Functional Genomics, Gene Arrays, and the Future of Pathology

Edward Gabrielson, M.D., Karin Berg, M.D., Ramaswamy Anbazhagan, M.D., Ph.D.

Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, Maryland

The human genome project has attracted a great deal of attention in recent years among the general public as well as the scientific community. Although it is likely to be a number of years before many of the expected benefits of the genomics revolution are realized, the impact of these scientific breakthroughs on diagnostic pathology is likely to become apparent relatively quickly. In particular, gene array technology, which allows gene expression measurements of thousands of genes in parallel, provides a powerful tool for pathologists seeking new markers for diagnosis. Several recent studies demonstrate how the gene array approach can not only recognize markers for known categories of neoplasia but also lead to recognition of different categories not previously appreciated. Although this approach shows great potential, the successful application of gene arrays to diagnostic problems will require thoughtful interpretation, just as immunochemical technologies require careful planning and analysis.

Measuring gene expression is not new to pathologists. For tissue diagnosis as well as for research, pathologists routinely use immunochemical methods to measure gene expression in tissues and cells. As an adjunct to microscopic examination and flow cytometry, immunochemical methods have significantly improved our ability to diagnose and classify diseases.

However, a major limitation to measuring gene expression with immunochemical methods (as well as with more specialized research methods such as Northern and Western blots and RNAse protection assays) is that analysis is generally restricted to serial measurements of individual candidate genes. Selection of these candidate genes requires some previous knowledge or suspicion of involvement in the situation being studied, and our previous knowledge or suspicions are unlikely to consider all of the important molecular components of a disease process. Furthermore, the number of genes that can be measured by these techniques is limited by the rate at which the assays can be performed and by the availability of suitable reagents. For example, screening various antisera for the ability to differentiate two diseases from one another is not only laborious but also cannot consider the vast numbers of genes for which antisera do not exist.

The candidate gene approach is also limited in that interactions among multiple different factors are not readily recognized. This limitation is illustrated by a recent retrospective study of Her2/Neu gene expression as a predictive factor for breast cancer, in which the investigators had also fortuitously used immunohistochemistry to measure the p53 tumor suppressor gene product in the same series of cases (1). Remarkably, increased staining for both Her2/Neu (a consequence of gene amplification) and p53 (a consequence of gene mutation and protein accumulation) were correlated with a generally unfavorable natural history but, more significantly, with an exceptionally favorable response to anthracycline chemotherapy. Increased staining for either molecule individually did not have such predictive value. Such situations, in which the contribution of one factor to an effect is dependent upon one or more other factors, will often go unrecognized when individual candidate genes are measured with conventional methods.

The Human Genome Project and Functional Genomics

The limitations of the candidate gene approach have become even more apparent as the Human Genome Project has been completing its task of defining all genes in the human genome. Approximately 35,000 human genes are currently believed

to exist, although most of them are only known in the form of partial gene sequences, known as expressed sequence tags (ESTs). The majority of ESTs have no known function, much less a name; yet many of these genes represented by ESTs undoubtedly have important roles in disease and could be useful markers for improved diagnosis.

Fortunately, several new high-throughput technologies have been developed that allow measuring expression of many genes in parallel, and these technologies will facilitate characterizing the roles of the numerous different genes in human disease. Most of the current high-throughput technologies measure mRNA levels rather than protein levels. Some of these technologies are technically complex, such as SAGE (serial analysis of gene expression; 2), subtractive hybridization (3), and cDNA fragment fingerprinting (4). Gene arrays, on the other hand, offer a technology that is accessible, cost efficient, and amenable to standardization.

Fundamentally, gene arrays consist of multiple spots of immobilized single-stranded DNA, to which labeled RNA samples can be hybridized and measured. Although gene array technology was conceived over 10 years ago (5), it has been only in recent years that numerous academic institutions and biotechnology concerns have committed significant resources to develop facilities for gene array analysis. A number of recent publications using this technology provide evidence that gene array analysis will become an important component of medical research and, more important, will have an impact on clinical aspects of medicine, including pathology.

cDNA Array and Oligonucleotide Array Technologies for Analysis of Gene Expression

Two types of gene arrays are commonly used today: cDNA gene arrays and oligonucleotide gene arrays. cDNA arrays are generally manufactured through robotic spotting of small quantities of PCR amplified segments of genes, cloned from cDNA libraries, onto glass slides or nylon membranes. After spotting, the DNA on the array is denatured and hybridized to a labeled sample. Two general options are available for sample labeling: radionucleotide labeling and fluorescent labeling. With traditional radionucleotide labeling, samples are individually labeled and hybridized to arrays, and signals are measured with radiographic film or a phosphorimager. With fluorescent labeling, two (or more) samples are labeled with differently colored flours, cohybridized to an array, and read at different excitation wavelengths (6). The fundamental principles underlying the alternative methods for

sample labeling and measuring gene expression are relatively straightforward and are diagramed in Figure 1.

Each of these methods for sample labeling has advantages and disadvantages. Two-color fluorescence allows an internal standard to be hybridized (as one of the two colors) to each array, facilitating comparisons of results from many different samples. This labeling method also has higher spatial resolution, allowing more than 10,000 genes to be spotted on a single glass slide. The major advantage of radiolabeling is the high level of sensitivity (7), allowing analysis of small tissue samples without amplification of the RNA sample. This is a critical factor for many types of tissue analysis, although more sensitive fluorescent labeling and detection methods are currently in development.
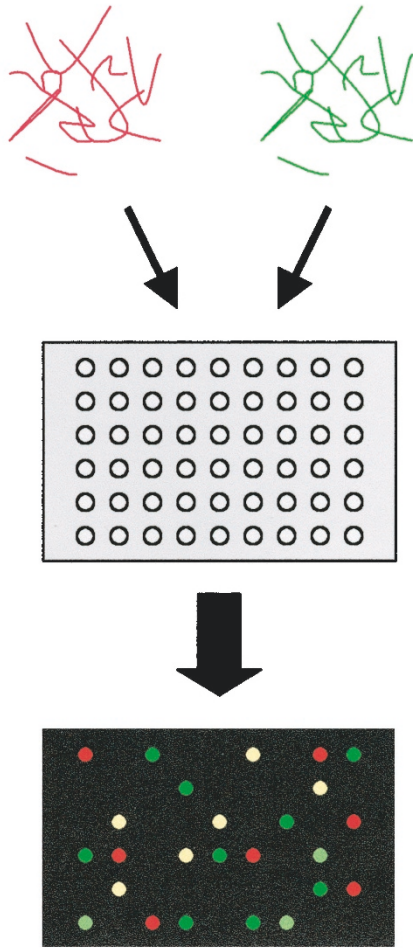
Oligonucleotide gene arrays are prepared by either spotting presynthesized oligonucleotides (typically 50 to 80 bases) or, alternatively, by synthesis of short oligonucleotides (on the order of 20 bases) directly on glass (8). The *in situ* synthesis technology allows very high resolution, multiple different sequences to represent each gene, and, at least in theory, a high level of internal controls. Furthermore, the process for manufacturing oligonuceotide arrays is readily standardized, providing relative precision in estimations of expression levels. Whereas spotted cDNA arrays are being custom manufactured in many research laboratories throughout the country, the *in situ* synthesis oligonucleotide arrays require a more complicated manufacturing process involving photolithography. Thus, oligonucleotide arrays are generally more expensive then spotted cDNA arrays, and currently have relatively stringent requirements for mRNA quantity and preservation.

Gene Discovery Using Gene Arrays

Perhaps the most obvious application of the high-throughput technologies for gene expression analysis is to screen thousands of genes for association with disease. Because large numbers of variables are typically screened with relatively small numbers of samples, the screening data by itself often lacks statistical validity. However, this approach can lead to the discovery of new candidate genes or markers that can subsequently be tested in a more rigorous manner as possible diagnostic markers for the disease or for their mechanistic role in the disease.

There are several examples of how cDNA arrays, often in combination with other methods, have discovered new markers or molecular players in disease. In one study, cDNA arrays were constructed to represent genes expressed in squamous cell lung
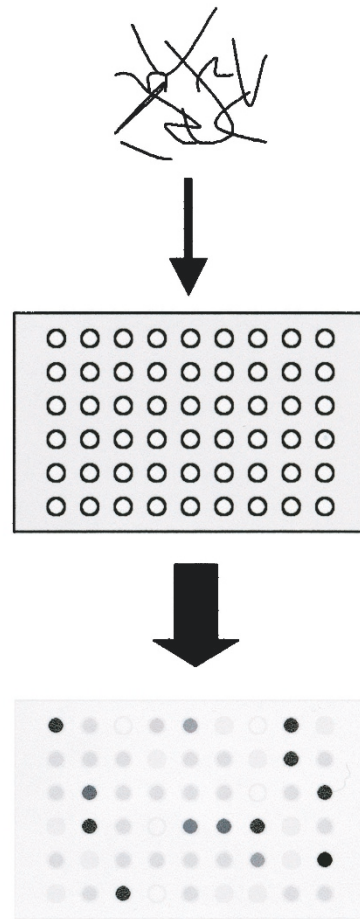
FIGURE 1. Principles of gene expression measurements using arrays. RNA samples are labeled in a reverse transcription reaction using fluorescent nucleotides (*left*) or radioactive nucleotides (*right*), and hybridized to arrays representing multiple different genes. For fluorescent labeling, two or more samples labeled with differently colored fluorescent markers (*e.g.*, red and green) are cohybridized to an array. Level of RNA for each gene in the sample is measured as intensity of fluorescence or radioactivity binding to the specific spot. With fluorescence labeling, relative levels of expressed genes in two samples can be directly compared with a single array.

cancers, and through analysis of expression in a large set of tumor samples, a set of genes that are consistently and selectively overexpressed in this type of cancer were identified (9). These genes will be further evaluated as possible therapeutic targets. In another "gene discovery" study, cDNA arrays were used to screen gene expression differences between hormone refractory and hormone sensitive prostate cancer cells, leading to identification of two genes that were overexpressed in the hormone refractory cells: insulin-like growth factor–binding protein 2 (IGFBP2) and 27-kd heat-shock protein (HSP27; 10). This finding was followed by immunohistochemical analysis of these two new candidates in 208 primary prostate cancers and 30 hormone-refractory local recurrences. High expression of IGFBP2 protein was found in 100% of the hormone-refractory clinical tumors, in 36% of the primary tumors, and in 0% of the benign prostatic specimens, and overexpression of HSP27 protein was demonstrated in 31% of the hormone-refractory tumors, in 5% of the primary tumors, and in 0% of the benign prostatic specimens. Thus, this investigation using cDNA arrays has led to the identification of promising markers for identifying prostate cancers with potential to develop androgen independence.

Functional Genomics and Cancer Classification

One of the most exciting applications of gene arrays is the classification of cancers (and potentially other diseases) by gene expression profiles. If different classes of neoplasia are already defined by some meaningful characteristic, such as tissue of origin or distinctive histopathological features,

gene arrays can be used to screen expression levels of many genes to find those that reliably distinguish well-characterized examples of the different classes of tumors. The best candidate markers can subsequently be used to correctly classify tumors that are currently difficult to categorize. Identifying gene expression profiles characteristic of predefined subsets of cancers is known as *class distinction.*

In many situations, however, our current classification structures are inadequate. For example, it is currently not possible to accurately predict which *in situ* epithelial proliferations have potential to develop into invasive cancers or to predict how most of the common forms of cancer will respond to radiation therapy or chemotherapy. Creating entirely new, clinically meaningful classification systems—or class discovery—represents a far more challenging problem than class distinction but is clearly an important goal for pathologists. For class discovery, previous distinguishing characteristics of different classes—and possibly not even the number of different classes—is unknown. In this situation, gene expression data is analyzed to find previously unrecognized subsets of tumors that share gene expression profiles. The gene expression profiles represent objective measures of the cellular phenotype and, if properly analyzed and categorized, can lead to an objective classification structure.

An early display of the utility of gene expression analysis for cancer classification came from a laboratory at the Massachusetts Institute of Technology in a study of acute lymphoblastic leukemia (ALL) and acute myelogenous leukemia (AML). Using oligonucleotide arrays from Affymetrix and novel informatics tools, this group showed how gene expression profiles could be used to distinguish these two types of leukemia and how, in a simulated situation blinded to the histological diagnosis, the two classes could have been "discovered" by the expression profiles alone (11).

An even more challenging example of class discovery was the identification of two molecularly distinct forms of diffuse large B-cell lymphoma (DLBCL) by groups from Stanford and the National Heart, Lung, and Blood Institute (12). In this instance, the two forms of DLBCL were identified on the basis of gene expression patterns indicative of different stages of B-cell differentiation. One type expresses genes characteristic of germinal center B cells (germinal center B-like DLBCL), and the second type expresses genes normally induced during *in vitro* activation of peripheral blood B cells (activated B-like DLBCL). Remarkably, this molecular classification has prognostic value independent of stratification by the usual clinical grading: germinal center B-like DLBCL patients generally respond fa-

vorably to therapy and have a prolonged survival, whereas activated B-like DLBCL patients respond poorly to therapy and succumb to the disease. Although morphology alone can be used to predict outcome for many cases of DLBCL, this gene expression data provides a framework for a more objective and comprehensive classification system that could be applied uniformly by pathologists everywhere.

Early progress has also been made in the development of classification schemes for solid tumors as well. A group from the National Human Genome Research Institute found two distinct groups of gene expression patterns in a set of pathologically similar melanoma, with a reduced expression of genes involved in cell motility in one of the subsets (13). Reduced motility of cells derived from these tumors was confirmed in cell culture systems, and although outcome data regarding the patients with the tumors is incomplete, there is a suggestion that the melanomas with higher motility have a worse clinical outcome.

In another cancer classification study, the Stanford group has shown that different subsets of breast cancer can be appreciated by gene expression profiling, including two subsets of estrogen receptor– (ER) negative cancers (14). One of these ER-negative subsets has expression of basal cell types of cytoskeletal genes, and the other is notable for high expression of the *Erb-B2* gene, suggesting that these subsets of ER-negative breast cancers have significantly different tumor biology. Although these investigations are still in relatively early stages, it is evident that gene expression profiles will lead to recognition of many more diagnostic categories than we can recognize by currently applied methods.

Toxicogenomics and Pharmacogenomics

Functional genomics will impact significantly on a number of other disciplines related to pathology. For example, many toxicologists are now considering the application of gene expression profiling, using cells exposed to test compounds, to reduce the use of animals as well as improve test results (15). Current animal testing protocols require months or years for results and carry a considerable expense, and thus many compounds are never subjected to testing in important areas such as carcinogenicity. With the use of gene arrays, it is hoped that profiles of gene expression changes induced by carcinogenic agents can be developed and then used to evaluate new compounds for similar effects.

Similarly, pharmacologists are using functional genomics to identify new molecular targets for therapy as well as to identify drug response genes that

correlate with response to therapy (16). The previously discussed example of Her2/Neu and p53 expression correlating with favorable response to anthracycline chemotherapy is one example in which molecular markers provide predictive information and not only prognostic information. It is likely that pathologists will increasingly be expected to provide data and interpretation regarding expected response of different cancers to different chemotherapeutic agents based on expression profiles of the tumors.

Challenges and Future Perspectives

Although functional genomics offers many new opportunities to advance our understanding of disease and improve diagnosis, many challenges remain. Analysis of high-throughput gene expression data requires creative, computationally intensive statistical approaches to deal with high-dimensional data sets. Furthermore, the noise and technical variability of the measurements can be significant and will require continued work from molecular biologists and biophysicists to improve and standardize the technology.

Yet another issue is the complexity of working with human tissues, a factor that is often not well appreciated by the molecular biologists conducting many of the early studies in this area. For example, many cancer tissues have more lymphocytes and stromal cells than neoplastic cells, and the analysis of gene expression specifically in the neoplastic cells of these heterogeneous tissues will not be straightforward. Laser capture microdissection (17) is one technique that can be applied to the purification of neoplastic cells, and recently a simple procedure for scraping nearly pure clusters of neoplastic cells from cut surfaces of tumor tissue was described (18). Application of appropriate methods for purifying tumor cells will require pathologists to take leadership roles in molecular classification projects.

It is not yet clear which methodologies will eventually prove to be optimal for measuring gene expression in the practice of pathology. Although much recent laboratory discovery has been based on the use of gene arrays, this technology, and problems associated with preservation of RNA, may prove to be cumbersome for routine diagnostic purposes. In addition to gene arrays, proteomics technologies are also rapidly being developed (19), and because protein is more stable than RNA, protein chips may be more practical than gene chips for comprehensive analysis of gene expression in tissue samples. Furthermore, it can be argued that protein levels are more closely related to cellular function than mRNA levels and that protein chips will provide a clearer picture of cellular phenotype.

On the other hand, it is quite possible that neither gene arrays nor protein chips will be used extensively for tissue diagnosis. On the basis of the experience of studies cited above, it appears that relatively small numbers of gene markers are ultimately important for making the distinctions that arrays helped to discover. Thus, it is quite possible that gene arrays (and related technologies in proteomics) will be used mainly for research and that individual gene products will be continue to be measured by immunohistochemistry or flow cytometry.

Although the way pathologists measure gene expression may ultimately come full circle—back to immunochemical methods—it is impossible to ignore the impact that the new technologies will have on the practice of pathology. It is virtually certain that new markers will be discovered at a faster pace using high-throughput approaches than traditional approaches and that pathologists willing to exploit these new technologies will have unprecedented opportunity to find new markers for improved diagnosis.

REFERENCES

1. Thor AD, Berry DA, Budman DR, Muss HB, Kute T, Henderson IC, *et al*. erbB-2, p53, and efficacy of adjuvant therapy in lymph node-positive breast cancer. J Natl Cancer Inst 1998; 90:1346–60.
2. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science 1995;270:484–7.
3. Chien Y, Becker DM, Lindsten T, Okamura M, Cohen DI, Davis MM. A third type of murine T-cell receptor gene. Nature 1984;312:31–5.
4. Ivanova NB, Belyavsky AV. Identification of differentially expressed genes by restriction endonuclease-based gene expression fingerprinting. Nucleic Acids Res 1995;23:2954–8.
5. Lennon GG, Lehrach H. Hybridization analyses of arrayed cDNA libraries. Trends Genet 1991;7:314–7.
6. Shalon D, Smith SJ, Brown PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res 1996;6:639–45.
7. Bertucci F, Bernard K, Loriod B, Chang YC, Granjeaud S, Birnbaum D, *et al*. Sensitivity issues in DNA array-based expression measurements and performance of nylon microarrays for small samples. Hum Mol Genet 1999;8:1715–22.
8. Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ. High density synthetic oligonucleotide arrays. Nat Genet 1999;21:20–4.
9. Wang T, Hopkins D, Schmidt C, Silva S, Houghton R, Takita H, *et al*. Identification of genes differentially over-expressed in lung squamous cell carcinoma using combination of cDNA subtraction and microarray analysis. Oncogene 2000; 19:1519–28.
10. Bubendorf L, Kolmer M, Kononen J, Koivisto P, Mousses S, Chen Y, *et al*. Hormone therapy failure in human prostate cancer: analysis by complementary DNA and tissue microarrays. J Natl Cancer Inst 1999;91:1758–64.
11. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, *et al*. Molecular classification of cancer: class

discovery and class prediction by gene expression monitoring. Science 1999;286:531–7.

12. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000;403:503–11.

13. Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, *et al.* Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 2000;406: 536–40.

14. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, *et al.* Molecular portraits of human breast tumours. Nature 2000;406:747–52.

15. Lovett RA. Toxicogenomics. Toxicologists brace for genomics revolution. Science 2000;289:536–7.

16. Weinstein JN. Pharmacogenomics—teaching old drugs new tricks. N Engl J Med 2000;343:1408–9.

17. Emmert-Buck MR, Bonner RF, Smith PD, Chuaqui RF, Zhuang Z, Goldstein SR, *et al.* Laser capture microdissection. Science 1996;274:998–1001.

18. Maitra A, Wistuba II, Virmani AK, Sakaguchi M, Park I, Stucky A, *et al.* Enrichment of epithelial cells for molecular studies. Nat Med 1999;5:459–63.

19. Pandey A, Mann M. Proteomics to study genes and genomes. Nature 2000;405:837–46.

Book Review

Knowles DM, editor: Neoplastic Hematopathology, Second Edition, 1957 pp, Philadelphia, Lippincott Williams & Wilkins, 2001 ($249.00).

Perusing this rather heavy book is an onerous task, but it cannot be (even in jest) compared with the time and effort that the editor and his collaborators must have spent in compiling this encyclopedic work. On the other hand, it is probably the only way that the rapidly expanding field of hematopathology could be surveyed and presented to all those who are interested in finding out what is worth knowing at this point in time. I should also add that there are not too many pathologists besides Dan Knowles who could have done the editing of such a Henke-Lubarschian *Handbuch* for the new century.

As indicated in the title, the book deals with leukemias, lymphomas, and related diseases affecting the hematopoietic system. In addition to chapters dedicated to specific hematologic entities, there are also extensive descriptions of basic facts, such as normal hematopoiesis, special techniques currently used for diagnostic purposes (*e.g.*, flow cytometry or cytogenetics), and *de rigueur* classifications of lymphomas. It seems that the contributors have been encouraged by the editor to provide as comprehensive coverage as possible, and that most of them have followed literally those instructions and taken their assignments very seriously. To a general pathologist like me, the text was often overbearing, and often way above my head. On balance I should immediately add that there are also many easily accessible facts, diagnostic pearls, useful tables, and diagrams. Nevertheless, I still came away with the impression that the book was primarily aimed at those who are practicing hematopathology much more intensely than this reviewer. This impression of mine was confirmed in discussions with a number of my hematopathology friends, almost all of whom had only compliments for the book. I also found out that some of them did not agree with all that was written in some of the chapters, especially those on lymphomas, but that is nothing more than the usual disagreement encountered in every rapidly expanding field of medicine. Furthermore, it would be impossible not to find some fault in a multi-authored book of this size. The positive remarks nevertheless outweighed heavily the negatives, and that was true at every level of expertise of my collocutors, which included among others hematologists, as well as pathologists, fellows, and residents.

From my readings and the discussion-based "field work," I concluded that this rather complex book will be used mostly by hematopathologists and clinical hematologists/oncologists, their fellows, and residents. As a most comprehensive source of current information on neoplastic blood diseases, it should, however, appeal also to general diagnostic pathologists, and I foresee that many of us will be using the Knowles book in practice. I do not know of any other book that covers hematopathology in such a detailed way.

Ivan Damjanov
University of Kansas School of Medicine
Kansas City, Kansas