Introduction to microarray analysis and tools

Module A: Approach to Microarray-based studies

Agnes Viale, Ph.D.
Genomics Core lab
MSKCC
Plan

I- Introduction and potential applications of array platform

II- Existing platforms
   Spotted arrays
   Affymetrix GeneChip
   Other commercially available platforms

III- Experimental design
   Projects
   Technical limitations
   Financial considerations

IV- Steps involved in data analysis

V- Validation
   Real time PCR
   Northern Blot
1998- The Genomic Era

Genome: Gene+ Chromosome

Genomics

Structural Genomics (1986)
scientific discipline of
mapping, sequencing, and
analyzing genomes.

Functional Genomics (1995?)
Analysis of genome function
Characterization of the “transcriptome”
and the “proteome”.

Pubmed records for Microarray

Shena M, Shalon D, Davis RW, Brown PO.
Quantitative monitoring of gene expression patterns with a complementary DNA microarray.
(48 cDNA clones- Yeast)

Shalon D, Smith SJ, Brown PO.
A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization.
(874 cDNA clones-Yeast)
Microarray-The technical foundations

1975: Ed Southern
Labeled nucleic acid could be used to interrogate nucleic acid attached to solid support
⇒ Southern Blot

~1980: 1) Filter-based screening of clones libraries
2) Gridded libraries, stored in microtitre plates and stamped onto filters in fixed positions
Microarray- Key Innovations

1- Use of non-porous solid support (Glass)
   => miniaturization
   => fluorescence based detection

2- Methods for high-density spatial synthesis for oligonucleotides

Oligonucleotides array platform
Microarray - Definitions

Solid support (Glass) covered with spots of biomolecules
High Throughput Comparative hybridization platform
Monitor thousands of genes in one single experiment

Applications:
- Gene expression profiling
- Comparative Genomic Hybridization (CGH)
- Large scale protein-DNA interaction
- Genotyping...
Microarray- Definitions

- **Spotted/printed arrays:**
  Deposition of biomolecules by contact or non contact process
  In-house or commercial (Agilent-Amersham-Nimblegen)

- **High Density Oligonucleotides or GeneChip arrays (Affymetrix)**

- **Commercial Market (11-2003)**

  - Agilent 12%
  - Affymetrix 85%
Lawsuits in microarray field
Potential applications

RNA analysis
  - Gene expression profiling
  - Splice variant analysis

DNA analysis
  - Comparative Genomic Hybridization (CGH)
  - Large scale protein-DNA interaction
  - SNP analysis
  - Resequencing

Protein analysis (proteomics)
  - Tissue microarray
  - Etc Etc Etc
RNA analysis

Spotted cDNA array

Spotted oligonucleotides arrays Genechip (Affymetrix)
Splice variant analysis

One Strategy: exon junctions specific oligonucleotides

Problem with current labeling/amplification protocols 3’ biased
Genomic DNA Analysis

Comparative genomic hybridization (CGH) array to study chromosomal abnormalities (gain/loss)

1. cDNA
2. Oligo (very high density)
3. BAC (1MB resolution)
Genomic DNA Analysis

1- BAC array CGH: Identification of deletion on chromosome 2

2- Identification of commonly deleted region

3- Candidate gene
Promoter Analysis-CpG island array

ChIP on a CHIP = Chromatin immunoprecipitation

Comments
1- Control: total chromatin or mock IP
2- Amplification is necessary when working with mammalian cells

CpG islands array Intergenic regions and 5’ ends of genes.
SNP and RESEQUENCING array

- **Single Nucleotide Polymorphism array:**
  - high throughput analysis of DNA variation (100,000 SNP array)
  - LD (linkage disequilibrium)
  - LOH (loss of heterozygocity) studies
    - => Mapping of candidate genes/markers

- **Resequencing array:**
  - Comparative sequencing of candidate regions identified in mapping experiments
  - Current version 30kb, next version 100kb
Proteomics

- Antibodies array
- Protein array
- Two-D gels
- High throughput mass spectrometry analysis
Tissue microarray

Used in molecular classification of cancers for validation of gene array data by immunohistochemistry on hundreds of tissues.
Plan

I- Introduction and potential applications of array platform

II- Existing platforms
   Spotted arrays
   Affymetrix GeneChip
   Other commercially available platforms

III- Experimental design
   Projects
   Technical limitations
   Financial considerations

IV- Steps involved in data analysis
   too many steps to write

V- Validation
   Real time PCR
   Northern Blot
   RNAse protection assay
Platforms

- Spotted Arrays (cDNA or oligo)
- Oligonucleotides Genechips (Affymetrix)

Advantages:
- Spotted Arrays: Flexibility, Cost
- Oligonucleotides Genechips: Reliability, Reproducibility

Disadvantages:
- Spotted Arrays: Reliability, Reproducibility
- Oligonucleotides Genechips: Flexibility, Cost
Platforms

<table>
<thead>
<tr>
<th>Platforms</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spotted Arrays (cDNA or oligo)</td>
<td>Reliability</td>
<td>Flexibility</td>
</tr>
<tr>
<td></td>
<td>Reproducibility</td>
<td>Cost</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oligonucleotides Genechips (Affymetrix)</td>
<td>Reproducibility</td>
<td>Reliability</td>
</tr>
<tr>
<td></td>
<td>Flexibility</td>
<td>Cost</td>
</tr>
</tbody>
</table>

Advantages: Flexibility, Cost
Disadvantages: Reliability, Reproducibility
Spotted array platform

- cDNA clones (>10,000)
  - Miniprep
  - PCR
  - PCR cleaning
  - Quality control on gel
  - Printing on glass slides

- 1st generation: 1998-2001

Liquid handling station
Spotted array platform

• 1st generation: 1998-2001
  - cDNA clones (>10,000)
    - Miniprep
    - PCR
    - PCR cleaning
    - Quality control on gel
  - Printing on glass slides

• Semi-automated
• Results archived in database
Spotted array platform

• 1st generation: 1998-2001
  cDNA clones (>10,000)
  Miniprep
  PCR
  PCR cleaning
  Quality control on gel
  Printing on glass slides

Arrayer
Pins
Slides
Plate of cDNA
Printing process

- **Contact arrayers** (Home made or commercial)
  - Split pin
  - Micro spotting pin
  - Pin and ring

- **Non-contact arrayers**
  - Ink-jet printers
  - Piezoelectric

1- The Pins
Printing process

1- The Pins

2- The plate +cDNA

3- Slides (on a tray)
Printing process
Spotting problems
Hybridization problems
Spotted array platform

• 1st generation: 1998-2001
  cDNA clones (>10,000)
  Miniprep
  PCR
  PCR cleaning
  Quality control on gel
  Printing on glass slides

• 2nd generation: 2001-....
  Oligonucleotides libraries (25 to 100 mers)
  Printing on glass slides

Multi step process
Requires high level of QC
TIME CONSUMING
Competitive hybridization

Sample A

CY5

Reverse transcription Labeling

Sample B

CY3

- Sample A specific/enriched
- Sample B specific/enriched
- Equal levels of expression
Image Analysis-Gridding process
Image Analysis-Gridding process
Fluorescence intensities are proportional to level of expression of a given gene in each condition
Spot 12: Cy3/Cy5 = 4869/2635 = 1.8 => up regulation of 1.8 fold for gene 12 in condition A
Platforms

Spotted Arrays
(cDNA or oligo)

Advantages
Flexibility
Cost
Disadvantages
Reliability
Reproducibility

Oligonucleotides
Genechips
(Affymetrix)

Advantages
Reliability
Reproducibility
Disadvantages
Flexibility
Cost
One PM+MM = probe pair
11 PM=MM= probe set
Oligo = 25 mers

11 Perfect Match (PM) oligonucleotides
11 MisMatch (MM) oligonucleotides
MM => negative control
Affy array manufacturing - Photolithographic process
Affymetrix Genechips
Non-competitive hybridization

Affymetrix GeneChip

SampleB → chipB → Row CHP fileB → Comparison file
SampleA → chipA → Row CHP fileA

Spotted array

Two samples → One chip → Ratio (=Fold change)
Sample Labeling

RT with OdT-T7

In Vitro Transcription
T7 RNA polymerase

Labeling: Biotinilated nucleotides
Amplification (400-1000)

Labeled cRNA

Hybridization
Image Analysis

Gridding → Quantification

Level of expression of a gene is proportional to the intensity of fluorescence

Complicated statistical mysterious controversial calculations

One number per probe set (.CHP file)
Technology improvements

Reduction of feature size => more genes/array
Current version of human array: 44,000 transcripts /array

1997-2003
18 µm
(500,000 features/array)

2003-?
11 µm
(1,100,000 features/array)

<5 µm
(5,000,000 features/array)

Whole genome array
In 96 wells format
Data comparability

- Human Affy arrays

1995 HuFL
6800 genes

1998 U95 set
A,B,C,D,E arrays
63,000 genes

2001 U133set
A,B arrays
44,000 transcripts

2003 U133 2.0

- Redesign the oligos
- Change probe set names
- Redesign the oligos
- Change probe set names
- Keep same name
- Change the manufacturing process

⇒ Difficult to compare data from one version to the next.
⇒ If possible: use one version of the array across one project
Other commercial platforms

- **Agilent**
  - Inkjet spotted arrays. In situ synthesis of oligonucleotides (whole genome on 1 array)

- **Nimblegen**
  - 85,000 oligo. In situ light modulated oligo synthesis (Digital micromirror Device or DMD). Until recently, not sold in US

- **Amersham Codelink**
  - Slides are coated with a 3-D surface chemistry comprised of a long-chain, hydrophilic polymer containing amine-reactive groups
    => Higher sensitivity
Plan

I- Introduction and potential applications of array platform
II- Existing platforms
   Spotted arrays
   Affymetrix GeneChip
   Other commercially available platforms

III- Experimental design
   Projects
   Technical limitations
   Financial considerations

IV- Steps involved in data analysis
   too many steps to write

V- Validation
   Real time PCR
   Northern Blot
   RNAse protection assay
Experimental Design

\[ \mu_{xyz} = \alpha + \beta_1x + \beta_2y + \beta_3z + \gamma_1xy + \gamma_2xz + \gamma_3yz + \delta xyz \]

- Possibility 1:
  \((x^0, y^0, z^0), (x^*, y^*, z^0), (x^0, y^0, z^*)\)
  \[\mu_{000} = \alpha \quad \mu_{010} = \alpha + \beta_2 \quad \mu_{100} = \alpha + \beta_1 \quad \mu_{001} = \alpha + \beta_3\]
  \[\beta_1 = \bar{\mu}_{100} - \bar{\mu}_{000} \quad \text{Var}(\beta_1) = \text{Var}(\bar{\mu}_{100} - \bar{\mu}_{000}) = 2\sigma^2\]

- Possibility 2:
  \((x^0, y^0, z^0), (x^*, y^*, z^0), (x^0, y^*, z^*)\)
  \((x^*, y^0, z^0), (x^*, y^*, z^0), (x^*, y^0, z^*)\)
  There are 2N observations for every Placebo and 2N for every Verum
  \[\mu_{000} = \alpha \quad \mu_{110} = \alpha + \beta_1 + \beta_2 + \gamma_1 \quad \mu_{101} = \alpha + \beta_1 + \beta_3 + \gamma_2 \quad \mu_{011} = \alpha + \beta_2 + \beta_3 + \gamma_3\]
  \[\beta_1 = (\bar{\mu}_{110} + \bar{\mu}_{101} - \bar{\mu}_{000} - \bar{\mu}_{011}) / 2 \quad \text{Set } \gamma_i = 0, \text{ is this justified} \quad \text{Var}(\beta_1) = 4\sigma^2 / 4 = \sigma^2\]

Ref:
Microarrays for an integrative genomics. (Kohane, Kho, Butte)
Design and analysis of comparative microarrayexperiments (Yang, Speed)
When to use microarray?

- **Basic research: “Fishing” experiments**
  Not to **TEST** hypotheses but to **GENERATE** hypotheses
  Comparison between two (or more) states
  Study of new system: Knock Out mouse; drug treatment etc

- **Molecular classification of cancer**
  Gene expression profiling provide an **alternative molecular diagnostic**
  Better diagnostic, prognostic => more appropriate treatment

  Identification of markers => elucidation of molecular mechanisms underlying diseases.
Experimental design

Choices you have to make

Parameters which will determine your choices

1. Replicates
2. Platform
3. Amount of RNA
4. Budget
Replicates

Should I do replicates?

How many replicates?

Biological replicates: YES

Fake replicates: NO

RNA-1 → Array 1
RNA-2 → Array 2
RNA-3 → Array 3

HoxA9_GFP(August2003) HoxA9-2

28 3 36
8 15 8
4

22181 all genes

Statistician’s answer: 5
Realistic answer: 3
Investigator’s choice: 1 or 2
RNA Quality

Degraded (or partially degraded RNA) = bad array data
- Wrong conclusions
- Waste your time chasing false positives

- Homogeneity in a sample set is VERY IMPORTANT
RNA quantity

1- Enough RNA to process without amplification
   1 microgram < RNA

2- Not enough RNA: linear amplification
   20 ng < RNA < 700 ng

Diagram:
- Total RNA
  - Double strand cDNA
  - cRNA
  
- RT
  
- IVT

1X

2X
RNA quantity

3- ?< RNA<20 ng:

Laser Capture microdissection
FACS

1X
RT
IVT

RT
IVT

RT
IVT

RT
IVT

Total RNA
Double strand cDNA
cRNA

Double strand cDNA
cRNA

Double strand cDNA
Labeled cRNA
Pooling vs. Amplification

1. Pooling
   - 3 arrays
   - 0.9\(\mu\)g => no amplification

2. 2X Amplification for each sample
   - 3 arrays
   - “Some statistical Power”

RNA = 300 ng
Financial considerations

Why are the answers different depending on who is answering the question about the number of replicates?

Statistician’s answer: 5
Realistic answer: 3
Investigator’s choice: 1 or 2

Affy arrays (New York state) $ 400/ whole genome (human mouse rat)
Labeling (average in US)= $250
=> $650/samples
Two conditions (KO and WT), in triplicates = $3,900

Expensive technology but with good experimental design, it provides data that cannot be generated in other ways and is worth the investment.
1. Right questions and right controls
2. Homogeneity among samples
3. Don’t be cheap

For a mathematical/statistical approach of experimental design:
• Microarrays for an integrative genomics. (Kohane, Kho, Butte)
• Design and analysis of comparative microarray experiments (Yang, Speed)
Plan

I- Introduction and potential applications of array platform
II- Existing platforms
III- Experimental design

IV- Steps involved in data analysis
   - Data set QC
   - Normalization
   - Feature (gene) filtering
   - Replicate analysis
   - Clustering
   - Statistical tests
   - Pathway

V- Validation
   - Real time PCR
   - Northern Blot
   - RNAse protection assay
**Data set QC**

**GeneChip built-in control 1**: % present genes

- Varies according to tissue (30 to 60%)
- Higher on last generation of arrays (more sensitive)
- Allows easy identification of outliers in a dataset

<table>
<thead>
<tr>
<th>ArrayType</th>
<th>Filename</th>
<th>Noise RawQ</th>
<th>Scale Factor</th>
<th>Background</th>
<th>Number Present</th>
<th>% Present</th>
</tr>
</thead>
<tbody>
<tr>
<td>HG-U133A</td>
<td>JS_U133A_SetD-A-.CHP</td>
<td>2.42</td>
<td>3.222</td>
<td>Avg: 61.23</td>
<td>10210</td>
<td>45.80%</td>
</tr>
<tr>
<td>HG-U133A</td>
<td>JS_U133A_SetD-A+.CHP</td>
<td>2.50</td>
<td>3.240</td>
<td>Avg: 60.43</td>
<td>10000</td>
<td>44.90%</td>
</tr>
<tr>
<td>HG-U133A</td>
<td>JS_U133A_SetD-B-.CHP</td>
<td>2.44</td>
<td>4.643</td>
<td>Avg: 65.01</td>
<td>9404</td>
<td>42.20%</td>
</tr>
<tr>
<td>HG-U133A</td>
<td>JS_U133A_SetD-B+.CHP</td>
<td>2.74</td>
<td>2.603</td>
<td>Avg: 73.12</td>
<td>10165</td>
<td>45.60%</td>
</tr>
<tr>
<td>HG-U133A</td>
<td>JS_U133A_SetD-C-.CHP</td>
<td>2.72</td>
<td>2.157</td>
<td>Avg: 65.20</td>
<td>10211</td>
<td>45.80%</td>
</tr>
<tr>
<td>HG-U133A</td>
<td>JS_U133A_SetD-C+.CHP</td>
<td>2.89</td>
<td>2.255</td>
<td>Avg: 73.21</td>
<td>10006</td>
<td>44.90%</td>
</tr>
<tr>
<td>HG-U133A</td>
<td>JS_U133A_SetD-D-.CHP</td>
<td>2.77</td>
<td>2.820</td>
<td>Avg: 71.02</td>
<td>9767</td>
<td>43.80%</td>
</tr>
<tr>
<td>HG-U133A</td>
<td>JS_U133A_SetD-D+.CHP</td>
<td>2.28</td>
<td>5.662</td>
<td>Avg: 54.54</td>
<td>8686</td>
<td>39.00%</td>
</tr>
<tr>
<td>HG-U133A</td>
<td>JS_U133A_SetD-E-.CHP</td>
<td>3.10</td>
<td>3.300</td>
<td>Avg: 83.54</td>
<td>9156</td>
<td>41.10%</td>
</tr>
<tr>
<td>HG-U133A</td>
<td>JS_U133A_SetD-E+.CHP</td>
<td>2.39</td>
<td>2.187</td>
<td>Avg: 56.12</td>
<td>9355</td>
<td>40.40%</td>
</tr>
</tbody>
</table>

**3X amplification**

<table>
<thead>
<tr>
<th>ArrayType</th>
<th>Filename</th>
<th>Noise RawQ</th>
<th>Scale Factor</th>
<th>Background</th>
<th>Number Present</th>
<th>% Present</th>
</tr>
</thead>
<tbody>
<tr>
<td>HG-U133A</td>
<td>MLe_U133A_C1.CHP</td>
<td>2.33</td>
<td>10.221</td>
<td>Avg: 55.61</td>
<td>8140</td>
<td>36.50%</td>
</tr>
<tr>
<td>HG-U133A</td>
<td>MLe_U133A_C2.CHP</td>
<td>2.43</td>
<td>5.836</td>
<td>Avg: 58.10</td>
<td>8626</td>
<td>38.70%</td>
</tr>
<tr>
<td>HG-U133A</td>
<td>MLe_U133A_C3.CHP</td>
<td>2.16</td>
<td>14.664</td>
<td>Avg: 51.92</td>
<td>6435</td>
<td>28.90%</td>
</tr>
<tr>
<td>HG-U133A</td>
<td>MLe_U133A_S1.CHP</td>
<td>2.37</td>
<td>6.432</td>
<td>Avg: 59.89</td>
<td>7220</td>
<td>32.40%</td>
</tr>
<tr>
<td>HG-U133A</td>
<td>MLe_U133A_S2.CHP</td>
<td>2.23</td>
<td>9.504</td>
<td>Avg: 53.16</td>
<td>7371</td>
<td>33.10%</td>
</tr>
</tbody>
</table>
### Data set QC

**GeneChip built-in control 2: 3’/5’ ratio for “house keeping” genes**

1 $<$ Signal 3’ probe/Signal 5’ probe $<$ 3

---

#### Housekeeping Controls:

<table>
<thead>
<tr>
<th>Probe Set</th>
<th>Sig(5’)</th>
<th>Det(5’)</th>
<th>Sig(M’)</th>
<th>Det(M’)</th>
<th>Sig(3’)</th>
<th>Det(3’)</th>
<th>Sig(all)</th>
<th>Sig(3'/5’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUMISGF3A/M97935</td>
<td>66.7</td>
<td>P</td>
<td>374.9</td>
<td>P</td>
<td>658.9</td>
<td>P</td>
<td>9.88</td>
<td></td>
</tr>
<tr>
<td>HUMRGE/M10098</td>
<td>455.6</td>
<td>P</td>
<td>244</td>
<td>P</td>
<td>577.2</td>
<td>P</td>
<td>1.27</td>
<td></td>
</tr>
<tr>
<td>GAPDH</td>
<td>24262.7</td>
<td>P</td>
<td>19308</td>
<td>P</td>
<td>19918.3</td>
<td>P</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td>b-ACTIN</td>
<td>14621.8</td>
<td>P</td>
<td>18055</td>
<td>P</td>
<td>15733</td>
<td>P</td>
<td>1.08</td>
<td></td>
</tr>
<tr>
<td>M27830</td>
<td>494.8</td>
<td>P</td>
<td>1132</td>
<td>P</td>
<td>150.4</td>
<td>A</td>
<td>0.3</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample Name</th>
<th>Percent Present</th>
<th>![ Signal 3’ probe/Signal 5’ probe ]&lt; 3</th>
<th>![ Signal 3’ probe/Signal 5’ probe ]&lt; 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>s1</td>
<td>39.40%</td>
<td>GAPDH 216.00 P 352.50 A 1880.40 P 816.30 A 8.71</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>B-ACTIN 63.00 A 537.70 A 7356.60 P 2652.42 A 116.80</td>
<td></td>
</tr>
<tr>
<td>s2</td>
<td>44.10%</td>
<td>GAPDH 211.30 P 544.40 P 3047.90 P 1267.87 P 14.42</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>B-ACTIN 42.20 A 539.10 P 7630.10 P 2737.13 P 180.95</td>
<td></td>
</tr>
<tr>
<td>s3</td>
<td>41.40%</td>
<td>GAPDH 189.10 P 747.30 P 6559.80 P 2498.71 P 34.70</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>B-ACTIN 102.50 A 248.00 A 13449.20 P 4599.90 P 131.16</td>
<td></td>
</tr>
</tbody>
</table>
Genes filtering

Filtering parameters
  - Fold change cut-off
  - Merge Replicate analysis
  - P-values (statistical test)

Advice 1: Most stringent filtering → Less stringent

Advice 2: Do not forget about biology
Replicate analysis

Exp: WT vs. KO- Triplicate for each condition
List of regulated genes?
Filtering on fold change (and Affy p-val- next module)

KO1 → WT1
KO2 → WT2
KO3 → WT3

3 comparisons

KO1 → WT1
KO2 → WT2
KO3 → WT3

9 comparisons
Replicate analysis

Exp: WT vs. KO- Triplicate for each condition
List of regulated genes?
Filtering on fold change (and Affy p-val- next module)

If samples were processed two by two, then you should match
Each KO with the corresponding WT

KO1 → WT1
KO2 → WT2
KO3 → WT3
Replicate Analysis

“9 comparisons” more stringent analysis than “3 comparisons”
“9 comparisons” included in the “3 comparisons”
Clustering

Clustering = grouping genes (or samples) with similar expression pattern.

Why clustering?

Groups of Genes Which Share Common Patterns of Expression May Share Common Transcriptional Regulation

Hierarchical
Partitional
- K-means
- SOM (Self Organizing Map)
- PCA (Principal Component Analysis)

Supervised clustering
Clustering

Principle

Real life situation: Signal to noise ratio issue

Nick Socci, Bioinformatics Core, MSKCC
Clustering

- Not always necessary

- No straightforward solution for clustering
  a. Choose an algorithm which make the fewest/simplest assumptions.
  b. At least know the assumptions the algorithm is making.
  c. Supervised vs unsupervised: may use an unsupervised algorithm but features selection supervised: may not be a bad idea to start.

- Interactive process between the person who will analyze the data and the biologist. This interaction should start at the experimental design level

Nick Socci, Bioinformatics Core, MSKCC
To test the **null hypothesis** : the hypothesis for each gene is that there is no difference in the mean gene expression intensities in the groups tested

=> a gene will have equal means across every group

- Rejection of the null hypothesis (i.e. acceptance of the alternative hypothesis) indicates that the means intensities are from two different populations.

- The value(s) returned by `ttest` is the P-value: indicates the probability of getting a mean difference between the groups as high as what is observed by chance. The lower the P-value, the more significant the difference between the groups.
## Statistical Tests

### When to use what?

<table>
<thead>
<tr>
<th>2 groups</th>
<th>More than 2 groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parametric (variances equal)</td>
<td>Student’s T-test</td>
</tr>
<tr>
<td>Parametric (variance not equal)</td>
<td>Welch T-test</td>
</tr>
<tr>
<td>Non Parametric</td>
<td>Wilcoxon-Mann Whitney test</td>
</tr>
</tbody>
</table>

- **To use if very few replicates or group without replicates**: Recommended for most cases.
- **Least assumption**: More than 5 replicates per group.
Pathway analysis

A pathway is a graphical representation of the interaction between gene products in a biological system.

Genes can be superimposed on the pathway, allowing you to view their expression levels in a biological context.
Pathway analysis

• **GenMapp** (<http://www.genmapp.org/>) (human, mouse, rat, yeast)

• **EASE**: the Expression Analysis Systematic Explorer (<http://david.niaid.nih.gov/david/ease.htm>)

• **Cytoscape** (<http://www.cytoscape.org>)

• **Pathway Processor**
  http://cgr.harvard.edu/cavalieri/pp.html (yeast, B. subtilis)

(Module: Functional interpretation of high-throughput data)
Data analysis-Conclusions

• Data QC => eliminate outliers samples
• Gene filtering => Help focusing on “interesting” genes
• Clustering => only if necessary
  => noise-dependant
• Statistical test => use the appropriate one
• Pathway analysis => Global picture of regulation
Plan

I- Introduction and potential applications of array platform
II- Existing platforms
III- Experimental design

IV- Steps involved in data analysis
   - Data set QC
   - Normalization
   - Feature (gene) filtering
   - Replicate analysis
   - Clustering
   - Statistical tests
   - Pathway

V- Validation
   - Real time PCR
   - Northern Blot
   - RNAse protection assay
   - ISH
Validation of microarray data

• Do I need to validate my data?
  – YES

• Why?
  – Because it will make you feel more confident about your results
  – Because the reviewers will ask for it

• How?

Oct.2004 survey: Do you validate your microarray data? If yes, how?

<table>
<thead>
<tr>
<th>Method</th>
<th>Response Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northern Analysis</td>
<td>31.60%</td>
</tr>
<tr>
<td>RNase Protection</td>
<td>8.40%</td>
</tr>
<tr>
<td><strong>Real-time PCR</strong></td>
<td><strong>87.40%</strong></td>
</tr>
<tr>
<td>In-situ hybridization</td>
<td>16.80%</td>
</tr>
<tr>
<td>None</td>
<td>9.50%</td>
</tr>
<tr>
<td>Other</td>
<td>8.40%</td>
</tr>
</tbody>
</table>

Total Respondents: 95
Northern-blot: RNA > 5μg

Cyclophilin
Average Difference (wt || ob) 1* || 778
Fold Change ~31.5

Fold Change ~31.5

Cyclophilin
Average Difference (wt || ob) 548 || 71*
Fold Change ~-7.7

Fold Change ~-7.7

Real Time PCR

- “Real time quantitative PCR.”
  Heid CA, Stevens J, Livak KJ, Williams PM.

- Very “popular”
- Request smaller amounts of RNA

- Real-time reverse-transcriptase (RT) PCR quantitates the initial amount of the template most specifically, sensitively and reproducibly, and is a preferable alternative to other forms of quantitative RT-PCR, which detects the amount of final amplified product.
- Based on the detection and quantification of a fluorescent reporter
Real Time PCR

Expensive assays:
- Taqman
- Molecular beacon
- Scorpions

Cheaper alternative
- SYBR-green: measures the amplicon production (including non-specific amplification and primer-dimer complex)
Real Time PCR

**Internal reference:**
All data should be “normalized” to a reference gene (HPRT, Actin, GAPDH, Cyclophilin…)

**Correlation between array data and Q-PCR data:**
Good to excellent even at fold changes<2
CONCLUSIONS

- Everyone should take advantage of this HYPOTHESES GENERATING Technology, ideal to “explore” a new biological system

- Expensive technology but with good experimental design, it provides data that cannot be generated in other ways

<table>
<thead>
<tr>
<th>Before microarray</th>
<th>After microarray</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thousands of days</td>
<td>One day</td>
</tr>
<tr>
<td>Thousands of experiments</td>
<td>One experiment</td>
</tr>
<tr>
<td>=&gt; Study of one gene</td>
<td>Study thousands of genes</td>
</tr>
</tbody>
</table>