Introduction to microarray technology and data analysis

Aron Eklund

eklund@cbs.dtu.dk

Cancer Systems Biology group
Center for Biological Sequence Analysis
DTU Systems Biology

Introduction to Systems Biology
February 16, 2015
What is the role of the microarray in systems biology?

• The gene expression microarray was the first tool to **efficiently** and **quantitatively** characterize the **global** state of a biological system.

• Being replaced by sequencing (RNA-seq)... but many of the same principles apply.
Learning objectives

1. Describe a *gene expression microarray* and what it measures

2. Explain the steps needed to generate gene expression microarray data

4. Describe at least one application of a gene expression microarray

5. Use R at a basic level

6. Explain and calculate a $\log_2$ ratio

7. Explain and calculate a $P$ value using the $t$ test
“Expression” of a gene

Gene expression = how much of a certain mRNA is present

Protein expression = how much of a certain protein is present
Why do we want to measure gene expression?

1. Biological / medical research
   - inference of the function of a gene
   - understanding gene regulation
   - biomarker discovery

2. Clinical diagnostics
   - predictive biomarkers for cancer
   - identifying cancers of unknown primary
Why do we want to measure gene expression (vs. protein expression)?

Because gene expression is cheaper / faster / easier

- Gene expression microarrays are based on sequence-specific hybridization to complementary sequences
  There is no such thing as a “complementary amino acid sequence”

- Gene expression microarrays can be fabricated in parallel
  Protein expression arrays depend on antibodies that must be individually created.
The DNA microarray (MICROscopio Array)

Thousands to millions of DNA “spots” in a few square centimeters

a. Cartridge
b. Glass slide

In situ photolithographic synthesis

Robotic spotting or in situ synthesis
The DNA microarray

Millions of single-stranded DNA “probes”, each with the same sequence:
TCAGCTAGCTAGTCGATGCTAGTCGACGGG

Each “spot” or “feature” is designed to detect

Millions of single-stranded DNA “probes” with the same sequence:
CCGATGCTAGTTCAGCTAGCTAGCGACGATA
Example microarrays

Affymetrix

Illumina (24-plex)
The gene expression microarray

Labeled target
RNA made from a biological sample

DNA probe
attached to the microarray substrate

Specific hybridization
between a complementary probe and target
detected by fluorescence
Gene expression profile

The results from a single sample run on a microarray:

<table>
<thead>
<tr>
<th>Identifier</th>
<th>Expression Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DDR1</td>
<td>10.404655</td>
</tr>
<tr>
<td>RFC2</td>
<td>6.250778</td>
</tr>
<tr>
<td>HSPA6</td>
<td>6.221254</td>
</tr>
<tr>
<td>PAX8</td>
<td>7.839859</td>
</tr>
<tr>
<td>GUCA1A</td>
<td>4.019081</td>
</tr>
<tr>
<td>UBA7</td>
<td>7.746138</td>
</tr>
<tr>
<td>THRA</td>
<td>5.535657</td>
</tr>
<tr>
<td>PTPN21</td>
<td>4.403469</td>
</tr>
<tr>
<td>CCL5</td>
<td>8.516195</td>
</tr>
<tr>
<td>CYP2E1</td>
<td>3.805621</td>
</tr>
<tr>
<td>EPHB3</td>
<td>6.256173</td>
</tr>
<tr>
<td>ESRRA</td>
<td>7.745937</td>
</tr>
<tr>
<td>CYP2A6</td>
<td>5.948766</td>
</tr>
<tr>
<td>GAS6</td>
<td>9.765667</td>
</tr>
<tr>
<td>MMP14</td>
<td>7.721172</td>
</tr>
<tr>
<td>TRADD</td>
<td>7.978667</td>
</tr>
<tr>
<td>FNTB</td>
<td>4.775533</td>
</tr>
<tr>
<td>PLD1</td>
<td>5.559748</td>
</tr>
</tbody>
</table>

An expression value for each “feature”

Typically 20,000 – 50,000 genes
Gene expression matrix

<table>
<thead>
<tr>
<th>Genes</th>
<th>Sample 01</th>
<th>Sample 02</th>
<th>Sample 03</th>
<th>Sample 04</th>
<th>Sample 05</th>
</tr>
</thead>
<tbody>
<tr>
<td>DDR1</td>
<td>10.404655</td>
<td>10.134845</td>
<td>10.089483</td>
<td>10.317264</td>
<td>10.754314</td>
</tr>
<tr>
<td>PAX8</td>
<td>7.839859</td>
<td>7.931699</td>
<td>7.846006</td>
<td>7.966344</td>
<td>7.821996</td>
</tr>
<tr>
<td>GUCA1A</td>
<td>4.019081</td>
<td>3.960418</td>
<td>4.119253</td>
<td>4.143439</td>
<td>4.022938</td>
</tr>
<tr>
<td>UBA7</td>
<td>7.746138</td>
<td>8.120003</td>
<td>7.595514</td>
<td>7.892662</td>
<td>8.025442</td>
</tr>
<tr>
<td>THRA</td>
<td>5.535657</td>
<td>5.262903</td>
<td>5.613080</td>
<td>5.284967</td>
<td>5.581082</td>
</tr>
<tr>
<td>PTPN21</td>
<td>4.403469</td>
<td>4.285923</td>
<td>5.090882</td>
<td>4.606929</td>
<td>4.657228</td>
</tr>
<tr>
<td>CCL5</td>
<td>8.516195</td>
<td>9.008406</td>
<td>5.616732</td>
<td>6.603475</td>
<td>5.941453</td>
</tr>
<tr>
<td>CYP2E1</td>
<td>3.805621</td>
<td>3.830757</td>
<td>3.923257</td>
<td>4.057170</td>
<td>3.917387</td>
</tr>
<tr>
<td>ESRRA</td>
<td>7.745937</td>
<td>7.778842</td>
<td>7.836441</td>
<td>7.678031</td>
<td>7.753971</td>
</tr>
<tr>
<td>CYP2A6</td>
<td>5.948766</td>
<td>8.556330</td>
<td>7.456848</td>
<td>6.466415</td>
<td>9.677679</td>
</tr>
<tr>
<td>MMP14</td>
<td>7.721172</td>
<td>7.846023</td>
<td>7.954670</td>
<td>8.010790</td>
<td>8.260653</td>
</tr>
<tr>
<td>TRADD</td>
<td>7.978667</td>
<td>7.985578</td>
<td>7.581977</td>
<td>7.648343</td>
<td>8.106099</td>
</tr>
<tr>
<td>FNTB</td>
<td>4.775533</td>
<td>4.913803</td>
<td>5.138638</td>
<td>5.264397</td>
<td>5.089515</td>
</tr>
<tr>
<td>PLD1</td>
<td>5.559748</td>
<td>11.580453</td>
<td>5.525487</td>
<td>5.579042</td>
<td>5.594873</td>
</tr>
</tbody>
</table>

which values can be compared here?
The process

Biological specimen

Expression profile
The process

1. Isolate RNA from a biological specimen
2. Create “labeled target” from the RNA
3. Let the labeled target hybridize (incubate) on the microarray
4. Wash off the unbound target
5. Scan the microarray
6. Analyze the data
Two-color labeled cDNA

(mostly outdated “simple” approach!)
The Eberwine protocol:
for “one-color / single-channel” microarrays

42 °C  
2 h  
+ Reverse Transcriptase

16 °C  
2 h  
+ RNase H  
+ DNA Polymerase

37 °C  
6 h  
+ RNA Polymerase  
+ Biotin-labeled nucleotides

mRNA/cDNA hybrid

double-stranded DNA (dsDNA)

amplified antisense RNA
Affymetrix fluidics station and scanner

Hybridize, incubate, rinse

Scan
Image processing and normalization

Normalization adjusts data to correct for systematic bias in intensity levels (e.g. due to different amounts of material)
<table>
<thead>
<tr>
<th>Genes</th>
<th>Sample 01</th>
<th>Sample 02</th>
<th>Sample 03</th>
<th>Sample 04</th>
<th>Sample 05</th>
</tr>
</thead>
<tbody>
<tr>
<td>DDR1</td>
<td>10.404655</td>
<td>10.134845</td>
<td>10.089483</td>
<td>10.317264</td>
<td>10.754314</td>
</tr>
<tr>
<td>PAX8</td>
<td>7.839859</td>
<td>7.931699</td>
<td>7.846006</td>
<td>7.966344</td>
<td>7.821996</td>
</tr>
<tr>
<td>GUCA1A</td>
<td>4.019081</td>
<td>3.960418</td>
<td>4.119253</td>
<td>4.143439</td>
<td>4.022938</td>
</tr>
<tr>
<td>UBA7</td>
<td>7.746138</td>
<td>8.120003</td>
<td>7.595514</td>
<td>7.892662</td>
<td>8.025442</td>
</tr>
<tr>
<td>THRA</td>
<td>5.535657</td>
<td>5.262903</td>
<td>5.613080</td>
<td>5.284967</td>
<td>5.581082</td>
</tr>
<tr>
<td>PTPN21</td>
<td>4.403469</td>
<td>4.285923</td>
<td>5.090882</td>
<td>4.606929</td>
<td>4.657228</td>
</tr>
<tr>
<td>CCL5</td>
<td>8.516195</td>
<td>9.008406</td>
<td>5.616732</td>
<td>6.603475</td>
<td>5.941453</td>
</tr>
<tr>
<td>CYP2E1</td>
<td>3.805621</td>
<td>3.830757</td>
<td>3.923257</td>
<td>4.057170</td>
<td>3.917387</td>
</tr>
<tr>
<td>ESRRA</td>
<td>7.745937</td>
<td>7.778842</td>
<td>7.836441</td>
<td>7.678031</td>
<td>7.753971</td>
</tr>
<tr>
<td>CYP2A6</td>
<td>5.948766</td>
<td>8.556330</td>
<td>7.456848</td>
<td>6.466415</td>
<td>9.677679</td>
</tr>
<tr>
<td>MMP14</td>
<td>7.721172</td>
<td>7.846023</td>
<td>7.954670</td>
<td>8.010790</td>
<td>8.260653</td>
</tr>
<tr>
<td>TRADD</td>
<td>7.978667</td>
<td>7.985578</td>
<td>7.581977</td>
<td>7.648343</td>
<td>8.106099</td>
</tr>
<tr>
<td>FNTB</td>
<td>4.775533</td>
<td>4.913803</td>
<td>5.138638</td>
<td>5.264397</td>
<td>5.089515</td>
</tr>
<tr>
<td>PLD1</td>
<td>5.559748</td>
<td>5.580453</td>
<td>5.525487</td>
<td>5.579042</td>
<td>5.594873</td>
</tr>
</tbody>
</table>
What is the question?

Typically, we use microarrays to compare between groups:

- Disease vs. normal
- Drug treatment vs. control
- Disease A vs. Disease B
- Good prognosis vs. bad prognosis
Comparing a gene between two groups

MBNL1

- fold change
- log$_2$ ratio
- $P$ value (Student’s $t$ test)

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$
Visualization of microarray data

Samples

Genes

“clustering”
Microarray data: what do the numbers mean

1. **No calibration**: Expression values cannot be readily translated into concentration

2. **Response curves are not parallel**: Comparison between one gene and another gene is almost meaningless

3. **Noisy at low levels** (on the log scale): Fold changes can be misleading.

4. **Saturation at high levels**: Cannot detect changes
Evolution of gene expression (mRNA) measurements

1. 1970s - Northern blot
   - one gene at a time, not very accurate

2. 1980s - Quantitative reverse transcriptase polymerase chain reaction (RT-PCR)
   - one gene at a time, very accurate

3. mid-1990s - DNA microarray
   - many genes at a time, medium accuracy

4. Late 2000s - RNA-seq
   - all genes, higher accuracy (?)
Other applications/types of microarrays

DNA microarray
  - **Gene expression**
  - Genotyping (SNP profile)
  - DNA copy number profiling (CGH, copy number)
  - ChIP-Chip

Protein microarrays

Tissue microarrays

Cellular microarrays

Compound microarrays
About the exercises
The data set in the exercise

**Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease.**


31 samples:
- 14 normal
- 5 presymptomatic
- 12 symptomatic

Huntington’s disease:
- neurological disorder
- polyglutamine expansion in the huntingtin gene

**Affymetrix HG-U133A arrays**

Why search for marker of disease *progression* (not diagnosis)?
- assess treatment efficacy
- surrogate endpoint in drug trials
Lunch break now

Exercises at 13:00