Biochemistry 412

DNA Microarrays & RNAi

April 7, 2009
DNA Microarrays
The development of DNA microarrays led to an explosion in mRNA profiling studies.

The literature on gene expression profiling grew exponentially over the past seven years (circles), as evident from the perfect doubling per year represented by the dashed line (data taken from [1]).

Two Main Types of DNA Microarrays

• Oligonucleotide microarrays. One type is manufactured by Affymetrix and is known as a Gene Chip™. It contains hundreds of thousands of ordered, single-stranded synthetic oligonucleotides that are typically 25 bases in length, which have been synthesized \textit{in situ} on a solid support. Each gene is generally probed by several (>15) oligonucleotides. DNA and RNA samples are labelled and fragmented before being hybridized to the array. Quantitative estimates of the transcript abundance can be obtained directly by averaging the signal from all the probes that belong to one gene.

• Spotted DNA microarrays. These usually contain ordered, double-stranded DNA molecules that were created by PCR. They correspond to either genomic or cDNA sequences that have been deposited onto glass slides. Usually there is one probe per gene. For technical reasons, the information that is obtained from these microarrays generally gives the relative concentration (ratio) of a given transcript when two conditions are compared. Samples of mRNA from two experiments are labelled with different dyes, pooled and hybridized to the microarray by competitive hybridization.

Oligonucleotide microarray

Spotted DNA microarray

Labelled cDNA

Labelled RNA

Labelled cRNA

How labeling is done to make probes for use with oligonucleotide microarrays

Fluorescence detection is made possible by “staining” the hybridized array with an avidin-fluorophore conjugate, which binds to the biotin moieties on the hybridized nucleic acids.

Affymetrix Gene Chips - *In Situ* Synthesis

![Diagram](image)

**Fig. 1.** Light-directed synthesis of oligonucleotides. A surface bearing photoprotected hydroxyls (X-O) is illuminated through a photolithographic mask ($M_1$), generating free hydroxyl groups in the photodeprotected regions. The hydroxyl groups are then coupled to a deoxynucleoside phosphoramidite ($5'$-photoprotected). A new mask pattern ($M_2$) is applied, and a second photoprotected phosphoramidite is coupled. Rounds of illumination and coupling are repeated until the desired set of products is obtained.

Note: 4N masks required to make an array of oligonucleotides, each of length N.

Note: this is the photolabile blocking group, “X”, indicated schematically in Figure 1.

Scheme I

Key feature: known oligo sequence at each "address" on the chip.

**Fig. 1 a,** Light directed oligonucleotide synthesis. A solid support is derivatized with a covalent linker molecule terminated with a photolabile protecting group. Light is directed through a mask to deprotect and activate selected sites, and protected nucleotides couple to the activated sites. The process is repeated, activating different sets of sites and coupling different bases allowing arbitrary DNA probes to be constructed at each site. **b,** Schematic representation of the lamp, mask and array.


See also: http://www.affymetrix.com/technology/index.affx
In situ synthesized (Affymetrix) microarrays:

Advantage - massive coverage of sequences possible

Disadvantage - expensive and not easily customized

Spotted arrays (oligonucleotides or dsDNA):

Advantage - flexibility and cost; can be “home-made”

Disadvantage - big initial investment; less standardization
BEADARRAY™ TECHNOLOGY OVERVIEW

How it works

1. Compatible with manual or robotic operation, our Array of Arrays format is designed to match the wells of standard microtiter plates.

2. Close up of the bead-containing ends of the fiber bundles arranged in a 96-well format.

3. Dipping the fiber optic bundle into a chemical solution etches a microscopic well at the end of each individual fiber in the bundle.

4. To form an array, we dip each fiber bundle into a pool of coated beads which self-assemble into the wells, one bead per well.

5. Up to 2000 different bead types, with each bead type containing oligonucleotides of a unique sequence, can be represented in each bundle, with a targeted twenty-five-fold average redundancy per bead type.

6. Hundreds of thousands of molecules of the same type coat each bead. We determine which bead type occupies which well via a proprietary decoding process.

7. The molecules in the sample hybridize, or bind to their complementary molecules on the coated bead.

Note:
Not all arrays have to be on chips...!
- Illumina, Inc.

Pooling and Assembly

Illumina’s initial BeadArray implementation contains approximately 50,000 beads in individual 3 micron wells at the end of each fiber strand.

Multiple-Color, Hybridization-Based Decoding

Individual bead types are color coded; fiber optic bundles are then “dipped” into the pool where they self-assemble to form the microarray.

Optical Fibers

Illumina’s optical fiber technology enables high-quality laser scanning and imaging.
Caveat... Caveat... Caveat...!!

- **Results from the different DNA microarray methods don’t always agree!**

- **And results for mRNA abundance differences don’t always agree with protein abundance data!**
Applications of DNA Microarrays

- Genotyping (cf. DNA sequencing lecture)
- mRNA profiling and “transcriptome” analysis
- Genome analysis (cancer and evolutionary studies)
- Genome-wide splicing analyses
- Etc.
Applications of DNA Microarrays

• Genotyping (cf. DNA sequencing lecture)

• mRNA profiling and “transcriptome” analysis

• Genome analysis (cancer and evolutionary studies)

• Genome-wide splicing analyses

• Etc.
Applications of DNA Microarrays

- Genotyping (cf. DNA sequencing lecture)
- mRNA profiling and “transcriptome” analysis
- Genome analysis (cancer and evolutionary studies)
- Genome-wide splicing analyses
- Etc.
Fig. 1. Flowchart of a GeneChip System microarray experiment. Once the nucleic acid sample has been obtained, target amplification and labeling result in a labeled sample. The labeled sample is then injected into the probe array and allowed to hybridize overnight in the hybridization oven. Probe array washing and staining occur on the fluidics station, which can handle four probe arrays simultaneously. The probe array is then ready to be scanned in the Affymetrix GeneChip scanner, where the fluorescence intensity of each feature is read. Data output includes an intensity measurement for each transcript or the detailed sequence or genotyping (SNP) information.

**Fig. 4.** Gene expression array design strategies. The different expression strategies for probe selection are represented. The gene sequence shown at the top represents an example of a target transcript. Rectangles represent exons, while the connecting lines represent introns. The 3′ IVT expression probes target sequences are at the extreme 3′ end and are adjacent to the poly(A) tail of the mRNA. This strategy is the most commonly used for commercial whole genome transcriptome designs. Exon array probe sets include probes that are within exon sequences. For tiling arrays, probes are placed sequentially throughout the genome at the same approximate distance from each other.
Comparative mRNA Analyses Using Spotted Microarrays (not Affymetrix GeneChips)

Note: this experiment shows competitive hybridization for a spotted array.

Table 1 (left). Aging-related changes in gene expression in gastrocnemius muscle. The extent to which caloric restriction prevented age-associated alterations in gene expression is denoted as either C (complete, >90%), N (none), or partial (20 to 90%, percentage effect indicated). The fold increase shown represents the average of all nine possible pairwise comparisons among individual mice determined by means of a specific algorithm (12). GenBank accession numbers are listed under ORF. A more comprehensive list that includes genes that did not fit into the six classes can be found at www1.genetics.wisc.edu/prolla/Prolla_Tables.html. Table 2 (right). Caloric restriction–induced alterations in gene expression. The data represent a comparison between 30-month-old CR-fed and control-fed mice. The gene expression alterations listed in this Table are diet related and do not include those representing prevention of age-associated changes (see Table 1). Additional CR-induced changes are posted at the aforementioned Web site.

<table>
<thead>
<tr>
<th>ORF</th>
<th>Δ Age (fold)</th>
<th>Gene Function</th>
<th>CR Prevention</th>
</tr>
</thead>
<tbody>
<tr>
<td>W08057</td>
<td>3.5</td>
<td>Heat Shock 27 kDa Protein</td>
<td>Chaperone</td>
</tr>
<tr>
<td>M17790</td>
<td>3.5</td>
<td>Serum Amyloid A (apol 4)</td>
<td>Unknown</td>
</tr>
<tr>
<td>A11926</td>
<td>3.5</td>
<td>Heat Shock 71 kDa Protein</td>
<td>Chaperone</td>
</tr>
<tr>
<td>L28777</td>
<td>2.6</td>
<td>GADD45a</td>
<td>DNA damage response</td>
</tr>
<tr>
<td>M74360</td>
<td>2.4</td>
<td>Aldolase Dihydrogenase</td>
<td>Aldolase function</td>
</tr>
<tr>
<td>A004982</td>
<td>2.2</td>
<td>Protease</td>
<td>Protease</td>
</tr>
<tr>
<td>L22482</td>
<td>2.2</td>
<td>HIC-5</td>
<td>Senescence and differentiation</td>
</tr>
<tr>
<td>X96963</td>
<td>2.2</td>
<td>rhB</td>
<td>Unknown</td>
</tr>
<tr>
<td>X65267</td>
<td>2.0</td>
<td>TNZ2</td>
<td>RNA metabolism</td>
</tr>
<tr>
<td>X65277</td>
<td>1.6</td>
<td>Ras1</td>
<td>JNK activator</td>
</tr>
<tr>
<td>A007177</td>
<td>3.8</td>
<td>Synaptic Vesicle Protein</td>
<td>Neurite extension</td>
</tr>
<tr>
<td>X65237</td>
<td>2.5</td>
<td>Neurophin-3</td>
<td>Reinnervation of muscle</td>
</tr>
<tr>
<td>X76197</td>
<td>2.2</td>
<td>AP-2 Beta</td>
<td>Neurogenesis</td>
</tr>
<tr>
<td>X96749</td>
<td>2.1</td>
<td>mtTGF</td>
<td>Differentiation</td>
</tr>
<tr>
<td>A01022</td>
<td>2.1</td>
<td>Dynactin</td>
<td>Transport</td>
</tr>
<tr>
<td>X65190</td>
<td>2.1</td>
<td>PAAS</td>
<td>Response to muscle injury</td>
</tr>
<tr>
<td>A109111</td>
<td>3.8</td>
<td>Mitochondrial Sarcoplasmic</td>
<td>ATP generation</td>
</tr>
<tr>
<td>A081886</td>
<td>2.0</td>
<td>Dihydropyridine-sensitive</td>
<td>Calcineurin Channel</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ORF</th>
<th>Δ Age (fold)</th>
<th>Gene Function</th>
<th>CR Prevention</th>
</tr>
</thead>
<tbody>
<tr>
<td>A001310</td>
<td>4.1</td>
<td>Mitochondrial LON Protease</td>
<td>Mitochondrial biogenesis</td>
</tr>
<tr>
<td>W56097</td>
<td>2.9</td>
<td>Alpha Enolase</td>
<td>Glycogenolysis</td>
</tr>
<tr>
<td>W09719</td>
<td>2.9</td>
<td>Alpha-Amylase-1</td>
<td>Carbohydrate metabolism</td>
</tr>
<tr>
<td>M84375</td>
<td>2.5</td>
<td>Phosphoprotein Phosphatase</td>
<td>Glycogen metabolism</td>
</tr>
<tr>
<td>A034842</td>
<td>2.1</td>
<td>ERV1</td>
<td>mdDNA maintenance</td>
</tr>
<tr>
<td>A106406</td>
<td>2.0</td>
<td>ATP Synthase A Chain</td>
<td>ATP synthesis</td>
</tr>
<tr>
<td>A0441796</td>
<td>2.0</td>
<td>IPP-2</td>
<td>Glycogen metabolism</td>
</tr>
<tr>
<td>L72482</td>
<td>2.0</td>
<td>PMMP5</td>
<td>Peroxisome assembly</td>
</tr>
<tr>
<td>Z48204</td>
<td>2.0</td>
<td>NADP Transhydrogenase</td>
<td>Glycerophosphate shunt</td>
</tr>
<tr>
<td>A077176</td>
<td>1.9</td>
<td>Glucose-6-Phosphate Isomerase</td>
<td>Glycogenolysis</td>
</tr>
<tr>
<td>M13396</td>
<td>1.9</td>
<td>Glycogen Dephosphorylase</td>
<td>Glycogen metabolism</td>
</tr>
<tr>
<td>A107752</td>
<td>2.0</td>
<td>EF-1-Gamma</td>
<td>Protein synthesis</td>
</tr>
<tr>
<td>J22031</td>
<td>2.8</td>
<td>20S Protease Subunit</td>
<td>Protein turnover</td>
</tr>
<tr>
<td>A001830</td>
<td>2.2</td>
<td>Ubiquitin Thioltransferase</td>
<td>Protein turnover</td>
</tr>
<tr>
<td>A1145829</td>
<td>2.1</td>
<td>20S Protease Component TBP1</td>
<td>Protein turnover</td>
</tr>
<tr>
<td>L00931</td>
<td>2.1</td>
<td>Ubc6 Ubiquitin Specific Protease</td>
<td>Protein turnover</td>
</tr>
<tr>
<td>J35741</td>
<td>2.0</td>
<td>Rhodanese</td>
<td>Mitochondrial protein folding</td>
</tr>
<tr>
<td>D86815</td>
<td>2.1</td>
<td>Protease Z Subunit</td>
<td>Protein turnover</td>
</tr>
<tr>
<td>D76440</td>
<td>2.0</td>
<td>Nacdin</td>
<td>Neuronal growth suppressor</td>
</tr>
<tr>
<td>X75014</td>
<td>2.7</td>
<td>Phox2 Homeodomain Protein</td>
<td>Throphic factor</td>
</tr>
<tr>
<td>M32540</td>
<td>2.1</td>
<td>GASS</td>
<td>Myelin protein</td>
</tr>
<tr>
<td>M14845</td>
<td>3.4</td>
<td>Calpain 1 Light Chain</td>
<td>Calpain effector</td>
</tr>
<tr>
<td>L34611</td>
<td>2.3</td>
<td>PTHR</td>
<td>Calcium homeostasis</td>
</tr>
<tr>
<td>A103356</td>
<td>2.2</td>
<td>Calmodulin</td>
<td>Calcium effector</td>
</tr>
<tr>
<td>D86916</td>
<td>1.4</td>
<td>Squaleenate Synthase</td>
<td>Cholesterol/phytanic acid synthesis</td>
</tr>
<tr>
<td>M21295</td>
<td>2.1</td>
<td>Staurospyr-CoA Desaturase</td>
<td>PUFA synthesis</td>
</tr>
<tr>
<td>U73744</td>
<td>2.1</td>
<td>HSP70</td>
<td>Chaperone</td>
</tr>
</tbody>
</table>

Note: caloric restriction gene chip experiment w/ mice.

<table>
<thead>
<tr>
<th><strong>Aging</strong></th>
<th><strong>Caloric restriction</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>↑ Stress response</strong></td>
<td><strong>↑ Protein metabolism</strong></td>
</tr>
<tr>
<td>Induction of:</td>
<td>Increased synthesis</td>
</tr>
<tr>
<td>Heat shock response</td>
<td>Increased turnover</td>
</tr>
<tr>
<td>DNA damage–inducible genes</td>
<td></td>
</tr>
<tr>
<td>Oxidative stress–inducible genes</td>
<td></td>
</tr>
<tr>
<td><strong>↓ Energy metabolism</strong></td>
<td><strong>↑ Energy metabolism</strong></td>
</tr>
<tr>
<td>Reduced glycolysis</td>
<td>Up-regulation of gluconeogenesis, and the pentose phosphate shunt</td>
</tr>
<tr>
<td>Mitochondrial dysfunction</td>
<td></td>
</tr>
<tr>
<td><strong>↑ Neuronal injury</strong></td>
<td><strong>↑ Biosynthesis</strong></td>
</tr>
<tr>
<td>Reinnervation</td>
<td>Fatty acid synthesis</td>
</tr>
<tr>
<td>Neurite extension and sprouting</td>
<td>Nucleotide precursors</td>
</tr>
<tr>
<td><strong>↓ Macromolecular damage</strong></td>
<td><strong>↓ Macromolecular damage</strong></td>
</tr>
<tr>
<td>Suppression of:</td>
<td>Inducible heat shock factors</td>
</tr>
<tr>
<td>Inducible heat shock factors</td>
<td>Inducible detoxification systems</td>
</tr>
<tr>
<td>Inducible detoxification systems</td>
<td>Inducible DNA repair systems</td>
</tr>
</tbody>
</table>

Technical Proficiency & Experimental Design are Key to Reproducibility

Figure 2 | Experimental design and reproducibility. a | General schematic of mouse brain gene expression experiments. Tissue is obtained by systematic dissection of the appropriate brain regions from at least two mice of the same age, sex and genetic background, which have been housed and handled identically. In this example, the hippocampus (arrows) was used. After obtaining tissue, total RNA is extracted and labelled complementary RNA is made for array hybridization. Each sample is hybridized to a separate array b | A comparison of the quantitative results for independent replicates for two different mice (C57BL/6). The correlation coefficient is very near 1.0 (0.994), and the number of genes that score as 'differentially expressed' based on a single comparison is small (9 of a possible 6854 in this example), indicating the high degree of reproducibility of the procedures, measurements and analyses. c | Analysis showing the low false-positive rate achieved when using stringent analysis criteria and independent replicates. Samples were prepared from dissected hippocampus from four C57BL/6 mice. When the results for mouse 1 and mouse 2 were compared, only nine genes scored as different. When mouse 3 was compared with mouse 4, only five genes were scored as different and there were no genes that scored as different in both of the independent comparisons. The criteria used were a 1.8-fold change or greater, a qualitative call of increased, marginally increased, decreased or marginally decreased, a signal change of 50 (after scaling to an average signal of 200), and a call of present in at least one of the samples using the standard Affymetrix GeneChip algorithms and software.

Lockhart & Barlow (2001)
However, if your technique isn't good enough, the data must be "normalized"…


**Figure 4** Log-log plot of untreated vs treated conditions. A simple log-log plot, which serves as an at-a-glance visualisation of the fold-differences in expression between any two conditions is shown. Conditions 1 and 2 are the average (arithmetic mean) for a population of untreated and treated mice respectively. Plotting in log-space avoids graphing problems usually associated with large ranges of data values. The central dotted line represents ‘no differentiation’, i.e. any gene on or near this line has no difference in expression between the two conditions. In order to enable us to distinguish individual dots more easily, we have also reduced the size of dots near this line, the most central ones are additionally coloured blue. The two solid lines represent the limits for twofold up-regulation and down-regulation respectively. Also shown in red is a fitted LOWESS (LOcally WEighted Scatterplot Smoother) line, which in this case only serves as a robust representation of central trends in the global dataset. It should be noted that after data normalisation, one usually expects most of the data points to be within the fold-limit lines, unless global expression changes between the two conditions in question are expected.
Applications of DNA Microarrays

- Genotyping (cf. DNA sequencing lecture)
- mRNA profiling and “transcriptome” analysis
- Genome analysis (cancer and evolutionary studies)
- Genome-wide splicing analyses
- Etc.
"CNV" = “copy number variant”

Figure 4 Individuals with a single-copy deletion CNV or a single recessive mutation can be healthy. However, the same single-copy deletion CNV could uncover a recessive mutation in certain individuals, leading to a clinically recognizable phenotype.

Microarrays Can Also be Used to Analyze Chromosomal Rearrangements

![Diagram showing various chromosomal abnormalities and their detection with different techniques.](image)

<table>
<thead>
<tr>
<th>Technique</th>
<th>Polyploid</th>
<th>Aneuploid</th>
<th>Interstitial Deletion</th>
<th>Reciprocal Translocation</th>
<th>Non-reciprocal Translocation</th>
<th>Amplification (double minutes)</th>
<th>Amplification (HSR)</th>
<th>Amplification (distributed insertions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banding</td>
<td>+</td>
<td>+</td>
<td>+/-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>SKY</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>CGH</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>LOH</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

**Figure 1.** Detection and mapping of chromosomal aberrations by different molecular and cytological techniques. Rearrangements involving two chromosomes of a diploid genome are shown. The ability of various techniques to detect the aberration is indicated in the table. Modified from Albertson et al. (67).

*CGH - “Comparative genome hybridization”*  
*LOH - “Loss of heterozygosity”*  

Use of hybridization of genomic DNA to tiling arrays to look for over- or under-abundance of specific genome regions

Note that what is being plotted is the log\textsubscript{2} of the signal (so a heterozygous deletion yields a “-1”).

Figure 2. Detection of copy number aberrations in tumor genomes by array CGH. (A) Chromosomal aberrations in cancer are likely to arise following inappropriate management of DNA damage or telomere dysfunction. Common aberrations include gene amplifications, non-reciprocal translocations and interstitial deletions. Amplifications may be visible cytogenetically as double minutes, chromosomes with homogeneously staining regions (hsr) or the amplified DNA may be distributed at multiple sites. The array CGH copy number profile of the amplified MYC in COLO320 is shown. The amplification level is \(\sim 70\) fold (log\textsubscript{2} ratio \(> 6\)). Breakage of a chromosome or a non-reciprocal translocation event may lead to low level copy number changes, as shown in the copy number profile of chromosome 1 from 600MPE. Homozygous deletions are indicated by log\textsubscript{2} ratio \(< -2\) and heterozygous deletions by log\textsubscript{2} ratio \(\sim -1\), as shown in the copy number profile for chromosome 16 in HCT116. The log\textsubscript{2} ratios are plotted on individual chromosomes according to the positions of the clones on the UCSC August 2001 freeze of the genome sequence (http://genome.ucsc.edu). Complete array data sets are available from Snijders et al. (2). (B) Whole genome array CGH copy number profiles. The spectrum of copy number aberrations seen in different tumor types varies as illustrated by the copy number profile of HCT116 (mismatch repair defective colon tumor cell line, with few copy number alterations), 600MPE (breast tumor cell line with amplification of CCND1), T47D (breast tumor cell line with many low level copy number changes) and HCC1937 (BRCA1 deficient breast cancer cell line, with many copy number changes). The log\textsubscript{2} ratios for each chromosome in order from 1p to Xpter are plotted in order according to the positions of the clones on the UCSC August 2001 freeze of the genome sequence (http://genome.ucsc.edu). Complete array data sets for HCT116, 600MPE and T47D are available from Snijders et al. (2).
Two final notes about DNA microarrays…

(Has their usefulness peaked??)
Even in 2009, microarrays still suffer from big standardization and reproducibility problems!

Repeatability of published microarray gene expression analyses

John P A Ioannidis¹⁻³, David B Allison⁴, Catherine A Ball⁵, Issa Coulibaly⁴, Xiangqin Cui⁴, Aedin C Culhane⁶⁻⁷, Mario Falchi⁸⁻⁹, Cesare Furlanello¹⁰, Laurence Game¹¹, Giuseppe Jurman¹⁰, Jon Mangion¹¹, Tapan Mehta⁴, Michael Nitzberg⁵, Grier P Page⁴⁻¹², Enrico Petretto¹¹⁻¹³ & Vera van Noort¹⁴

Given the complexity of microarray-based gene expression studies, guidelines encourage transparent design and public data availability. Several journals require public data deposition and several public databases exist. However, not all data are publicly available, and even when available, it is unknown whether the published results are reproducible by independent scientists. Here we evaluated the replication of data analyses in 18 articles on microarray-based gene expression profiling published in Nature Genetics in 2005–2006. One table or figure from each article was independently evaluated by two teams of analysts. We reproduced two analyses in principle and six partially or with some discrepancies; ten could not be reproduced. The main reason for failure to reproduce was data unavailability, and discrepancies were mostly due to incomplete data annotation or specification of data processing and analysis. Repeatability of published microarray studies is apparently limited. More strict publication rules enforcing public data availability and explicit description of data processing and analysis should be considered.
And...

Ultra-high throughput, next generation DNA sequencing methods (see 1/27/09 and 2/13/09 lectures), which can characterize the sequences of millions of nucleic acid molecules in parallel and have other attractive features (for example, the sequence info!), may supplant microarrays as the method of choice for many applications.
However, before we leave the topic of DNA microarrays, let’s look at one more data slide....
The human genome is highly transcribed, even where there are no known genes...

What’s with all this extra RNA? To provide the “raw material” for si/miRNAs??

Fig. 2. High-resolution maps of four regions within DGCR of chromosome 22 (2q11.2). For each map, the contigs predicted by the DGCR array for 6 of the 11 cell lines analyzed is presented. Below the array map are cartoons derived from the Sanger hand-curated map of this region (19) or from GenBank. Selected regions suggested by the array map were further analyzed with the use of RT-PCR.

The sequenced products from these analyses are mapped below the Sanger and ESTs maps (A) DGCR6 gene region (Golden Path (GP) sequence 15,833,950–15,840,390), (B) DGCR2 region (GP sequence 15,959,850–16,057,850), (C) SLC25A1 gene and flanking region (GP sequence 16,098,590–16,107,090), and (D) DGCR5 exon1 region (GP sequence 15,898,300–15,905,040).

RNA Interference (RNAi) (see also miRNA, siRNA, micRNA, shRNA, etc.)
The Discovery of the RNA Interference (RNAi) Phenomenon

1. Gene-specific inhibition of expression by anti-sense nucleic acids was discovered in the 1980’s (Inouye, 1988).

2. Guo and Kemphues (1995) showed that, in some cases for C. elegans genes, the sense strand inhibited just as well as the anti-sense strand (!? - why careful controls are always wise!).

3. Three years later, Mello and Fire (1998) tested whether both the sense and the anti-sense strand together would inhibit or cancel other out. They hit the jackpot: the dsRNA that they inadvertently created inhibited homologous mRNA expression much more strongly than either the sense or the anti-sense strands alone (!!??).

>>> RNAi may be a very ancient defense mechanism that evolved in eukaryotic organisms to protect their cells against viruses.
And eight years after that....
RNA silencing directed by foreign dsRNA. In this model, all triggers that induce RNA silencing operate through a dsRNA intermediate giving rise to the formation of siRNAs. While inversely orientated transgenes, in vitro prepared dsRNA, and viruses are direct sources of dsRNA, an additional step is needed to explain the production of dsRNA from multicopy transgenic arrays or highly transcribed single-copy transgenes. This might occur via read-through transcription, either from an endogenous promoter at the site of integration or as a result of head-to-head organization of transcription units in the multicopy array, giving rise to dsRNA directly or to antisense RNA that can pair with the sense to form dsRNA. Alternatively, sense (and possibly antisense) aberrant RNAs are converted into dsRNA via de novo RNA synthesis, reminiscent of virus replication.

Figure 2  Two-step model to explain RNAi. In the first step, dsRNA is diced by an ATP-dependent ribonuclease (later found to be Dicer) into short interfering RNAs (siRNAs): duplexes of 21–23 nucleotides bearing two-nucleotide 3’ overhanging ends. These siRNAs are subsequently transferred to a second enzyme complex, designated RISC for RNAi-induced silencing complex, which contains an endoribonuclease that is distinct from Dicer. The siRNA guides the endonuclease—the antisense strand of the siRNA is perfectly complementary to the target mRNA—to the target mRNA leading to its destruction. The position of the cleavage site in the target is within the sequences covered by the siRNA (near the center) and is determined by the 5’ end of the guiding molecules (34, 35).

Micro-RNAs (e.g., so-called stRNAs) are also involved in endogenous gene regulation.

Note that the inhibition of expression is exerted by binding at the 3'-UTR!
A model for RNA-induced silencing complex (RISC) assembly

Note: perfect basepairing

Note: slight sequence mismatch

Note: the proteins that combine with the siRNA/miRNAs are members of the “Argonaute” family.

The RNAi silencing phenomenon may be part of an ancient cellular anti-virus defense mechanism, since many viruses generate virus-specific dsRNA that could trigger the response (see also interferon). Consistent with this is the fact that RNAi has been found in a large range of metazoan organisms from plants to fruit flies to worms to humans.

On the other hand, RNAi silencing may stem from an even more ancient mechanism of RNAs regulating RNAs, the machinery of which was later “recruited” (via evolution and selection) to take on an additional antivirus role.
Some data regarding endogenous miRNAs

**A. thaliana**
- 20 conserved families
- 90 genes
- 72 genes known to have a role in plant development

**C. elegans**
- 65 conserved families
- 100 - 120 genes found (so far)

**H. sapiens**
- Approximately 130 gene families
- > 400 known genes

Source:
D. Bartel, Oct. 2005
The ENCODE Project: ENCYclopedia Of DNA Elements

ENCODE Overview

The National Human Genome Research Institute (NHGRI) launched a public research consortium named ENCODE, the Encyclopedia Of DNA Elements, in September 2003, to carry out a project to identify all functional elements in the human genome sequence. The project started with two components - a pilot phase and a technology development phase.

The pilot phase tested and compared existing methods to rigorously analyze a defined portion of the human genome sequence (See: ENCODE Pilot Project). The conclusions from this pilot project were published in June 2007 in Nature and Genome Research. The findings highlighted the success of the project to identify and characterize functional elements in the human genome. The technology development phase also has been a success with the promotion of several new technologies to generate high throughput data on functional elements.

With the success of the initial phases of the ENCODE Project, NHGRI funded new awards in September 2007 to scale the ENCODE Project to a production phase on the entire genome along with additional pilot-scale studies. Like the pilot project, the ENCODE production effort is organized as an open consortium and includes investigators with diverse backgrounds and expertise in the production and analysis of data (See: ENCODE Participants and Projects). This production phase also includes a Data Coordination Center to track, store and display ENCODE data along with a Data Analysis Center to assist in integrated analyses of the data. All data generated by ENCODE participants will be rapidly released into public databases (See: Accessing ENCODE Data) and available through the project’s Data Coordination Center.

- Read about the ENCODE Pilot Project.
Fig. 1. Functional genomic elements being identified by the ENCODE pilot phase

What if you want to design an RNAi molecule to do “knockdown” experiments or for therapeutic purposes?

**Empirical rules for designing siRNAs with proper specificity and efficacy**

<table>
<thead>
<tr>
<th>Box 2</th>
<th><strong>Considerations for generating an effective and specific siRNA</strong></th>
</tr>
</thead>
</table>
| A small interfering RNA (siRNA) is defined as successful when it can provide effective and specific gene silencing (>90% reduction in protein levels) when it is used at a concentration of 1–20 nM invitro. Recent studies have important implications for the design of functional siRNAs for mammalian RNAi.
On the basis of these studies, the table depicts mechanism-based rules that when applied to siRNA design are expected to show maximum suppression of target mRNA expression at the lowest possible concentration of siRNA. |
| **Criteria** | **Probable reason** |
| **Biophysical, thermodynamic and structural considerations** | |
| Overall low to medium G+C content (30–50%) | Facilitates interaction with RISC and unwinding |
| Low internal stability at the 5’ antisense strand | Promotes antisense-strand selection by RISC |
| High internal stability at the 5’ sense strand | Blocks sense-strand selection by RISC |
| Absence of internal repeats or palindromes | Increases the concentration of functional, stable hairpins |
| A-form helix between siRNAs and target mRNA | Enhances RNA–RNA interactions and promotes cleavage |
| **Base preferences at specific positions in the sense strand** | |
| Presence of an A at position 3 and 19+ of sense strand | Promotes antisense-strand selection by RISC |
| Absence of a G or C at position 19 of sense strand | Promotes antisense-strand selection by RISC |
| Presence of a U at position 10 of sense strand | Promotes RISC mediated cleavage of mRNA and dissociation of the RISC–siRNA complex |
| Absence of a G at position 13 of sense strand | Promotes efficient unwinding |
| **Enhancing specificity of siRNA-mediated gene silencing** | |
| Perform stringent homology searches | Minimizes potential nonspecific gene silencing |
| Avoid low-stringency sequence interactions between siRNA and 3'UTR | Minimizes potential nonspecific gene silencing |

*The base preference for A at position 19 reflects the same bias that is observed for microRNA (miRNA) precursors. For example, most miRNAs contain a U at position 1 (corresponding to A in position 19 of the siRNA (small interfering RNA) sense strand). RISC, RNA-induced silencing complex; UTR, untranslated region.

RISC-associated helicase can unwind less well from this end because of higher stability.

RISC-associated helicase can unwind better from this end because of lower stability.

**Figure 2 | The generation of effective siRNA.** A small interfering RNA (siRNA) is a 21–23-nucleotide (nt) dsRNA that contains: a 19-nt duplexed region, symmetric 2–3-nt 3’ overhangs, and 5’-phosphate (P) and 3’-hydroxyl (OH) groups. The positions of each nucleotide in the 19-nt duplexed region of the sense strand are shown. On the basis of recently established design criteria, an effective siRNA has high stability at the 5’ terminus of the sense strand (blue box), lower stability at the 5’ antisense terminus (orange box) and at the cleavage site (purple box). In addition, the sequence-specific preferences at the following positions on the sense strand are important: the presence of an A at position 19, an A at position 3, a U at position 10 (BOX 2 lists other parameters). RISC, RNA-induced silencing complex.

An RNAi movie:

http://www.nature.com/focus/rnai/animations/rnai_revised_500x280.mov

[Note: see also the “RNA Silencing” poster (downloaded as a pdf file from Science 309, 1518 (2005)]
Problem:

How to make **stable** RNAi-mediated gene knockdowns?

*and*

...also would be nice to be able to do tissue- and/or developmental stage-specific regulation.
Heritable and inducible genetic interference by double-stranded RNA encoded by transgenes

Nektarios Tavernarakis¹, Shi Liang Wang¹, Maxim Dorovkov², Alexey Ryazanov² & Monica Driscoll¹

Double-stranded RNA interference (RNAi) is an effective method for disrupting expression of specific genes in Caenorhabditis elegans and other organisms¹-⁵. Applications of this reverse-genetics tool, however, are somewhat restricted in nematodes because introduced dsRNA is not stably inherited⁵. Another difficulty is that RNAi disruption of late-acting genes has been generally less consistent than that of embryonically expressed genes, perhaps because the concentration of dsRNA becomes lower as cellular division proceeds or as developmental time advances¹. In particular, some neuronally expressed genes appear refractory to dsRNA-mediated interference. We sought to extend the applicability of RNAi by in vivo expression of heritable inverted-repeat (IR) genes. We assayed the efficacy of in vivo-driven RNAi in three situations for which heritable, inducible RNAi would be advantageous: (i) production of large numbers of animals deficient for gene activities required for viability or reproduction; (ii) generation of large populations of phenocopy mutants for biochemical analysis; and (iii) effective gene inactivation in the nervous system. We report that heritable IR genes confer potent and specific gene inactivation for each of these applications. We suggest that a similar strategy might be used to test for dsRNA interference effects in higher organisms in which it is feasible to construct transgenic animals, but impossible to directly or transiently introduce high concentrations of dsRNA.

¹Department of Molecular Biology and Biochemistry, Rutgers, The State University of New Jersey, and ²Department of Pharmacology, Robert Wood Johnson Medical School, University of Medicine and Dentistry of New Jersey, Piscataway, New Jersey, USA. Correspondence should be addressed to M.D. (e-mail: driscoll@mbd.rutgers.edu).
Construction of the hairpin expression vector

Fig. 1 Strategy for generation of heritable and inducible RNAi. a, A strategy for \textit{in vivo} RNAi. A strong heat shock-inducible promoter was fused to a direct IR gene. Upon heat shock of transgenic animals harbouring this gene, transcripts were generated which are predicted to fold back in a uni-molecular reaction to generate double-stranded RNA in all cells that express the heat-shock gene. The size of the single-stranded loop that occurs after foldback is not known. b, Construction of inducible IR genes. Exon-rich genomic DNA (or cDNA) was amplified using two primers that introduced unique restriction sites at the fragment ends. One restriction site was used to generate the IR and was ultimately situated at the inversion point (IP). The other restriction site (designated as end) was used to join the IR to the vector. Amplified fragments were digested with the enzyme situated at the IP restriction site (IPRS) and ligated together. Digestion at the end restriction site (ERS) enabled the fragment to be cloned into a similarly digested, CIAP-treated \textit{C. elegans} expression vector. We used vector pPD49.78 (ref. 22), which includes the \textit{hsp16-2} promoter and the 3' UTR of muscle myosin \textit{unc-54}. 
Endogenously-expressed hairpin RNA actually sometimes works better than injected dsRNAi
RNAi “Knockdown” Experiments are Very Useful for Studying the Functions of Genes *In Vivo*

**Figure 1** Double-stranded RNA can be introduced experimentally to silence target genes of interest. In plants, silencing can be triggered, for example, by engineered RNA viruses or by inverted repeat transgenes. In worms, silencing can be triggered by injection or feeding of dsRNA. In both of these systems, silencing is systemic and spreads throughout the organism. **a.** A silencing signal moves from the veins into leaf tissue. Green is green fluorescent protein (GFP) fluorescence and red is chlorophyll fluorescence that is seen upon silencing of the GFP transgene. **b.** *C. elegans* engineered to express GFP in nuclei. Animals on the right have been treated with a control dsRNA, whereas those on the left have been exposed to GFP dsRNA. Some neuronal nuclei remain fluorescent, correlating with low expression of a protein required for systemic RNAi. **c.** HeLa cells treated with an ORC6 siRNA and stained for tubulin (green) and DNA (red). Depletion of ORC6 results in accumulation of multinucleated cells. Stable silencing can also be induced by expression of dsRNA as hairpins or snap-back RNAs. **d.** Adult *Drosophila* express a hairpin homologous to the white gene (left), which results in unpigmented eyes compared with wild type (right).
SYSTEMATIC GENOME-WIDE SCREENS OF GENE FUNCTION

Anne E. Carpenter and David M. Sabatini

By using genome information to create tools for perturbing gene function, it is now possible to undertake systematic genome-wide functional screens that examine the contribution of every gene to a biological process. The directed nature of these experiments contrasts with traditional methods, in which random mutations are induced and the resulting mutants are screened for various phenotypes. The first genome-wide functional screens in *Caenorhabditis elegans* and *Drosophila melanogaster* have recently been published, and screens in human cells will soon follow. These high-throughput techniques promise the rapid annotation of genomes with high-quality information about the biological function of each gene.
Figure 3 | Examples of scorable phenotypes from various screens. In addition to studies in which several classical phenotypes were scored by eye in *Caenorhabditis elegans*\(^{44,45,47}\), visual screens have begun to use microscopy. Yeast genome-wide visual screens have identified genes that are important for sporulation (by microscopy of unstained cells)\(^{93}\), cell-cycle progression (by microscopy of unstained cells)\(^{94}\), bud site selection (by fluorescence microscopy of calcofluor-stained cells)\(^{68}\), mitochondrial morphology (by fluorescence microscopy of antibody-stained cells)\(^{29}\) and endocytosis (by fluorescence microscopy of lucifer-yellow-stained cells)\(^{29}\). In addition, the subcellular localization of about two-thirds of *Saccharomyces cerevisiae* genes has been determined by tagging genes with green fluorescent protein (GFP) or an epitope\(^{96,97}\).  

**a, b** | A partial-genome screen in *C. elegans*, which used time-lapse microscopy of whole worms, identified genes that are required for proper cell division\(^{89}\). RNAi against C16A3.9, a 40S ribosomal protein, resulted in several female pronuclei rather than one (arrowheads).  

**c, d** | A genome-wide screen of yeast deletion mutants identified genes that are required for normal shape and size, including phenotypes that are classified as ‘elongated’\(^{14}\).  

**e, f** | A full-genome RNAi screen in *C. elegans*, which examined the pattern and intensity of Nile-red staining of fat storage droplets, identified genes that are involved in fat metabolism\(^{98}\). RNAi against choline/ethanolamine phosphotransferase reduced fat staining.  

**g, h** | A partial-genome RNAi screen in *Drosophila*-cultured cells identified genes that are involved in cytoskeletal organization, viability, attachment, cell-cycle progression and cytokinesis\(^{99}\). RNAi against fizzy, a protein that is involved in cyclin catabolism, produced cells with an increased frequency of mitotic spindles.


Carpenter & Sabatini (2004)
But how to deliver the RNAi for therapeutic effect...??
Alternatives for gene silencing in vivo: gene knockouts vs. various RNAi approaches

- **a** Knockout
  - Time to create: years
  - Duration of action: indefinite
  - ES cells

- **b** Transgenic RNAi
  - Time to create: months
  - Duration of action: indefinite
  - ES cells

- **c** Local RNAi
  - Time to create: weeks
  - Duration of action: virus: weeks to indefinite; siRNAs: weeks
  - Virus producing hairpin RNA

- **d** Systemic RNAi
  - Time to create: days
  - Duration of action: virus: weeks to indefinite; siRNAs: weeks
  - siRNA

Figure 6 | *In vivo mammalian gene silencing*. The figure outlines several methods of gene silencing in the mouse, and compares their time of preparation and duration of action. ES, embryonic stem; RNAi, RNA interference; siRNA, small interfering RNA.