

## PART 1: Basic Concepts in Cancer Genetics

### Chapter 21.1: Gene Expression Profiling in Cancer

Gregory J. Riggins, Patrice J. Morin

#### Abstract

**Note :** *This chapter last appeared in OMMBID Mark II, released in January 2006. It is now archived and should not be considered current.*

1. Gene expression profiling is a powerful new approach for viewing the expression of many genes simultaneously in different types of malignant or normal cells. Using computational approaches, differentially expressed genes or informative patterns of expressed genes are mined from large data sets produced by new expression profiling technology. This technology yields the opportunity to classify tumors by gene expression and to locate genes of diagnostic or therapeutic importance.
2. DNA array technology uses thousands of DNA fragments arrayed on a solid surface in order to probe many messenger ribonucleic acid (mRNA) levels in one experiment. Both oligonucleotides and portions of complementary deoxyribonucleic acid (cDNA) are used as hybridization probes on the arrays. Clustering and other statistical based algorithms are used to locate patterns of gene expression of importance when analyzing large numbers of RNA samples.
3. Serial Analysis of Gene Expression (SAGE) is a sequencing based technology that provides an in-depth quantitative assessment of gene expression. SAGE works by counting transcripts and storing digital values electronically, providing absolute gene expression levels that make historical comparisons and databasing facile. It is useful for studying small numbers of tissue or cellular samples derived from well-controlled experiments.
4. Gene expression profiling techniques have been used to obtain global gene expression patterns from several common malignancies and corresponding normal tissues. These studies highlight the potential of gene expression profiling in cancer taxonomy, and in the identification of molecular targets for diagnosis and therapy.
5. Gene expression profiling has been used in the dissection of specific oncogenic molecular pathways, including the p53 tumor-suppressor pathway, the APC/ $\beta$ -catenin pathway, and numerous *in vitro* and *in vivo* models of angiogenesis and cancer drug resistance.

The ability to determine gene expression levels from thousands of genes simultaneously has recently transformed many aspects of cancer research. Large-scale gene expression profiling provides a powerful means to create an overall view of how the genome provides instructions to the cell. Ultimately, the genetic background, mutations, environment, and history of the cell all impact on mRNA and subsequent protein expression. Unlike the positional-cloning approaches that during the last decade revealed the genes mutated during oncogenesis, gene expression profiling does not directly reveal cancer-causing genes, but the pattern of genes used by the malignant (or normal) cell. These patterns, and the differentially expressed genes found within these patterns, have a variety of important uses for improved clinical correlation or therapy design. This chapter reviews the major mRNA profiling techniques and how they are applied to the study of cancer.

Technology advances make research advances possible. Just as the invention of the first compound microscope allowed biologists to view cellular patterns in tissues, the recent advent of gene expression technologies allows the biologist to observe molecular patterns in cells. Although protein levels are the

ultimate goal for many uses, large numbers of protein levels cannot be assayed as rapidly as RNA, and this chapter is mostly limited to reviewing RNA profiling methods. High-throughput RNA expression-profiling techniques can be broadly divided into two categories: methods that count transcripts using DNA sequencing and methods that are based on DNA hybridization.

Sequencing is used for profiling to count transcripts from a cDNA library. The relative levels of mRNA can be preserved when reverse-transcribed into cDNA and cloned into a collection of plasmids forming a cDNA library. Automated DNA sequencing<sup>1, 2</sup> makes it possible to infer RNA levels of many genes by sequencing cDNA molecules, or fragments of cDNAs, in large numbers.

Alternatively, one can discriminate between different mRNA transcripts by hybridization to a nucleic acid probe of known sequence. Hybridization can be used to either detect mRNA levels by arraying the gene-specific probes on a solid surface, or to subtract sequences from a sample followed by detection of the remaining sequence. By containing many probes arrayed into a small area, DNA Arrays<sup>3-5</sup> can be used to detect thousands of genes simultaneously. Advanced computational methods enable biologists to look for genes with significant differences in expression, or those genes that cluster according to a particular feature.

Collectively all of the expressed mRNA transcripts from a cell are known as the “transcriptome.” Significant portions of the transcriptome can be assayed for a given cell population, but the large data sets that are produced by these technologies create new challenges. Expression profiling requires the ability to database and effectively interpret this information. Sophisticated computational and statistical approaches, either new or derived from approaches formally applied to the physical sciences, are now required to interpret the datasets. Other bioinformatics approaches are necessary to draw on the vast and growing archive of information available through public databases or the biomedical literature. Correlating expression levels in malignant cells with the information derived from the recently sequenced human genome is a particularly important example. The genome sequence provides a catalyst for “functional genomic” approaches, such as RNA expression profiling.

Although gene expression profiling techniques are far from reaching their full potential, there have already been several important applications of the technology. Expression profiling has been a useful means for finding the cancer-related genes whose expression levels are dependent on a known oncogene or by the loss of a tumor suppressor. RNA profiling has also been successfully applied to locating cell-type specific genes or gene-expression changes that depend on environmental factors, including drug or hormonal exposures. Even without knowing the function of the genes detected by profiling, the overall pattern of genes has been used to help classify tumors by malignant potential or response to therapy. Expression profiling holds the promise of revealing a much more complete picture of the molecular pathways within the malignant cell. However, this data is only the first step for better understanding, diagnosis, and treatment of cancer.

## TECHNIQUES FOR EXPRESSION PROFILING

Many techniques have been developed to find those transcripts whose expression level changes between two samples. The first techniques to be widely used to find differentially expressed transcripts were subtractive hybridization and differential display methods. Both could identify transcripts but do not have the same capacity to assay multiple samples like DNA arrays, nor do they provide an in-depth transcriptome characterization of sequencing-based techniques. For this reason, DNA arrays and SAGE are currently the most widely used for transcript profiling of malignant cells. This is, however, a rapidly evolving field. The overview of the features of common RNA profiling techniques (Table 21.1-1) will likely

require significant updating in the not so distant future.

**Table 21.1-1: Key Features of Common RNA Profiling Techniques**

	Differential Display	DNA Arrays	SAGE
Basis of assay	cDNA fragments compared on gel	Hybridization to spotted DNA	Sequence ligated cDNA tags
Detection method	Electrophoresis of labeled fragments	Optical imaging of hybridization signal	Automated sequencer
Gene identification	Excise band and sequence	DNA probes preidentified	Match SAGE tag to database(s)
Transcript quantification	Comparison of band intensities	Analogue fluorescent signal from DNA spot	Digital counts of SAGE tags
Probe requirements	Starting RNA only	Requires set of arrayed DNA probes	Starting RNA only.
Starting amounts of RNA (approx.)	>5 µg of total RNA	>5 µg of total RNA	>1 µg of total RNA
Number of RNA samples that can be processed per month	Few	Many	Few
Number of genes assayed per sample	Few	Equal to the number of genes on the array	Most all genes expressed above the detection limit
Sensitivity of transcript detection	Higher levels easier to detect	~10 mRNA copies/cell	Dependent on number of tags sequenced

### Subtraction Methods

Various methods have been derived to find transcripts that are differentially expressed between two different cell populations.<sup>6</sup> Subtractive hybridization is used to produce a cDNA library that has sequences that are present in one sample of RNA, but not another.<sup>7, 8</sup> A typical example is to subtract tumor mRNA from normal mRNA (or *vice versa*) to find transcripts that may have been deleted or amplified in the process of tumor formation. The general approach is to hybridize in solution the two samples (normally cDNA) that are to be subtracted. An excess of one sample (the “driver”) hybridizes to most all the unwanted common sequences from the other sample (the “tester” or “tracer”). Typically the driver is labeled in such a way that molecules containing one or both strands in common with the driver are removed or otherwise not cloned. The remaining cDNA consisting mostly of tracer can be cloned to form a library for further analysis such as sequencing.

Other subtractive methods used to find differentially expressed genes include suppression subtractive hybridization<sup>9</sup> and representation difference analysis (RDA).<sup>10</sup> These newer techniques incorporate polymerase chain reaction (PCR) amplification steps in order to work from smaller quantities of starting material. RDA is an effective way to compare two sets of DNA by hybridization and subtraction, frequently either genomic DNA or cDNA. Overall, the subtractive techniques have been used to locate many important cancer-related genes, but these approaches necessitate a pair-wise analysis of samples and a time-consuming cloning step that make them unsuitable for automated high-throughput gene expression profiling.

### Differential Display

In 1992, differential display was described as a method to locate differentially expressed transcripts.<sup>11, 12</sup> Differential display works by first producing a set of cDNA fragments that have been identically prepared from each RNA sample, usually based on restriction enzyme digestion of the cDNA or by producing PCR products with arbitrary primers. Next, the fragments are resolved on a gel, producing a characteristic pattern of bands for each sample. The bands from each sample are compared to reveal those bands that differ in intensity between lanes. The cDNA fragment within the band can be excised for further analysis.

The advantage of differential display is that it is performed by one person using equipment available in most molecular biology laboratories. The disadvantage of this technique is that in order to identify most genes they must be excised and sequenced—requiring significant labor for the gene identification step. Also, the technique can be prone to false positives that arise from various factors, including PCR-induced amplification biases. Although there are many successful variations of the differential display approach,<sup>13</sup> differential display approaches do not allow for rapid and efficient identification of expression levels *en masse* that makes it suitable as a transcript “profiling” technique. It has, however, been very useful for identifying differentially expressed genes.

### DNA Arrays

A method to detect nucleic acids of a specific sequence supported by a solid surface was developed over 25 years ago by Edwin M. Southern.<sup>14</sup> In 1992, cDNA fragments were arrayed on a solid surface in large numbers and used for parallel gene expression profiling.<sup>5</sup> The idea of large-scale transcript profiling captured the imagination of scientists starting in the mid 1990s, when methods were used to miniaturize DNA arrays;<sup>3, 4, 15</sup> introducing “chip technology” to biological research. DNA arrays have enormous potential and have an implicit promise that a reliable, low-cost, and standardized format for gene expression profiling will eventually be available to cancer and other researchers. By means of introduction, this section describes basic concepts and readers interested in applying chip technology should consult relevant publications<sup>16–22</sup> and Web sites (Table 21.1-2).

**Table 21.1-2: Human Transcript Profiling Databases and Resources**

Web Site	URL	Description
cDNA Library sequencing		
Body Map	<a href="http://bodymap.ims.u-tokyo.ac.jp">bodymap.ims.u-tokyo.ac.jp</a>	Expression resources for normal tissues based on cDNA library sequencing.

CGAP cDNA xProfiler	<a href="http://cgap.nci.nih.gov/CGAP/Tissue%20s/xProfiler">cgap.nci.nih.gov/CGAP/Tissue s/x Profiler</a>	Expression between pools of cDNA libraries can be compared based on extensive database.
DNA Arrays		
Affymetrix	<a href="http://www.affymetrix.com">www.affymetrix.com</a>	Vendor of expression chips, and other products for expression profiling via DNA arrays.
Brown Lab Homepage	<a href="http://cmgm.stanford.edu/pbrown">cmgm.stanford.edu/pbrown</a>	Contains useful information on custom cDNA arrays.
Developmental Therapeutics	<a href="http://www.dtp.nci.nih.gov">www.dtp.nci.nih.gov</a>	Microarray and drug response data for NCI 60 cell lines.
Genexpress - CNRS	<a href="http://idefix.upr420.vjf.cnrs.fr/EXPR/">idefix.upr420.vjf.cnrs.fr/EXPR/</a>	Expression profile of 5,058 human genes by cDNA array.
Microarray Project	<a href="http://www.nhgri.nih.gov/DIR/Microarray/">www.nhgri.nih.gov/DIR/Microarray/</a>	Protocols, descriptions and resources for cDNA Microarray technology from the NHGRI.
Molecular Oncology and Development	<a href="http://chroma.mbt.washington.edu/mod_www/">chroma.mbt.washington.edu/mod_www/</a>	Protocols and links for DNA arrays from Hood lab at University of Washington.
Molecular Pattern Recognition	<a href="http://waldo.wi.mit.edu/MPR">waldo.wi.mit.edu/MPR</a>	Protocols, links, software and downloads for DNA arrays from Whitehead/MIT.
UCSD Array Science	<a href="http://array.ucsd.edu">array.ucsd.edu</a>	Information and Bioinformatics Tools for expression information.
SAGE		
SAGEmap	<a href="http://www.ncbi.nlm.nih.gov/SAGE">www.ncbi.nlm.nih.gov/SAGE</a>	Large RNA expression database from CGAP based on SAGE profiles of malignant and normal cells.
SAGEnet (Johns Hopkins)	<a href="http://www.sagenet.org">www.sagenet.org</a>	SAGE database, protocols, references and links.
Genzyme Molecular Oncology	<a href="http://www.genzyme.com/sage/welcome.htm">www.genzyme.com/sage/welcome.htm</a>	SAGE information and applications for commercial users of the technology.
Other		
Cancer Genome Anatomy Project (CGAP)	<a href="http://cgap.nci.nih.gov/">cgap.nci.nih.gov/</a>	CGAP homepage with links to expression databases and cancer research resources.

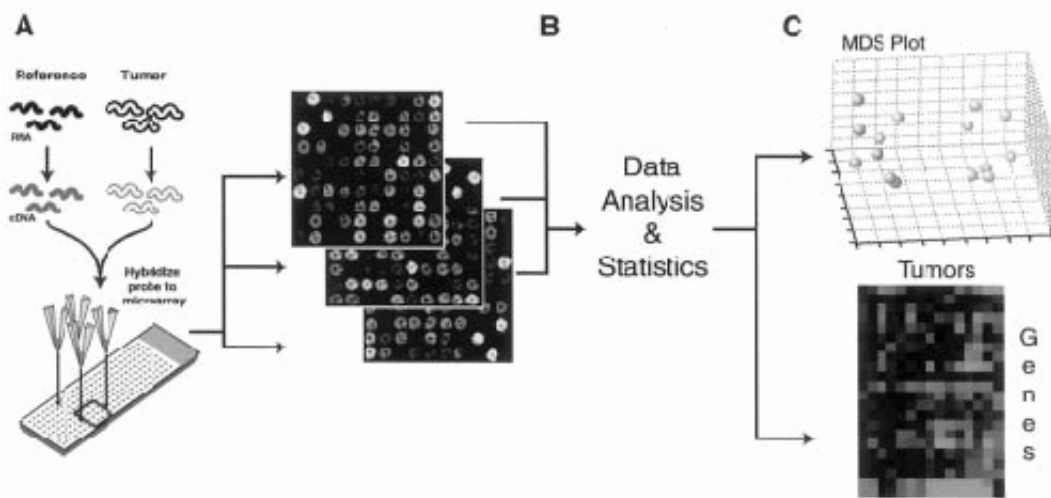
Digital Gene Expression Displayer (DGED)	<a href="http://cgap.nci.nih.gov/CGAP/Tissues/GXS">cgap.nci.nih.gov/CGAP/Tissues/GXS</a>	CGAP tool that compares gene expression between pools of SAGE and/or cDNA libraries.
Gene Expression Omnibus (GEO)	<a href="http://www.ncbi.nlm.nih.gov/geo">www.ncbi.nlm.nih.gov/geo</a>	NCBI repository and comparison interface for all types of expression data.
Tissue Microarray Project	<a href="http://www.nhgri.nih.gov/DIR/CGB/TMA">www.nhgri.nih.gov/DIR/CGB/TMA</a>	Protocols and information on tissue microarrays from the NHGRI.

### cDNA Arrays.

There are many variations of DNA arrays, but they can be viewed in two groups: those that array a fragment of cDNA and those that array a shorter synthetic DNA oligonucleotide. Arraying cDNAs on a membrane for hybridization with a labeled sample was the first DNA array approach (filter arrays), and it is still widely used today.<sup>5, 23</sup> Typically, hundreds to thousands of cDNA fragments are amplified by PCR and spotted densely onto a membrane. An RNA or cDNA test sample is then radioactively labeled and hybridized to the targets on the membrane. Expression levels are accessed by the signal intensity produced by the amount of radioactivity hybridized to each probe on the membrane. Several molecular biology companies sell membranes to researchers for use in their studies and services for doing the hybridization and/or analyzing the results.

The technology of cDNA arrays took another leap forward when researchers at Stanford University started spotting cDNA onto glass slides at densities much greater than what could be achieved with nylon membranes.<sup>4</sup> The introduction of cDNA “microarrays” has opened the minds of scientists to the possibility that gene expression patterns could be routinely measured.<sup>24</sup> Robotics are employed for making these arrays that can reproducibly spot well over 5000 cDNAs on a single slide.<sup>25</sup> An additional advancement is the use of two-color hybridization (Fig. 21.1-1). Two different-colored fluorescent probes, typically red and green, are made from the test and control sample and hybridized to the same array. Each spot on the array is measured in terms of the expression ratio between probes, rather than an absolute level of expression. This approach helps to normalize array-to-array variations in hybridization or printing and provides a more accurate means of comparing expression between chips. However, this approach does result in the loss of the absolute expression levels since a ratio is being measured. A variety of commercial enterprises make and sell cDNA arrays or services. Additionally most universities have, or are, developing some type of service that provides cDNA arrays technology to investigators.

Fig. 21.1-1:



Approach for expression profiling using a two-color cDNA chip. A. Two RNA samples are converted to cDNA and are labeled with different fluorescent dyes; the tumor sample is labeled red and the normal reference is labeled green. The labeled cDNA is hybridized to the DNA on the microarray. B. Each DNA spotted on the array is a cDNA fragment that represents a gene. The relative ratio of the gene at each spot is determined by the color after hybr...

### Oligonucleotide Arrays.

Oligonucleotides built on a glass support by photolithography and phosphoramidite DNA synthesis chemistry are commonly known as "DNA Chips."<sup>15, 26</sup> This process builds a chip for DNA analysis in a method that is analogous to the mass production of semiconductor chips for the electronics industry. DNA Microchip technology has been developed and commercialized primarily by Affymetrix Corporation. Normally about 20 different oligonucleotides of approximately 20 base pairs (bp) in length are used to represent each gene on mRNA expression chips. The oligonucleotide sequences that represent a particular gene are chosen carefully using algorithms that have been designed to minimize cross-hybridization between different genes. After hybridization with a specially prepared and fluorescently labeled cDNA probe, the chip is read using a laser scanner. Currently, Affymetrix has produced a series of chips that will cover up to a total of 12,000 different known genes plus 48,000 sequences derived from expressed sequence tag (EST) clusters. Thus far, oligonucleotide chips have delivered a more standardized product than the cDNA spotted chips, along with a preoptimized and working infrastructure for array analysis, but at a cost. Although costs for chips have declined, they are still considerably more than glass slide systems.

Oligonucleotides for use in DNA arrays are not limited to Affymetrix chips. Longer oligonucleotides, typically more than 50 bp, can be arrayed by robotic spotting onto glass slides and used in place of PCR fragments amplified from a cDNA template. The choice of arrayed material and support is usually based on what is locally available. It is expected that as the technology continues to evolve, market competition will produce one or more dominant DNA array technologies that will deliver reliability and convenience at a reasonable cost.

### cDNA Library Sequencing

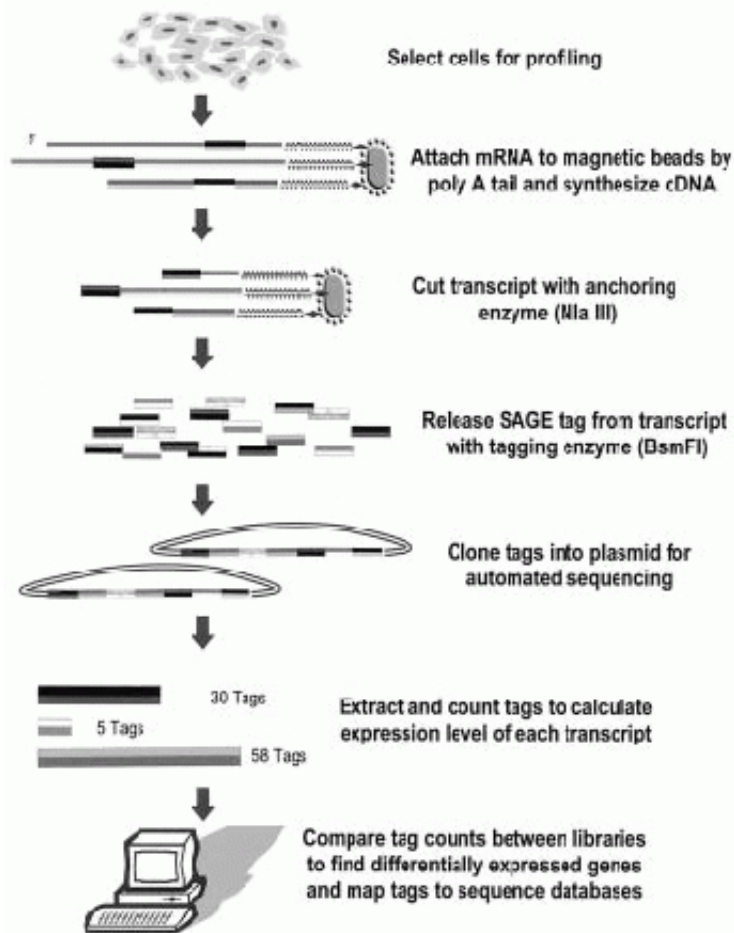
Large-scale sequencing of cDNA libraries was first proposed as a rapid means to access transcribed regions from the human genome.<sup>27</sup> Random transcribed sequences generated by cDNA library sequencing are known as expressed sequence tags (ESTs). The Merck/Washington University EST project made one of the first large-scale efforts to disseminate EST sequence data.<sup>28</sup> The Cancer Genome Anatomy Project (CGAP)<sup>29–31</sup> succeeded this effort with its Tumor Gene Index, contributing more than one million ESTs from normal, premalignant, and malignant cells. The data from these projects has greatly reduced the time and effort necessary for many gene-cloning projects, but also serves to reveal which tissues express which transcripts.

Counting transcripts by EST sequencing is a very accurate way of accessing the fractional representation of each transcript, but it is a very expensive and laborious approach. Consequently, expression levels derived from EST data are normally derived from the large public EST sequencing projects. EST-based expression data can be accessed from many of these projects via the World Wide Web as described in the Bioinformatics section below (see Table 21.1-2 for Web sites). The main advantage of this data is that it is free and easily accessed. The main disadvantages are that the individual experimenter cannot practically generate his own EST data and that the level of detection is low, because often only a few thousand transcripts are assayed for each tissue or cell type, out of the tens of thousands expressed. One must also keep in mind that cDNA libraries used to generate EST data are frequently normalized or subtracted, and that data derived from these libraries can only reveal the presence of a transcript and not quantitative expression levels.

### Serial Analysis of Gene Expression

SAGE was first developed in 1995,<sup>32</sup> as a means for efficient counting of mRNA transcripts in large numbers.<sup>33–38</sup> SAGE increases the number of genes that can be counted per sequencing reaction, as compared to cDNA library sequencing, by minimizing the portion of the transcript sequenced. The method (Fig. 21.1-2) works by cloning and sequencing a 10-bp portion of the cDNA at a defined position near the 3' end of the transcript. These 10 base pairs, normally next to the last *Nla* III restriction site, are known as the transcript "tags." The transcript tags from a particular RNA sample are linked together and are cloned into a sequencing vector forming a SAGE library. Automated sequencing then produces tag sequences rapidly in large numbers by the sequencing of many clones simultaneously. Typically, more than 50,000 transcript tags can be counted, with about 2000 sequencing reactions. Although sequencing costs increase proportionally with the number of tags assayed, automated sequencing has increased in efficiency and speed. The SAGE transcript profile from various types of cells can be archived on a computer database and electronically compared to find statistically significant differences in gene expression between cell types. The gene responsible for the differentially expressed tag is identified using informatics or, in rare instances, cloned using the tag sequence. The majority of tags can be matched to a list of possibilities extracted from transcript databases such as the cDNA portion of GenBank,<sup>34</sup> the EST clusters forming the NCBI UniGene database,<sup>39</sup> and coding sequence extracted from the human genome sequence.

Fig. 21.1-2:



Approach for expression profiling using SAGE. Gene expression is quantified in a population of cells by isolating a transcript tag from the expressed genes. These tags are paired into ditags, ligated to form concatamers, and cloned into a sequencing vector for efficient counting on an automated sequencer. Tag counts from each tissue type are stored electronically and used for comparison to other cell populations. A relative fraction of each ...

Because SAGE counts transcripts by sequencing and avoids the errors inherent in hybridization-based assays, it is often regarded as a very accurate means for expression profiling. SAGE transcript levels are expressed as a fraction of the total transcripts counted, not relative to another experiment or a housekeeping gene, avoiding error-prone normalization between experiments. The absolute nature of SAGE data makes cumulative data sets useful and historical comparisons valid.<sup>39, 40</sup> An additional strength of SAGE is that it determines expression levels directly from an RNA sample. It is not necessary to have a gene-specific fragment of DNA arrayed to assay each gene. This allows SAGE to identify genes that are not included in an array,<sup>41</sup> and avoids the infrastructure necessary to create and read large DNA arrays. This flexibility has a downside. The number of samples that can be processed using SAGE is small as compared to DNA arrays because it takes 2 weeks or more of skilled labor to construct a SAGE library. The potential to analyze hundreds of samples by SAGE for a single experiment is not a practical

option for the technology in its present form. However, when an in-depth and quantitative profile is desired for a small number of samples, the extra work involved in creating a SAGE library can be justified. To date, SAGE has been successful for determining the differentially expressed transcripts in well-controlled experimental systems.<sup>41–48</sup> This type of data generated by SAGE is often complementary to a typical use of DNA arrays in cancer research for a wide survey of many patient tumor samples.

### Follow-up Techniques

After a gene expression profile has been obtained on a set of RNA samples, it is desirable to experimentally confirm the expression differences and to extend the analysis to other samples. Normally, a small set of interesting genes is identified by using DNA arrays or SAGE, but several different techniques are more efficient for assaying this smaller set of interesting genes. In addition, each gene expression technique has inherent errors and an independent method is required for validating the original expression levels.

Northern blotting has been the gold standard for gene expression analysis for many years. Because the transcript being assayed is identified by both molecular weight and by a long hybridization probe, there is normally a low error rate. Although northern blotting is a time-consuming approach, it is still a useful way to confirm profiling data for a limited number of genes.<sup>49</sup> When a good antibody is available for the gene of interest, a western blot or immunohistochemistry are reliable methods for confirming expression changes. This approach is advantageous, particularly when the end point is knowledge of protein levels rather than mRNA levels.

Real-time PCR, sometimes called quantitative or fluorescent PCR, has gained popularity for rapid follow-up and confirmation of profiling data.<sup>50, 51</sup> Expression determination by real-time PCR is based on continuous fluorescent monitoring of PCR products<sup>52–54</sup> from a cDNA template. Under the right conditions, the number of cycles required to PCR amplify a product to a certain level is directly proportional to the amount as starting template. Different real-time PCR systems are available from at least four molecular biology vendors. Each of these systems has software for plotting and analyzing fluorescent-labeled PCR products' accumulation for the determination of starting concentration. Normally a serially diluted known sample is used for a standard curve to interpolate concentrations of unknown samples.

There are a variety of methods for detecting the accumulation of PCR products during real-time PCR. A simple method is to incorporate a fluorescent dye directly into the PCR product during amplification. A double-stranded DNA binding dye, SYBR green I (Molecular Probes, Eugene, OR), is effective for this purpose.<sup>52, 53</sup> To increase specificity of PCR product detection, additional oligonucleotide can be employed in the assays that hybridize to an internal portion of the PCR product. There are a variety of systems for this purpose marketed by different vendors: TaqMan Assay (PE Biosystems), Hybridization Probes (Roche), and Molecular Beacons (Stratagene). Real-time PCR allows for a quick and low-cost assessment of the expression pattern of several genes in many tumors and can be automated. It is becoming a popular method for the follow-up of profiling data.

To look at protein levels of many samples simultaneously a tissue microarray system has been developed.<sup>55–57</sup> This system allows for up to 1000 small tissue samples, made from a narrow gauge biopsy needle, to be arrayed in a single block of tissue. This block of tissue can then be used to produce hundreds of slides that can be probed by immunohistochemistry or other means. In this way, a standard set of the same samples can be probed for expression levels for many different genes. A digital imaging system is used to record and read the data. Although, robotics are now employed to array the tissues,

many good quality samples must be collected and oriented for biopsy in the region of interest oriented by a pathologist. The results must also be scored in some fashion by signal intensity, done manually at this point in the technology's development. Finally, a good antibody is needed for each gene of interest that will work in the normally available formalin-fixed tissue. However, this approach has the potential to make gene expression correlations with a vast archive of preserved tumor material.

## BIOINFORMATICS

High-throughput gene expression profiling has produced large data sets that require specialized tools for effective handling and analysis. The earlier RNA expression technologies—subtractive hybridization and differential display—were designed to locate small numbers of differentially expressed genes. With the advent of large-scale gene expression profiling techniques, such as DNA arrays and SAGE, a need has arisen to analyze gene expression level measurements in numbers that are orders of magnitude higher than previous.<sup>58</sup> Although DNA arrays and SAGE have different data formats, there are still elements in common for the data analysis. There is a need to be able to archive and sort through multiple measurements, set statistical confidence levels, recognize patterns within the data, and retrieve information rapidly regarding the detected genes. The large volume of data being generated primarily by DNA arrays has created gene expression informatics challenges that have produced some creative solutions in this rapidly expanding field of research. Perhaps good news for the individual cancer investigator is an increasing amount of publicly available gene expression data. Databases and data-mining approaches are making global expression profiles available to virtually any investigator.

### DNA Array Data Analysis

The majority of large-scale gene expression data is currently being generated in the form of images from fluorescent-labeled high-density DNA arrays. While there are many variations particular to each system, there are some features in common. The image from each spotted DNA on the array must be processed to yield a numeric gene expression level (or ratio between samples). Quality control is key at this step; poorly spotted arrays or other technical artifacts can be excluded at this stage. This initial image analysis step is critical and there are several available options for the signal processing required.<sup>59–61</sup> Most DNA array analysis software has some provision for normalizing the data, based on overall intensity, to account for differences in the probe amount, or other variables altering overall spot detection.

A popular method of analyzing gene expression information from DNA arrays is to cluster the data in a dendrogram or “tree” format.<sup>62, 63</sup> Clustering programs, now available from commercial sources as well, are based on a variety of statistical algorithms including self-organizing maps and K-means.<sup>64</sup> The genes can be clustered by their expression response to experimental conditions, or the RNA samples used for the array experiment can be clustered by overall similarity of expression patterns. Microarray experiments may address the similarity of expression patterns based on tumor type, response to drugs, response to environmental conditions or a variety of other variables.

### SAGE Data Analysis

For handling SAGE data, most investigators rely on the SAGE software generated by the Johns Hopkins SAGE group. This software extracts tag sequences from raw sequence data and tabulates the counts in a database. The software also will make comparisons between libraries of tags and calculate the statistical significance of differences based on Monte-Carlo simulations.<sup>34</sup> Additionally, the software helps create a relational database by extracting tags, gene name and gene information from the sequence database. The program uses this information to match tags to known genes or ESTs. Additional tag-to-gene mapping information can be downloaded from the NCBI from the SAGEmap Web site (Table 21.1-2). This

information that is used by the SAGE software is freely available to noncommercial users of the technology and can be obtained via SAGEnet (Table 21.1-2).<sup>8</sup>

### Databases and Data Mining

Many of the complex expression patterns generated by gene profiling are being deposited on publicly accessible Web sites (Table 21.1-2) or are commercially available. This data can be very valuable when planning experiments, or for making correlations with potential cancer-related genes.<sup>51, 65</sup> It is, therefore, important to access the public expression information for a variety of reasons, but mining the best data and adapting the results for a particular application can be challenging.

To identify genes based on mRNA expression in cDNA libraries, a variety of public databases are available (Table 21.1-2). These databases range from a display of the number of sequences observed; to more sophisticated statistical approaches that assign confidence levels to differentially expressed genes. The CGAP specializes in creating databases and resources for cancer research,<sup>29-31</sup> and has cDNA-based expression information from malignant, premalignant, and normal cells.

Because SAGE involves digital counts of transcript numbers that are compared electronically, SAGE data naturally lends itself to large-scale collaborative projects and the formation of databases. SAGE libraries constructed at different times or in different laboratories can be accurately compared, resulting in a powerful cumulative database. To complement EST data and to provide a more efficient means for archiving quantitative expression profiles, CGAP adopted SAGE technology starting in 1998, and can be accessed through the SAGEmap Web site.<sup>39, 40</sup> From a total of 171,692 sequencing runs, more than 3.4 million valid transcript tags have been processed from 84 different malignant and normal cell types. Online tools built specifically to handle SAGE data<sup>39, 40</sup> allow users to make statistical-based comparisons between libraries to find differentially expressed genes by using the xProfiler, or by downloading data for local analysis. SAGE tags can be mapped to UniGene clusters via SAGEmap, making the identification of a gene from a differentially expressed tag easier. The SAGE data generated through this project is also used to create a Digital Northern tool, where the expression level of a particular gene can be determined for each of the tissues used to make SAGE libraries. Expression comparisons based on SAGE have the additional advantage that no normalization to a housekeeping gene or to a reference standard is necessary, because absolute levels of transcripts are compared.

Although there has been significant data generated using DNA arrays, there are challenges in integrating the different expression profiles generated by varying technology platforms. Most DNA array data that is currently publicly available is through laboratory homepages (see Table 21.1-2 for some links) or as electronic supplements to the journal publications mentioned throughout this chapter.

### GENE EXPRESSION PROFILING IN CANCER: DIAGNOSIS AND THERAPY

The expression of the vast majority of the genes remains unchanged during the complex process of tumorigenesis. Indeed, a pioneering study found that the expression of no more than 1–1.5 percent of the genes was significantly altered in colon cancer, as compared to normal colon tissue.<sup>34</sup> Nonetheless, the analysis of human cancer with the techniques described above typically identify hundreds of genes differentially expressed between normal tissues and malignant specimens. Clearly, prioritization rules according to the identity of the genes uncovered or according to specific sets of criteria must be implemented. Efficient techniques for the validation of these genes have been described in the previous section. In addition to the identification of specific differentially expressed genes, global expression profiles can be used for tumor taxonomy. This section will illustrate these principles through the

description of some of the most important results of gene profiling obtained in common malignancies.

### Colon Cancer

By using SAGE, one study analyzed more than 300,000 transcripts derived from colorectal cancers, pancreatic cancer, and normal colon epithelium. While all the abundant transcripts (more than five copies per cell) represented 75 percent of the mRNA mass, the rare transcripts were responsible for much of the diversity of gene expression: 86 percent of all the different genes were expressed at less than five copies per cell. Interestingly, and perhaps unexpectedly, most of the genes were expressed at similar levels between normal and cancer cells. Indeed, in the case of normal and neoplastic colon, only 548 genes were differentially expressed (less than 1.5 percent of the transcripts present in a given cell). Many genes elevated in cancer represented products known to be involved in growth and proliferation, while genes found in the normal colon were often related to differentiation. Importantly, many of the individual genes found to be differentially regulated may represent targets for mechanism-based therapy or biomarkers for diagnosis. In another study, an Affymetrix oligonucleotide array containing 6500 genes was used to investigate 40 colon tumors and normal colon tissues.<sup>66</sup> Using two-way clustering, clusters of gene with similar patterns of gene expression were identified. Some of these clusters may represent the activation of molecular pathways relevant to colon tumorigenesis.

### Breast Cancer

Gene expression in breast cancer has been monitored by using differential display, cDNA arrays, and SAGE in a variety of experimental systems. In a SAGE study of normal and neoplastic breast tissue, at least 50,000 transcripts were analyzed from 4 libraries and highly differentially expressed genes were identified. Small custom arrays were used to validate the genes identified. Claudin-7 was found up-regulated more than a hundredfold in 85 percent and 60 percent of the primary and metastatic tumors, respectively. While differences in gene expression levels can be subtle in other diseases, many genes appear to be vastly differentially regulated in cancer. Similarly, another study used a combination of differential display and cDNA arrays to gain a better understanding of gene expression patterns in breast cancer.<sup>67</sup> This study identified 700 genes differentially expressed between normal and cancer cells, and a cDNA array containing 107 of these genes was constructed. Most of the genes highly expressed in normal cells, and down-regulated in cancer cells, represented genes important for cell adhesion, communication, and maintenance of cell shape. In contrast, most of the genes elevated in cancer were those encoding enzymes involved in metabolism, macromolecular synthesis, and disruption of the extracellular matrix. By using the custom cDNA array, clusters of genes were identified that were associated with relevant clinical parameters, such as estrogen receptor (ER) status, stage, and tumor size. Overall, gene expression patterns allowed the clustering of breast tumors into two major groups that differed in their ER status.

Other studies with large cDNA arrays contributed significantly to our understanding of gene expression in breast cancer.<sup>68, 69</sup> Many clusters of genes with related expression patterns were identified. For example, an interferon (IFN)-regulated gene cluster and a proliferation cluster (correlated with mitotic index) were found. Importantly, gene expression clusters corresponding to noncancer tissue such as stroma, lymphocytes, and endothelium were also recognized. These issues are important when primary tumors are analyzed because the gene expression profiles represent a complex environment of interacting tissues. While the large arrays failed to distinguish multiple tumor categories, a smaller, more focused array containing 496 genes clustered the tumors into two groups according to their ER status,<sup>69</sup> in a fashion reminiscent of the differential display study described above.<sup>67</sup> The small array further divided ER-negative breast cancers into two groups. It is unclear whether these two categories may have divergent clinical characteristics, but this experiment emphasizes the power of these techniques in cancer taxonomy. These results suggest the possibility that gene expression patterns may be used effectively for diagnosis and therapeutic decisions in breast cancer. In yet another study, laser capture microdissection,

an important validation tool, was used in combination with cDNA arrays for the identification of genes relevant to breast tumorigenesis.<sup>70</sup>

Germ-line mutations in BRCA1 and BRCA2 confer a significant risk of breast and ovarian cancer. Using a large cDNA array, it was recently shown that BRCA1 and BRCA2 tumors could be distinguished from each other and from sporadic breast cancer on the basis of gene expression profiles.<sup>71</sup> Indeed, all the tumors with BRCA1 mutation, and 14 of 15 without the mutation, were appropriately recognized in the BRCA1 classification. Similarly, accurate classification was obtained with BRCA2 tumors. A total of 176 genes were found differentially regulated between BRCA1 and BRCA2 tumors. Interestingly, BRCA1 tumors exhibited increased expression of genes involved in response to cellular stress. A sporadic tumor, which clustered with BRCA1 tumors, proved to exhibit hypermethylation of the BRCA1 promoter. Gene profiling may thus help in the identification of breast cancer genetic status, including the identification of BRCA1 or BRCA2-like phenotype. These different categories may be useful in patient management, as patients with BRCA1-like tumors may require more rigorous follow-up.

### Ovarian Cancer

The vast majority of ovarian cancers are diagnosed in late stages and a major emphasis of functional genomics approaches is to identify biomarkers. Schummer et al. constructed a cDNA array consisting of 21,500 randomly selected transcripts from ovarian cDNA libraries.<sup>72</sup> The vast majority of genes were expressed at similar levels in ovarian cancer and ovarian surface epithelium, the presumed normal counterpart of ovarian cancer. However, they were able to identify cDNAs that were expressed more than 2.5-fold in at least 50 percent of the tumors. These clones also had low levels of expression in nonovarian tissues. Many of these cDNAs were novel and corresponded to ESTs, and others had previously been implicated in various cancers. One candidate, HE4, a protease inhibitor, emerged as a promising candidate because it was highly up-regulated in many ovarian tumors and was found at low levels in other tissues. These findings were subsequently confirmed and extended in a SAGE study of ovarian cancer.<sup>73</sup> An analysis was done of 385,000 transcripts from 10 different ovarian libraries, and differentially expressed genes were identified using a strict set of criteria. Selected genes had to be high in all three primary ovarian cancers and low in all three nonmalignant specimens. Twenty-seven genes were identified that met these criteria and that were overexpressed more than tenfold in ovarian tumors. Interestingly, a majority of those genes were predicted to encode membrane or secreted proteins, making them candidates for biomarkers or tumor targeting. Many of these secreted genes encoded protease inhibitors. Another study using a combination of cDNA-RDA and cDNA arrays also found a large number of genes encoding secreted products to be elevated in ovarian cancer.<sup>74</sup>

### Prostate Cancer

Prostate cancer originally responds to hormone therapy but typically becomes refractory to this therapy and develops into an androgen-independent tumor.<sup>75</sup> The elucidation of the molecular mechanisms accompanying this phenomenon has begun, but our understanding is still incomplete. To monitor gene expression changes that are associated with hormone-independent growth, the androgen-independent prostate cancer xenograft model CWR22-R and its parental androgen-dependent xenograft CWR22 were analyzed by cDNA microarray.<sup>76, 77</sup> Hybridization to a large cDNA array (10,000 clones) revealed that the expression of 160 genes was altered in CWR22 upon androgen removal. The pattern of gene expression changes suggested that the CWR22 cells were undergoing growth arrest upon androgen removal. Interestingly, the majority of these genes were expressed at similar levels between CWR22 and CWR22-R, suggesting that CWR22-R had adapted to growth without androgen and had reentered the cell cycle.<sup>77</sup> Comparison of genes differentially expressed between CWR22 and CWR22-R allow the identification of genes that may be crucial in the progression from androgen dependence to androgen independence.<sup>76, 77</sup> Some of these genes were found to be involved in thyroid hormone receptor

signaling. IGFBP2 and HSP27 were also found elevated in CWR22-R and were validated through the use of tissue microarrays.<sup>76</sup>

## Leukemia

The classification of acute leukemias has long relied on the identity of the precursors. Lymphoid precursors give rise to acute lymphoblastic leukemia (ALL) while myeloid precursor give rise to acute myeloid leukemia (AML). The treatment regimen for these classes is distinct and accurate classification of these tumors can have a significant impact on survival. A cDNA array consisting of 6817 genes was used in order to determine whether global patterns of gene expression could be used to distinguish various classes in leukemias.<sup>78</sup> The original data set, consisting of 38 bone marrow samples (27 ALL, 11 AML), demonstrated that a large number of genes appeared to be correlated with the AML-ALL class distinction. The 50 genes most closely correlated with class distinction were chosen and a class predictor algorithm was developed. In cross-validation analysis by using the original 38 bone marrow samples, 36 of these samples were correctly assigned to the clinical category (AML or ALL). The 50-gene predictor correctly predicted the tumor class in 29 of 34 additional acute leukemias. Interestingly, the number of genes included for prediction was not crucial, as the same results were obtained with predictors containing anywhere from 10 to 200 genes.

An important issue was whether gene expression profiling could be used to determine these classes without a priori knowledge of their existence. This is important because many cancers (prostate, for example) have a variable response to treatment, but cannot readily be divided into classes using current methods. Using self-organizing maps (SOMs), it was possible to identify two categories of acute leukemia that essentially fell along the known ALL-AML classes.<sup>78</sup> Gene profiling could thus identify the two classes of leukemia without previous biological information. Astonishingly, considering the number of clinical specimens studied, SOM could further stratify the acute leukemia classes into four clusters. A first cluster corresponded to AML, a second cluster to T-lineage ALL, and two additional clusters corresponded to B-lineage ALL. The AML, T-cell ALL, and B-cell ALL are the most important clinical distinctions among acute leukemias. These studies demonstrate that gene profiling can accurately identify new classes of cancers (class discovery) and assign tumors to known classes (class prediction). Unfortunately, clinical outcome was not strongly correlated with a particular expression signature. In any event, because leukemic cells can easily be obtained as relatively pure population, these findings may have immediate and important clinical application.

## Lymphoma

Diffuse large B-cell lymphoma (DLBCL) is the most common non-Hodgkin lymphoma subtype. DLBCLs are highly heterogeneous, but attempts at further subclassification have failed. A cDNA array containing 17,856 clones was constructed from various lymphoid cell cDNA libraries.<sup>79</sup> DLBCL exhibited a distinct and complex pattern of gene expression and displayed a lymph node signature. Importantly, reclustering the tumors by using genes of the germinal center (GC) B-cell cluster yielded two subtypes: the GC B-like DLBCLs and the activated-B-like DLBCLs. The expression of no single gene correlated well with the new subtypes, but only the analysis of the patterns of a large number of genes could identify these novel groups. Interestingly, these novel subtypes exhibited marked differences in prognosis. Indeed, 76 percent of GC B-like DLBCL patients were alive after 5 years, while only 16 percent of the activated B-like DLBCL patients were still alive after the same period. Gene profiling thus provides a new classification scheme for DLBCL that define prognostic categories. The molecular and clinical differences are significant and suggest that B-like DLBCL and activated B-like DLBCL may represent distinct diseases. Although this last example represents a clear case in which molecular signature involving large number of genes can be of use clinically, there are also examples of gene profiling identifying individual genes for diagnosis. For example, a recent study used Atlas cDNA arrays (Clontech) to identify the gene *clusterin* as a marker for

anaplastic large-cell lymphomas.<sup>80</sup>

### Melanoma

Thirty-one melanomas and 7 controls were hybridized to a cDNA array that contained probes for nearly 7000 genes.<sup>81</sup> Although no classification schemes for melanoma existed, the gene expression data and hierarchical clustering analysis subdivided the tumors into two groups of 12 and 19. These two groups were analyzed for association with several clinical parameters, such as age and survival, but no associations were found. However, the larger cluster of tumors was predicted, from its expression signature, to consist of tumors with reduced motility and invasiveness. Indeed, these two groups showed differential responses in their ability to migrate into scratch wounds, contract collagen gels, and form tubular networks. Although the analysis did not show association with known clinical parameters, it nonetheless enabled the classification of melanoma into distinct and important classes related to the motility of the tumor cells, and identified genes that may play a role in the invasive ability of this cancer. Further analyses may allow the identification of optimized treatment for each of the classes or other parameters of clinical relevance for melanoma patients.

In another study, melanoma cells were selected for high metastatic potential *in vitro* and analyzed using cDNA arrays.<sup>82</sup> Several genes involved in extracellular matrix assembly were elevated, including RhoC, which single-handedly enhanced metastasis when overexpressed in melanoma cells. A better understanding of gene expression in highly metastatic cells may lead to improve therapeutic strategies aimed at preventing invasion and metastasis. cDNA arrays and other gene profiling methods will undoubtedly continue to play a major role in this endeavor.

### Brain Cancer

Most expression profiling for brain tumors has been applied to glioblastoma multiforme (GBM). DNA arrays,<sup>83–85</sup> SAGE,<sup>40, 86</sup> and tissue arrays<sup>55</sup> have all been applied to the study of the genes expressed in GBM and normal neural tissue. Even if the biological implications of the revealed patterns are not yet clear, there are practical uses for this data. One example is the use of large-scale expression data to find potential tumor markers or antigens for GBM.<sup>51</sup> It is also likely that the pattern of expression will be useful for the classification of brain tumors, including the molecularly heterogeneous GBM classification.<sup>87</sup>

Brain tumors other than GBM have been studied by expression profiling. The major malignant pediatric brain tumor, medulloblastoma, has been studied by SAGE.<sup>88</sup> Detailed SAGE expression profiles are also available for medulloblastomas and a variety of gliomas at the CGAP SAGEmap database.<sup>40</sup>

### NCI60

A series of 60 cancer cell lines of various histologic origins, known as the NCI60, forms the basis of the National Cancer Institute's cancer drug-screening program.<sup>89</sup> Gene expression in these lines was studied by using a cDNA microarray consisting of approximately 8000 different genes.<sup>90</sup> Except for breast and non-small cell lung carcinoma cell lines, the gene expression patterns clustered the lines according to their presumed histologic origin. The patterns of gene expression in the different tissue were thus sufficiently conserved in the cell lines to be grouped together although it is clear that the establishment of cancer lines is accompanied by changes in gene expression patterns. The clustering of the cell lines depended on the exact genes included in the analysis and other studies have shown that cell lines are significantly different from the tissue of origin in colon<sup>66</sup> and ovarian cancer.<sup>73, 74</sup> In any event, analysis of the 60 cell lines allowed the identification of coordinately regulated cluster of genes. The clusters could be labeled according to the genes present in the cluster (proliferation cluster, interferon cluster) or to the

patterns of expression of these genes (epithelial cluster, melanoma cluster). Much information might be gained concerning the microenvironment of tumors by comparing expression patterns between primary tumors and their corresponding *in vitro* cultures or cell lines.

The findings with the NCI60 described above validate the use of cell lines for *in vitro* manipulation such as treatment with hormone or chemotherapeutic drugs. Indeed, the same 60 cell lines were clustered according to the growth inhibitory activity (GI<sub>50</sub>) of 1400 compounds.<sup>91</sup> The cell lines no longer clustered according to their tissue of origin, but according to their drug response. When a subset of these drugs with known mechanisms was used for analysis, several clusters corresponding to mechanisms of action emerged. This could clearly help to identify mechanism of action for unknown drugs. For example, 5-FU appeared with the RNA synthesis inhibitors, suggesting that the main activity of 5-FU may be as an RNA synthesis inhibitor. Further analysis allowed the identification of associations between clusters of genes and clusters of drugs. These relationships may help to identify a genetic basis for certain drug action. For example, an inverse relationship was found between dihydropyrimidine dehydrogenase (DPYD) and 5-FU potency. DPYD is a rate-limiting enzyme in 5-FU degradation. Most cell lines expressing low levels of DPYD were sensitive to 5-FU. DPYD may become useful as a prognosis marker.

### Endothelial Cells

Endothelial cells provide the blood supply to solid tumors and are therefore highly relevant to the process of tumorigenesis. A better understanding of angiogenesis may thus provide tools in the fight against cancer. SAGE was used to identify genes differentially expressed *in vivo* between endothelial cells derived from normal and malignant colorectal tissue.<sup>92</sup> The study showed that at least 79 different genes are significantly differentially expressed between these tissues, including 46 that were specifically elevated in tumor-associated endothelial cells. On the basis of these results, it was concluded that neoplastic and normal endothelium are fundamentally different at the molecular level, suggesting that these differences may be clinically relevant. Nine SAGE tags elevated in the tumor corresponded to novel, uncategorized genes. These genes were named tumor endothelial marker (TEM), and designated TEM-1 to TEM-9. Further experiments confirmed the tumor endothelium-specific expression of these genes, not only for colorectal tumors, but also for other major tumor types. These TEMs, or other genes identified in this study, may become targets of antiangiogenic therapies.

Subtractive hybridization techniques and cDNA arrays have also been used for studying the process of angiogenesis.<sup>93, 94</sup> Overall, many known and novel genes have been implicated in this process. These candidates await testing as targets for therapeutic interventions.

### Gene Profiling Techniques in the Identification of Targets of Specific Oncogenic Molecular Pathways

A main application of techniques such as differential display, SAGE, and cDNA microarrays has been the identification of downstream targets of specific pathways. For example, SAGE was used to identify many genes whose expression is believed to mediate p53-induced apoptosis.<sup>41</sup> Many of these genes were novel and predicted to encode proteins involved in oxidative stress, providing a new paradigm for the mechanism of p53-mediated apoptosis. Similarly, SAGE was used to identify downstream targets of the APC/ $\beta$ -catenin pathway, a pathway activated in the vast majority of colon cancer.<sup>45, 46</sup> c-MYC and PPAR $\Delta$  were both identified as direct transcriptional targets of the TCF- $\beta$ -catenin transcription complex and provided important mechanistic insights into colon tumorigenesis.

cDNA arrays have also been used to identify genes relevant to specific cancer pathways. For example, superoxide dismutase was identified as a target of estrogen derivatives that could kill leukemia cells.<sup>95</sup> In a different approach, ER-responsive breast cancer cells were treated with estrogen and analyzed by SAGE for expression changes leading to the identification of many, possibly useful, estrogen-regulated genes.<sup>48</sup> Differential display was used in the identification of genes involved in Ras transformation.<sup>96, 97</sup> Drug resistance has also been studied extensively by gene profiling and genes relevant to cisplatin and taxol resistance have been identified.<sup>98, 99</sup> There are no doubts that gene profiling techniques will play a major role in the dissection of the myriad of molecular pathways important in human cancer. The examples above represent a minute fraction of the efforts that have already been dedicated toward this goal.

DOI Reference Number: <http://dx.doi.org/10.1036/ommbid.31>

## REFERENCES

1. Smith LM, Fung S, Hunkapiller MW, Hunkapiller TJ, Hood LE: The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus Synthesis of fluorescent DNA primers for use in DNA sequence analysis. *Nucleic Acids Res.* **13**:2399, 1985. PMID 4000959
2. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C: , et al Fluorescence detection in automated DNA sequence analysis. *Nature.* **321**:674, 1986. PMID 3713851
3. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M: , et al Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol.* **14**:1675, 1996. PMID 9634850
4. Schena M, Shalon D, Davis RW, Brown PO: Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science.* **270**:467, 1995. PMID 7569999
5. Gress TM, Hoheisel JD, Lennon GG, Zehetner G, Lehrach H: Hybridization fingerprinting of high-density cDNA-library arrays with cDNA pools derived from whole tissues. *Mamm Genome.* **3**:609, 1992. PMID 1450511
6. Carulli JP, Artinger M, Swain PM, Root CD, Chee L, Tulig C, Guerin J: , et al High throughput analysis of differential gene expression. *J Cell Biochem Suppl.* **31**:286, 1998.
7. Swendeman SL, La Quaglia MP: cDNA subtraction hybridization A review and an application to neuroblastoma. *Semin Pediatr Surg.* **5**:149, 1996. PMID 8858760
8. Sagerstrom CG, Sun BI, Sive HL: Subtractive cloning past, present, and future. *Annu Rev Biochem.* **66**:751, 1997. PMID 9242923
9. Diatchenko L, Lau YF, Campbell AP, Chenchik A, Moqadam F, Huang B, Lukyanov S: , et al *Suppression subtractive hybridization.* : 6025, 1996. PMID 8650213
10. Lisitsyn N, Wigler M: Cloning the differences between two complex genomes. *Science.* **259**:946, 1993. PMID 8438152
11. Liang P, Pardee AB: Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science.* **257**:967, 1992. PMID 1354393
12. Welsh J, Chada K, Dalal SS, Cheng R, Ralph D, McClelland M: Arbitrarily primed PCR fingerprinting of RNA. *Nucleic Acids Res.* **20**:4965, 1992. PMID 1383934
13. Matz MV, Lukyanov SA: Different strategies of differential display areas of application. *Nucleic Acids Res.* **26**:5537, 1998. PMID 9837980
14. Southern EM: Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J Mol Biol.* **98**:503, 1975. PMID 1195397
15. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SP: Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc Natl Acad Sci U S A.* **91**:5022, 1994. PMID 8197176
16. Bowtell DD: Options available—from start to finish—for obtaining expression data by microarray. *Nat Genet.* **21**:25, 1999. PMID 9915497
17. Kurian KM, Watson CJ, Wyllie AH: DNA chip technology. *J Pathol.* **187**:267, 1999. PMID 10398077

18. Johnston M: Gene chips Array of hope for understanding gene regulation. *Curr Biol.* **8**:R171, 1998. PMID 9501061
19. Wilgenbus KK, Lichter P: DNA chip technology ante portas. *J Mol Med.* **77**:761, 1999. PMID 10619436
20. De Benedetti VM, Biglia N, Sismondi P, De Bortoli M: DNA chips The future of biomarkers. *Int J Biol Markers.* **15**:1, 2000. PMID 10763133
21. Wang J: From DNA biosensors to gene chips. *Nucleic Acids Res.* **28**:3011, 2000. PMID 10931914
22. Lee PS, Lee KH: Genomic analysis. *Curr Opin Biotechnol.* **11**:171, 2000. PMID 10753760
23. Augenlicht LH, Wahrman MZ, Halsey H, Anderson L, Taylor J, Lipkin M: Expression of cloned sequences in biopsies of human colonic tissue and in colonic carcinoma cells induced to differentiate in vitro. *Cancer Res.* **47**:6017, 1987. PMID 3664505
24. Brown PO, Botstein D: Exploring the new world of the genome with DNA microarrays. *Nat Genet.* **21**:33, 1999. PMID 9915498
25. Cheung VG, Morley M, Aguilar F, Massimi A, Kucherlapati R, Childs G: Making and reading microarrays. *Nat Genet.* **21**:15, 1999. PMID 9915495
26. Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ: High density synthetic oligonucleotide arrays. *Nat Genet.* **21**:20, 1999. PMID 9915496
27. Adams MD, Soares MB, Kerlavage AR, Fields C, Venter JC: Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nat Genet.* **4**:373, 1993. PMID 8401585
28. Williamson AR: The Merck Gene Index project. *Drug Discov Today.* **4**:115, 1999. PMID 10322263
29. Strausberg RL, Buetow KH, Emmert-Buck MR, Klausner RD: The cancer genome anatomy project Building an annotated gene index. *Trends Genet.* **16**:103, 2000. PMID 10689348
30. Strausberg RL, Dahl CA, Klausner RD: New opportunities for uncovering the molecular basis of cancer. *Nat Genet.* **15**:415, 1997. PMID 9140408
31. Riggins GJ, Strausberg R: Genome and genetic resources from the Cancer Genome Anatomy Project. *Hum Mol Genet.* **10**:663, 2001. PMID 11257097
32. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: Serial analysis of gene expression. *Science.* **270**:484, 1995. PMID 7570003
33. Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE, Hieter P: , et al Characterization of the yeast transcriptome. *Cell.* **88**:243, 1997. PMID 9008165
34. Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, Vogelstein B: , et al Gene expression profiles in normal and cancer cells. *Science.* **276**:1268, 1997. PMID 9157888
35. Lal A, Sui I-M, Riggins G: Serial analysis of gene expression Probing transcriptomes for molecular targets. *Curr Opin Mol Ther.* **1**:720, 1999.
36. Powell J: SAGE. The serial analysis of gene expression. *Methods Mol Biol.* **99**:297, 2000. PMID 10909091
37. Velculescu VE, Vogelstein B, Kinzler KW: Analysing uncharted transcriptomes with SAGE. *Trends Genet.* **16**:423, 2000. PMID 11050322
38. Madden SL, Wang CJ, Landes G: Serial analysis of gene expression From gene discovery to target identification. *Drug Discov Today.* **5**:415, 2000. PMID 10931659
39. Lash AE, Tolstoshev CM, Wagner L, Schuler GD, Strausberg RL, Riggins GJ, Altschul SF: SAGEmap A public gene expression resource. *Genome Res.* **10**:1051, 2000. PMID 10899154
40. Lal A, Lash AE, Altschul SF, Velculescu V, Zhang L, McLendon RE, Marra MA: , et al A public database for gene expression in human cancers. *Cancer Res.* **59**:5403, 1999. PMID 10554005
41. Polyak K, Xia Y, Zweier JL, Kinzler KW, Vogelstein B: A model for p53-induced apoptosis. *Nature.* **389**:300, 1997. PMID 9305847
42. Madden SL, Galella EA, Zhu J, Bertelsen AH, Beaudry GA: SAGE transcript profiles for p53-dependent growth regulation. *Oncogene.* **15**:1079, 1997. PMID 9285562

43. Hermeking H, Lengauer C, Polyak K, He TC, Zhang L, Thiagalingam S, Kinzler KW: , et al 14-3-3 sigma is a p53-regulated inhibitor of G2/M progression. *Mol Cell*. **1**:3, 1997. PMID 9659898
44. Yu J, Zhang L, Hwang PM, Rago C, Kinzler KW, Vogelstein B: Identification and classification of p53-regulated genes. *Proc Natl Acad Sci U S A*. **96**:14517, 1999. PMID 10588737
45. He TC, Chan TA, Vogelstein B, Kinzler KW: PPARdelta is an APC-regulated target of nonsteroidal anti-inflammatory drugs. *Cell*. **99**:335, 1999. PMID 10555149
46. He TC, SA, Rago C, Hermeking H, Zawel L, da Costa LT, Morin PJ, Vogelstein B, Kinzler KW: Identification of c-MYC as a target of the APC pathway. *Science*. **281**:1509, 1998. PMID 9727977
47. Inadera H, Hashimoto S, Dong HY, Suzuki T, Nagai S, Yamashita T, Toyoda N: , et al WISP-2 as a novel estrogen-responsive gene in human breast cancer cells. *Biochem Biophys Res Commun*. **275**:108, 2000. PMID 10944450
48. Charpentier AH, Bednarek AK, Daniel RL, Hawkins KA, Laflin KJ, Gaddis S, MacLeod MC: , et al *Effects of estrogen on global gene expression*. : 5977, 2000. PMID 11085516
49. Taniguchi M, Miura K, Iwao H, Yamanaka S: Quantitative Assessment of DNA microarrays—comparison with northern blot analyses. *Genomics*. **71**:34, 2001. PMID 11161795
50. Rajeevan MS, Vernon SD, Taysavang N, Unger ER: Validation of array-based gene expression profiles by real-time (kinetic) RT-PCR. *J Mol Diagn*. **3**:26, 2001. PMID 11227069
51. Loding WT, Lal A, Siu IM, Loney TL, Wikstrand CJ, Marra MA, Prange C: , et al Identifying potential tumor markers and antigens by database mining and rapid expression screening. *Genome Res*. **10**:1393, 2000. PMID 10984457
52. Wittwer CT, Herrmann MG, Moss AA, Rasmussen RP: Continuous fluorescence monitoring of rapid cycle DNA amplification. *Biotechniques*. **22**:130, 1997. PMID 8994660
53. Morrison TB, Weis JJ, Wittwer CT: Quantification of low-copy transcripts by continuous SYBR Green I monitoring during amplification. *Biotechniques*. **24**:954, 1998. PMID 9631186
54. Wittwer CT, Ririe KM, Andrew RV, David DA, Gundry RA, Balis UJ: The LightCycler A microvolume multisample fluorimeter with rapid temperature control. *Biotechniques*. **22**:176, 1997. PMID 8994665
55. Kononen J, Bubendorf L, Kallioniemi A, BÅrlund M, Schraml P, Leighton S, Torhorst J: , et al Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat Med*. **4**:844, 1998. PMID 9662379
56. Moch H, Schraml P, Bubendorf L, Mirlacher M, Kononen J, Gasser T, Mihatsch MJ: , et al High-throughput tissue microarray analysis to evaluate genes uncovered by cDNA microarray screening in renal cell carcinoma. *Am J Pathol*. **154**:981, 1999. PMID 10233835
57. Schraml P, Kononen J, Bubendorf L, Moch H, Bissig H, Nocito A, Mihatsch MJ: , et al Tissue microarrays for gene amplification surveys in many different tumor types. *Clin Cancer Res*. **5**:1966, 1999. PMID 10473073
58. Bassett DE, Eisen MB, Boguski MS: Gene expression informatics—it's all in your mine. *Nat Genet*. **21**:51, 1999. PMID 9915502
59. Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM: Expression profiling using cDNA microarrays. *Nat Genet*. **21**:10, 1999. PMID 9915494
60. Ermolaeva O, Rastogi M, Pruitt KD, Schuler GD, Bittner ML, Chen Y, Simon R: , et al Data management and analysis for gene expression arrays. *Nat Genet*. **20**:19, 1998. PMID 9731524
61. Strehlow D: Software for quantitation and visualization of expression array data. *Biotechniques*. **29**:118, 2000. PMID 10907086
62. Eisen MB, Spellman PT, Brown PO, Botstein D: Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*. **95**:14863, 1998. PMID 9843981
63. Michaels GS, Carr DB, Askenazi M, Fuhrman S, Wen X, Somogyi R: Cluster analysis and data visualization of large-scale gene expression data. *Pac Symp Biocomput*. **1**:42, 1998.
64. Sherlock G: Analysis of large-scale gene expression data. *Curr Opin Immunol*. **12**:201, 2000. PMID 10712947
65. Scheurle D, DeYoung MP, Binniger DM, Page H, Jahanzeb M, Narayanan R: Cancer gene

- discovery using digital differential display. *Cancer Res.* **60**:4037, 2000. PMID 10945605
66. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A.* **96**:6745, 1999. PMID 10359783
  67. Martin KJ, Kritzman BM, Price LM, Koh B, Kwan CP, Zhang X, Mackay A: , et al Linking gene expression patterns to therapeutic groups in breast cancer. *Cancer Res.* **60**:2232, 2000. PMID 10786689
  68. Perou CM, Jeffrey SS, van de Rijn M, Rees CA, Eisen MB, Ross DT, Pergamenschikov A: , et al Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc Natl Acad Sci U S A.* **96**:9212, 1999. PMID 10430922
  69. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR: , et al Molecular portraits of human breast tumours. *Nature.* **406**:747, 2000. PMID 10963602
  70. Sgroi DC, Teng S, Robinson G, LeVangie R, Hudson JR, Elkahoul AG: In vivo gene expression profile analysis of human breast cancer progression. *Cancer Res.* **59**:5656, 1999. PMID 10582678
  71. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P: , et al Gene-expression profiles in hereditary breast cancer. *N Engl J Med.* **344**:539, 2001. PMID 11207349
  72. Schummer M, Ng WV, Bumgarner RE, Nelson PS, Schummer B, Bednarski DW, Hassell L: , et al Comparative hybridization of an array of 21,500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas. *Gene.* **238**:375, 1999. PMID 10570965
  73. Hough CD, Sherman-Baust CA, Pizer ES, Montz FJ, Im DD, Rosenshein NB, Cho KR: , et al Large-scale serial analysis of gene expression reveals genes differentially expressed in ovarian cancer. *Cancer Res.* **60**:6281, 2000. PMID 11103784
  74. Ismail RS, Baldwin RL, Fang J, Browning D, Karlan BY, Gasson JC, Chang DD: Differential gene expression between normal and tumor-derived ovarian epithelial cells. *Cancer Res.* **60**:6744, 2000. PMID 11118061
  75. Klotz L: Hormone therapy for patients with prostate carcinoma. *Cancer.* **88**:3009, 2000. PMID 10898345
  76. Bubendorf L, Kolmer M, Kononen J, Koivisto P, Mousses S, Chen Y, Mahlamaki E: , et al *Hormone therapy failure in human prostate cancer.* : 1758, 1999. PMID 10528027
  77. Amler LC, Agus DB, LeDuc C, Sapinoso ML, Fox WD, Kern S, Lee D: , et al Dysregulated expression of androgen-responsive and nonresponsive genes in the androgen-independent prostate cancer xenograft model CWR22-R1. *Cancer Res.* **60**:6134, 2000. PMID 11085537
  78. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H: , et al *Molecular classification of cancer.* : 531, 1999. PMID 10521349
  79. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC: , et al Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature.* **403**:503, 2000. PMID 10676951
  80. Wellmann A, Thieblemont C, Pittaluga S, Sakai A, Jaffe ES, Siebert P, Raffeld M: Detection of differentially expressed genes in lymphomas using cDNA arrays Identification of clusterin as a new diagnostic marker for anaplastic large-cell lymphomas. *Blood.* **96**:398, 2000. PMID 10887098
  81. Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M: , et al Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature.* **406**:536, 2000. PMID 10952317
  82. Clark EA, Golub TR, Lander ES, Hynes RO: Genomic analysis of metastasis reveals an essential role for RhoC. *Nature.* **406**:532, 2000. PMID 10952316
  83. Huang H, Colella S, Kurrer M, Yonekawa Y, Kleihues P, Ohgaki H: Gene expression profiling of low-grade diffuse astrocytomas by cDNA arrays. *Cancer Res.* **60**:6868, 2000. PMID 11156382
  84. Sallinen SL, Sallinen PK, Haapasalo HK, Helin HJ, Helen PT, Schraml P, Kallioniemi OP: , et al Identification of differentially expressed genes in human gliomas by DNA microarray and tissue chip techniques. *Cancer Res.* **60**:6617, 2000. PMID 11118044

85. Ljubimova JY, Khazenon NM, Chen Z, Neyman YI, Turner L, Riedinger MS, Black KL: Gene expression abnormalities in human glial tumors identified by gene array. *Int J Oncol.* **18**:287, 2001. PMID 11172594
86. Gunnarsen JM, Spirkoska V, Smith PE, Danks RA, Tan SS: Growth and migration markers of rat C6 glioma cells identified by serial analysis of gene expression. *Glia.* **32**:146, 2000. PMID 11008214
87. Caskey LS, Fuller GN, Bruner JM, Yung WK, Sawaya RE, Holland EC, Zhang W: Toward a molecular classification of the gliomas Histopathology, molecular genetics, and gene expression profiling. *Histol Histopathol.* **15**:971, 2000. PMID 10963139
88. Michiels EMC, Oussoren E, Van Groenigen M, Pauws E, Bossuyt MM, Voute PA, Baas F: Genes differentially expressed in medulloblastoma and fetal brain. *Physiol Genomics.* **1**:83, 1999. PMID 11015565
89. Stinson SF, Alley MC, Kopp WC, Fiebig HH, Mullendore LA, Pittman AF, Kenney S: , et al Morphological and immunocytochemical characteristics of human tumor cell lines for use in a disease-oriented anticancer drug screen. *Anticancer Res.* **12**:1035, 1992. PMID 1503399
90. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V: , et al Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet.* **24**:227, 2000. PMID 10700174
91. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW: , et al A gene expression database for the molecular pharmacology of cancer. *Nat Genet.* **24**:236, 2000. PMID 10700175
92. St Croix B, Rago C, Velculescu V, Traverso G, Romans KE, Montgomery E, Lal A: , et al Genes expressed in human tumor endothelium. *Science.* **289**:1197, 2000. PMID 10947988
93. Kahn J, Mehraban F, Ingle G, Xin X, Bryant JE, Vehar G, Schoenfeld J: , et al Gene expression profiling in an in vitro model of angiogenesis. *Am J Pathol.* **156**:1887, 2000. PMID 10854212
94. Glienke J, Schmitt AO, Pilarsky C, Hinzmann B, Weiss B, Rosenthal A, Thierauch KH: Differential gene expression by endothelial cells in distinct angiogenic states. *Eur J Biochem.* **267**:2820, 2000. PMID 10785405
95. Huang P, Feng L, Oldham EA, Keating MJ, Plunkett W: Superoxide dismutase as a target for the selective killing of cancer cells. *Nature.* **407**:390, 2000. PMID 11014196
96. Ohnami S, Matsumoto N, Nakano M, Aoki K, Nagasaki K, Sugimura T, Terada M: , et al Identification of genes showing differential expression in antisense K- ras-transduced pancreatic cancer cells with suppressed tumorigenicity. *Cancer Res.* **59**:5565, 1999. PMID 10554036
97. Edamatsu H, Kaziro Y, Itoh H: LUCA15, a putative tumour suppressor gene encoding an RNA-binding nuclear protein, is down-regulated in ras-transformed Rat-1 cells. *Genes Cells.* **5**:849, 2000. PMID 11029660
98. Johnsson A, Zeelenberg I, Min Y, Hilinski J, Berry C, Howell SB, Los G: Identification of genes differentially expressed in association with acquired cisplatin resistance. *Br J Cancer.* **83**:1047, 2000. PMID 10993653
99. Duan Z, Feller AJ, Penson RT, Chabner BA, Seiden MV: Discovery of differentially expressed genes associated with paclitaxel resistance using cDNA array technology Analysis of interleukin (IL) 6, IL-8, and monocyte chemotactic protein 1 in the paclitaxel-resistant phenotype. *Clin Cancer Res.* **5**:3445, 1999. PMID 10589757