Challenges in Developing a Molecular Characterization of Cancer

Jonathan R. Pollack, Matt van de Rijn, and David Botstein

DNA microarrays are wildly used to measure gene expression across thousands of genes in parallel. Recently, considerable efforts have been made to utilize this technology to improve our understanding of cancer and to identify novel diagnostic markers and therapeutic targets. Here, we detail some of the challenges in developing a molecular characterization of cancer and in translating these new discoveries towards clinical utility. 

Semin Oncol 29:280-285. Copyright 2002, Elsevier Science (USA). All rights reserved.

DNA MICROARRAYS consist of thousands of individual DNA elements, each representing a different gene, synthesized or deposited in an array of rows and columns onto a solid support surface (eg, silicon wafer or glass slide). The DNA elements, typically oligonucleotides or cDNAs, serve as probes to query levels of cognate nucleic acid in complex hybridization mixtures. DNA microarrays were first applied widely to the measurement of gene expression, permitting the quantification of mRNA levels across hundreds or thousands of genes simultaneously (Fig 1). DNA microarrays have since been used to profile gene expression in human cancer, uncovering biologically and clinically significant patterns of gene expression.

The microarray format has more recently been utilized for the detection of other biomolecules. For example, microarrays comprising spotted genomic DNA fragments or cDNAs have been used to measure genomic DNA copy number alterations (gene amplification and deletion) in tumors. Microarrays comprising spotted antibodies have been used to measure protein levels. Here, we focus on the multiple challenges in using DNA microarrays to profile gene expression in human cancer. Another highly parallel method for quantifying gene expression, serial analysis of gene expression (SAGE), has been reviewed elsewhere.

Challenges in Developing a Molecular Characterization of Cancer

Specimen Selection

Profiling gene expression requires that tumor specimens be fresh, rapidly frozen, or otherwise preserved in such a way as to maintain the integrity of the mRNA. Formalin-fixed, paraffin-embedded specimens, the mainstay of pathology department archives, have not proved thus far to be suitable specimens for microarray analysis. Additionally, specimens are more useful if associated with detailed pathologic and clinical data. However, because the need for properly frozen/preserved specimens has become apparent only recently, many suitable specimens are being collected prospectively and hence have no substantial clinical follow-up as yet.

One of the most significant technical challenges today relates to the size of tumor specimens. In most protocols, tumor specimens of at least 0.2 g have been required to obtain sufficient mRNA for microarray analysis. This size would preclude the analysis of many valuable clinical specimens, including surgical core biopsies and fine-needle aspirates. Robust methods based on the amplification of mRNA or hybridization signal are being developed to enable analysis of these smaller specimens.

Tumor Heterogeneity

Tumors are variably heterogeneous at the cellular level. In addition to tumor cells, tumors contain nontumor cell types, including stromal, inflammatory, and endothelial cells. Tumor cells themselves also display heterogeneity, and possibly only a small fraction represent the self-renewing "stem cell" population.

In order to determine a molecular characterization of cancer, it is therefore necessary to identify the gene expression signatures of tumor cells within specimens. This can be accomplished by...
physical isolation of tumor cells, for example, by microdissection, laser capture microdissection, or disaggregation followed by cell sorting. Alternatively, this may be accomplished by "virtual dissection," using computational analyses to identify tumor-specific patterns of gene expression from an undissected specimen (Fig 2).

Data Analysis

Each DNA microarray hybridization may provide gene expression measurements for more than 20,000 different human genes. Profiling 50 different tumors would therefore generate more than 1 million gene expression measurements. Organizing and analyzing the large amount of microarray data collected has become a formidable challenge.

Many standard statistical approaches have been applied to the analysis of microarray data, including, for example, analysis of variance (ANOVA) methods. Additionally, several newer tools have been developed and applied, including hierarchical clustering, k-means clustering, and self-organizing maps.

In general, "unsupervised" methods permit analysis of microarray data independent of known pathologic/clinical specimen parameters. One such unsupervised method, hierarchical clustering (Fig 3), has proven widely useful in delineating biologically significant patterns of gene expression, identifying new tumor subtypes, and inferring gene function (Figs 2 and 4). "Supervised" methods use pathologic/clinical labels to explicitly identify genes associated with a given diagnosis, treatment response, etc. Most often, both types of methods are usefully applied to the analysis of microarray data.

Interpreting Patterns of Gene Expression

Following hierarchical clustering, or similar analyses, patterns of gene expression become evident in the microarray data. The next challenge lies in the interpretation of the various gene expression patterns. A variety of tools have been employed to assist in this endeavor. The presence of characterized genes can be used to infer the significance of gene expression patterns; for example, immunoglobulin expression suggests a B-cell signature (Fig 2). In situ hybridization and immunohistochemistry can assist in determining the cell types contributing to particular gene expression patterns. Importantly, profiling gene expression in a variety of cultured cell types, including cancer cells, normal cell types likely to be present within tumors, cultured cells exposed to a variety of physiologic stimuli, as well as cell culture models of development, differentiation, and cell cycle progression, can provide a useful reference for interpreting patterns of gene expression in tumor samples.

![Fig 2. "Virtual dissection" of tumor specimen. Computational algorithms, including hierarchical clustering (shown here), may be used to identify the gene expression signatures of the various cell types present in a heterogeneous tumor specimen. Figure courtesy of Charles M. Perou.](image-url)
CHALLENGES IN TRANSLATING DISCOVERIES INTO DIAGNOSTIC AND THERAPEUTIC UTILITY

Embarrassment of Riches

In performing a microarray characterization of cancer, it often quickly becomes apparent that many genes, perhaps several dozen, behave in biologically or clinically interesting ways, meriting further evaluation as candidate diagnostic markers or therapeutic targets. A major challenge is to prioritize the many candidate genes for future investigations. Strong therapeutic candidates include genes with plausible pathogenetic roles, and genes encoding protein types associated with historic success in the development of small-molecule inhibitors (e.g., G-protein-coupled receptors). Strong diagnostic and therapeutic candidates also include genes found associated with underlying DNA copy number alteration, which suggests a direct pathogenetic role, and genes encoding secreted/membrane-bound proteins. The latter are especially attractive as they may be ideal targets for an antibody-directed therapeutic delivery of radioactive or cytotoxic agents.

High-Throughput Validation and Evaluation of Diagnostic Potential

A related challenge involves increasing the speed and efficiency of validating and further evaluating the many candidate genes. While DNA microarray experiments characterize gene expression across many thousand genes, typically only a small number (~10 to 50) of specimens are examined. Findings must be validated on a larger collection of clinical specimens. Additionally, because of the requirement for intact mRNA, specimens for microarray analysis are often prospectively collected, and therefore lack sufficient clinical follow-up for the evaluation of potential prognostic tumor markers.

Tissue microarrays provide an important new tool for meeting many of these challenges. Tissue microarrays comprise a high-density array of small (~0.6 mm) tissue cores from several hundred different specimens, permitting highly parallel immunohistochemistry, in situ hybridization, and fluorescence in situ hybridization (FISH) analyses (Fig 5). Importantly, because the arrayed cores are cut from formalin-fixed, paraffin-embedded tissue blocks, suitable specimens have been banked in pathology departments for many years, providing (with appropriate institutional review board approval and patient informed consent) the prolonged clinical follow-up necessary for the evaluation of prognostic utility. While in situ hybridizations to quantify mRNA in tissue can be performed with readily available nucleic acid probes, they are technically more challenging than immunohistochemistry. However, unless an antibody is already available, immunohistochemistry for the evaluation of candidate genes requires the more time-consuming and costly production of new antibodies.
An alternative tool for the validation of genes and evaluation of candidate diagnostic markers is quantitative reverse-transcriptase polymerase chain reaction (RT-PCR). New "real-time" methods, such as TaqMan, permit the high-throughput (96- or 384-well format) quantification of mRNA levels. However, as with in situ hybridization, the extent to which mRNA from formalin-fixed, paraffin-embedded tissue samples is sufficiently intact for these analyses is quite variable.

**Diagnostic Platforms**

For a given tumor, or tumor subtype, it may turn out that a single gene demonstrates diagnostic utility. Alternatively, it may turn out that only the pattern of expression across many genes (eg, 20 to 50 genes, or more) provides the relevant diagnostic information. Particularly in the latter case, it will be a challenge to incorporate new markers into routine diagnostic testing within a clinical laboratory.

One possibility is that smaller DNA microarrays, perhaps specialized disease-specific arrays, will become a useful platform for diagnostic testing. Quantitative RT-PCR, as suggested above,
represents another possible platform for mRNA quantification in that it also allows examination of many separate genes in a single experiment. However, these assay formats will require that clinical laboratories become proficient in routine isolation and analysis of mRNA, which is more challenging than DNA-based diagnostics. Alternatively, immune reagents for the relevant diagnostic genes can be developed; however, it also remains to be determined the extent to which gene expression changes will be mirrored by alterations in protein levels. In addition, the immunohistochemical examination of more than about five markers on any given tumor specimen may be impractical in the clinical laboratory. Another possibility, somewhat further in the future perhaps, is a higher-throughput diagnostic system based on protein determinations, possibly in an array format.12

Determining Gene Function

DNA microarrays permit the high-throughput measurement of gene expression, providing a wealth of candidate genes with relevance to the etiology of cancer, and its diagnosis and treatment. However, DNA microarrays only provide an indirect inference of gene function (eg, see Fig 4). As knowledge of gene function is important in both furthering our understanding of cancer, and in prioritizing candidate therapeutic targets, clearly new techniques to speed the assessment of gene function are of great importance.

Two array-based methods that promise to speed investigations of gene function are (1) protein arrays,26 which permit the determination of protein-protein interactions, enzymatic targets, and small-molecule binding, and (2) cell microarrays,27 which permit the functional analysis of many transfected gene products in parallel. High-throughput, miniaturized, microfluidic assays for the direct measurement of enzymatic activities are also likely to play a role in the future. RNA interference (RNAi),28 wherein short, duplex RNAs complementary to gene coding sequences are used to silence specific gene expression, promises to accelerate functional studies in cell culture. Finally, a growing collection of transgenic and "knock-out" mouse models will also facilitate in vivo functional analyses.

CONCLUSION

DNA microarrays are providing new opportunities for the detailed molecular characterization of cancer. Some of the challenges, including the collection of appropriate specimens, handling tumor heterogeneity, and analyzing and interpreting the vast amount of data generated, are being addressed. However, a new set of challenges is emerging in translating new discoveries towards clinical utility, including prioritizing and evaluating the large numbers of candidate genes, selecting diagnostic platforms, and assessing gene function. While tissue microarrays have provided important utility, clearly additional new methods will need to be developed to meet the many challenges that lie ahead.

REFERENCES