Challenges in Developing a Molecular Characterization of Cancer

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

DNA microarrays are widely 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.

DNA MICROARRAYS[1,2] 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.[3-7]

The microarray format has more recently been utilized for the detection of other biomolecules. For example, microarrays comprising spotted genomic DNA fragments[8,9] or cDNAs[10,11] 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.[12] 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.[13]

From the Departments of Pathology and Genetics, Stanford University School of Medicine, Stanford, CA.

Address reprint requests to Jonathan R. Pollack, MD, PhD, Department of Pathology, Stanford University School of Medicine, CCSR Building, Room 3245A, 269 Campus Dr, Stanford, CA 94305-5176.

0093-7754/02/2903-0009$35.00/0

doi:10.1053/sonc.2002.32903

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[14] 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.[15]

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

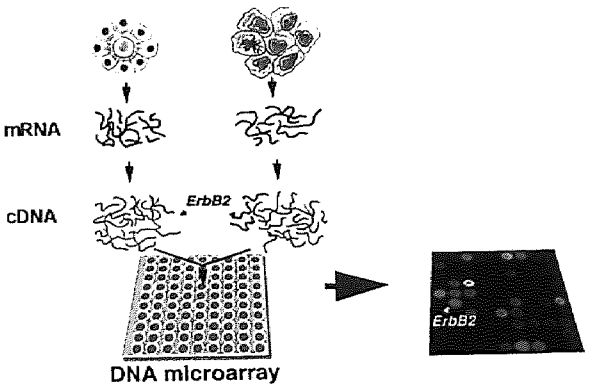

Fig 1. Gene expression profiling using a cDNA microarray. To quantify mRNA levels, mRNAs from two different samples are differentially fluorescently labeled and cohybridized to a cDNA microarray. For each gene spot on the microarray, the ratio of fluorescence reflects the relative abundance of the corresponding mRNA between the two samples.

physical isolation of tumor cells, for example, by microdissection, laser capture microdissection,[16] 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,[17] including hierarchical clustering,[18] k-means clustering, and self-organizing maps.[19]

In general, "unsupervised" methods permit analysis of microarray data independent of known pathologic/clinical specimen parameters. One such unsupervised method, hierarchical clustering[18] (Fig 3), has proven widely useful in delineating biologically significant patterns of gene expression, identifying new tumor subtypes, and inferring gene function[6,7,20,21] (Figs 2 and 4). "Supervised" methods[22] 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.

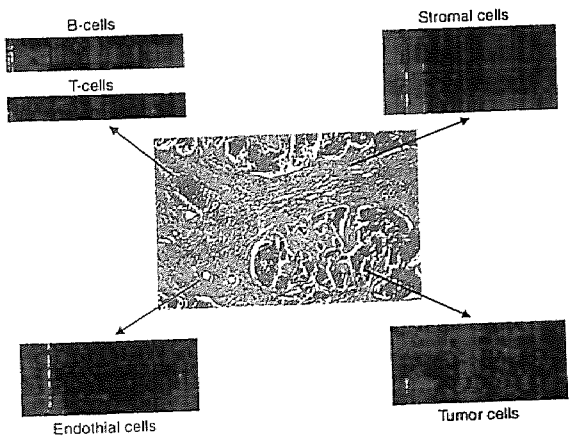

Fig 2. "Virtual dissection" of tumor specimen. Computational algorithms, including hierarchical clustering (shown here), may be used to identify the gene expression signatures of the various cell types present in a heterogeneous tumor specimen. Figure courtesy of Charles M. Perou.

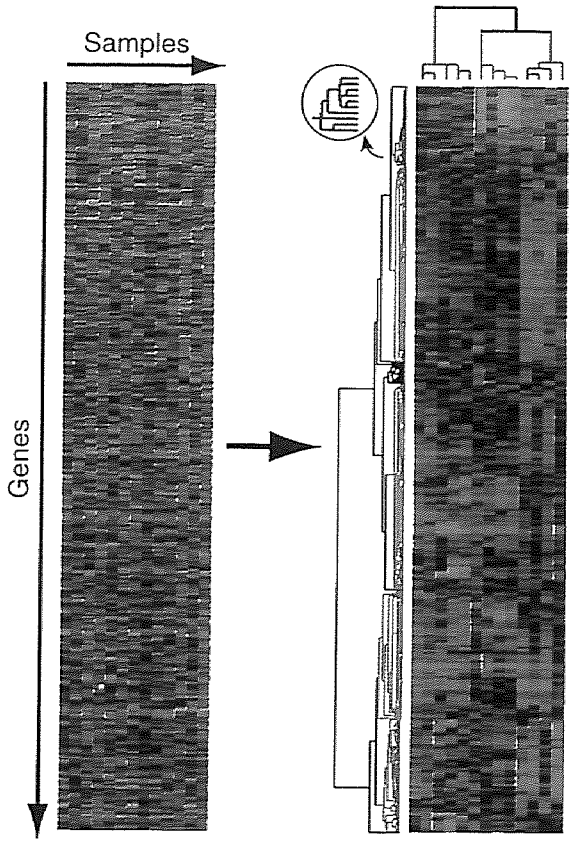

Fig 3. Hierarchical clustering analysis. Hierarchical clustering is an unsupervised method for the organization and display of microarray data. On the left is the input table of gene expression information. Each row represents a different gene on the microarray, and each column represents a different array/sample; fluorescence ratios are depicted in pseudocolor-scale. On the right, hierarchical clustering has been applied to rearrange the rows and columns such that genes with similar expression profiles are clustered together, as are samples with similar expression profiles. Dendrograms (trees) describes the relationships among genes (inset) and among samples. Non-random gene expression "patterns" become evident from the analysis.

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 (eg, G-protein–coupled receptors). Strong diagnostic and therapeutic candidates also include genes found associated with underlying DNA copy number alteration,[10] which suggests a direct pathogenetic role, and genes encoding secreted/membrane-bound proteins.[23] 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[24] 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 flourescence 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.

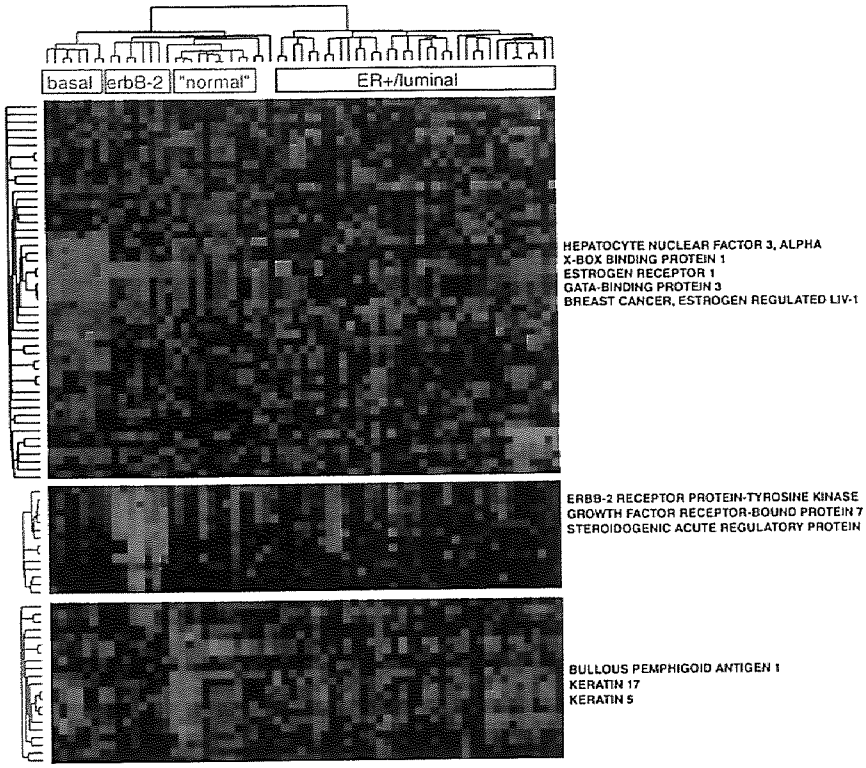

Fig 4. Microarray analysis of breast cancer reveals tumor subtypes. Hierarchical clustering applied to a series of breast tumors suggests at least four distinct breast tumor subtypes: basal, erbB-2, normal-like, and estrogen receptor–positive (ER+)/luminal. The genes that define each subtype reflect underlying biological differences and suggest candidate diagnostic markers. Possible functions of uncharacterized genes can be inferred by their proximity to named genes of known function. Selected genes are indicated. Modified with permission.[7] © *Nature*.

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,[25] 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,

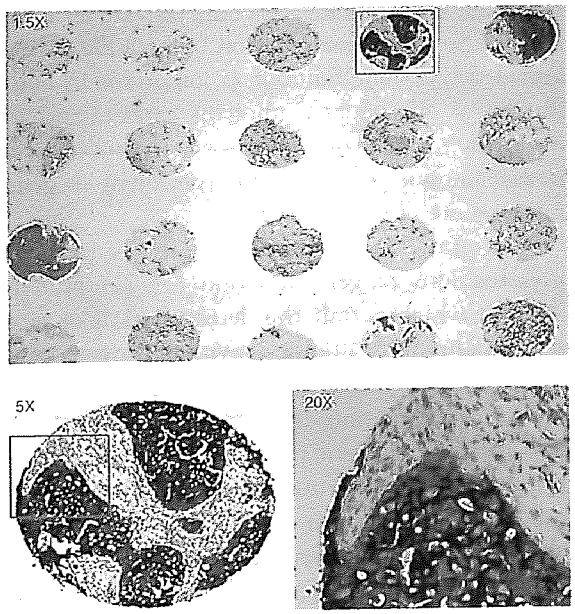

Fig 5. Tissue microarray. Consisting of an array of tissue cores from several hundred different clinical specimens, tissue microarrays permit the high-throughput evaluation of candidate genes by immunohistochemistry, in situ hybridization, or FISH. Here, an array of 0.6-mm breast cancer tissue cores is stained with an antibody to HER2/neu (erbB-2); viewed at 1.5X, 5X, and 20X magnification.

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

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

2. Chee M, Yang R, Hubbell E, et al: Accessing genetic information with high-density DNA arrays. Science 274:610-614, 1996

3. DeRisi J, Penland L, Brown PO, et al: Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet 14:457-460, 1996

4. Alon U, Barkai N, Notterman DA, et al: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 96:6745-6750, 1999

5. 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

6. 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

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

8. Solinas-Toldo S, Lampel S, Stilgenbauer S, et al: Matrix-based comparative genomic hybridization: Biochips to screen for genomic imbalances. Genes Chromosomes Cancer 20:399-407, 1997

9. 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

10. Pollack JR, Perou CM, Alizadeh AA, et al: Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat Genet 23:41-46, 1999

11. Heiskanen MA, Bittner ML, Chen Y, et al: Detection of gene amplification by genomic hybridization to cDNA microarrays. Cancer Res 60:799-802, 2000

12. Haab BB, Dunham MJ, Brown PO: Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. Genome Biol 2:research0004.1-004.13, 2001

13. Polyak K, Riggins GJ: Gene discovery using the serial analysis of gene expression technique: Implications for cancer research. J Clin Oncol 19:2948-2958, 2001

14. Van Gelder RN, von Zastrow ME, Yool A, et al: Amplified RNA synthesized from limited quantities of heterogeneous cDNA. Proc Natl Acad Sci USA 87:1663-1667, 1990

15. Reya T, Morrison SJ, Clarke MF, et al: Stem cells, cancer, and cancer stem cells. Nature 414:105-111, 2001

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

17. Sherlock G: Analysis of large-scale gene expression data. Curr Opin Immunol 12:201-205, 2000

18. 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

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

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

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

22. 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

23. Diehn M, Eisen MB, Botstein D, et al: Large-scale identification of secreted and membrane-associated gene products using DNA microarrays. Nat Genet 25:58-62, 2000

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

25. Gibson UE, Heid CA, Williams PM: A novel method for real time quantitative RT-PCR. Genome Res 6:995-1001, 1996

26. MacBeath G, Schreiber SL: Printing proteins as microarrays for high-throughput function determination. Science 289:1760-1763, 2000

27. Ziauddin J, Sabatini DM: Microarrays of cells expressing defined cDNAs. Nature 411:107-110, 2001

28. Elbashir SM, Harborth J, Lendeckel W, et al: Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. Nature 411:494-498, 2001