Microarray

Key components
- Array
- Probes
- Detection system

Normalisation

Data-analysis - ratio generation
MICROARRAY

Measures Gene Expression

Global - Genome wide scale

Why Measure Gene Expression?

What information does it tell us?
Mammalian cells, same organism, identical DNA

Muscle fiber (cell)

Fat cells

Neuron (nerve) cell
PROTEINS

Different cells contain different proteins
or
Different quantities of specific proteins

Different proteins have diverse functions
  structural
  enzymes (biochemical reactions)
  reproduction
  protect (stress)

How do different cells with identical DNA contain different proteins?
Gene expression - Protein synthesis

DNA (template) → Transcription factors → Transcription → gene

Promoter → transcription → mRNA (transcribed gene)

nucleus → transcription

cytoplasm → translation

Proteins synthesised that have functional roles in the cell

Measuring mRNA levels tells us what proteins cells are making
Individual genes encode specific proteins

Important for:

cell differentiation/development
specialise cells for specific functions

Stress responses

pathogens (antibodies)
drought (protect confer tolerance)
In plants, Rice / Arabidopsis,

25-30,000 genes (~60% expressed at once)

Energy requirement for making proteins

Tight regulation to synthesise only required proteins
Protein requirements of a cell change constantly

Gene expression is constantly changing
developmental stages
stress biotic (pathogens) or abiotic (water, light etc)

Expressed gene (up / down)

Non-expressed gene (can be activated)

typical of stress response genes/proteins
normal conditions : not required : no/low expression
stress : required for tolerance/repair: strong induction
Gene expression analysis
typically, compare two or more conditions

Plant examples

Tissue specificity:
Root Vs leaf or leaf Vs flower

Development
seed Vs seedling

Stress
Pathogen challenge Vs healthy (control)
Drought Vs healthy (control)

Compare relative abundance of transcripts
not measuring absolute values ie: number of transcripts
Summary

Measure changes in gene expression

measuring dynamic state of a cell/organism

identify genes that are differentially expressed (up/down) in response to specific conditions

function in specific processes

↑ gene transcription > ↑ protein > functional role in process

↓ gene transcription > protein ↓ > not required
Implications for genetic engineering

Over-express an endogenous gene

or

Transform gene from other organism

Increase stress tolerance/disease resistance
Microarray

Global genome analysis of gene transcription

Commonly used at the beginning of a research program to identify genes expressed in response to specific conditions

Relatively new (10 years)

Most common technique to study global gene expression

Entire genome (-30K genes) in a single experiment.
Restricted to sequenced genes

In Arabidopsis and rice the full genome sequence has been determined

Commercially manufactured slides can be purchased that contain the entire genome: 30,000 genes
Two distinct platforms

different design / manufacturers

Single Channel
    test either treatment or control
    eg: Affymetrix

Two Channel
    contains both treatment and control

Fundamentally operate by the same principle
All microarray systems have common key components:

The **array**, which contains thousands of nucleic acid (NA) sequences or ‘targets’ spotted onto small chip

One or more labeled samples or ‘probes’ that are hybridised to the array

A **detection system** that quantifies the hybridisation signal
Microarray: component overview

2 channel

mRNA

Labeled ‘probes’

‘Target’ array

Detection system

Cy3-dUTP → Label cDNA with fluorophores → Cy5-dUTP

Microarray hybridization → Laser excitation → Image overlay

16-bit image

No differential expression

Up-regulated

Down-regulated
The Array

small chip (size microscope slide)

contains tens of thousands of specific fixed nucleotide sequences that represent specific genes

each sequence must be unique to a specific gene

Only genes represented on the array can be measured
Array target design

1 ATGGTACCAG CAGTAAATCC TCCTACCACA TCCAACCACG TGGCAGTAAT
51 CGGTGCTGGA GCAGCCGGAC TCGTGGCTGC TCGAGAGCTC CGCCGTGAGG
101 GCCACTCTCT GCTTGGTTTC GAACGTGGGA ACCACATCGG AGGCGTGTGG
151 GCTTACACGC CAAACGTTGA ACCCGACCCC CTTAGCATCG ACCCGACCCG
201 ACCCGTAAAC CATTCGAGCC TCTACAGTTC TCTCCGAACC ATCATCCCAC
251 AAGAATGCAT GGGTTTCACT GATTTCCCGT TTTCGACCAG ACTCGAAAAC
301 GGGTCTAGAG ATCCGAGAAG ACATCCAGGT CATAGTGAGG TTCTTGCTTA
351 CCTGAGAGAC TTTGTGAGGG AGTTCAAGAT TGAGGAGATG ATTCGTTTCG
401 AGACGGAGGT TGTGAGGGTT GAGCAGGCGG GGGAAAATCC CAAAAAGTGG
451 AGAGTCAAGT CTAGAAATTT CGGTGATATC TCCGACGAGA TTTATGACGC
501 CGTAGTTGTT TGTAACGGTC ACTATACAGA GCCTCGTCAT GCTCTAATAC

Full length coding sequence
of gene (550nt)

Only segment of gene used to make target
Spotted sequence must be unique to the specific gene

Requires careful planning,
many genes share similar sequences
Affymetrix (single channel)

Very popular, high accuracy
Each gene represented by around 10 pairs on an array
perfect match -PM
miss match –MM (non-specific)
Disadvantage
use twice as many arrays (expensive)
‘Probe’ preparation

Example Experiment:
Sample 1: Treatment group Exposed to pathogen

Sample 2: Control group

Total RNA extraction, purify mRNA

mRNA (2% of total)

contains the genetic information of genes being expressed
RNA and gene transcription

Complementary DNA synthesis from mRNA

During cDNA synthesis a tag is incorporated to which the fluorescent/dye label is attached

Have two RNA/cDNA pools – represent TREATMENT AND CONTROL
‘Probe’/ sample labelling

A common labelling method is fluorescent dyes, Cy3 & Cy5 -
They achieve:

- High sensitivity of detection
- Detection of different dyes simultaneously (2 channel)
- Avoid toxicity of radioactive labels

Label different samples with different specific dyes to distinguish the specific binding of each group within the same experiment.

Single channel: only need one label
‘Probe’/ sample labelling

**Cy 5**
- Treatment group
- Control group

**Fluorescent labels**
- Cy 3
- Tagged mRNA/cDNA
- Control group

**Cy 5**
- Labeled cDNA
- Control group

**Cy 3**
- Treatment group

AAAAA- AAAAA- AAAAA- AAAAA-
Hybridisation

Array ‘target’

Labelled ‘probe’ (mRNA, labelled cDNA)

Add labelled ‘probe’ and cover

Hybridise overnight in humid environment

Wash - remove unbound probe so only specifically bound labelled probe remains

Scan
Hybridisation

The labelled cDNA will bind specifically to complementary target sequences (nucleotides) present on the array

Increased transcript = more binding = increased signal
Identification of differentially expressed genes

Two channel

After hybridisation

each ‘target’ or dot on chip will bind various quantities of the two labelled ‘probes’ which are representative of the relative amount of mRNA present in each sample

Each target dot will emit 2 fluorescent signals, the intensity of which is related to the amount of bound probe

No saturation of the target - Not a competitive reaction, the amount bound will be related to the relative amounts present
An example:

- Treatment group, exposed to pathogen,
- Activation / up-regulation of defence related genes e.g. PR-1
- More PR-1 mRNA/cDNA present in treatment group, cf. to control
- Result: more binding and more fluorescent signal.
Labeling and hybridisation

Measure Fluorescence in 2 channels red/green

Analyze the data to identify patterns of gene expression
A two-color hybridization

Control gp
4 copies of gene A, 1 copy of gene B

Treatment gp
4 copies of gene A, 4 copies of gene B

After Hybridization
Cy3/Cy5
For genes A, B,.....

Analysis

Gene A
Gene B

Labeled probes

Array chip
Critical steps in the generation & analysis of data

Array

Excitation

Emission

Cy3: 532nm

Cy3: 550 - 600nm

Cy5: 635nm

Cy5: 655 - 695nm

Gene Pix

Overlay Images

Single channel > only one signal / slide
Each spot represents a specific NA sequence unique to a particular gene

Eg:
Treatment group: green Cy3
Control group: red Cy5

Yellow:
Similar amounts of both the Cy3 and 5 labelled probes bound

Green:
Relatively more of the Cy3 labelled probe bound (up-regulated)

Red:
Relatively more of the Cy5 labelled probe bound (down-regulated)
Addressing or gridding

- The basic structure/layout of a microarray image is determined by the arrayer.

- File associated with each version of an array indicates specific location of spots.

- Gridding the spots on the image results in matching the known spot layout with the scanned image data.

- Software program knows the specific location of genes (NA sequences) on the slide and can thus associate the emitted signal from specific spots with a specific gene name or sequence identifier.
Automatic spot-finding

Automatic spot finding algorithms are common

Grids must be aligned with spots so that the software “knows” exactly what area of the slide to read the signal for any specific sequence
Results

Once spots are aligned, software generates quantitative numerical results.

Results in the generation of masses of data (excel file):
- 30,000 genes x 10x pairs
- each gene generates around 50 different columns of data
- Mean Foreground and background emissions for each spot
  also......

RAW DATA OUTPUT FILES
- .CEL files Affymetrix

Must be processed
Normalisation

First step in analysis

Standardises data from a number of microarrays into common scale

Allows meaningful comparisons between slides

Options

- background corrections (subtract)
- standardisation between slides
- PM-MM adjustment (non-specific binding)
- computes average value of all probes in the probe-set
Important to normalise all related slides together

Eg: Time course study

3 x 0h control replicates / 3 x 0h treatment replicates
3 x 1h control replicates / 3 x 1h treatment replicates
3 x 6h control replicates / 3 x 6h treatment replicates etc

Accounts for technical variability between slides
Simplified Normalisation over view

Rep1 control
Rep1 treatment

General technical variability

Slightly less cDNA added (user error, tech limitations)
dye older
slide older

Differences in emitted signal due technical variability's

Normalisation aims to compensate for these
## Table

<table>
<thead>
<tr>
<th>GENE</th>
<th>T1</th>
<th>C1</th>
<th>ratio</th>
<th>log2 RATIO</th>
<th>N C1 (x1.34)</th>
<th>T1</th>
<th>N C1</th>
<th>ratio</th>
<th>log2 RATIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1300.00</td>
<td>1000.00</td>
<td>1.30</td>
<td>0.38</td>
<td>1330.00</td>
<td>1300.00</td>
<td>1330.00</td>
<td>0.98</td>
<td>-0.03</td>
</tr>
<tr>
<td>B</td>
<td>2600.00</td>
<td>2000.00</td>
<td>1.30</td>
<td>0.38</td>
<td>2660.00</td>
<td>2600.00</td>
<td>2660.00</td>
<td>0.98</td>
<td>-0.03</td>
</tr>
<tr>
<td>C</td>
<td>5200.00</td>
<td>4000.00</td>
<td>1.30</td>
<td>0.38</td>
<td>5320.00</td>
<td>5200.00</td>
<td>5320.00</td>
<td>0.98</td>
<td>-0.03</td>
</tr>
<tr>
<td>D</td>
<td>650.00</td>
<td>500.00</td>
<td>1.30</td>
<td>0.38</td>
<td>665.00</td>
<td>650.00</td>
<td>665.00</td>
<td>0.98</td>
<td>-0.03</td>
</tr>
<tr>
<td>E</td>
<td>800.00</td>
<td>200.00</td>
<td>4.00</td>
<td>2.00</td>
<td>266.00</td>
<td>800.00</td>
<td>266.00</td>
<td>3.01</td>
<td>1.59</td>
</tr>
<tr>
<td>F</td>
<td>7800.00</td>
<td>6000.00</td>
<td>1.30</td>
<td>0.38</td>
<td>7980.00</td>
<td>7800.00</td>
<td>7980.00</td>
<td>0.98</td>
<td>-0.03</td>
</tr>
<tr>
<td>AV SIG</td>
<td>3058.33</td>
<td>2283.33</td>
<td>1.34</td>
<td>0.42</td>
<td>3036.83</td>
<td>3058.33</td>
<td>3036.83</td>
<td>1.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>

### Note

Average signal intensity/ratio should be same across arrays
Normalisation

- dye bias
- efficiency of dye incorporation
## Ratio generation – log2 transformation

<table>
<thead>
<tr>
<th>gene</th>
<th>treatment</th>
<th>control</th>
<th>RATIO</th>
<th>RATIO (log base 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1000</td>
<td>1000</td>
<td>1</td>
<td>no change</td>
</tr>
<tr>
<td>B</td>
<td>2000</td>
<td>1000</td>
<td>2</td>
<td>UP 2x</td>
</tr>
<tr>
<td>C</td>
<td>500</td>
<td>1000</td>
<td>0.5</td>
<td>DOWN 2x</td>
</tr>
<tr>
<td>D</td>
<td>4000</td>
<td>1000</td>
<td>4</td>
<td>UP 4x</td>
</tr>
<tr>
<td>E</td>
<td>1000</td>
<td>4000</td>
<td>0.25</td>
<td>DOWN 4x</td>
</tr>
</tbody>
</table>

### RATIO vs. RATIO LOG2

<table>
<thead>
<tr>
<th>RATIO</th>
<th>RATIO LOG2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>16</td>
<td>4</td>
</tr>
<tr>
<td>32</td>
<td>5</td>
</tr>
</tbody>
</table>
Let $A_i$: Green intensity
$B_i$: Red intensity

$R_i = \frac{A_i}{B_i}$

$\log_2 R_i = \log_2 \left( \frac{A_i}{B_i} \right)$

**Gene1:** $R_1 = 4$, $\log_2 R_1 = 2$

**Gene2:** $R_2 = 0.25$, $\log_2 R_2 = -2$

**Ratio 1 = no difference = $\log_2 0$**

**Advantages of log transformation:**

Treat up-regulated and down-regulated genes symmetrically