

DNA Microarrays in Clinical Cancer Research

Raymond Wadlow and Sridhar Ramaswamy*

Massachusetts General Hospital Cancer Center & Harvard Medical School, USA

Abstract: The recent sequencing of the human genome, coupled with advances in biotechnology, is enabling the comprehensive molecular "profiling" of human tissues. In particular, DNA microarrays are powerful tools for obtaining global views of human tumor gene expression. Complex information from tumor "expression profiling" studies can, in turn, be used to create novel molecular cancer diagnostics. We discuss the utility of DNA microarray-based tumor profiling in clinical cancer research, highlight some important recent studies, and identify future avenues of research in this evolving field.

INTRODUCTION

Cancer medicine is in the midst of a revolution that is being driven by an increasing understanding of the human genome and advances in molecular biotechnology. This revolution promises to transform clinical practice from population-based risk assessment and empirical treatment to a predictive, individualized model based on the molecular classification of disease and targeted therapy. The expectation is, of course, that personalized approaches to clinical care will increase the efficacy of treatment while decreasing its toxicity and cost.

Cancer is a complex disease. Our taxonomy of cancer, which currently is based mostly on histopathology, includes more than 200 distinct entities arising from diverse cell types. In addition, tumors have somatic mutations and epigenetic changes, many of which are specific to individual neoplasms; these molecular abnormalities influence the expression of genes that control a tumor's growth, invasiveness, metastatic potential, and responsiveness or resistance to chemotherapy. The genetic complexity of cancer probably explains the clinical diversity of histologically similar tumors, but it has been difficult to study this diversity with traditional methods, which are best suited to investigating one gene at a time. In particular, this complexity also confounds the evaluation of new treatment approaches in clinical oncology, since clinically homogeneous patient populations often represent molecularly heterogeneous patient subsets.

Recently, tools have become available for the "molecular profiling" of both tumor and host at the DNA, RNA, protein, and metabolite levels, thus allowing for the comprehensive identification of molecular changes associated with the cancer phenotype. Techniques include genomic re-sequencing, single nucleotide polymorphism (SNP) genotyping, array-based comparative genomic

hybridization (aCGH), DNA microarray-based gene expression profiling, and serum / tissue proteomics. These approaches generate large amounts of data that can then be deciphered to identify molecular patterns that are useful for the molecular sub-typing of cancer patients. In particular, the use of DNA microarrays for the comprehensive analysis of RNA expression (expression profiling) in human tumor samples holds much promise by enabling the quantitative measurement of complex multi-gene expression patterns. We discuss here the utility of DNA microarrays in clinical cancer research and identify future directions and challenges.

MICROARRAY-BASED TUMOR PROFILING

DNA Microarrays

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. Microarrays have been used extensively 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.

With cDNA arrays, polymerase chain reaction products of cDNA clone inserts representing genes of interest are spotted systematically on nitrocellulose filters or glass slides [1]. Spotted arrays are constructed using cDNA collections (i.e., libraries) that can be focused on genes expressed in a particular context or cell type (e.g., the "lymphochip," which contains genes known to be important in lymphocyte biology [2]). 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

*Address correspondence to this author at the Center for Cancer Research, Massachusetts General Hospital, Building 149, Room 7214, 13th Street, Charlestown, MA 02129, USA; E-mail: sridhar@mgh.harvard.edu

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 [3]. 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 [4]. 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 has made probe design easier. 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).

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.

Clinical Samples

Microarray experiments require between 10 and 40 mg of high-quality RNA, corresponding roughly to a 100-mm³ 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 [5]. Unfortunately, methods do not yet exist for reliably obtaining sufficient RNA from formalin-fixed tissues for these types of experiments although this is an

active area of research. 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 (i.e., 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 (e.g., 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 have focused on the malignant components of this heterogeneous cellular mix using a variety of microdissection techniques. Laser capture microdissection (LCM) allows for the isolation of individual cells from a tumor section and has been used to isolate cancer cell RNA for microarray studies [6]. However, it is difficult to obtain adequate amounts of high-quality RNA for expression profiling with this technique alone. Recently, whole-genome RNA amplification has been coupled with LCM to successfully perform cell-type specific gene expression profiling in human tumors [7]. Further refinement of this and other approaches to isolating and profiling 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 non-malignant cells may also be informative. For these reasons, both

approaches will likely prove complementary and should be encouraged.

Sources of Variation

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 [8]. 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 often be addressed by replicating sample preparation or array hybridization, although this is often not possible with small tissue samples [9]. While most high-throughput expression profiling centers have informal criteria for what constitutes bad data, there are no generally accepted guidelines.

Data Analysis

Gene expression studies pose many challenges for data organization, storage, and analysis [10, 11]. 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.

While a detailed discussion is beyond the scope of this present work, the computational analysis of gene expression data has largely centered on two approaches. Unsupervised learning, or clustering, involves the aggregation of a diverse collection of data into clusters based on different features in a data set [12]. For example, one could sort 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 interesting clinical or biologic differences.

In contrast, supervised learning incorporates the knowledge of class label information to make distinctions of interest [13]. 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, using one of many available machine-learning algorithms (e.g. k-nearest neighbors, neural network, support vector machine) 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 classify 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 [13,14].

DNA MICROARRAYS IN CLINICAL CANCER RESEARCH

New cancer drugs are traditionally evaluated for efficacy in clinically defined cancer types independent of molecular information about tumor or host. 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. Recent work has suggested that likelihood of tumor progression and response to treatment are encoded in a primary tumor's pattern of gene expression. An exciting prospect therefore is the coupling of gene expression-based tumor profiling with clinical studies of traditional cytotoxic and newer molecularly-targeted chemotherapy.

Microarrays can be used to identify molecularly homogeneous patient subsets with similar natural histories, thus decreasing the number of patients necessary to identify statistically significant drug

effects early in the clinical investigation of a new agent. Similarly, microarrays could be used to find molecular correlates of drug response in individuals in the face of modest drug effects across a study population. These markers would then be used to prospectively identify molecular subtypes of patients most likely to respond to that agent. Again, far fewer patients would be required for subsequent clinical trials to prove efficacy, streamlining the drug development process. Finally, both these approaches might identify patients who either have aggressive forms of disease or who will likely be refractory to particular treatments, thus raising the possibility that they may be offered alternatives earlier in their course of treatment. While this is a very active area of research, and a comprehensive review of all relevant studies is clearly beyond the scope of the present discussion, we highlight several recent studies that have established important basic principles on which to build. A summary of DNA-microarray based clinical oncology studies is presented in Table 1.

Molecular Diagnosis

The use of expression profiling for cancer diagnosis was originally demonstrated using oligonucleotide microarrays to study the expression of 6,817 human genes in 72 acute leukemia samples [13]. Using unsupervised learning, leukemia samples were 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 more difficult to distinguish using this approach. In addition, class discovery and prediction both 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 and prediction is not possible without accurate class labels. On the heels of these findings, recent studies by others have centered on the comprehensive analysis of human leukemias using DNA microarrays to identify gene expression correlates of known, leukemia-specific

genetic translocations and novel leukemia subsets [15].

Gene expression profiling has also been used for the molecular diagnosis of solid tumors. Perou *et al.* reported the molecular classification of 65 breast adenocarcinoma specimens from 42 individuals [16]. Hierarchical cluster analysis defined separate subtypes in this highly heterogeneous tumor class, based on patterns of gene expression. One subtype was known (Erb-B2 cancers), and three others were previously unknown (estrogen receptor-positive / luminal-like cancers, basal-like cancers, and normal breast-like). A unique feature of this study was the presence of 20 primary tumors that were biopsied before and after a 16-week course of single-agent 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 intervening chemotherapy or metastatic evolution. More recently, this group extended its findings to a larger set of tumors and reported that the previously characterized estrogen receptor-positive / luminal epithelial group could be further divided into at least two subgroups, each with a distinctive expression profile. The clinical significance of these molecularly-defined breast cancer subsets remains an open question, but it does appear that patients with basal-like tumors have a significantly worse clinical prognosis compared to patients with the other breast cancer subtypes [17].

Molecular Risk Stratification

Recent data support the idea that the natural history of a primary tumor is encoded in its gene expression profile. For example, breast cancer is a clinically heterogeneous disease and despite much effort to identify clinical measures of risk, methods to accurately predict an individual's clinical course are currently lacking. While lymph-node status at diagnosis is the most important measure for future recurrence and overall survival, it is a surrogate that is imperfect at best. About a third of patients with no detectable lymph-node involvement, for example, will develop recurrent disease within ten years. Adjuvant chemotherapy or hormonal therapy reduces the risk of distant metastases by approximately one-third; however, 70-80% of patients receiving this treatment would have survived without it [18]. Van't Veer *et al.* applied DNA microarrays to primary tumors from 117 patients with lymph node negative breast cancer, who received minimal treatment after surgery, to identify a 70-gene expression signature predictive of a short interval to distant metastases [19]. In a subsequent validation study on 295 patients (which included the original cohort of 117 patients), this gene expression profile was an independent prognostic factor that outperformed currently used clinical parameters in predicting disease outcome [20]. These results were the first to suggest that

Table 1. DNA microarray-based oncology clinical studies.

First Author	Year	Journal	Tumor Type	Sample Number	Array	Data Analysis & supervised	Clinical Endpoint	Primary Findings	Secondary Findings	Additional Comments
Golub	1999	Science	ALL & AML	72	Oligo	Unsupervised & supervised		Identified gene expression correlates of clinically-defined AML and ALL subclasses		Interpretation required prior knowledge of class labels
Alizadeh	2000	Nature	Diffuse Large B-cell Lymphoma	96	cDNA	Unsupervised	Overall survival with standard multi-agent chemotherapy	Diffuse large B-cell lymphoma is comprised of at least two distinct subgroups with gene expression profiles resembling either normal germinal center B cells or activated peripheral blood B cells	Patients with germinal center B cell-like DLBCL have significantly better survival than those with activated B-like DLBCL	
Perou	2000	Nature	Breast	42	cDNA	Unsupervised		Breast tumors can be classified into ER+/luminal-like, basal-like, Erb-B2+, and normal breast-like subtypes defined by gene expression patterns	Primary tumor gene expression patterns are stable despite exposure to chemotherapy	Matched primary tumors and lymph node metastases (n = 2) have globally similar gene expression
van't Veer	2002	Nature	Breast	117	Oligo	Supervised	Relapse-free survival	A "poor prognosis" gene expression signature in primary tumors predicts disease relapse in patients with node-negative breast cancer	This poor prognosis signature outperforms currently used clinical parameters in predicting disease outcome	
Yeoh	2002	Cancer Cell	ALL	360	Oligo	Unsupervised & supervised	Disease relapse after treatment	Gene expression profiles differ among important genetic subgroups of ALL	Gene expression profiles identify patients destined to fail therapy within some genetic subgroups	
Ramaswamy	2003	Nature Genetics	Various solid tumors	355	Oligo	Supervised & unsupervised	Metastasis, disease recurrence, overall survival	A 17-gene signature in primary tumors correlates with development of metastatic disease across a broad range of solid tumor types	Metastasis-associated gene expression may arise from both malignant and stromal elements within a tumor	
Chang	2003	Lancet	Breast	30	Oligo	Supervised	Response to neo-adjuvant chemotherapy	Gene expression pattern predicts short-term response to neoadjuvant taxotere		Small validation set (n=6)
Ma	2004	Cancer Cell	Breast (ER+)	80	Oligo	Supervised	Breast cancer recurrence after adjuvant tamoxifen	A two-gene expression ratio (HOXB13:IL17 R) correlates with breast cancer recurrence on adjuvant tamoxifen		Small validation set (n=20)
Holleman	2004	NEJM	ALL	271	Oligo	Supervised	In vitro drug resistance, disease relapse after treatment	A gene expression score predicts in vitro chemosensitivity and correlates with treatment outcome in pediatric ALL		Approach more difficult to apply to solid tumors

primary tumor gene expression profiles can be used to molecularly stratify breast cancer patients according to risk of progression. They also suggested the possibility that such profiles could be used to spare "low-risk" patients the toxicity of unnecessary adjuvant chemotherapy. This gene expression signature is currently undergoing prospective evaluation [21]. Open questions include the degree to which this signature, derived from a mixed cohort of patients, is truly independent of tumor estrogen receptor status and whether this signature can truly be used to make treatment decisions regarding withholding of adjuvant chemotherapy in early stage breast cancer [22, 23]. These findings have spurred a flurry of similar studies in a variety of solid and hematologic malignancies, although many preliminary observations need to be rigorously validated.

While most microarray-based cancer studies to date have focused on single tumor types, we recently identified a 17-gene primary tumor gene expression signature through the comparison of primary and metastatic tumors from a wide variety of sites [24]. This signature identifies primary cancers with a greater likelihood of metastasis in a variety of common solid tumor types. These findings, along with those from Van't Veer *et al.* suggest that molecular programs associated with metastasis exist at the earliest stages of tumorigenesis. This contrasts with the commonly held view that metastasis results solely from molecular changes arising late in a tumor's natural history [25]. These findings, if validated, also suggest the exciting possibility that generic profiles might exist for molecular risk stratification regardless of tumor type. It remains unclear whether similar profiles exist which can predict the sites and tempo of tumor progression and spread.

Treatment Outcome Prediction

Investigators have also demonstrated the utility of using pretreatment gene expression profiling to determine treatment outcome. In a retrospective study of 38 patients with diffuse large B-cell lymphoma (DLBCL), Alizadeh *et al.* clustered tumor-derived cDNA microarray data to define previously unappreciated subtypes of this lymphoma [26]. 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 (CHOP), they then correlated treatment outcome with these two subsets. Although overall 5-year survival in all patients 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. Although a small study, this work was the first to demonstrate expression-based correlates of treatment outcome. These findings have been subsequently pursued in a larger patient cohort and are actively being translated into molecular tools for treatment planning (see discussion below), suggesting that expression profiles might be useful for outcome prediction in lymphoma patients beyond currently available clinical criteria [27, 28].

Recently, Holleman *et al.* tested acute lymphocytic leukemia cells from 173 children for sensitivity *in vitro* to four commonly used cytotoxic drugs in the treatment of B-cell acute lymphoblastic leukemia [29]. The cells were also profiled using DNA microarrays to identify differentially expressed genes that correlated with sensitivity or resistance to prednisolone (33 genes), vincristine (40 genes), asparaginase (35 genes), or daunorubicin (20 genes). A combined gene-expression score of resistance to the four drugs was significantly and independently related to treatment outcome in multivariate analysis. Results were then confirmed in an independent population of 98 patients treated with the same drugs. Importantly, predictive genes might offer clues to both the underlying biology of drug sensitivity to these agents as well as new targets for drug discovery. Future studies will need to explore the utility of stratifying childhood ALL patients to alternate modalities of therapy using differing predictive cancer drug-resistance gene expression profiles.

Interestingly, Chang *et al.* have also explored the use of gene expression profiling for predicting chemosensitivity in breast cancer. By obtaining pre- and post-treatment biopsies of breast tumors during a short course of neo-adjuvant chemotherapy taxotere, they reported a gene expression profile predictive of short-term clinical response. While this profile was independently tested in six patients, much larger studies will be required to define the true utility of this approach in defining molecular correlates of long-term drug response [30].

Gene expression profiling has also been used to identify markers correlated with progression in the face of adjuvant hormonal therapy for breast cancer. Ma *et al.* applied oligonucleotide arrays to estrogen receptor-positive tumors from early-stage breast cancer patients treated with adjuvant tamoxifen after surgery [7]. Remarkably, they were able to identify two genes, HOXB13 and IL17 R, whose primary tumor expression patterns were predictive of longer disease-free survival with adjuvant treatment. The predictive value of these markers was then validated in a small, independent patient set (n = 20). They also demonstrated that HOXB13 is over-expressed in ductal carcinoma *in situ* and invasive ductal carcinoma compared to normal breast epithelium, suggesting that these molecular changes occur early

in the natural history of tumors. While exciting, further work is needed to prospectively validate these findings.

Clinical Translation

Microarrays are wonderful tools for discovery, allowing researchers to obtain unbiased surveys of gene expression in tissue samples, but some have questioned their direct clinical application for individualized diagnosis and treatment planning. Microarrays have an appreciable failure rate and occasionally show significant inter-replicate and inter-batch variability in measurement. A second concern is that in a single sample, microarrays measure thousands of variables, most of which are irrelevant to the clinical endpoint under investigation. Complex statistical and computational tools are thus required to extract informative patterns from raw microarray data. Current technology also requires snap-frozen tissue for microarray-based gene-expression profiling. It is usually possible for established tumor banks to provide small frozen specimens for the initial discovery of clinically useful gene-expression profiles, but validation studies are often limited by the availability of tissue, since tumor specimens are generally fixed in formalin rather than frozen. These limitations pose considerable obstacles to the routine use of microarrays in the clinical laboratory, where tests must be highly reliable and easy to interpret.

One strategy for translating microarray profiles into clinical tests is first to identify small, diagnostic gene-expression profiles with microarrays and then to validate the clinical usefulness of these genes either retrospectively or prospectively with the use of a simple, robust, conventional assay such as the quantitative reverse-transcriptase polymerase chain reaction (RT-PCR). This particular strategy is based on the assumption that there are small gene sets (that are amenable to multiplexed PCR assays) for all interesting diagnostic distinctions. Although this assumption might not always be valid, it appears to be reasonable at first glance. A major virtue of this approach is that potentially useful gene signatures, initially discovered in frozen tissue with microarrays, can be validated with multiplexed quantitative RT-PCR in formalin-fixed, paraffin-embedded tissue sections, which are the global standard for pathological studies.

Lossos *et al.* used this strategy to identify and validate a gene-expression signature specific to diffuse large-B-cell lymphoma, composed of genes culled from multiple lymphoma DNA microarray studies, that predicts the response to standard combination chemotherapy with cyclophosphamide, doxorubicin, vincristine, and prednisone (CHOP) [28]. Their six-gene, PCR-based diagnostic test provides information that is independent of the International Prognostic Index and that adds to it as a clinical measure of the likely treatment outcome in patients with diffuse large-B-cell lymphoma. Similarly, Ma *et al.* distilled complex microarray-based gene expression

information from their studies into a 2-gene signature associated with benefit from tamoxifen treatment in early stage breast cancer, which was subsequently validated using RT-PCR assays in an independent cohort of 20 patients using formalin-fixed, paraffin-embedded tissue. [7]. Thus, both groups were able to move from unbiased, genome-scale surveys of gene expression in human tumors to the creation and initial validation of novel diagnostic tools that should fit easily into clinical practice and might refine currently available measures used for predicting treatment response. These tools stratify patient populations based on high, medium, or low likelihood of response to treatment. The evaluation of larger cohorts of patients might permit the development of probabilistic models for more accurate prediction of the likelihood of a response to chemotherapy in an individual patient.

Limitations

Despite early suggestions that DNA microarrays can be used for the subtyping of cancers into molecularly homogeneous groups, these studies also have a number of limitations. Most have been performed with relatively small patient cohorts and findings must now be validated in larger, independent patient populations. Moreover, molecular correlates of treatment response might simply be identifying patients who are generally refractory to all potential treatment rather than to the specific agent being studied. Most studies have also grouped patients into "better" and "worse" prognosis groups, but individual patients lie along a continuous spectrum with regard to disease progression and response to treatment. An outstanding challenge, largely unmet, is highly accurate, individualized diagnosis. Finally, it can be particularly challenging to move beyond molecular patterns and correlations to detailed molecular understanding of mechanisms of transformation, metastasis, and chemoresponsiveness. Newer analytic methods, in combination with evolving technologies for DNA and protein profiling, will likely be required to develop detailed, global views of molecular cancer biology.

FUTURE DIRECTIONS

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 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. 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. Presumably, genetic markers that correlate with different phenotypes or clinical outcomes will also be useful for understanding the molecular basis of disease progression, although more work is needed in this realm.

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. Genome resequencing, microarray-based comparative genomic hybridization, 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.

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 academic and 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. For example, spotted array data is usually reported as ratios between experimental and control expression values and cannot be easily compared with oligonucleotide microarray data that is not. Multiple expression profiling technologies require more sophisticated methods for data comparison and integration.

Toxicogenomics

Once an active lead compound is discovered, the next step in the drug development pipeline involves an assessment of its potential toxicity. Advances in combinatorial chemistry have allowed for activity testing of thousands of chemical entities using high-throughput screens, but toxicology studies are still largely performed with animal models using traditional biomarkers. These studies typically require weeks to months as well as large amounts of the investigational compound. Furthermore, the process of lead discovery is essentially divorced from toxicology studies, such that tremendous resources can be spent optimizing an agent's activity only to have it abandoned due to unacceptable toxicity. As a result, toxicology profiling has become a major rate-limiting step in drug development.

The emerging field of toxicogenomics integrates toxicology into the lead discovery process. Gene expression profiling is being used in model organisms to catalogue changes in gene expression associated with specific drugs, mechanisms, and toxicities [31] (See the article by Searfoss *et al.* in this issue). Mechanisms of toxicity might be suggested by comparing gene expression profiles from cells treated with a compound to profiles of drugs with known toxicities in a reference compendium. As these databases grow, so will the ability to elucidate the role of individual genes in drug response and toxic effects. These types of studies might thus decrease the cost and accelerate the development of new drugs by shifting the discovery of potential toxicities to the earlier, pre-clinical phase of development. Barriers to use of these approaches include the uncertainty of using short-term gene expression changes in model systems as a predictive tool for long-term toxic effects in humans. In addition, the pharmaceutical industry and regulatory agencies need to work together to establish strict criteria for distinguishing between exploratory data and data that can be used appropriately to assess the safety profile of an investigational agent. Nonetheless, the continued

accumulation and validation of toxicogenomic data promises to streamline the drug development process.

Microarrays in the Clinic

Microarrays are often viewed as screens to identify markers for more traditional diagnostic approaches such as immunohistochemistry. However, immunohistochemistry is generally non-quantitative, identification of antibodies can be laborious, and multiplexing is not easy. More sophisticated and high-throughput validation methods are required, and PCR-based approaches have been discussed above. 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 downstream analysis of relevant genes. As commercially available, low-cost, technically robust and 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.

CONCLUSION

Expression profiling is driving the movement towards the comprehensive molecular subtyping of cancers. Why should these developments interest the clinical investigator? Although clinical diagnostic tests have traditionally taken a backseat to therapeutic agents in cancer medicine, change is at hand. In principle, it should be possible to create molecular diagnostic tools that can predict the response of all human tumors to single agents or combination chemotherapy, thereby allowing for precise, individualized matching of molecular diagnosis with treatment. Moreover, early studies in cancer genomics have focused on microarray-based RNA-expression profiling, teaching us how to grapple with complex biologic data sets in the post-genomic era. Lessons learned from this initial experience are already informing our use of newer techniques for obtaining system-wide molecular views of cancer. These advances have profound implications for the development of new cancer drugs, the design of clinical trials, and the planning of treatment during routine patient care. So, what can the practicing oncologist expect in the future? Diagnosis by molecular 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 with specific molecular

features for clinical trials. Although many challenges remain ahead, whole genome approaches such as DNA microarrays are starting to change the face of clinical cancer research.

REFERENCES

- [1] Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) *Science*, **270**, 467-470.
- [2] Alizadeh, A., Eisen, M., Davis, R.E., Ma, C., Sabet, H., Tran, T., Powell, J.I., Yang, L., Marti, G.E., Moore, D.T., Hudson, J.R. Jr., Chan, W.C., Greiner, T., Weisenburger, D., Armitage, J.O., Lossos, I., Levy, R., Botstein, D., Brown, P.O. and Staudt, L.M. (1999) *Cold Spring Harb. Symp. Quant. Biol.*, **64**, 71-78.
- [3] Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. and Brown, E.L. (1996) *Nat. Biotechnol.*, **14**, 1675-1680.
- [4] Hughes, T.R., Mao, M., Jones, A.R., Burchard, J., Marton, M.J., Shannon, K.W., Lefkowitz, S.M., Ziman, M., Schelter, J.M., Meyer, M.R., Kobayashi, S., Davis, C., Dai, H., He, Y.D., Stephaniants, S.B., Cavet, G., Walker, W.L., West, A., Coffey, E., Shoemaker, D.D., Stoughton, R., Blanchard, A.P., Friend, S.H. and Linsley, P.S. (2001) *Nat. Biotechnol.*, **19**, 342-347.
- [5] Huang, J., Qi, R., Quackenbush, J., Dauway, E., Lazaridis, E. and Yeatman, T. (2001) *J. Surg. Res.*, **99**, 222-227.
- [6] Emmert-Buck, M.R., Bonner, R.F., Smith, P.D., Chuaqui, R.F., Zhuang, Z., Goldstein, S.R., Weiss, R.A. and Liotta, L.A. (1996) *Science*, **274**, 998-1001.
- [7] Ma, X.J., Wang, Z., Ryan, P.D., Isakoff, S.J., Barmettler, A., Fuller, A., Muir, B., Mohapatra, G., Salunga, R., Tuggle, J.T., Tran, Y., Tran, D., Tassin, A., Amon, P., Wang, W., Enright, E., Stecker, K., Estepa-Sabal, E., Smith, B., Younger, J., Balis, U., Michaelson, J., Bhan, A., Habin, K., Baer, T.M., Brugge, J., Haber, D.A., Erlander, M.G. and Sgroi, D.C. (2004) *Cancer Cell*, **5**, 607-616.
- [8] Churchill, G.A. (2002) *Nat. Genet.*, **32** Suppl., 490-495.
- [9] Lee, M.L., Kuo, F.C., Whitmore, G.A. and Sklar, J. (2000) *Proc. Natl. Acad. Sci. USA*, **97**, 9834-9839.
- [10] Ermolaeva, O., Rastogi, M., Pruitt, K.D., Schuler, G.D., Bittner, M.L., Chen, Y., Simon, R., Meltzer, P., Trent, J.M. and Boguski, M.S. (1998) *Nat. Genet.*, **20**, 19-23.
- [11] Quackenbush, J. (2001) *Nat. Rev. Genet.*, **2**, 418-427.
- [12] Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) *Proc. Natl. Acad. Sci. USA*, **95**, 14863-14868.
- [13] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. (1999) *Science*, **286**, 531-537.
- [14] Tusher, V.G., Tibshirani, R. and Chu, G. (2001) *Proc. Natl. Acad. Sci. USA*, **98**, 5116-5121.
- [15] Yeoh, E.J., Ross, M.E., Shurtleff, S.A., Williams, W.K., Patel, D., Mahfouz, R., Behm, F.G., Raimondi, S.C., Relling, M.V., Patel, A., Cheng, C., Campana, D., Wilkins, D., Zhou, X., Li, J., Liu, H., Pui, C.H., Evans, W.E., Naeve, C., Wong, L. and Downing, J.R. (2002) *Cancer Cell*, **1**, 133-143.
- [16] Perou, C.M., Sorlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S.X., Lonning, P.E., Borresen-Dale, A.L., Brown, P.O. and Botstein, D. (2000) *Nature*, **406**, 747-752.
- [17] Sorlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Thorsen, T., Quist, H., Matese, J.C., Brown, P.O., Botstein, D., Eystein Lonning, P. and Borresen-Dale, A.L. (2001) *Proc. Natl. Acad. Sci. USA*, **98**, 10869-10874.
- [18] Cole, B.F., Gelber, R.D., Gelber, S., Coates, A.S. and Goldhirsch, A. (2001) *Lancet*, **358**, 277-286.
- [19] van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R. and Friend, S.H. (2002) *Nature*, **415**, 530-536.
- [20] van de Vijver, M.J., He, Y.D., van't Veer, L.J., Dai, H., Hart, A.A., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C.,

- Marton, M.J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E.T., Friend, S.H. and Bernards, R. (2002) *N. Engl. J. Med.*, **347**, 1999-2009.
- [21] Tuma, R.S. (2004) *J. Natl. Cancer Inst.*, **96**, 648-649.
- [22] Eden, P., Ritz, C., Rose, C., Ferno, M. and Peterson, C. (2004) *Eur. J. Cancer*, **40**, 1837-1841.
- [23] Gruvberger, S.K., Ringner, M., Eden, P., Borg, A., Ferno, M., Peterson, C. and Meltzer, P.S. (2003) *Breast Cancer Res.*, **5**, 23-26.
- [24] Ramaswamy, S., Ross, K.N., Lander, E.S. and Golub, T.R. (2003) *Nat. Genet.*, **33**, 49-54.
- [25] Poste, G. and Fidler, I.J. (1980) *Nature*, **283**, 139-146.
- [26] Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson, J. Jr., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown, P.O. and Staudt, L.M. (2000) *Nature*, **403**, 503-511.
- [27] Rosenwald, A., Wright, G., Chan, W.C., Connors, J.M., Campo, E., Fisher, R.I., Gascoyne, R.D., Muller-Hermelink, H.K., Smeland, E.B., Giltnane, J.M., Hurt, E.M., Zhao, H., Averett, L., Yang, L., Wilson, W.H., Jaffe, E.S., Simon, R., Klausner, R.D., Powell, J., Duffey, P.L., Longo, D.L., Greiner, T.C., Weisenburger, D.D., Sanger, W.G., Dave, B.J., Lynch, J.C., Vose, J., Armitage, J.O., Montserrat, E., Lopez-Guillermo, A., Grogan, T.M., Miller, T.P., LeBlanc, M., Ott, G., Kvaloy, S., Delabie, J., Holte, H., Krajci, P., Stokke, T. and Staudt, L.M. (2002) *N. Engl. J. Med.*, **346**, 1937-1947.
- [28] Lossos, I.S., Czerwinski, D.K., Alizadeh, A.A., Wechser, M.A., Tibshirani, R., Botstein, D. and Levy, R. (2004) *N. Engl. J. Med.*, **350**, 1828-1837.
- [29] Holleman, A., Cheok, M.H., den Boer, M.L., Yang, W., Veerman, A.J., Kazemier, K.M., Pei, D., Cheng, C., Pui, C.H., Relling, M.V., Janka-Schaub, G.E., Pieters, R. and Evans, W.E. (2004) *N. Engl. J. Med.*, **351**, 533-542.
- [30] Chang, J.C., Wooten, E.C., Tsimelzon, A., Hilsenbeck, S.G., Gutierrez, M.C., Elledge, R., Mohsin, S., Osborne, C.K., Chamness, G.C., Allred, D.C. and O'Connell, P. (2003) *Lancet*, **362**, 362-369.
- [31] Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., Kidd, M.J., King, A.M., Meyer, M.R., Slade, D., Lum, P.Y., Stepaniants, S.B., Shoemaker, D.D., Gachotte, D., Chakraburty, K., Simon, J., Bard, M. and Friend, S.H. (2000) *Cell*, **102**, 109-126.