Transcriptomics

June 9, 2008

Systems biology and the omics cascade

Peter Nilsson
nipe@kth.se

Dept. of Proteomics
and
Dept. of Gene Technology

School of Biotechnology
KTH – Royal Institute of Technology
Albanova University Center
Transcriptome

"The complete set of transcripts and their relative levels of expression in a particular cell or tissue type under defined conditions."

"The full complement of all activated genes, mRNAs, or transcripts in a particular cell at a particular time"

"A transcriptome is a collection of all the gene transcripts present in a given cell."

mRNA abundance

**High:**
<100 different transcripts
100-1000/cell
~20% of the transcriptome

**Intermediate:**
several hundred different transcripts
10-100/cell
~30% of the transcriptome

**Low**
>10,000 different transcripts
0.1-10/cell
~50% of the transcriptome
Noncoding RNAs
- Recently very much focus, interesting research and new findings

2001

NON-CODING RNA GENES AND THE MODERN RNA WORLD

Sean R. Eddy
Non-coding RNA (ncRNA) genes produce functional RNA molecules rather than encoding proteins. However, almost all means of gene identification assume that genes encode proteins, so even in the era of complete genome sequences, ncRNA genes have been effectively invisible. Recently, several different systematic screens have identified a surprisingly large number of new ncRNA genes. Non-coding RNAs seem to be particularly abundant in roles that require highly specific nucleic acid recognition without complex catalysis, such as in directing post-transcriptional regulation of gene expression or in guiding RNA modifications.

Box 1 | Abbreviations for different classes of non-coding RNA
- rRNA
  Functional RNA — essentially synonymous with non-coding RNA
- miRNA
  MicroRNA — putative translational regulatory gene family
- ncRNA
  Non-coding RNA — all RNAs other than mRNA
- rRNA
  Ribosomal RNA
- siRNA
  Small interfering RNA — active molecules in RNA interference
- snRNA
  Small nuclear RNA — includes spliceosomal RNAs
- snmRNA
  Small non-mRNA — essentially synonymous with small ncRNAs
- snoRNA
  Small nucleolar RNA — most known snoRNAs are involved in rRNA modification
- srRNA
  Small temporal RNA — for example, lin-4 and let-7 in Caenorhabditis elegans
- tRNA
  Transfer RNA
cDNA synthesis

Labelled cDNA

1. Total or poly(A) RNA
2. Synthesize cDNA, incorporating amino-allyl/aminohexyl nucleotides
3. Hydrolyze RNA, neutralize, and purify with low-elution volume purification columns
4. Add monofunctional NHS-reactive fluorescent dyes
5. Purify with low-elution volume purification columns and hybridize to microarray
Disease/tissue/sample specific gene expression

Identify relative differences in expression of gene products

Analyze the relative amount of genes that are transcribed at a certain timepoint or condition
Comparison of gene activity
What is an array?

Array = impressive ordered collection
Microarray = “Mikromatris” in Swedish
>31,000 microarray articles in PubMed

(June 9)

Annual increase
Reports

Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray

Mark Schena (1), Dari Shalon (1), Ronald W. Davis (2), Patrick O. Brown (3)

A high-capacity system was developed to monitor the expression of many genes in parallel. Microarrays prepared by high-speed robotic printing of complementary DNAs on glass were used for quantitative expression measurements of the corresponding genes. Because of the small format and high density of the arrays, hybridization volumes of 2 microliters could be used that enabled detection of rare transcripts in probe mixtures derived from 2 micrograms of total cellular messenger RNA. Differential expression measurements of 45 Arabidopsis genes were made by means of simultaneous, two-color fluorescence hybridization.

M. Schena and R. W. Davis, Department of Biochemistry, Beckman Center, Stanford University Medical Center, Stanford, CA 94305, USA.
D. Shalon and P. O. Brown, Department of Biochemistry and Howard Hughes Medical Institute, Beckman Center, Stanford University Medical Center, Stanford, CA 94305, USA.
(1) These authors contributed equally to this work.
(2) Present address: Synteni, Palo Alto, CA 94303, USA.
(3) To whom correspondence should be addressed. E-mail: pbrown@cmgm.stanford.edu
Microarray publications from 1996

Schena M.
Genome analysis with gene expression microarrays.

Shalon D, Smith SJ, Brown PO.
A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization.

Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW.
Parallel human genome analysis: microarray-based expression monitoring of 1000 genes.

DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM.
Use of a cDNA microarray to analyse gene expression patterns in human cancer.

Nelson N.
Microarrays pave the way to 21st century medicine.
Fragment-based DNA printing
Array of PCR products

Purified PCR products are spotted by an arrayer, on a glass slide, with a positively charged coating.

The DNA is snap-dried and UV-crosslinked to the surface.


Schena (1995), Science, 270: 467
Methods for creating a DNA array

I) In situ (on-chip) synthesis of oligonucleotides
   Photolithography (Affymetrix)
       Light-directed oligonucleotide synthesis on chip.
       Adapted from semiconductor industry
   Inkjet technology (Agilent)
       Adapted from the technique used in ink-jet printers.
       Piezoelectric effect: a charge on a narrow tube containing nucleotides, forces out a small drop onto a coated glass slide.
       The oligonucleotides are synthesized drop-by-drop.

II) Spotting of long DNA fragments
   Fragment based cDNA printing (Stanford University)
       Array of spotted PCR products.
       For gene expression analysis.
   Spotting of prefabricated oligonucleotides
       Usually 50-70 nt longmers
   Contact or non-contact printing
Photolithographic *in situ* synthesis of high density oligonucleotide arrays

\[ 4 \times n \text{ cycles} \]
\[ n = \text{length of oligonucleotides, } n \leq 20 \]
Affymetrix GeneChip
Capacity of >400,000 oligos / cm²
Spotted probes on chip

- Buy clones
- Create clone libraries
- Plasmid preparation
- Sequence
- Create unigene collection

**Oligonucleotides**
(50-70 nt)
(design)
buy

**PCR**

- Purify
- Cloned cDNA library
- Genomic DNA
- Gene database

PCR with vector-specific primers
PCR with gene-specific primers
Oligonucleotide synthesis
Example of Oligo Sets for Microarray Applications

Mammalian
Homo sapiens (human)
Mus musculus (mouse)
Rattus norvegicus (rat)
Bos taurus (bovine)
Canis familiaris (dog)
Sus scrofa (pig)

Bacterial
Bacillus Genus (anthrax)
Bordetella pertussis
Campylobacter coli
Campylobacter jejuni
Chlamydia pneumoniae
Escherichia coli
Haemophilus influenzae
Helicobacter pylori
Lactobacillus sakei
Listeria monocytogenes
Mycobacterium tuberculosis
Rhizobium leguminosarum
Salmonella Genus
Sinorhizobium meliloti
Streptococcus mitis
Streptococcus pneumoniae

Animal
Caenorhabditis elegans (nematode)
Danio rerio (zebrafish)
Drosophila melanogaster (fruit fly)
Gallus gallus (chicken)
Xenopus tropicalis (Pepad frog)

Yeast and Fungi
Candida albicans
Magnaporthe grisea
Saccharomyces cerevisiae

Plant
Arabidopsis thaliana
Lycopersicon esculentum (tomato)
Zea mays (maize)
Medicago truncatula (barrel medic)
Oryza sativa (rice)
Prunus persica (peach)
Pisum sativum (garden pea)
Vitis vinifera (grape)

Additional Genomes
Plasmodium falciparum (malaria)
Rhodopirellula baltica (pirellula)
Different microarray printers

~400 pl is deposited
Robotics for microarray production

GeSIM Nanoplotter 2, 8-pins, non-contact

BioRobotics MicroGrid II, 64-pins

Genetix Qarray2, 48-pins

Genetix Qarray, 24-pins

GMS 417, 4-pins

Robotic workstations

Scanner
# DNA microarray production for internal projects and academic collaborations

<table>
<thead>
<tr>
<th>Organism</th>
<th>Type</th>
<th>Array Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>cDNA, oligo, genomic (CGH)</td>
<td>46 k, 34 k, 11 k</td>
</tr>
<tr>
<td>Poplar</td>
<td>cDNA</td>
<td>25 k</td>
</tr>
<tr>
<td>Mouse</td>
<td>cDNA, (stem cells) oligo</td>
<td>20 k, 35 k</td>
</tr>
<tr>
<td>Rat</td>
<td>oligo</td>
<td>27 k</td>
</tr>
<tr>
<td>Chicken</td>
<td>cDNA</td>
<td>14 k</td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>gene spec. PCR, gene spec. PCR (promoters) oligo</td>
<td>22 k, 10 k, 29 k</td>
</tr>
<tr>
<td>Drosophila</td>
<td>oligo</td>
<td>15 k</td>
</tr>
<tr>
<td>C. elegans</td>
<td>oligo</td>
<td>20 k</td>
</tr>
<tr>
<td>Yeast</td>
<td>oligo</td>
<td>6,4 k</td>
</tr>
<tr>
<td>E. coli</td>
<td>oligo</td>
<td>6,0 k</td>
</tr>
<tr>
<td>Plasmodium falciparum</td>
<td>oligo</td>
<td>7,1 k</td>
</tr>
<tr>
<td>Sulfolobus sulfataricus</td>
<td>gene spec. PCR</td>
<td>3,0 k</td>
</tr>
<tr>
<td>Sulfolobus acidocaldarius</td>
<td>gene spec. PCR</td>
<td>2,8 k</td>
</tr>
<tr>
<td>Helicobacter pylori</td>
<td>oligo</td>
<td>1,9 k</td>
</tr>
<tr>
<td>Campylobacter jejuni</td>
<td>oligo</td>
<td>1,6 k</td>
</tr>
<tr>
<td>Streptococcus pneumoniae</td>
<td>oligo</td>
<td>2,0 k</td>
</tr>
<tr>
<td>Bartonella henselae</td>
<td>gene spec. PCR</td>
<td>1,9 k</td>
</tr>
<tr>
<td>Bartonella grahamii</td>
<td>oligo</td>
<td>4,4 k</td>
</tr>
</tbody>
</table>

[www.ktharray.se](http://www.ktharray.se)
cDNA microarray technology

**Probe**
- DNA probes (gene fragments)
- Robotic printing

**Target**
- Sample mRNA
- Reference mRNA
- Labelled cDNA
- Hybridize to microarray

**Laser 1**
- “red” raw data

**Laser 2**
- “green” raw data

**Image analysis**

**Microarray / DNA chip**
Image analysis

Before auto align

After auto align

Calculate ratios

Background subtraction
Fold change vs log$_2$ ratios

### Fold change

**sample intensity/ref intensity**

<table>
<thead>
<tr>
<th>Name</th>
<th>0</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>1</td>
<td>8</td>
<td>12</td>
<td>16</td>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>4</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>F</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.25</td>
<td>0.25</td>
<td>0.1</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>H</td>
<td>1</td>
<td>0.5</td>
<td>0.33</td>
<td>0.25</td>
<td>0.33</td>
<td>0.5</td>
</tr>
<tr>
<td>I</td>
<td>1</td>
<td>4</td>
<td>8</td>
<td>4</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>J</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>K</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>L</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>M</td>
<td>1</td>
<td>0.33</td>
<td>0.25</td>
<td>0.25</td>
<td>0.33</td>
<td>0.5</td>
</tr>
<tr>
<td>N</td>
<td>1</td>
<td>0.125</td>
<td>0.0625</td>
<td>0.0625</td>
<td>0.0833</td>
<td>0.125</td>
</tr>
</tbody>
</table>

### log$_2$ ratios

**log$_2$ (sample intensity/ref intensity)**

<table>
<thead>
<tr>
<th>Name</th>
<th>0</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0</td>
<td>3</td>
<td>3.58</td>
<td>4</td>
<td>3.58</td>
<td>3</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>1.58</td>
<td>2</td>
<td>2</td>
<td>1.58</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-2</td>
<td>-2</td>
<td>-3.32</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>1</td>
<td>1.58</td>
<td>2</td>
<td>1.58</td>
<td>1</td>
</tr>
<tr>
<td>H</td>
<td>0</td>
<td>-1</td>
<td>-1.60</td>
<td>-2</td>
<td>-1.60</td>
<td>-1</td>
</tr>
<tr>
<td>I</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>J</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>K</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.58</td>
<td>1.58</td>
</tr>
<tr>
<td>L</td>
<td>0</td>
<td>1</td>
<td>1.58</td>
<td>2</td>
<td>1.58</td>
<td>1</td>
</tr>
<tr>
<td>M</td>
<td>0</td>
<td>-1.60</td>
<td>-2</td>
<td>-2</td>
<td>-1.60</td>
<td>-1</td>
</tr>
<tr>
<td>N</td>
<td>0</td>
<td>-3</td>
<td>-3.59</td>
<td>-4</td>
<td>-3.59</td>
<td>-3</td>
</tr>
</tbody>
</table>

**Graphs**

- **Left graph**: Fold change vs HOURS.
- **Right graph**: log$_2$ ratios vs HOURS.
Normalization

X: intensities
Y: ratio

The objective of normalization is to scale the spot intensity measurement so that the ratio of intensities is approximately equal to the ratios of gene expression.
No normalization
Experimental design
How many replicates are needed and how should they be used?

Different levels of replicates

- Individual animals/cell-lines/etc
- Repeated sampling of the same animal
- Repeated RNA extraction
- Repeated amplification of the mRNA
- Repeated labelling & hybridisation
- Repeated measurement of each gene on each slide (duplicate spots)

Controls...
...biological variation
...technical variation
Analysis of variance (ANOVA) for gene expression data
Significance testing
Don’t look only on fold change, take also variance into account
**Microarray data mining**

**How to visualize data**

- **Hierarchical clustering**
  - First, find the two genes, with the most similar behaviour.
  - Join these together into a cluster. Next, join the next two most similar objects (gene or cluster), forming a new cluster.
  - Continue until there is only one object left, a single cluster containing all genes.
Clustering; grouping of similar expression patterns

Profile plots
Guilt by association

<table>
<thead>
<tr>
<th>Ph</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

mean

174 genes
98 ND
17 tubulin (out of 25)

42 genes
16 ND
12 tubulin

Downregulated genes
Upregulated genes
Equally expressed genes
Expression Profiling of Adrenocortical Tumors suggests a Molecular Signature of Malignancy
Differential Gene Expression in Wood-formation
Expression profiling in metabolic pathways over the tissue sections

UDP-Glucose Pathway

Hertzberg et al. (2001) Proc Natl Acad Sci USA
DE genes based on foldchange and B top

<table>
<thead>
<tr>
<th>GeneOntology</th>
<th>Percentage</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>defense response</td>
<td>9.0%</td>
<td>13</td>
</tr>
<tr>
<td>response to pest/pathogen/parasite</td>
<td>6.9%</td>
<td>10</td>
</tr>
<tr>
<td>regulation of transcription</td>
<td>5.6%</td>
<td>8</td>
</tr>
<tr>
<td>transcription, DNA-dependent</td>
<td>5.6%</td>
<td>8</td>
</tr>
<tr>
<td>intracellular transport</td>
<td>4.9%</td>
<td>7</td>
</tr>
<tr>
<td>phosphorylation</td>
<td>4.9%</td>
<td>7</td>
</tr>
<tr>
<td>protein biosynthesis</td>
<td>4.9%</td>
<td>7</td>
</tr>
<tr>
<td>protein amino acid phosphorylation</td>
<td>4.2%</td>
<td>6</td>
</tr>
<tr>
<td>cytoplasm organization and biogenesis</td>
<td>3.5%</td>
<td>5</td>
</tr>
<tr>
<td>protein transport</td>
<td>3.5%</td>
<td>5</td>
</tr>
<tr>
<td>cellular defense response</td>
<td>2.8%</td>
<td>4</td>
</tr>
<tr>
<td>ion transport</td>
<td>2.8%</td>
<td>4</td>
</tr>
<tr>
<td>nuclear organization and biogenesis</td>
<td>2.8%</td>
<td>4</td>
</tr>
<tr>
<td>protein kinase cascade</td>
<td>2.8%</td>
<td>4</td>
</tr>
<tr>
<td>DNA packaging</td>
<td>2.1%</td>
<td>3</td>
</tr>
<tr>
<td>blood coagulation</td>
<td>2.1%</td>
<td>3</td>
</tr>
<tr>
<td>small GTPase mediated signal transduction</td>
<td>2.1%</td>
<td>3</td>
</tr>
</tbody>
</table>

Get Unclassifieds 63.9% (92)
Biology

Cancer

Drug

Inflammation

Molecular signature of cancer

Drug effects

Molecular signature of inflammatory sites

Other sources of information
....
(metabolomic, proteomics data)
Public databases for microarray data

- **USA**
- **Europe**

**ArrayExpress at the EBI**

ArrayExpress is a public repository for microarray data, which is aimed at storing well annotated data in accordance with MGED recommendations.

- **Query Database**
- **Login To Database**
- **NEW: FTP access**
- **Try gene queries in prototype data warehouse**
- **Submissions**
- **Help & Documentation**
  - FAQ
  - Statistics

**Current Content Overview**

- Experiments: 896
- Arrays: 381
- Protocols: 4674
- Hybridizations: 26550

**CIBEX**

CIBEX is a gene expression database system in compliance with MIAME, which is a standard that the MGED Society has developed for reporting and data produced in microarray experiments at different laboratories worldwide.
Open source data analysis software

The R Project for Statistical Computing
www.r-project.org

BioConductor is an open-source and open-development software project
www.bioconductor.org

KTH-package for microarray data analysis
www.ktharray.se

TIGR Microarray Data Analysis System (MIDAS) is a microarray data quality filtering and normalization tool that allows raw experimental data to be processed through various data normalizations, filters, and transformations via a user-designed analysis pipeline. Currently implemented normalization and data analysis algorithms include intensity normalization, Lowess (Loess) normalization, fold-change consistency checking, replicates analysis, intensity-dependent z-score filtering (slice analysis), etc. MIDAS is implemented in Java language and thus a platform-independent application. It requires JDK v1.3 or higher. Refer to the included manual for details.

MADAM
Microarray experiments produce large amounts of data for even the simplest of experiments. In order to analyze data from many experiments that data must be stored in an accessible form, such as in a database. MADAM (MicroArray Data Manager) is a java-based application designed to load and retrieve microarray data to and from a database (also supplied with the software). MADAM provides data entry forms, data report forms and additional applications necessary to maintain microarray data for further analysis. Madam requires JRE 1.3.1.

TIGR MultiExperiment Viewer (MEV) is a Java application designed to allow the analysis of microarray data to identify patterns of gene expression and differentially expressed genes. Numerous normalization, clustering and distance algorithms have been implemented, along with a variety of graphical displays to best present the results. MEV was written to be flexible and expandable, and support a variety of input and output formats. MEV requires version 1.2 or higher of Sun's JRE and JSDK package.

TIGR Spottinder is a software tool designed for Microarray image processing using the TIFF image files generated by most microarray scanners. TIGR Spottinder was written in C++ for PCs running Windows NT/2000/ME/XP.
Microarray-based tool box

DNA

RNA

Protein

Carbohydrates

Genome

Transcriptome

Proteome

Antibody array

Genotyping array

cDNA/oligo array

Protein array

Serum array
Principle of SAGE
Serial Analysis of Gene Expression

• convert mRNA to cDNA
• Fragmentize

• capture mRNA
• add linker
• remove from bead

• ligate two tags (15+15 bp)
• amplify with PC

• ligate many di-tags (1000 bp)
• clone into a plasmid

• Sequence >10,000 tags
• identify and count the tags
Microbead-based expression profiling
**BeadArray™ Technology Overview**

**Overview**
Illumina's BeadArray Technology is based on 3-micron silica beads that self assemble in microwells on either of two substrates: fiber optic bundles or planar silica slides. When randomly assembled on one of these two substrates, the beads have a uniform spacing of ~5.7 microns. Each bead is covered with hundreds of thousands of copies of a specific oligonucleotide that act as the capture sequences in one of Illumina's assays (Figure 1).

**Direct Hybridization Assay**

**Direct Hybridization Assay Overview**
Biotin Labeled cRNA

Address Probe 50 bases

Decoder Oligo 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

Decoder hyb 1

Decoder hyb 2

Decoder Hybridization 1 Decoder Hybridization 2
qPCR
quantitative reverse-transcription PCR (Q-RT-PCR)
real time PCR (RT-PCR)

Often used to verify microarray data, gene-by gene
Summary

Transcriptome
Microarrays
  analysis of whole coding genome
  fabrication and use
  spotted vs in-situ synthesized
  transcript profiling
Microarray data analysis
  experimental design
  image analysis
  normalization
  significance test of differentially expressed genes
  clustering and visualization
  data mining
  open source software
Other methods
  Bead-based arrays
  SAGE
  qPCR