MICROARRAYS: CHIPPING AWAY AT THE MYSTERIES OF SCIENCE AND MEDICINE

National Center for Biotechnology Information <www.ncbi.nlm.nih.gov/About/primer/microarrays.html>

With only a few exceptions, every cell of the body contains a full set of chromosomes and identical genes. Only a fraction of these genes are turned on, however, and it is the subset that is "**expressed**" that confers unique properties to each cell type. "**Gene expression**" is the term used to describe the transcription of the information contained within the **DNA**, the repository of genetic information, into messenger RNA (mRNA) molecules that are then translated into the proteins that perform most of the critical functions of cells. Scientists study the kinds and amounts of mRNA produced by a cell to learn which genes are expressed, which in turn provides insights into how the cell responds to its changing needs. Gene expression is a highly complex and tightly regulated process that allows a cell to respond dynamically both to environmental stimuli and to its own changing needs. This mechanism acts as both an "**on/off" switch** to control which genes are expressed in a cell as well as a "**volume control**" that increases or decreases the level of expression of particular genes as necessary.

The proper and harmonious expression of a large number of genes is a critical component of normal growth and development and the maintenance of proper health. Disruptions or changes in gene expression are responsible for many diseases.

Enabling Technologies

Biomedical research evolves and advances not only through the compilation of knowledge but also through the development of new technologies. Using traditional methods to assay gene expression, researchers were able to survey a relatively small number of genes at a time. The emergence of new tools enables researchers to address previously intractable problems and to uncover novel potential targets for therapies. Microarrays allow scientists to analyze expression of many genes in a single experiment quickly and efficiently. They represent a major methodological advance and illustrate how the advent of new technologies provides powerful tools for researchers. Scientists are using microarray technology to try to understand fundamental aspects of growth and development as well as to explore the underlying genetic causes of many human diseases.

DNA Microarrays: The Technical Foundations

Two recent complementary advances, one in knowledge and one in technology, are greatly facilitating the study of gene expression and the discovery of the roles played by specific genes in the development of disease. As a result of the Human Genome Project, there has been an explosion in the amount of information available about the DNA sequence of the human genome. Consequently, researchers have identified a large number of novel genes within these previously unknown sequences. The challenge currently facing scientists is to find a way to organize and catalog this vast amount of information into a usable form. Only after the functions of the new genes are discovered will the full impact of the Human Genome Project be realized.

The second advance may facilitate the identification and classification of this DNA sequence information and the assignment of functions to these new genes: the emergence of **DNA microarray technology**. A microarray works by exploiting the ability of a given mRNA molecule to bind specifically to, or **hybridize** to, the DNA template from which it originated. By using an array containing many DNA samples, scientists can determine, in a single experiment, the expression levels of hundreds or thousands of genes within a cell by measuring the amount of mRNA bound to each site on the array. With the aid of a computer, the amount of mRNA bound to the spots on the microarray is precisely measured, generating a profile of gene expression in the cell.

A **microarray** is a tool for analyzing gene expression that consists of a small membrane or glass slide containing samples of many genes arranged in a regular pattern.

Why Are Microarrays Important?

Microarrays are a significant advance both because they may **contain a very large number of genes** and because of their **small size**. Microarrays are therefore useful when one wants to survey a large number of genes quickly or when the sample to be studied is small. Microarrays may be used to assay gene expression within a single sample or to compare gene expression in two different cell types or tissue samples, such as in healthy and diseased tissue. Because a microarray can be used to examine the expression of hundreds or thousands of genes at once, it promises to revolutionize the way scientists examine gene expression. This technology is still considered to be in its infancy; therefore, many initial studies using microarrays have represented simple surveys of gene expression profiles in a variety of cell types. Nevertheless, these studies represent an important and necessary first step in our understanding and cataloging of the human genome.

As more information accumulates, scientists will be able to use microarrays to ask increasingly complex questions and perform more intricate experiments. With new advances, researchers will be able to infer probable functions of new genes based on similarities in expression patterns with those of known genes. Ultimately, these studies promise to expand the size of existing gene families, reveal new patterns of coordinated gene expression across gene families, and uncover entirely new categories of genes. Furthermore, because the product of any one gene usually interacts with those of many others, our understanding of how these genes coordinate will become clearer through such analyses, and precise knowledge of these inter-relationships will emerge. The use of microarrays may also speed the identification of genes involved in the development of various diseases by enabling scientists to examine a much larger number of genes. This technology will also aid the examination of the integration of gene expression and function at the cellular level, revealing how multiple gene products work together to produce physical and chemical responses to both static and changing cellular needs.

What Exactly Is a DNA Microarray?

DNA Microarrays are small, solid supports onto which the sequences from thousands of different genes are immobilized, or attached, at fixed locations. The supports themselves are usually **glass microscope slides**, the size of two side-by-side pinky fingers, but can also be **silicon chips** or **nylon membranes**. The DNA is printed, spotted, or actually synthesized directly onto the support.
The American Heritage Dictionary defines "**array**" as "to place in an orderly arrangement". It is important that the gene sequences in a microarray are attached to their support in an orderly or fixed way, because a researcher uses the location of each spot in the array to identify a particular gene sequence. The spots themselves can be DNA, cDNA, or **oligonucleotides**.

An **oligonucleotide**, or oligo as it is commonly called, is a short fragment of a single-stranded DNA that is typically 5 to 50 nucleotides long.

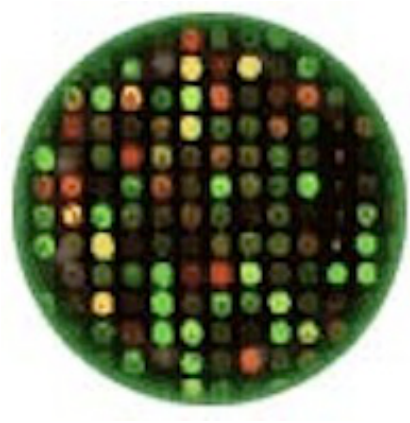
Designing a Microarray Experiment: The Basic Steps

One might ask, how does a scientist extract information about a disease condition from a dime-sized glass or silicon chip containing thousands of individual gene sequences? The whole process is based on **hybridization probing**, a technique that uses fluorescently labeled nucleic acid molecules as "**mobile probes**" to identify **complementary molecules**, sequences that are able to base-pair with one another. Each single-stranded DNA fragment is made up of four different nucleotides, adenine (A), thymine (T), guanine (G), and cytosine (C), that are linked end to end. Adenine is the **complement** of, or will always pair with, thymine, and guanine is the complement of cytosine. Therefore, the complementary sequence to G-T-C-C-T-A will be C-A-G-G-A-T. When two complementary sequences find each other, such as the immobilized target DNA and the mobile probe DNA, cDNA, or mRNA, they will lock together, or **hybridize**.

Now, consider two cells: cell type 1, a healthy cell, and cell type 2, a diseased cell. Both contain an identical set of four genes, **A, B, C,** and **D**. Scientists are interested in determining the expression profile of these four genes in the two cell types. To do this, scientists isolate mRNA from each cell type and use this mRNA as templates to generate cDNA with a "**fluorescent tag**" attached. Different tags (red and green) are used so that the samples can be differentiated in subsequent steps. The two labeled samples are then mixed and incubated with a microarray containing the immobilized genes **A, B, C,** and **D**. The labeled molecules bind to the sites on the array corresponding to the genes expressed in each cell.

A DNA Microarray Experiment
1. Prepare your DNA chip using your chosen target DNAs.
2. Generate a hybridization solution containing a mixture of fluorescently labeled cDNAs.
3. Incubate your hybridization mixture containing fluorescently labeled cDNAs with your DNA chip.
4. Detect bound cDNA using laser technology and store data in a computer.
5. Analyze data using computational methods.

After this hybridization step is complete, a researcher will place the microarray in a "reader" or "**scanner**" that consists of some lasers, a special microscope, and a camera. The fluorescent tags are excited by the laser, and the microscope and camera work together to create a digital image of the array. These data are then stored in a computer, and a special program is used either to calculate the red-to-green fluorescence ratio or to subtract out background data for each microarray spot by analyzing the digital image of the array. If calculating ratios, the program then creates a table that contains the ratios of the intensity of red-to-green fluorescence for every spot on the array. For example, using the scenario outlined above, the computer may conclude that both cell types express **gene A** at the same level, that cell 1 expresses more of **gene B**, that cell 2 expresses more of **gene C**, and that neither cell expresses **gene D**. But remember, this is a simple example used to demonstrate key points in experimental design. Some microarray experiments can contain up to 30,000 target spots. Therefore, the data generated from a single array can mount up quickly.



The Colors of a Microarray
Reproduced with permission from the Office of Science Education, the National Institutes of Health.

In this schematic: **GREEN** represents Control DNA, where either DNA or cDNA derived from normal tissue is hybridized to the target DNA. **RED** represents Sample DNA, where either DNA or cDNA is derived from diseased tissue hybridized to the target DNA. **YELLOW** represents **a** combination of Control and Sample DNA, where both hybridized equally to the target DNA. **BLACK** represents areas where neither the Control nor Sample DNA hybridized to the target DNA.

Each spot on an array is associated with a particular gene. Each color in an array represents either healthy (control) or diseased (sample) tissue. Depending on the type of array used, the location and intensity of a color will tell us whether the gene, or mutation, is present in either the control and/or sample DNA. It will also provide an estimate of the expression level of the gene(s) in the sample and control DNA.

Types of Microarrays

There are three basic types of samples that can be used to construct DNA microarrays, two are

genomic and the other is "**transcriptomic**", that is, it measures mRNA levels. What makes them different from each other is the kind of immobilized DNA used to generate the array and, ultimately, the kind of information that is derived from the chip. The target DNA used will also determine the type of control and sample DNA that is used in the hybridization solution.

I. Changes in Gene Expression Levels

Determining the level, or volume, at which a certain gene is expressed is called **microarray expression analysis**, and the arrays used in this kind of analysis are called "**expression chips**". The immobilized DNA is cDNA derived from the mRNA of known genes, and once again, at least in some experiments, the control and sample DNA hybridized to the chip is cDNA derived from the mRNA of normal and diseased tissue, respectively. If a gene is overexpressed in a certain disease state, then more sample cDNA, as compared to control cDNA, will hybridize to the spot representing that expressed gene. In turn, the spot will fluoresce red with greater intensity than it will fluoresce green. Once researchers have characterized the expression patterns of various genes involved in many diseases, cDNA derived from diseased tissue from any individual can be hybridized to determine whether the expression pattern of the gene from the individual matches the expression pattern of a known disease. If this is the case, treatment appropriate for that disease can be initiated.

As researchers use expression chips to detect **expression patterns**— whether a particular gene(s) is being expressed more or less under certain circumstances—expression chips may also be used to examine changes in gene expression over a given period of time, such as within the **cell cycle**. The cell cycle is a molecualr network that determines, in the normal cell, if the cell should pass through its life cycle. There are a variety of genes involved in regulating the stages of the cell cycle. Also built into this network are mechanisms designed to protect the body when this system fails or breaks down because of mutations within one of the "**control genes**", as is the case with **cancerous cell growth**. An expression microarray "experiment" could be designed where cell cycle data are generated in multiple arrays and referenced to time "zero". Analysis of the collected data could further elucidate details of the cell cycle and its "**clock**", providing much needed data on the points at which gene mutation leads to cancerous growth as well as sources of therapeutic intervention.

In the same way, expression chips can be used to develop new drugs. For instance, if a certain gene is overexpressed in a particular form of cancer, researchers can use expression chips to see if a new drug will reduce overexpression and force the cancer into remission. Expression chips could also be used in disease diagnosis as well, e.g., in the identification of new genes involved in environmentally triggered diseases, such as those diseases affecting the immune, nervous, and pulmonary/respiratory systems.

II. Genomic Gains and Losses

DNA repair genes are thought to be the body's frontline defense against mutations and, as such, play a major role in cancer. Mutations within these genes often manifest themselves as lost or broken chromosomes. It has been hypothesized that certain chromosomal gains and losses are related to cancer progression and that the patterns of these changes are relevant to clinical prognosis. Using different laboratory methods, researchers can measure gains and losses in the copy number of chromosomal regions in tumor cells. Then, using mathematical models to analyze these data, they can predict which chromosomal regions are most likely to harbor important genes for tumor initiation and disease progression. The results of such an analysis may be depicted as a hierarchical treelike branching diagram, referred to as a "**tree model of tumor progression**".

Researchers use a technique called microarray Comparative Genomic Hybridization **(CGH)** to look for genomic gains and losses or for a change in the number of copies of a particular gene involved in a disease state. In microarray CGH, large pieces of genomic DNA serve as the target DNA, and each

spot of target DNA in the array has a known chromosomal location. The hybridization mixture will contain fluorescently labeled genomic DNA harvested from both normal (**control**) and diseased (**sample**) tissue.

Therefore, if the number of copies of a particular target gene has increased, a large amount of sample DNA will hybridize to those spots on the microarray that represent the gene involved in that disease, whereas comparatively small amounts of control DNA will hybridize to those same spots. As a result, those spots containing the disease gene will fluoresce red with greater intensity than they will fluoresce green, indicating that the number of copies of the gene involved in the disease has gone up.

III. Mutations in DNA

When researchers use microarrays to detect mutations or polymorphisms in a gene sequence, the target, or immobilized DNA, is usually that of a single gene. In this case though, the target sequence placed on any given spot within the array will differ from that of other spots in the same microarray, sometimes by only one or a few specific nucleotides. One type of sequence commonly used in this type of analysis is called a **Single Nucleotide Polymorphism**, or **SNP**, a small genetic change or variation that can occur within a person's DNA sequence. Another difference in **mutation microarray analysis**, as compared to expression or CGH microarrays, is that this type of experiment only requires genomic DNA derived from a normal sample for use in the hybridization mixture.

Once researchers have established that a SNP pattern is associated with a particular disease, they can use SNP microarray technology to test an individual for that disease expression pattern to determine whether he or she is susceptible to (at risk of developing) that disease. When genomic DNA from an individual is hybridized to an array loaded with various SNPs, the sample DNA will hybridize with greater frequency only to specific SNPs associated with that person. Those spots on the microarray will then fluoresce with greater intensity, demonstrating that the individual being tested may have, or is at risk for developing, that disease.

In Brief: Microarray Applications	
Microarray type	**Application**
CGH	Tumor classification, risk assessment, and prognosis prediction
Expression analysis	Drug development, drug response, and therapy development
Mutation/Polymorphism analysis	Drug development, therapy development, and tracking disease progression

NCBI and Microarray Data Management

Why is it necessary to have a uniform system that will manage and provide a disbursement point for microarray data? Consider the amount of data that can potentially be generated using a single microarray chip. Suppose that chip contains 30,000 spots of target DNA. Researchers interpreting the data generated by that chip would need to know the biological identity of each target—what gene is where; the biological properties of the control and sample DNA; the experimental conditions and procedures used in setting up the experiment; and finally, the results. Although experiments such as these will undoubtedly push forward our current understanding of gene expression and regulation, many new challenges are presented in terms of data tracking and analysis.

What Is GEO?

As we have just alluded, microarray technology is one of the most recent and important experimental breakthroughs in molecular biology. Today, proficiency in generating data is fast

overcoming the capacity for storing and analyzing it. Much of this information is scattered across the Internet or is not even available to the public. As more laboratories acquire this technology, the problem will only get worse. This avalanche of data requires standardization of storage, sharing, and publishing techniques.

To support the public use and dissemination of **gene expression data**, NCBI has launched the Gene Expression Omnibus, or **GEO**. GEO represents NCBI's effort to build an expression data repository and online resource for the storage and retrieval of gene expression data from any organism or artificial source. Many types of gene expression data, such as those types discussed in this primer, are accepted and archived as a public dataset.

Developing MAML: Reading Off the Same Platform

Microarray Markup Language, developed by the "**MAML**" working group of **MGED**, the Microarray Gene Expression Database, is a first attempt to provide a standard platform for submitting and analyzing the enormous amounts of microarray expression data generated by different laboratories around the world. The goal of this group, which includes NCBI investigators, is to facilitate the adoption of standards for DNA-array experiment annotation and data representation, as well as the introduction of standard experimental controls and data normalization methods. The underlying goal is to facilitate the establishment of gene expression data repositories, the comparability of gene expression data from different sources, the interoperability of different gene expression databases, and data analysis software.

MAML proposes a framework for describing information about a DNA-array experiment and a data format for communicating this information, including details about:

- **Experimental design**: the set of the hybridization experiments as a whole
- **Array design**: each array used and each spot on the array
- **Samples**: samples used, the extract preparation, and labeling
- **Hybridizations**: procedures and parameters
- **Measurements**: images, quantitation, and specifications
- **Controls**: types, values, and specifications

MAML is independent of the particular experimental platform and provides a framework for describing experiments done on all types of DNA arrays, including spotted and synthesized arrays, as well as oligo and cDNA arrays. What's more, MAML provides format to represent microarray data in a flexible way, which allows analysis of data obtained from not only any existing microarray platforms but also many of the possible future variants, including protein arrays.

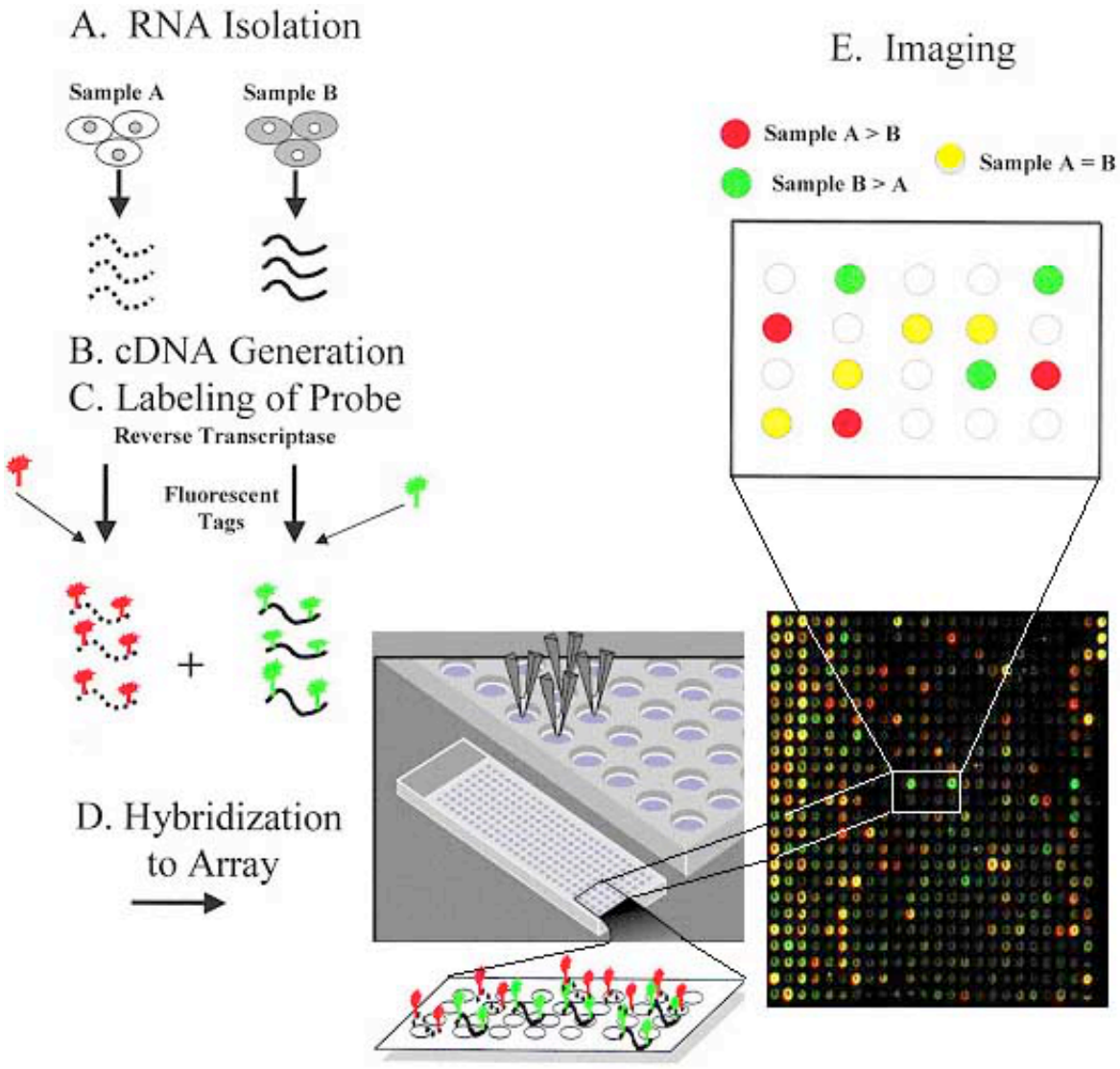
Although the data in GEO are not currently provided in MAML format, it is NCBI's goal to have the data delivered in a number of formats, including MAML, soon to be replaced by a more recent version called **MAGEML** (MicroArray Gene Expression Markup Language).

The Benefits of GEO and MAML

- By storing vast amounts of data on gene expression profiles derived from multiple experiments using varied criteria and conditions, GEO will aid in the study of functional genomics—the development and application of global experimental approaches to assess gene function
- GEO will facilitate the cross-validation of data obtained using different techniques and technologies and will help set benchmarks and standards for further gene expression studies
- By making the information stored in GEO publicly available, the fields of bioinformatics and functional genomics will be both promoted and advanced
- That such experimental data should be freely accessible to all is consistent with NCBI's legislative mandate and mission: to develop new information technologies to aid in the understanding of fundamental molecular and genetic processes that control health and disease

The Promise of Microarray Technology in Treating Disease

Now that you understand the concept behind array technology, picture this: a hand-held instrument that a physician could use to quickly diagnose cancer or other diseases during a routine office visit. What if that same instrument could also facilitate a personalized treatment regimen, exactly right for you? Personalized drugs. Molecular diagnostics. Integration of diagnosis and therapeutics. These are the long-term promises of microarray technology. Maybe not today or even tomorrow, but someday. For the first time, arrays offer hope for obtaining global views of biological processes—simultaneous readouts of all the body's components—by providing a systematic way to survey DNA and RNA variation. NCBI, by continuing its efforts to provide a standard format for microarray data and to provide free, universal access to that data, will help the scientific community in making those promises realities.



The microarray is scanned with a laser beam, first at one wavelength to collect fluorescence data representing one probe, then is scanned at a second wavelength to collect data representing the second probe. A computer compares the amount of fluorescence at each spot on the microarray for each probe. Through the use of computer software, the ratio of fluorescence is obtained and correlated with the clone address so the investigator knows which gene (spot on the slide) was expressed more in treated tissue as compared to control tissue.