Genomics III

Gene Expression Measurements:
DNA Microarrays
Measuring Gene Expression

• To measure the amount of a specific mRNA, one needs a specific probe – nucleic acid of complementary sequence
• One needs a means to detect the binding of the probe to the target
• RNA is very susceptible to nucleases – usually copied to cDNA before probing
Techniques for Measuring Gene Expression

• Northern blot
  - The old-fashioned way
  - Gel electrophoresis with a large, sequence-specific probe

• DNA microarray
  - The new-fangled way
  - Get profiles of expression of many genes at a time

• RT-PCR
  - Amplifies small amounts of starting material
  - Semi-quantitative at best

• Real-time PCR
  - Sensitive and quantitative
Northern blot

- Separation by gel electrophoresis
- Detection/quantitation by labeled cDNA probe (radio- or fluorescence)

From www1.qiagen.com

Formaldehyde-agarose gel (major bands are rRNA)

Northern for GAPDH

From www1.qiagen.com
DNA Microarrays

- Probe the expression of thousands of genes simultaneously, using probes immobilized on a microscope slide (or the equivalent)

http://filebox.vt.edu/cals/cses/maroof/mglab/Microarray.html
DNA Microarray Overview

**Making Microarrays**
- Design microarray layout and controls
- Spotting
- Arrays
- QC

**Sample Preparation**
- Experimental design
- RNA samples
- Labeled cDNA
- QC

**Data Analysis**
- Hybridization
- Imaging
- Background correction
- Normalization
- Data mining
Types of Microarray Experiments

- **Diagnostic/Prognostic**
  - Want to classify samples according to gene expression profile
    - Cancer sub-typing
    - Drug responsiveness

- **Biological Dynamics**
  - Transcriptional “program”
  - Unraveling gene networks
Experimental Design Issues

- **Samples**
  - Patient, in vivo, in vitro?
  - Homogeneous or mixed-cell population?

- **Controls**
  - Time-matched controls
  - Initial condition

- **Normalization**
  - Housekeeping genes - specific or generalized
  - Sample RNA content and purity

- **What are you trying to find?**
  - Gene expression dynamics
  - Differentially expressed genes
  - Sample classifiers
Spotted Microarray Basics

Sample Preparation

Control sample
Extract RNA, make cDNA
Label with Cy3
Co-hybridize to array

Test sample
Extract RNA, make cDNA
Label with Cy5

Hybridization

Control sample
Test sample
CO-hybridize to array

Data Analysis

Test gene expression

Control gene expression

Expression ratio

Background correction; normalization

Gene expression

<table>
<thead>
<tr>
<th>geneA 3.9</th>
<th>geneB 0.2</th>
<th>...</th>
<th>geneN 4.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>geneA 2.6</td>
<td>geneB 1.0</td>
<td>...</td>
<td>geneN 2.0</td>
</tr>
<tr>
<td>geneA 1.5</td>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Extract RNA, make cDNA
Label with Cy3
Label with Cy5
Affymetrix Gene Chip Arrays

- Market leader
- Technology based on sets of matched and mismatched oligo probes
- Biotin-cRNA vs. cDNA
- One sample hybridized at a time
- Tight controls on instrumentation, software, etc.

www.affymetrix.com
## Probe Specifications

### Critical Specifications for GeneChip® Human Genome Arrays

<table>
<thead>
<tr>
<th></th>
<th>Human Genome U133 Plus 2.0 Array</th>
<th>Human Genome U133A 2.0 Array</th>
<th>Human Genome U133 Set</th>
<th>Human Genome Focus Array</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of arrays in set</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Number of transcripts</td>
<td>~47,400</td>
<td>18,400</td>
<td>~39,000</td>
<td>~8,500</td>
</tr>
<tr>
<td>Number of genes</td>
<td>38,500</td>
<td>14,500</td>
<td>~33,000</td>
<td>~8,400</td>
</tr>
<tr>
<td>Number of probe sets</td>
<td>&gt;54,000</td>
<td>&gt;22,000</td>
<td>&gt;45,000</td>
<td>&gt;8,700</td>
</tr>
<tr>
<td>Feature size</td>
<td>11 µm</td>
<td>11 µm</td>
<td>18 µm</td>
<td>18 µm</td>
</tr>
<tr>
<td>Oligonucleotide probe length</td>
<td>25-mer</td>
<td>25-mer</td>
<td>25-mer</td>
<td>25-mer</td>
</tr>
<tr>
<td>Probe pairs/sequence</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
</tbody>
</table>

Control sequences included:
- **Hybridization controls**
  - **Poly-A controls**
  - **Normalization control set**
  - **Housekeeping/Control genes**
    - 100 probe sets GAPDH, beta-Actin, ISGF-3 (STAT1)
    - 100 probe sets GAPDH, beta-Actin, ISGF-3 (STAT1)
    - 100 probe sets GAPDH, beta-Actin, ISGF-3 (STAT1)

Detection sensitivity: 1:100,000*  
*As measured by detection of pre-labeled transcripts derived from human cDNA clones in a complex human background.

[www.affymetrix.com](http://www.affymetrix.com)
Microarrayers (Robots)

- Vacuum for drying pins
- Water for washing pins
- Blank glass plate for blotting pins
- 384 well plate containing DNA clones for spotting
- About 100 microscope slides for making DNA microarrays
- Robot arm
- 48 pins for printing
- Temperature and humidity control
Microarrays (photolithography)
RNA Purification

1. Stabilize sample in RNA later RNA Stabilization Reagent.
2. Disrupt sample and lyse with GITC-containing buffer (Buffer RLT).
3. Homogenize sample to shear genomic DNA and reduce viscosity of lysate.
4. Add ethanol to adjust binding conditions.
5. Apply sample to RNeasy spin column for adsorption of RNA to membrane.
6. Remove contaminants with special wash spins (Buffers RW1 and RPE).
7. Elute ready-to-use RNA in water.

Qiagen RNeasy Mini Handbook
Hybridization

- Need temperature and humidity control
- Diffusion-limited
- Large surface-volume ratio: careful for uniformity
Image Analysis

- Two-color TIFF images (not really red and green)
- Local or global background subtraction

Cy5

Cy3
Normalization

- Red and Green have different intensities
- Not necessarily related by simple ratio
- M-A plot: \( M = \log_2(R/G), \ A = \frac{1}{2}[\log_2(R \times G)] \)
- Lowess = local fitting function
Normalization (cont’d)

- Can use *rank invariant* set of housekeeping genes

Other Normalizations and Filtering

• So much data – need to look hard to make sure there are no systematic biases.
• E.g., look for spatial biases in R/G
<table>
<thead>
<tr>
<th>GENE</th>
<th>0 h</th>
<th>1 h</th>
<th>4 h</th>
<th>8 h</th>
</tr>
</thead>
<tbody>
<tr>
<td>M20259 cds RATFNE2 Rat fibronectin gene, exons 2b and 3a</td>
<td>661.75</td>
<td>54.95</td>
<td>339.2</td>
<td>444.75</td>
</tr>
<tr>
<td>O72795 SUTR welcomes Rattus norvegicus SPARc mRNA, untranslated region, partial sequence</td>
<td>43.7</td>
<td>205.3</td>
<td>122.65</td>
<td>73.65</td>
</tr>
<tr>
<td>U02612 cds RN32612 Rattus norvegicus fibronectin (fn-1) gene, partial cds</td>
<td>3670.45</td>
<td>1025</td>
<td>2144.25</td>
<td>3637.65</td>
</tr>
<tr>
<td>Y01277 Rat gene encoding cytoplasmic beta-actin (5.5 M, 0.3 represents transcript regions 5 prime, Middle, and 3 prime</td>
<td>456.6</td>
<td>508.2</td>
<td>1539.6</td>
<td>466.7</td>
</tr>
<tr>
<td>X70389 R. norvegicus mRNA for pro alpha 1 collagen type III /cos (0.1911) /gb = X70936 /gb = X7915 /gb = RN.3427 /len = 2182</td>
<td>67.75</td>
<td>216.65</td>
<td>96.6</td>
<td>96.6</td>
</tr>
<tr>
<td>X78997 R. norvegicus (Sprague Dawley) mRna for cadherin /gb = (327, 2810) /gb = X79997 /gb = 505662 /gb = Rn.3800 /len = 31</td>
<td>233.65</td>
<td>364.45</td>
<td>270.4</td>
<td>31</td>
</tr>
<tr>
<td>Y01217 Rat gene encoding cytoplasmic beta-actin (5.5 M, 0.3 represents transcript regions 5 prime, Middle, and 3 prime</td>
<td>308.65</td>
<td>359.35</td>
<td>1272.25</td>
<td>324.1</td>
</tr>
<tr>
<td>Acid nuclear phosphoprotein 32 (leucine rich)</td>
<td>130.35</td>
<td>150.65</td>
<td>364.2</td>
<td>126.4</td>
</tr>
<tr>
<td>AF020618 Rattus norvegicus progression elevated gene 3 protein mRNA, complete cds</td>
<td>550.55</td>
<td>568</td>
<td>478.2</td>
<td>485.15</td>
</tr>
<tr>
<td>AJ133902 RND13902 Rattus norvegicus mRNA for GAS7 protein</td>
<td>126.1</td>
<td>151.9</td>
<td>184.3</td>
<td>160.6</td>
</tr>
<tr>
<td>Hras-revertant gene 107</td>
<td>164.2</td>
<td>217.25</td>
<td>386.2</td>
<td>158.36</td>
</tr>
<tr>
<td>L13636 RATRHS Rattus rattus insulin-induced growth-respons protein (HRS) mRNA, complete cds</td>
<td>1079.25</td>
<td>994.15</td>
<td>485.96</td>
<td>658.67</td>
</tr>
<tr>
<td>L20262 RATRSFREG Rattus norvegicus (clone 59) FSH-regulated protein mRNA</td>
<td>384.1</td>
<td>251</td>
<td>265.65</td>
<td>351.4</td>
</tr>
<tr>
<td>M10416 Rat nerve growth factor-induced (NGFI-A) gene, complete cds /gb = (352, 1876) /gb = M10416 /gb = 205333 /gb = Rn.</td>
<td>645.55</td>
<td>502.15</td>
<td>421.3</td>
<td>555.5</td>
</tr>
<tr>
<td>N-myoc downstream-regulated gene 2</td>
<td>458.65</td>
<td>574.55</td>
<td>429.25</td>
<td>452.7</td>
</tr>
<tr>
<td>X51615 RCC16 R rattus RNA for connexin protein Cx6</td>
<td>85.15</td>
<td>17.1</td>
<td>75.06</td>
<td>113.6</td>
</tr>
<tr>
<td>Y00369mRNA RINCMYC Rat c-myc oncogene and flanking regions</td>
<td>513.55</td>
<td>448.5</td>
<td>116.4</td>
<td>619.45</td>
</tr>
<tr>
<td>AF063333 Rattus norvegicus cytochrome P450 2E1 (CYP2E1) mRNA, partial cds</td>
<td>597.05</td>
<td>398.25</td>
<td>200.6</td>
<td>193.8</td>
</tr>
<tr>
<td>D38688 exon RATGCT1B2F Rat DNA for UDP glucuronosyltransferase, exon 1</td>
<td>1293.2</td>
<td>466.7</td>
<td>796.65</td>
<td>1039.6</td>
</tr>
<tr>
<td>E01270 SUTR welcomes Rattus norvegicus SPARc mRNA, untranslated region, partial sequence</td>
<td>251.35</td>
<td>289.65</td>
<td>2413.15</td>
<td>535.26</td>
</tr>
<tr>
<td>E00774cdsa DNA encoding soluble NADPH-cytochrome P450 reductase</td>
<td>219.65</td>
<td>35.35</td>
<td>404.0</td>
<td>1244.7</td>
</tr>
<tr>
<td>E0154cds DNA encoding soluble NADPH-cytochrome P450 reductase</td>
<td>3689.65</td>
<td>3133.4</td>
<td>2555.45</td>
<td>3682.1</td>
</tr>
<tr>
<td>M19335 RATCYF456 Rat cytochrome P450 mRNA, complete cds</td>
<td>1625.65</td>
<td>751.65</td>
<td>913.65</td>
<td>1441.2</td>
</tr>
<tr>
<td>AF068880 Rattus norvegicus beta defensin-1 mRNA, complete cds</td>
<td>271.45</td>
<td>177</td>
<td>242.1</td>
<td>245.1</td>
</tr>
<tr>
<td>Albumin</td>
<td>154.35</td>
<td>69.4</td>
<td>202</td>
<td>170.06</td>
</tr>
<tr>
<td>ApoA-2-macroglobulin</td>
<td>9928.45</td>
<td>3436.4</td>
<td>8835.15</td>
<td>15254.2</td>
</tr>
<tr>
<td>DCL2/adeno virus E1B 19 KDa-interacting protein 3-like</td>
<td>1307.55</td>
<td>1755.55</td>
<td>2534.35</td>
<td>1003.65</td>
</tr>
</tbody>
</table>
Data Organization: Clustering

- Clusters have minimum distance or maximum similarity among their members
Example: Hepatocyte response to IL-6

Jayaraman et al. Tissue Eng., 2005
Distance Measures: Minkowski Metric

Suppose two objects \( x \) and \( y \) both have \( p \) features:

\[
x = (x_1 x_2 \cdots x_p)
\]

\[
y = (y_1 y_2 \cdots y_p)
\]

The Minkowski metric is defined by

\[
d(x, y) = r \left( \sum_{i=1}^{p} |x_i - y_i|^r \right)^{1/r}
\]

Most common, \( r = 2 \) (Euclidean distance)

\[
d(x, y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}
\]
Similarity Measures: Correlation Coefficient

\[ s(x, y) = \frac{\sum_{i=1}^{p} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{p} (x_i - x)^2 \times \sum_{i=1}^{p} (y_i - y)^2}} \]

averages: \( \bar{x} = \frac{1}{p} \sum_{i=1}^{p} x_i \) and \( \bar{y} = \frac{1}{p} \sum_{i=1}^{p} y_i \).

\(|s(x, y)| \leq 1|
Similarity Measures: Correlation Coefficient

Gene A

Gene B

Expression Level

Expression Level

Time

Time

s ≈ -1

s ≈ +1

s ≈ 0
The distance between two clusters is defined as the distance between

- Single-Link Method / Nearest Neighbor
- Complete-Link / Furthest Neighbor
- Their Centroids.
- Average of all cross-cluster pairs.
Single Link Method Using Euclidean Distance

Distance Matrix
k-means clustering
Self-organizing maps
What does clustering give you?

- Fundamental modes or patterns of expression
- Groups of genes that *may* be co-regulated
  - Sense of which biological networks are being utilized
  - Can search for upstream signal
Identifying prevalent expression patterns (gene clusters)

Normalized Expression

Time-point 1

Time-point 2

Time-point 3

Normalized Expression

Time-point

Normalized Expression

Time-point
Evaluate Cluster contents

**Genes**
- gpm1
- HTB1
- RPL11A
- RPL12B
- RPL13A
- RPL14A
- RPL15A
- RPL17A
- RPL23A
- TEF2
- YDL228c
- YDR133C
- YDR134C
- YDR327W
- YDR417C
- YKL153W
- YPL142C

**GO functional category**
- Glycolysis
- Nuclear Organization
- Ribosome
- Translation
- Unknown
Other Applications of Microarrays

• Sequencing
• Measuring mRNA abundance
• Measuring mRNA decay rates
• Identifying transcription factor binding sites
Baker et al.

A. RT/PCR Analysis of Total RNA from Cell Pellets and Monolayer Cultured Hepatocytes

- Annexin II
- β-actin
- CYP2C23
- GST π
- ST1C1
- 18S

Pellet 4h 12h 24h 48h 72h

B. Affymetrix Analysis of Total RNA from Cell Pellets and Monolayer Cultured Hepatocytes

<table>
<thead>
<tr>
<th>Gene</th>
<th>Pellet</th>
<th>4H</th>
<th>12H</th>
<th>24H</th>
<th>48H</th>
<th>72H</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annexin II</td>
<td>1</td>
<td>1.2</td>
<td>8.5</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>β-actin</td>
<td>1</td>
<td>1.4</td>
<td>2.0</td>
<td>2.8</td>
<td>2.2</td>
<td>2.3</td>
</tr>
<tr>
<td>CYP2C23</td>
<td>1</td>
<td>-1.2</td>
<td>-1.1</td>
<td>-2.8</td>
<td>-20</td>
<td>-20</td>
</tr>
<tr>
<td>GST π</td>
<td>1</td>
<td>-1.6</td>
<td>-1.8</td>
<td>-1.3</td>
<td>-2</td>
<td>7</td>
</tr>
<tr>
<td>ST1C1</td>
<td>1</td>
<td>-1.6</td>
<td>-2.8</td>
<td>-20</td>
<td>-20</td>
<td>-20</td>
</tr>
</tbody>
</table>

Data presented as fold change from hepatocyte pellet as determined using Affymetrix Gene Chip software.
Selected Clusters

A.

Relative Gene Expression

<table>
<thead>
<tr>
<th>Hours post plating</th>
<th>Gene ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>M22670</td>
<td>Alpha-2-macroglobulin</td>
</tr>
<tr>
<td>12</td>
<td>D88066</td>
<td>UDP-glucuronosyltransferase</td>
</tr>
<tr>
<td>24</td>
<td>L32132</td>
<td>Lipopoly saccharide binding protein</td>
</tr>
<tr>
<td>48</td>
<td>M23566</td>
<td>Alpha-2-macroglobulin</td>
</tr>
<tr>
<td>72</td>
<td>2869284</td>
<td>Cytokeratinos</td>
</tr>
<tr>
<td>4</td>
<td>XI3083</td>
<td>Alpha-2-macroglobulin</td>
</tr>
<tr>
<td>12</td>
<td>AF044551</td>
<td>PS protein</td>
</tr>
<tr>
<td>24</td>
<td>CD16</td>
<td>Alpha-1-acid glycoprotein</td>
</tr>
<tr>
<td>48</td>
<td>rec_AA915582</td>
<td>Alpha-2-macroglobulin</td>
</tr>
<tr>
<td>72</td>
<td>rec_AA946503</td>
<td>EST</td>
</tr>
<tr>
<td>4</td>
<td>L13089</td>
<td>Annexin II</td>
</tr>
<tr>
<td>12</td>
<td>M126919</td>
<td>Albino A</td>
</tr>
<tr>
<td>24</td>
<td>rec_AA866976</td>
<td>EST</td>
</tr>
<tr>
<td>48</td>
<td>X06901</td>
<td>Vascular alpha actin</td>
</tr>
</tbody>
</table>

B.

Relative Gene Expression

<table>
<thead>
<tr>
<th>Hours post plating</th>
<th>Gene ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>M87588</td>
<td>Estrogen sulfotransferase</td>
</tr>
<tr>
<td>12</td>
<td>V01235</td>
<td>Liver fatty acid binding protein</td>
</tr>
<tr>
<td>24</td>
<td>M201314</td>
<td>CYP2E1</td>
</tr>
<tr>
<td>48</td>
<td>AF0370712</td>
<td>Carboxic anhydrase III</td>
</tr>
<tr>
<td>72</td>
<td>N91485</td>
<td>Dual specificity phosphatase Mkp-3</td>
</tr>
<tr>
<td>4</td>
<td>D31662</td>
<td>Kidney regucalcin</td>
</tr>
<tr>
<td>12</td>
<td>M31363</td>
<td>Hydroxysteroid sulfotransferase</td>
</tr>
<tr>
<td>24</td>
<td>M18335</td>
<td>P450</td>
</tr>
<tr>
<td>48</td>
<td>rec_AA817987</td>
<td>Hydroxysteroid sulfotransferase</td>
</tr>
<tr>
<td>72</td>
<td>L25939</td>
<td>N-hydroxy-2-acetylamino-thioene (STC1)</td>
</tr>
<tr>
<td>4</td>
<td>M30282</td>
<td>Plasma kalrein</td>
</tr>
<tr>
<td>12</td>
<td>S76489</td>
<td>Estrogen sulfotransferase</td>
</tr>
<tr>
<td>24</td>
<td>rec_AA803325</td>
<td>EST</td>
</tr>
<tr>
<td>48</td>
<td>J0728</td>
<td>P450e</td>
</tr>
<tr>
<td>72</td>
<td>rec_AA818122</td>
<td>Hydroxysteroid sulfotransferase</td>
</tr>
<tr>
<td>4</td>
<td>K07721</td>
<td>P450(1)</td>
</tr>
<tr>
<td>12</td>
<td>X79181</td>
<td>CYP2C11</td>
</tr>
<tr>
<td>24</td>
<td>S46785</td>
<td>Insulin-like growth factor binding protein</td>
</tr>
<tr>
<td>48</td>
<td>M14775</td>
<td>P450</td>
</tr>
<tr>
<td>72</td>
<td>K03243</td>
<td>Phosphoenolpyruvate carboxykinase (GTP)</td>
</tr>
<tr>
<td>4</td>
<td>M13234</td>
<td>P450e</td>
</tr>
</tbody>
</table>
CYT P450 expression
Up-regulation of glycolysis