DNA Microarrays in Clinical Oncology

By Sridhar Ramaswamy and Todd R. Golub

Abstract: **Aberrant gene expression is critical for tumor initiation and progression. However, we lack a comprehensive understanding of all genes that are aberrantly expressed in human cancer. Recently, DNA microarrays have been used to obtain global views of human cancer gene expression and to identify genetic markers that might be important for diagnosis and therapy. We review clinical applications of these novel tools, discuss some important recent studies, identify promising avenues of research in this emerging field of study, and discuss the likely impact that expression profiling will have on clinical oncology.**

J Clin Oncol 20:1932-1941. ***© 2001 by American Society of Clinical Oncology.***

CANCER IS A genetic malady, mostly resulting from acquired mutations and epigenetic changes that influence gene expression.[1,2] Accordingly, a major focus in cancer research is identifying genetic markers that can be used for precise diagnosis or therapy. Over the last half-century, investigators have used reductionism to discover such markers through the study of simple genetic changes like balanced chromosomal translocations.[3] For example, fundamental insights into the nature of the *bcr-abl* gene translocation product resulted in the precise molecular classification of chronic myelogenous leukemia and recently led to the development of the molecularly targeted tyrosine kinase inhibitor STI571 (Gleevec; Novartis, East Hanover, NJ) for the treatment of this disease.[4,5]

Ninety percent of human cancers, however, are epithelial in origin and display marked aneuploidy, multiple gene amplifications and deletions, and genetic instability, making resulting downstream effects difficult to study with traditional methods.[6] Because this complexity probably explains the clinical diversity of histologically similar tumors, a comprehensive understanding of the genetic alterations present in all tumors is required.

The initial sequencing of the human genome,[7,8] coupled with technologic advances, now make it possible to embrace the genetic complexity of common human cancers in a global fashion. Tools are currently available, or are being developed, for the identification of all changes that take place in cancer at the DNA, RNA, and protein levels. In particular, the use of DNA microarrays for the comprehensive analysis of RNA expression (expression profiling) in human tumor samples holds much promise. The uses of DNA microarrays for fundamental biomedical research have recently been reviewed elsewhere.[9] We discuss clinical applications of DNA microarrays and identify future directions and challenges in applying these new tools to cancer medicine.

OVERVIEW OF MICROARRAY-BASED CANCER GENE EXPRESSION PROFILING

Gene expression studies in human cancer can identify genetic markers of malignant transformation. Traditionally, such studies were limited to examining a few genes at a time. However, different methods are now available for large-scale gene expression analysis. Each has both advantages and disadvantages. For example, differential display,[10] serial analysis of gene expression,[11] and representational differential analysis[12] have all proven useful for identifying genes expressed in human tumors. Although these methods are powerful, they are technically difficult, require large-scale DNA sequencing, and only allow for the study of a few different biologic samples at one time.

In contrast, DNA microarray-based gene expression profiling relies on nucleic acid hybridization and the use of nucleic acid polymers, immobilized on a solid surface, as probes for complementary gene sequences.[13] Expression profiling techniques have been used to simultaneously monitor the expression of thousands of genes from human tumor samples. They are relatively easy to use and can be applied to large numbers of samples in parallel. Although a number of competing microarray technologies exist, two platforms (cDNA and oligonucleotide microarrays) are currently used by a majority of investigators and both are effective (Fig 1).

With cDNA arrays, polymerase chain reaction products of cDNA clone inserts representing genes of interest are spotted systematically on nitrocellulose filters or glass

From the Departments of Adult Oncology and Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston; and Whitehead Institute/MIT Center for Genome Research, Cambridge, MA.

Submitted June 1, 2001; accepted December 19, 2001.

Address reprint requests to Todd R. Golub, MD, Dana-Farber Cancer Institute, 44 Binney St, D640, Boston, MA 02115; email: golub@genome.wi.mit.edu.

© 2002 by American Society of Clinical Oncology.

0732-183X/02/2007-1932/$20.00

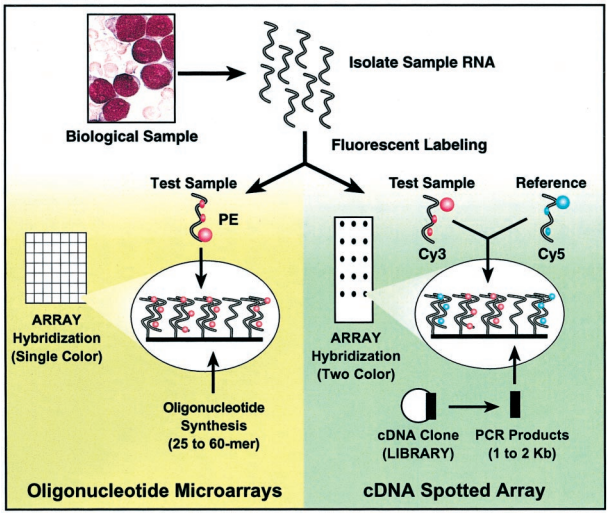

Fig 1. Oligonucleotide versus cDNA microarrays. Oligonucleotide microarrays: direct synthesis or deposition of oligonucleotides onto solid surface and single-color readout of gene expression from a test sample. CDNA microarrays: deposition of polymerase chain reaction products from cDNA libraries onto a solid surface and simultaneous, two-color readout of gene expression in test and reference samples.

slides.[14] Spotted arrays are constructed using cDNA collections (ie, libraries) that can be focused on genes expressed in a particular context or cell type (eg, the lymphochip, which contains genes known to be important in lymphocyte biology). The primary benefit of spotted arrays is that they can be made by individual investigators, are easily customizable, and do not require a priori knowledge of cDNA sequence because clones can be used and then sequenced later if of interest. Practically speaking, however, managing large clone libraries can be a daunting task for most laboratories, and making high-quality arrays can be difficult.

Oligonucleotide microarrays differ in a number of important ways. Oligonucleotide probes for different genes can be deposited or synthesized directly on the surface of a silicon wafer in a patterned manner.[15] Oligonucleotides offer greater specificity than cDNAs, because they can be tailored to minimize chances of cross-hybridization, and sequences up to 60 nucleotides have been used effectively.[16] Major advantages of this approach include uniformity of probe length and the ability to discern splice variants. Until recently, the design of specific oligonucleotides has been limited by sequence availability, but the initial sequencing of the human genome should make probe design easier in the future. Another advantage particular to the commonly used Affymetrix GeneChip system (Affymetrix, Santa Clara, CA) is the ability to recover samples after hybridization to a chip. This allows for a single biologic sample to be sequentially hybridized to multiple arrays, a considerable advantage when dealing with limited biologic material.

The hybridization of a test sample to an array can be detected in one of two ways. cDNA microarrays are commonly queried simultaneously with cDNAs derived from experimental and reference RNA samples that have been differentially labeled with two fluorophores to allow for the quantification of differential gene expression, and expression values are reported as ratios between two fluorescent values. Alternatively, the Affymetrix oligonucleotide system uses a single color fluorescent label, where experimental mRNA is enzymatically amplified, biotin-labeled for detection, hybridized to the wafer, and detected through the binding of a fluorescent compound (streptavidin-phycoerythrin) (Fig 1).

Advances in chip technology or design and decreasing costs are making affordable, commercially available whole genome arrays commonplace. The major challenge now is the effective application of these tools to clinical questions. Outlined below are a number of experimental considerations that must be kept in mind before embarking on such clinical studies.

Biologic Material

Microarray experiments require between 10 and 40 μg of high-quality RNA, corresponding roughly to a 100-mm^3 piece of tissue. Ideally, whole-tumor specimens should be snap-frozen in liquid nitrogen within half an hour of surgical resection and stored at −80°C or colder to prevent RNA degradation. However, this recommendation is guided in part by practicality because changes in some mRNA species have been noted even a few minutes after surgical manipulation and devascularization of tissue.[17] Unfortunately, methods do not yet exist for obtaining sufficient RNA from formalin-fixed tissues for these types of experiments. These requirements thus pose certain challenges. Biopsy specimens available for study tend to be small, increasingly so with earlier detection of certain cancer types and minimally invasive biopsy methods (ie, fine-needle aspiration). RNA quality varies dramatically in specimens from established tumor banks. In addition, clinical information can be difficult to obtain in a retrospective fashion, because of incomplete record keeping and patient confidentiality issues.

Currently, these considerations present major limitations in most clinical settings. There is a critical need for the prospective identification, collection, and storage of high-quality tissue that is broadly available to qualified investigators. Ideally, collected tissues should be linked to clinical information in the context of ongoing clinical trials, while

safeguarding patient confidentiality. Such resources would allow for correlative studies of tumor gene expression profiles and natural history, response to therapy, survival, and other clinically meaningful end points.

Tumor Sampling

Tumors are heterogeneous mixtures of different cell types, including malignant cells with varying degrees of differentiation, stromal elements, blood vessels, and inflammatory cells. Two tumors with similar clinical stages can vary markedly in grade and in relative proportions of different elements (eg, prostatic adenocarcinoma). Tumors of different grades might potentially differ in gene expression, and different markers can be expressed either by malignant cells or by other cellular elements. Because this heterogeneity can complicate the interpretation of gene expression studies, sample selection is an important issue.

The most obvious method for sample selection involves careful histopathologic examination of specimens before microarray analysis. In addition, numerous groups are attempting to focus on the malignant components of this heterogeneous cellular mix using a variety of microdissection techniques. Laser capture microdissection allows for the isolation of individual cells from a tumor section and has been used to isolate cancer cell RNA for microarray studies.[18,19] However, it is difficult to obtain adequate amounts of high-quality RNA for expression profiling with this technique, thus limiting its utility. Further refinement of this and other approaches to isolating pure-cell populations should be encouraged. However, a theoretical limitation of focusing only on malignant tumor components relates to the growing appreciation that tumor-stroma, tumor-endothelial, and tumor-immune cell interactions play critical roles in tumor progression. Expression signatures from nonmalignant cells may also be informative. For these reasons, we currently favor using whole tumors enriched in malignant cells.

Variability

Multiple sources of variation that must be understood in evaluating any microarray experiment include the following: (1) varying cellular composition among tumors, (2) genetic heterogeneity within tumors due to selection and genomic instability, (3) differences in sample preparation, (4) nonspecific cross-hybridization of probes, and (5) differences between individual microarrays.

In general, biologic variation is the major source of variation in gene expression experiments. Increasing the sample number can help in understanding the range of biologic variation in an experiment. Variation due to technical factors can be addressed by replicating sample preparation or array hybridization.[20] Although most high throughput expression profiling centers have informal criteria for what constitutes bad data, there are no generally accepted guidelines.

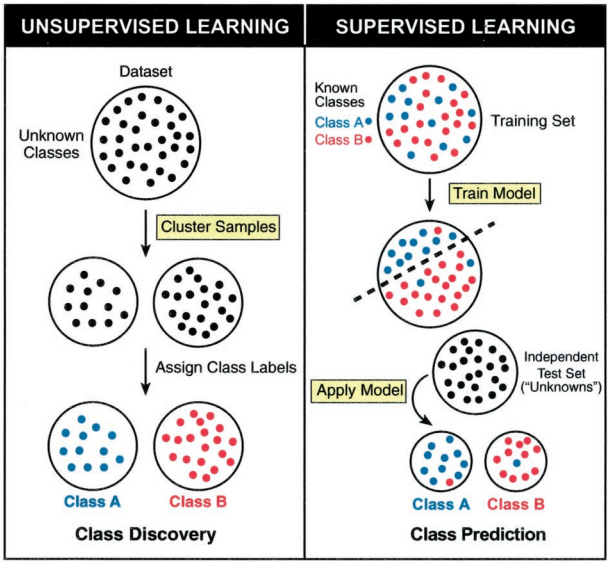

Fig 2. Unsupervised versus supervised learning. Unsupervised learning: multiple tumor samples are clustered into groups based on overall similarity of their gene expression profiles. This approach is useful for discovering previously unappreciated relationships. Supervised learning: multiple tumor samples from different known classes are used to train a model capable of classifying unknown samples. This model is then applied to a test set for class label assignment.

Data Analysis

Gene expression studies pose many challenges for data organization, storage, and analysis.[21,22] Present technology allows for the evaluation of nearly the entire genome from a single biologic sample. Databases are required for efficient storage and retrieval of this information, but most biomedical laboratories are not set up to handle this type of data. Furthermore, there are no standards for the design and implementation of expression databases. These limitations presently make it difficult to compare datasets generated in different laboratories.

To date, the computational analysis of gene expression data has centered on two approaches (Fig 2). Unsupervised learning, or clustering, involves the aggregation of a diverse collection of data into clusters based on different features in a data set.[23,24] For example, one could divide a group of people into clusters based on any combination of eye color, waist size, or height. Similarly, one can gather data about the various expressed genes in a collection of tumor samples

and then cluster the samples as best as possible into groups based on the similarity of their aggregate expression profiles. Alternatively, one could cluster genes across all samples, to identify genes that share similar patterns of expression in varying biologic contexts. Such approaches have the advantage of being unbiased and allow for the identification of structure in a complex data set without making any a priori assumptions. However, because many different relationships are possible in a complex data set, the predominant structure uncovered by clustering may not necessarily reflect clinical or biologic distinctions of interest.

In contrast, supervised learning incorporates the knowledge of class label information to make distinctions of interest. A training data set is used to select those features that best make a distinction. These features are then applied to an independent test data set to validate the ability of selected features to make that distinction. For example, one could select a subset of expressed genes that are best able to distinguish between two cancer types and build a computational model that uses these selected genes to sort an independent, unlabelled collection of those tumor types into the two groups of interest. However, supervised learning is dependent on accurate sample labels, which can be an issue given the limitations of histopathologic cancer diagnosis.

Sometimes, results from unsupervised and supervised learning on a single data set can overlap, but this does not have to be the case. An important issue with either analytic approach is that of statistical significance of observed correlations. A typical microarray experiment yields expression data for thousands of genes from a relatively small number of samples, and gene-class correlations, therefore, can be revealed by chance alone. This issue can be addressed by collecting more samples for each class studied, but this is often difficult with clinical cancer samples. Another approach is to perform exploratory data analysis on an initial data set and apply findings to an independent test set. Findings confirmed in this fashion are less likely a result of chance. Permutation testing, which involves randomly permuting class labels and determining gene-class correlations, has also been used to determine statistical significance. Observed gene-class correlations that are stronger than those seen in permuted data are considered statistically significant.[25,26]

CLINICAL APPLICATION OF MICROARRAYS

A number of different tumor types have been studied using DNA microarrays. Most reports use expression profiling as a screen to identify differentially expressed genes in malignant tissue,[27-39] and space constraints prevent a full discussion of all studies performed to date. However, several themes have emerged from a few studies that suggest a clinical use for these tools, apart from biologic investigation, as discussed below.

Cancer Diagnosis

The use of expression profiling for cancer diagnosis was recently demonstrated using oligonucleotide microarrays to study the expression of 6,817 human genes in 72 acute leukemia samples.[40] Using unsupervised learning, leukemia samples are neatly clustered into the known subsets of acute myelogenous leukemia (AML) and acute lymphocytic leukemia (ALL) solely on the basis of gene expression. In addition, using supervised learning, gene sets that are differentially expressed in AML and ALL were used to correctly classify a group of unknown samples into the correct categories, again solely on the basis of gene expression. Significantly, many markers that were both known, such as myeloperoxidase and terminal transferase, and unknown, were useful for making this distinction.

Although the distinction between AML and ALL generally is not clinically difficult using modern histopathology and cell surface phenotypes, this study provided strong evidence that tumor expression profiles can be used for cancer classification. However, it also raised a number of questions. AML and ALL are derived from distinct cellular precursors likely accounting for the robust expression signatures that distinguish these two cancers. More highly related cancers might be difficult to distinguish using this approach. In addition, class discovery in this case required prior biologic knowledge of AML and ALL to make sense of the observed clusters. The interpretation of new classes discovered with clustering is more difficult in the absence of known biologic or clinical correlates.

More recently, Armstrong et al[41] used both unsupervised and supervised learning to establish the globally distinct nature of mixed-lineage leukemia, a leukemia subset with a decidedly unfavorable prognosis that is defined by a chromosomal translocation involving the mixed-lineage leukemia gene. Importantly, molecular markers differentially expressed by this leukemia compared with both ALL and AML, such as the receptor tyrosine kinase FLT3, immediately suggest novel strategies for molecularly targeted treatment in this treatment-refractory cancer.

Expression-based class discovery has also been used for solid tumors. Bittner et al[42] studied 31 patients with malignant melanoma, for which there are no molecularly defined subsets. Cluster analysis defined two putative subsets, and they were able to define marker genes that were differentially expressed between these two subsets. They noted that one of these gene sets was differentially expressed in uveal melanoma cell lines with more aggressive tissue invasion

potential, as measured by the in vitro formation of primitive tubular networks. This suggested that primary cancers might be less or more invasive depending on the subset to which they belong. Using a variety of in vitro assays, they demonstrated that subset membership was indeed associated with differing tissue invasion potential. This study demonstrated that class discovery is possible in the absence of prior knowledge and that such findings can be validated using cancer cell line models. Unfortunately, the patients in this study had uniformly poor prognosis, and it was not possible to define patient subsets with different natural histories. Future work should determine whether these new subsets describe tumors with distinct natural histories.

Two additional studies have explored expression-based breast cancer classification. Perou et al[43] reported the molecular classification of 65 breast adenocarcinoma specimens from 42 individuals. Hierarchical cluster analysis defined three separate subtypes in this highly heterogeneous tumor class, based on patterns of gene expression. One subtype was known (*Erb-B2+* cancers), and two others were previously unknown (estrogen receptor–positive/luminal-like cancers and basal-like cancers). The clinical significance of the two novel cancer subsets remains an open question. A unique feature of this study was the presence of 20 primary tumors that were biopsied before and after a 16-week course of doxorubicin chemotherapy and two primary/lymph node metastases pairs. Using clustering, they showed that paired samples are more highly related to each other than to tumors from other individuals, despite chemotherapy or metastatic evolution. This study also identified gene expression correlates of different cellular features of these tumors. For example, a number of genes known to play roles in cellular proliferation were coordinately expressed by different tumors, and their expression could be correlated with mitotic index. They also identified eight independent gene clusters that seemed to reflect contributions of specific cell types present within tumors such as endothelial cells or B lymphocytes. This finding is of considerable interest, because some investigators have proposed whole tumor studies followed by computational techniques to infer the transcriptional fingerprint of each cellular component of a tumor. Such in silico studies have the potential to reveal the complex molecular and cellular interactions that drive tumor growth without the need for separation of tumor components. More recently, this group has extended its findings to a larger set of tumors.[44]

Hedenfalk et al[45] used expression profiling to study seven spontaneous and 15 hereditary breast adenocarcinomas with mutations in either *BRCA1* or *BRCA2*. Using supervised learning, they were able to identify a number of differentially expressed genes between *BRCA1*-mutated and *BRCA2*-mutated tumors and use these genes to accurately categorize these samples. Cyclin D1, an important cell cycle regulator known to be overexpressed in certain breast cancers, was one of the genes with increased expression in *BRCA2* mutation–positive tumors, and this finding was confirmed using immunohistochemistry. Interestingly, one spontaneous tumor was classified as having a *BRCA1*-mutated phenotype. Direct sequencing of the *BRCA1* gene in this patient showed no mutation, but the promoter of this gene showed aberrant methylation resulting in silencing of gene expression. Because epigenetic events can be important in oncogenesis, this intriguing finding points to the use of expression profiling for identifying such events in the absence of germline information.

Most published studies have applied expression profiling to single cancer types. However, recent efforts have focused on developing multiclass classifiers capable of distinguishing between multiple common human malignancies. This approach holds much promise for the uniform, molecular, and database-driven classification of all human tumors.[46-48] For example, we have trained a multiclass molecular classifier capable of predicting the identity of primary and metastatic cancers from 14 different tumor classes with high accuracy. Interestingly, poorly differentiated cancers have dramatically different gene expression profiles compared with their well-differentiated counterparts and cannot be classified accurately, suggesting that these tumors are distinct entities. These findings imply that the expression-based diagnosis of some clinically problematic samples (ie, metastases) is feasible, whereas other histopathologically difficult samples (ie, poorly-differentiated cancers) might elude classification by site of origin.

In addition, there is much clinical heterogeneity that is not explained using traditional histopathologic distinctions, such as site of origin. An alternate approach is to ignore traditional distinctions and try to classify tumors using distinct gene expression patterns. For example, Hanahan and Weinberg[49] have proposed that cancer arises from the cooperative dysfunction of distinct biologic pathways responsible for a variety of physiologic functions such as growth, programmed cell death, or angiogenesis. An intriguing question is whether a molecular taxonomy of cancer such as this exists, based purely on tumor gene expression rather than on histopathologic appearance. This type of taxonomy might reveal unexpected relationships between individual tumors. For example, using expression-based molecular descriptions, an individual colonic adenocarcinoma might prove to be more related to a given pancreatic adenocarcinoma rather than to another colorectal tumor based on molecular patterns. Additionally, the presence or absence of different molec-

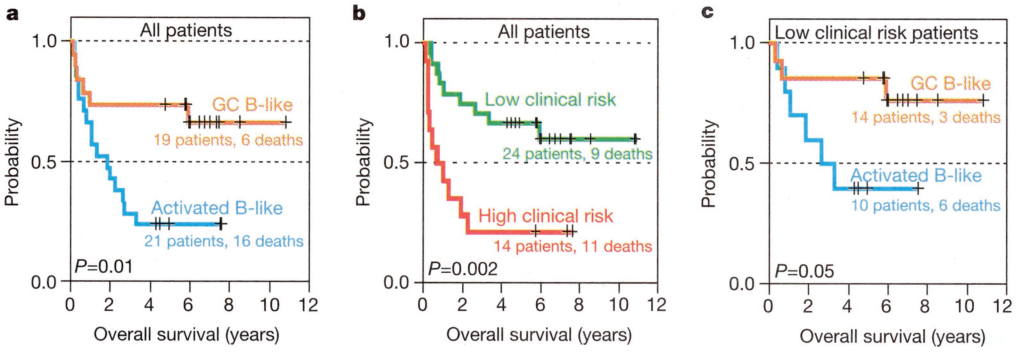

Fig 3. Kaplan-Meier plots. DLBCL subsets identified by (a) gene expression profiling and (b) according to their International Prognostic Index (IPI) scores. Low clinical risk: IPI 0-2, high clinical risk: IPI 3-5. (c) Low clinical risk DLBCL patients grouped on the basis of their gene expression profiles. Reprinted with permission from Nature 403:503-511 copyright 2000, Macmillan Magazines, Ltd.

ular pathways might correlate with differential outcome within individual classes.

Outcome Prediction

Presently, it is difficult to predict whether chemotherapy will be effective for individual patients. DNA microarrays offer the opportunity to ask whether tumor expression profiles can be used to predict chemosensitivity. The NCI60 panel of 60 cancer cell lines is used extensively at the National Cancer Institute as a screen for drug sensitivity. These lines have been treated with more than 70,000 agents, one at a time and independently. Scherf et al[50] attempted to correlate gene expression and drug sensitivity patterns for 118 drugs with known mechanisms of action in the NCI60 panel using clustering. They described correlations between the expression of certain genes with sensitivity or resistance of the NCI60 panel to several drugs. For example, dihydropyrimidine dehydrogenase expression, the rate-limiting enzyme in fluorouracil metabolism, was inversely correlated with sensitivity to fluorouracil. More recently, Staunton et al[51] used supervised learning to demonstrate that statistically significant prediction of chemosensitivity is possible for some compounds using this NCI60 cell-line system.

These findings are intriguing but their interpretation is complicated by a number of issues. Cell lines were necessarily used for these studies but are highly selected entities. Cell-line studies also ignore the potential role of the tumor microenvironment in drug resistance. Additionally, host pharmacokinetics and pharmacodynamics can govern in vivo responses to chemotherapy, and posttranslational mechanisms of drug resistance are not directly measured with DNA microarrays. It remains unclear whether in vitro drug sensitivity correlations can be validated in clinical studies. It is likely that whole-tumor expression profiling will only partially predict chemotherapy responses. However, when coupled with germ-line analysis of sequence variation within genes important for drug metabolism, expression profiling might provide important information regarding the susceptibility of a tumor to a given drug assuming optimal delivery of the agent.

Nevertheless, early attempts to predict treatment outcome in cancer patients seem encouraging. Investigators have demonstrated the utility of using pretreatment gene expression profiling to determine prognosis. In a retrospective study of 38 patients with diffuse large B-cell lymphoma (DLBCL), Alizadeh et al[52] clustered cDNA microarray data to define new subtypes of this lymphoma. They found that these subtypes differentially express genes that correlate with either an activated peripheral-blood B-cell (AB) or a normal germinal center B-cell (GCB) phenotype. Because all patients were uniformly treated with anthracycline-based chemotherapy, they then correlated treatment outcome with these two subsets. Although overall 5-year survival was 52%, 76% of GCB DLBCL patients were alive at 5 years compared with 16% of AB DLBCL patients. They also demonstrated that expression profiling can add value to existing clinical prognostic indices. In considering 24 patients with low-risk DLBCL tumors, as defined by the International Prognostic Index ([IPI] score 0 to 2), the AB subtype was again at higher risk of dying despite standard treatment in comparison with those with the GCB subtype (Fig 3). Although a small study, this work was the first to demonstrate expression-based correlates of outcome. Shipp et al[53] have also reported the use of expression profiling to substratify IPI low- and low-intermediate–risk DLBCL patients into subgroups with markedly differing survival (5-year overall survival, 75% *v* 32%).[53] These findings confirm that expression profiles should be useful for outcome prediction in lymphoma patients beyond currently available clinical criteria.

Expression-based outcome prediction is now being explored in a variety of other cancer types.[54-57] For example, Dhanasekaran et al[58] used cDNA arrays to examine the gene expression profiles of 50 normal and neoplastic prostate specimens and identify markers that were differentially expressed in tumor versus normal tissue. Hepsin, a transmembrane protease, and pim-1, a serine/threonine kinase, were both found to be overexpressed in prostate cancer at the RNA level. They then used tissue microarrays to validate this marker at the protein level in 738 prostate tissue and tumor samples. Interestingly, absence of pim-1 expression in prostatic adenocarcinoma was correlated with a greater risk of prostate-specific antigen failure after radical prostatectomy independent of Gleason score.

In addition, Pomeroy et al[59] studied the use of gene expression profiling for classification and outcome prediction in childhood medulloblastoma, a tumor type with highly variable responses to chemotherapy and radiation. Using gene expression profiles from 60 similarly treated patients for whom biopsies were obtained before treatment, a classifier capable of predicting outcome was generated. Patients who were predicted to be survivors had a 5-year overall survival of 80% compared with 17% for patients predicted to have poor outcome. Notably, this outcome predictor outperformed all other available measures, including clinical stage and *Trk-C* status, a single gene marker that has prognostic value.[60]

Because a major determinant of poor outcome is metastatic spread, MacDonald et al[61] used expression profiling of 23 nonmetastatic and metastatic medulloblastoma specimens to identify 85 genes that are differentially expressed between these tumor states. They found that platelet-derived growth factor receptor α and members of the downstream *RAS*/mitogen-activated protein kinase signal transduction pathway are upregulated in metastatic tumors. Importantly, inhibition of signaling through this receptor pathway inhibited tumor cell migration in vitro, again pointing to the power of expression profiling in clinical specimens for revealing novel molecular treatment targets.[61]

FUTURE CHALLENGES

Comprehensive Cancer Profiling

Despite early progress, cancer expression studies have examined relatively small numbers of clinical specimens, and there has not been sufficient time to reproduce many findings in this new field. Recent reports demonstrate the use of expression profiling for addressing important questions in clinical oncology, but many future challenges remain, including large-scale profiling across the spectrum of tumor class, stage, and grade.

Future studies in expression-based cancer classification should be coupled with clinically meaningful end points, such as survival. Presumably, genetic markers that correlate with different phenotypes or clinical outcomes will be useful for both prognostication and understanding the molecular basis of disease progression. Prospective clinical studies will be required to fully explore the possibility that all cancers can be divided into molecularly defined subtypes using expression profiles with variable natural history and response to treatment.

For most studies, the availability of sufficient numbers of patient samples is presently a limiting factor. Future work will require large numbers of tumors annotated with clinical information and might also include microdissected specimens. Given the costs inherent in such an undertaking and the rarity of certain clinical specimens, this makes performing definitive large studies difficult. Large-scale, cooperative expression profiling efforts, suitably linked with existing clinical trials groups, might represent attractive alternatives. Data generated from such a pooled effort could be made publicly available and would allow for systematic molecular diagnosis, classification, and prognostication. Ideally, these studies should be coupled with ongoing efforts to understand molecular changes that are present at the DNA and protein levels in malignant tissue. Microarray-based comparative genomic hybridization, tissue microarrays, and emerging proteomic technologies are high-throughput methods that hold much promise, and studies that integrate such approaches with gene expression profiling should yield truly comprehensive molecular profiles of human cancer.[62-65]

Data Mining

Despite initial sequencing of the human genome, we still have only a rudimentary knowledge of the physiologic roles of most genes. This represents a significant bottleneck in linking gene expression profiles to molecular mechanisms of transformation. There is a need for integrated databases, with complete annotation, comprehensive gene descriptions, and links to relevant genetic and proteomic information. In addition, as expression studies are performed in various species, integration of this information should prove as illuminating as inter-species gene sequence comparisons. Such databases will allow for an understanding of gene expression in the context of all other available biologic information. Although a number of commercial sources have started to create such databases, there is much room for improvement.

As expression profiling technologies mature, the identification of statistically significant patterns from relatively sparse and noisy data sets remains a major challenge.

Although sophisticated data-mining techniques are already being used to analyze expression data, most of these techniques achieve robust performance with a large number of samples and a small number of variables. However, gene expression data sets generally contain small numbers of samples, many profiled genes, and multiple sources of variation. Future advances will require adapting analytic and statistical techniques to this type of data.

Another important area relates to the integration of data sets generated in different laboratories using different profiling technologies. Many human cancer studies involve valuable or rare clinical specimens and are difficult to repeat. Ideally, one should be able to compare expression data sets obtained in any center, at any time, using any platform. However, this goal remains unrealized. Spotted array data is usually reported as ratios between experimental and control expression values and cannot be easily compared with oligonucleotide microarray data. Multiple expression profiling technologies require more sophisticated methods for data comparison and integration.

Drug Development

A major goal is the use of expression profiles to accelerate and refine the clinical evaluation of chemotherapy drugs. New drugs are traditionally evaluated for efficacy in clinically defined cancer types irrespective of mechanisms of transformation. However, a common theme is the low response rates seen in many early clinical studies. This often leads to the branding of a new agent as ineffective. An exciting prospect is the coupling of gene expression profiling with clinical studies of new agents. Such efforts could potentially turn negative studies into positive studies through the identification of gene expression correlates of drug responsiveness and resistance in individual patients. These markers might then be used to prospectively identify populations of patients likely to respond to the agent. Far fewer patients would be required for subsequent clinical trials to prove efficacy, streamlining the drug development process.

Microarrays in the Clinic

Microarrays are often viewed as screens to identify markers for traditional diagnostics, such as immunohistochemistry, for routine clinical use. However, immunohistochemistry is generally nonquantitative, identification of antibodies can be laborious, and multiplexing is not easy. More sophisticated and high-throughput validation methods are required. An alternative view would be to actually use microarrays in the clinic. This would require either custom arrays for different indications or whole genome analysis of every sample coupled with an analysis of relevant genes. As commercially available, low-cost, technically simple arrays and easy-to-use analytic software become available, their routine clinical use can be explored. In addition, the resulting data could populate large expression databases that would serve as growing, centralized, and standardized references to which new cancer samples could be compared. The feasibility of routine clinical use of microarrays, however, has yet to be established.

In conclusion, expression profiling is ripe for application to a multitude of clinical problems. So, what can the practicing oncologist expect in the future? Diagnosis by tumor gene expression database is conceivable. Encoded in this information would be a pathogenetic description of a tumor, its likely natural history, and its chemosensitivity. Additionally, new drug development and evaluation will likely be accelerated both through the identification of novel molecular targets and through the selection of patients for clinical trials with specific tumor gene expression profiles. Although many challenges remain ahead, whole genome approaches are poised to change the face of clinical oncology.

ACKNOWLEDGMENT

We thank all past and present members of the Cancer Genomics Group, Whitehead/MIT Center for Genome Research for many valuable and ongoing discussions.

REFERENCES

1. Knudson AG: Chasing the cancer demon. Ann Rev Genet 34:1-19, 2000

2. Ponder BA: Cancer genetics. Nature 411:336-341, 2001

3. Rowley JD: A new consistent chromosomal abnormality in chronic myelogenous leukemia identified by quinacrine fluorescence and Giemsa staining. Nature 243:290-293, 1973

4. Druker BJ, Talpaz M, Resta DJ, et al: Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. N Engl J Med 344:1084-1086, 2001

5. Druker BJ, Sawyers CL, Kantarjian H, et al: Activity of a specific inhibitor of the BCR-ABL tyrosine kinase in the blast crisis of chronic myeloid leukemia and acute lymphoblastic leukemia with the Philadelphia chromosome. N Engl J Med 344:1038-1042, 2001

6. Gray JW, Collins C: Genome changes and gene expression in human solid tumors. Carcinogenesis 21:443-452, 2000

7. International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome. Nature 409:860-921, 2001

8. Venter JC, Adams MD, Myers EW, et al: The sequence of the human genome. Science 291:1304-1351, 2001

9. Young RA: Biomedical discovery with DNA arrays. Cell 102:9-15, 2000

10. Liang P, Pardee AB: Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science 257:967-971, 1992

11. Velculescu VE, Zhang L, Vogelstein B: Serial analysis of gene expression. Science 276:1268-1272, 1995

12. Diatchenko L, Lau YFC, Campbell AP, et al: Suppression subtractive hybridization: A method for generating differentially regulated or tissue-specific cDNA probes and libraries. Proc Natl Acad Sci USA 93:6025-6030, 1996

13. Southern E, Mir K, Shchepinov M: Molecular interactions on microarrays. Nat Genet 21:5-9, 1999 (suppl 1)

14. Schena M, Shalon D, Davis RW, et al: Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467-470, 1995

15. Lockhart DJ, Dong H, Byrne MC, et al: Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 14:1675-1680, 1996

16. Hughes TR, Mao M, Jones AR, et al: Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat Biotech 19:342-347, 2001

17. Huang J, Qi R, Quackenbush J, et al: Effects of ischemia on gene expression. J Surg Res 99:222-227, 2001

18. Emmert-Buck M, Bonner RF, Smith PD, et al: Laser capture microdissection. Science 274:998-1001, 1996

19. Kitahara O, Furukawa Y, Tanaka T, et al: Alterations of gene expression during colorectal carcinogenesis revealed by cDNA microarrays after laser-capture microdissection of tumor tissues and normal epithelia. Cancer Res 61:3544-3549, 2001

20. Lee ML, Kuo FC, Whitmore GA, et al: Importance of replication in microarray gene expression studies: Statistical methods and evidence from replicative cDNA hybridizations. Proc Natl Acad Sci USA 97:9834-9839, 2000

21. Ermolaeva O, Rastogi M, Pruitt KD, et al: Data management and analysis for gene expression arrays. Nat Genet 20:19-23, 1998

22. Quackenbush J: Computational analysis of microarray data. Nat Rev Genet 2:418-427, 2001

23. Eisen MB, Spellman PT, Brown PO, et al: Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95:14863-14868, 1998

24. Tamayo P, Slonim DK, Mesirov J, et al: Interpreting patterns of gene expression with self-organizing maps: Methods and applications to hematopoietic differentiation. Proc Natl Acad Sci USA 96:2907-2912, 1999

25. Golub TR, Slonim DK, Tamayo P, et al: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 286:531-537, 1999

26. Tusher VG, Tibshirani R, Chu G: Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 98:5116-5121, 2001

27. Sgroi DC, Teng S, Robinson G, et al: In vivo gene expression profile analysis of human breast cancer progression. Cancer Res 59:5656-5661, 1999

28. Thykjaer T, Workman C, Kruhoffer M, et al: Identification of gene expression patterns in superficial and invasive human bladder cancer. Cancer Res 61:2492-2499, 2001

29. Notterman DA, Alon U, Sierk AJ, et al: Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide microarrays. Cancer Res 61:3124-3130, 2001

30. Virtaneva K, Wright FA, Tanner SM, et al: Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. Proc Natl Acad Sci USA 98:1124-1129, 2001

31. Welsh JB, Zarrinkar PP, Sapinoso LM, et al: Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. Proc Natl Acad Sci USA 98:1176-1181, 2001

32. Okabe H, Satoh S, Kato T, et al: Genome-wide analysis of gene expression in human hepatocellular carcinomas using cDNA microarray: Identification of genes involved in viral carcinogenesis and tumor progression. Cancer Res 61:2129-2137, 2001

33. Watson MA, Perry A, Budhjara V, et al: Gene expression profiling with oligonucleotide microarrays distinguishes World Health Organization grade of oligodendrogliomas. Cancer Res 61:1825-1829, 2001

34. Stratowa C, Loffler G, Lichter P, et al: CDNA microarray gene expression analysis of B-cell chronic lymphocytic leukemia proposes potential new prognostic markers involved in lymphocyte traficking. Int J Cancer 91:474-480, 2001

35. Sallinen SL, Sallinen PK, Haapasalo HK, et al: Identification of differentially expressed genes in human gliomas by DNA microarray and tissue chip techniques. Cancer Res 60:6617-6622, 2000

36. Huang H, Colella S, Kurrer M, et al: Gene expression profiling of low-grade diffuse astrocytomas by cDNA arrays. Cancer Res 60:6868-6874, 2000

37. Smid-Koopman E, Blok LJ, Chada-Ajwani S, et al: Gene expression profiles of human endometrial cancer samples using a cDNA-expression array technique: Assessment of an analysis method. Br J Cancer 83:246-251, 2000

38. Khan J, Simon R, Bittner M, et al: Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. Cancer Res 58:5009-5013, 1998

39. Rickman DS, Bobek MP, Misek DE, et al: Distinctive molecular profiles of high-grade and low-grade gliomas based on oligonucleotide microarray analysis. Cancer Res 61:6885-6891, 2001

40. Golub TR, Slonim DK, Tamayo P, et al: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 286:531-537, 1999

41. Armstrong SA, Staunton JE, Silverman LB, et al: MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat Genet 30:41-47, 2002

42. Bittner M, Meltzer P, Chen Y, et al: Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 406:536-540, 2000

43. Perou CM, Sorlie T, Eisen MB, et al: Molecular portraits of human breast tumors. Nature 406:747-752, 2000

44. Sorlie T, Perou CM, Tibshirani R, et al: Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA 98:10869-10874, 2001

45. Hedenfalk I, Duggan D, Chen Y, et al: Gene-expression profiles in hereditary breast cancer. N Engl J Med 344:539-548, 2001

46. Ramaswamy S, Tamayo P, Rifkin R, et al: A uniform approach to molecular cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci U S A 98:15149-15154, 2001

47. Su AI, Welsh JB, Sapinoso LM, et al: Molecular classification of human carcinomas by use of gene expression signatures. Cancer Res 61:7388-7393, 2001

48. Khan J, Wei JS, Ringner M, et al: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 7:673-679, 2001

49. Hanahan D, Weinberg RA: The hallmarks of cancer. Cell 100:57-70, 2000

50. Scherf U, Ross DT, Waltham M, et al: A gene expression database for the molecular pharmacology of cancer. Nat Genetics 24:236-244, 2000

51. Stauton JE, Slonim DK, Coller HA, et al: Chemosensitivity prediction by gene expression profiling in cancer cell lines. Proc Natl Acad Sci USA 98:10787-10792, 2001

52. Alizadeh AA, Eisen MB, Davis RE, et al: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403:503-511, 2000

53. Shipp MA, Ross KN, Tamayo P, et al: Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med 8:68-74, 2002

54. Takahashi M, Rhodes DR, Furge KA, et al: Gene expression profiling of clear cell renal cell carcinoma: Gene identification and prognostic classification. Proc Natl Acad Sci USA 98:9754-9759, 2001

55. West M, Blanchette C, Dressman H, et al: Predicting clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci USA 98:11462-11467, 2001

56. Bhattacharjee A, Richards WG, Staunton J, et al: Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci USA 98: 13790-13795, 2001

57. Garber ME, Troyanskaya OG, Schluens K, et al: Diversity of gene expression in adenocarcinomas of the lung. Proc Natl Acad Sci USA 98:13784-13789, 2001

58. Dhanasekaran SM, Barrette TR, Ghosh D, et al: Delineation of prognostic biomarkers in prostate cancer. Nature 412:822-826, 2001

59. Pomeroy S. Tamayo P, Gaasenbeek, et al: Classification and outcome prediction of embryonal tumors of the central nervous system. Nature 415:436-442, 2002

60. Segal RA, Goumnerova LC, Kwon YK, et al: Expression of the neurotrophin receptor TrkC is linked to a favorable outcome in medulloblastoma. Proc Natl Acad Sci USA 91:12867-12871, 1994

61. MacDonald TJ, Brown KM, LaFleur B, et al: Expression profiling of medulloblastoma: PDGFRA and the RAS/MAPK pathway as therapeutic targets for metastatic disease. Nat Genetics 29:143-152, 2001

62. Pinkel D, Segraves R, Sudar D, et al: High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 20:207-211, 1998

63. Kononen J, Bubendorf R, Kallioniemi A, et al: Tissue microarrays for high-throughput molecular profiling of tumor specimens. Nat Med 4:844-847, 1998

64. Idekar T, Thorsson V, Ranish JA, et al: Integrated genomic and proteomic analysis of a systematically perturbed metabolic network. Science 292:929-934, 2001

65. Monni O, Barlund M, Mousses S, et al: Comprehensive copy number and gene expression profiling of the 17q23 amplicon in human breast cancer. Proc Natl Acad Sci USA 98:5711-5716, 2001