “total RNA”
  – only 1-2% is mRNA
  – remainder is rRNA, tRNA, etc.
• RNA extracted from tissue is often very heterogeneous (many cells and cell types) – watch selectivity

Laser Capture Microdissection

• Cells of interest are visually selected and exposed to an IR laser, which adheres them to a transfer film
RNA Labeling

• Common source of systematic error (freshness, contaminants)

• Direct labeling
  – fluorescent nucleotides are incorporated during reverse transcription (“first strand”)

• Indirect labeling
  – reactive nucleotides (aminoallyl-dUTP) are incorporated during RT; first strand product is mixed with reactive fluorescent dyes that bind to amino group

Direct Labeling

Cy5

Cy3-ATP
Indirect Labeling

• Reverse Transcriptase
  - Aminoallyl-dUTP, dGTP, dATP, dCTP
  - NH₂ NH₂ NH₂

• Coupling Reaction
  - N-hydroxysuccinimide-activated fluorescent dye

- aminoallyl-dUTP

Hybridization

• Stringency of hybridization is affected by ions, detergents, formamide, temperature, time...
• Hybridization may be an important source of systematic error
• Automated hybridization systems exist; value is debatable
How Many Replicates?

**Table 5. Misclassification percentages for different combinations of replicates**

<table>
<thead>
<tr>
<th>Classification</th>
<th>Outcome</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(1, 2)</th>
<th>(1, 3)</th>
<th>(2, 3)</th>
<th>(1, 2, 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>False positive, %</td>
<td>8.3</td>
<td>1.4</td>
<td>9.0</td>
<td>1.0</td>
<td>2.1</td>
<td>0.7</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>False negative, %</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.3</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>Misclassified, %</td>
<td>8.7</td>
<td>1.4</td>
<td>9.0</td>
<td>1.4</td>
<td>2.4</td>
<td>0.7</td>
<td>0.7</td>
<td></td>
</tr>
</tbody>
</table>

- Substantial error when only one array analyzed, standard is to use 3 replicates

What Types of Replicates?

**Biological replicates**
- Treatment
- Biological replicate

**Technical replicates**
- Dye
- Technical replicate

**Array**
- Duplicate spot

Biological replication is most important because it includes all of the potential sources for error.
Microarray Production

- Probe design and selection
- Printing
- RNA extraction
- Labeling
- Hybridization and washing
  - Scanning
  - Data analysis