Gene Expression Profile Analysis by DNA Microarrays: Promise and Pitfalls

Hadley C. King, BA
Animesh A. Sinha, MD, PhD

Increasing numbers of human diseases, both acquired and genetic, are being considered to be based at least in part on alterations in DNA sequence. For most diseases, inheritance and acquisition are likely to be complex and polygenic. The efforts of the Human Genome Project to elucidate the structural genetic background by identifying the chromosomal positions and genomic organization of between approximately 30,000 and 35,000 human genes are nearly complete. Based on this structural knowledge, a byproduct should be a better "scaffolding" to help link specific genes to susceptibility to various human diseases. However, to understand how the products of these genetic linkages work together to orchestrate the initiation and progression of particular complex diseases, there will be a need to apply a functional genetic rather than a structural genetic approach.

Until recently, functional genetic studies have generally been of limited scope, only able to elucidate the role of 1 or a few genes at a time in 1 system. Information on the specificity and relative abundance of expression products has traditionally been obtained by techniques such as RNA Northern blot hybridization and ribonuclease protection assays. Somewhat more sophisticated methods, such as differential display and Serial Analysis of Gene Expression, have been used to screen larger numbers of complementary DNA (cDNA) clones. However, technical limitations render these techniques nonconducive to large-scale genetic survey.

To this end, a powerful new technology is emerging, using hybridization to nucleotide arrays, the so-called gene chips. This technological intersection of biology and computers enables the reliable screening of a vast number of genes simultaneously and is amenable to automation. On a nylon membrane or glass surface, genespecific cDNAs can be spotted, or oligonucleotides can be synthesized in situ by a combination of photolithography and oligonucleotide chemistry. This permits simultaneous monitoring of the expression of thousands of genes in a single step. Individual chips can be customized to include any chosen set of fully or partially characterized genomic or expressed sequences. Chips can monitor over 50,000 unique sequences. The power of these chips lies in the potential for comparative expression...
sion studies in diseased vs normal samples, and in documenting changes at different stages during the natural course of the disease or in response to treatment. It provides the researcher with a new arsenal to analyze underlying pathomechanisms on a grand scale and also to review the rationale of therapeutic concepts.

However, despite the enormous potential of this revolutionary technology, there are several issues and possible pitfalls that attenuate the power of microarrays. First, the definition of normal in expression comparisons is neither precise nor unambiguous. Second, the heterogeneity of the tissues being studied complicates the meaning of the expression profiles. Third, the statistically valid comparability of arrays is an unresolved problem. Fourth, the vast quantities of data create a logistical logjam for analysis, presentation, and archiving. Finally, confirmational studies are needed to corroborate the biological significance of microarray data (FIGURE 1).

TROUBLE WITH NORMAL

The standard normal vs diseased tissue type of comparison, which is the basic design foundation of profiling studies, may be more quicksand than bedrock. Normal is not so easy to define—neither is diseased. Gene expression in normal tissue is likely to be dependent on several factors involving patient and sample variation. These factors will also have an impact on expression profiles of diseased tissue.

Patient Variation: Ethnicity, Sex, Age, Genetic Background, Disease States

The ethnicity, sex, age, and genetic background of a patient are likely to affect the gene expression profiles of many tissues to varying extents. A simple example is provided by the expression profiles of genes involved in scalp and body hair follicle activity, which can be expected to vary over a normal range under the influence of all of these sources of patient variation. The effects of these parameters on gene expression are likely to be subtle but pervasive, not fully understood at this time, and quite problematic for defining normal.

The presence of disease in a subject who is the source of tissue for control purposes, presents further potential variabilities. For example, there may be a significant difference in the conclusions reached by 2 similar microarray expression profiling studies. One may compare genes expressed in a patient’s diseased lung tissue with those expressed in normal, nondiseased lung tissue from the same patient, and another may compare genes expressed in the same patient’s diseased lung tissue with those expressed in normal, nondiseased lung tissue from a healthy control or normal individual. Moreover, it is also possible that seemingly unrelated disease states may influence gene expression at distant sites. For instance, the presence of diabetes in 1 of 2 renal cancer patients may complicate the direct comparison of renal tissues.

Sample Variation: Proximity to Disease, Anatomic Location, and Developmental Range

Yet another complication derives from the proximity of the normal tissue used as a control for the diseased tissue. Tissue adjacent to an area of disease may not be normal despite absence of evidence of disease clinically or under the light microscope. Normal-appearing tissue near a tumor could, for example, be genotypically altered or exhibit an altered gene expression profile. Moreover, factors such as the degree of disease-associated inflammation may have a significant impact on gene expression profiles. Other bystander effects, epiphenomena, or secondary disease processes could all play important roles in determining expression profiles within these adjacent, so-called normal tissues. These factors must be considered in the choice of normal.

The precise location within a particular organ may be another important factor that affects gene expression. For example, just as location relative to the urethra may influence expression profiles in the prostate, skin from the nose, back, and palm are certain to have different expression profiles as well, despite all being from the same organ. Thus, site and specific anatomic location must also be taken into account in a description of normal.

It must also be kept in mind that the definition of normal actually represents a dynamic state. All tissues, which are composed of early and late-stage cells, have a normal developmental range. For example, normal epithelium in prostatic ducts ranges from atrophic to resting to hyperplastic, and each has a unique pattern of gene expression. A 3-dimensional analytic approach is a strategy that has been used to address some of these concerns about defining normal. Cole et al used a 3-dimensional model to characterize the entire prostate gland in their study of gene expression profiles in prostate cancer. In this study, whole-mount prostates—
Disease-Related Variation

Of course, many of the parameters that affect normal expression profiles (patient ethnicity, age, sex, and genetic background, location within an organ, and developmental stage) will also affect disease expression profiles. Disease heterogeneity, including subtype, activity, severity, stage, and previous as well as current treatments, also may have a significant impact on gene expression.

Categorizing and subgrouping patients on entry into a study may be useful to control for as many of these factors as possible. However, there may be problems surrounding attempts to define microarray-based categorization on the basis of another imperfect categorization system, such as histology, as these groups are sometimes arbitrary or inconsistently designated. Nevertheless, determining whether gene expression profiles correlate with existing clinical or histological categories can provide new insights into the meaning of these categories as can new methods of classifying cancers or other diseases into specific diagnostic categories based on their gene expression signatures. Several studies have been able to establish expression-based criteria (class predictors) for existing categories and then use these new criteria to categorize new cases (class prediction). Global profiling may also allow the development of new classification systems based on gene expression alone (class discovery).

Thus, when possible, it will be of value to profile a range of normal and diseased cell populations from a number of patients to distinguish between differences in expression that are relevant to the disease process and those reflecting the biological spectrum of the normal tissue or that have occurred for reasons unrelated to the disease. The significance of this distinction is further appreciated when taking into account the vast quantity of data generated from microarrays and the potential for confounding interpretation from the inclusion of differential expression unrelated to disease processes.

It is worth noting, however, that the issues of patient and sample variability are not unique to microarray experiments. In fact, microarray experiments, in contrast to classic single-gene experiments, may actually provide the tools for identifying this heterogeneity. For example, DNA microarrays have been used to study physiological variation in gene expression on a genomic scale in 60 cell lines derived from diverse tumor tissues. Cluster analysis allows the identification of prominent features in gene expression patterns that appear to reflect molecular signatures of the tissue from which the cells originated.

HETEROGENEOUS CELL POPULATIONS

A further complication encountered with expression profiles is that any given tissue is composed of several cell types, members of which are likely to be within a spectrum of dynamic functional states. For example, a simple punch biopsy of the skin may contain keratinocytes, melanocytes, Langerhans cells, Merkel cells, adipocytes, smooth muscle cells of arterioles, striated muscle cells of the panniculus carnosus, blood cells including immune system cells, and cellular elements of blood vessels, nerves, hair follicles, sebaceous glands, and sweat glands. Moreover, cells from each of these populations will be at various stages of development and levels of activation, performing different functions and responding to disease processes or treatments in different ways and to varying extents. The result is a highly heterogeneous sampling of cells, each expressing a unique set of genes. An expression profile generated from a microarray study of the RNA in such a sample will thus represent merely a snapshot of the genes expressed by a plethora of cells at a moment in time. Such extensive cellular heterogeneity complicates the ability to draw conclusions about specific processes occurring within a tissue specimen. An illustrative example is provided by Stanton et al, who used microarrays to identify genes differentially expressed during myocardial infarction. The expression profiles they studied represented transcripts from cell populations as diverse as immune system cells, which migrated to the infarct region and are responsible for the inflammatory response, cardiac myocytes within the ischemic area undergoing apoptosis and necrosis, fibroblasts undergoing proliferation and participating in the formation of scar tissue to replace the infarct, and cardiac myocytes undergoing hypertrophy to compensate for the loss of cells in the infarct area. The issue of such cellular heterogeneity was avoided by categorizing the differentially expressed genes into functional categories to look for patterns indicative of cardiac remodeling without attempting to attribute specific transcripts to specific cell types. For gene expression studies involving samples with mixed cellular populations, further investigation, such as with in situ messenger RNA (mRNA) hybridization, may be necessary to localize the transcripts before conclusions can be drawn about the roles of specific genes in specific cell types during the disease process.

Laser Capture Microdissection

An ingenious but technically delicate approach to the study of complex bio-
logical samples has become possible with the development of laser capture microdissection (FIGURE 2).34 This technique allows for the rapid and accurate procurement of cells from specific areas of tissue under direct microscopic visualization, and thus makes the molecular genetic analysis of defined populations in their native tissue environment possible.35 Sgroi et al36 demonstrated the feasibility of combining laser capture microdissection with high-throughput cDNA arrays. They showed that in vivo subpopulations of malignant cells from multiple stages of breast cancer progression could be separated from nonmalignant populations, and their expression profiles could subsequently be analyzed using microarrays.

The potential is powerful. Specimens could be separated into tissue layers; for example, separating a skin biopsy into epidermis, dermis, and hypodermis. Tissues could be further differentiated into specific structural components, such as dermis into blood vessels, adipose, arrector pili, and sebaceous glands. Structures could be separated into defined cell types, such as blood vessels into endothelial cells, erythrocytes, and lymphocytes. Cell types could even be separated into marker-defined subtypes, such as lymphocytes into CD4 and CD8 cells. Expression profiles from refined and defined structures and cell types likely would be extremely valuable in the study of disease.

Potential aside, there are significant limitations to this technology at the present time. The standard protocols for fixing and embedding tissue samples from surgical resections were not designed to be compatible with microarray experiments, with or without laser capture microdissection. Typically, tissue suspected of being important for diagnosis and staging is processed through aldehyde-based fixatives, such as formalin, which damage mRNA integrity.37 If frozen tissue is available, mRNA can be recovered and studied from dissected cell populations. However, frozen tissue sections are technically difficult to prepare, the histology is often severely compromised, and the tissue available may contain only a limited portion of the lesion.14

Moreover, the sample amounts generated from laser capture microdissection can be small, even as miniscule as a single cell.38 Consequently, the yields of RNA are low. Arrays have a threshold for the quantity of molecular starting material: at least 5 to 15 µg for oligonucleotide arrays and between 2 and 100 µg for cDNA arrays, depending on the manufacturer, the source of the RNA, and the use of signal amplification.39-40 Studies that have successfully integrated laser capture microdissection with microarray technology have used samples of approximately 1 × 10⁴ to 1 × 10⁵ cells with 95% to 98% homogeneity as determined by microscopic visualization.36-41 If needed, amplification techniques may be used to generate sufficient genetic material for microarray hybridizations.42 Laser capture microdissection is an intriguing technology, but time will tell whether its potential is realized.

Although some biological issues related to gene expression may be complicated by the presence of heterogeneous cell populations in studied samples, it is also true that some biological conditions can be understood only in the context of these heterogeneous cell populations. The nature of global gene expression experiments is to uncover differences between 2 biological samples, including those differences based on diverse cell populations. For example, to appreciate a disease that is characterized by an inflammatory infiltrate, it must be understood that the inflammatory infiltrate is part of the disease and is part of the difference between diseased and nondiseased tissue. Thus, the isolation of specific cell populations for study is not necessarily required or even desirable in all instances.

**MAKING MICROARRAYS COMPARABLE**

Ideally, microarray experiments should be comparable both within and between laboratory or manufacturing systems, but obtaining consistent and comparable data is a critical challenge for microarray-based expression analysis. Major sources for the observed variability of microarray data include the normal physiological gene expression variations in different samples and the noise introduced in the microarray assay process.42

**Physiological Gene Expression Variation**

Inextricably linked to the issues of patient and sample variability and tissue heterogeneity discussed above, is the

©2001 American Medical Association. All rights reserved.
problem of normal gene expression variations and how to distinguish these variations from significant disease-associated changes. Few studies have systematically investigated physiological expression changes, but data from in situ hybridizations suggest that normal variance for many tightly regulated tissue-specific genes can be within 20% to 30%. However, there can be as much as 2- to 4-fold random fluctuations for many genes in yeast. Affymetrix (Santa Clara, Calif) guidelines have suggested that for most of the "housekeeping" genes in human tissues, which are likely to be less tightly regulated, differences of less than 4-fold are probably not biologically significant. Consequently, a significant portion of microarray data variability for high- or medium-abundance mRNAs may simply be due to normal expression variations. Several previous studies have used arbitrary 2-fold change criteria to define significant expression change. However, the 2-fold threshold has been shown to be statistically invalid even for duplicate experiments. In a recent study that used cDNA microarrays to profile gene expression in samples of normal skin from breast-reduction surgeries, 71 of 4400 genes were found to demonstrate variability in expression greater than 2 SDs from the mean of each gene. These included genes coding for transport proteins, gene transcription, cell-signaling proteins, and cell-surface proteins. Thus, physiological variation should be considered in the analysis and interpretation of microarray data. More stringent criteria for defining significant expression change may be useful.

Noise in the Microarray Assay Process

For the tightly regulated (mostly low abundance) mRNA species, inconsistencies introduced at any stage of the microarray-based assay process may play a major role in data variability. Due to the miniaturization and the large number of genes involved, it is difficult to maintain consistent processing conditions for each sequence across multiple assays, and obtaining accurate absolute signals is unlikely. Noise may be introduced by slide heterogeneities, printing irregularities (eg, pin-to-pin variations), and spotting volume fluctuations. Some of the systematic variations may be reduced by the inclusion of controls, but random fluctuation at various manufacturing stages cannot be completely controlled and can accumulate quickly in a complicated assay.

In certain microarray systems, 2 samples are competitively hybridized to 1 array using different fluoros for labeling. In other systems, there is only single-sample hybridization. A 2-color system might be expected to be more reliable since variations in spot size or amount of cDNA probe on the chip should not affect the signal ratio (both signals are derived from the same spot). However, this only holds true if signals are well above the background in both detection channels. In fact, the signal level for most of the tightly regulated genes will likely be close to the background level. In addition, background level on a slide can also vary significantly from spot to spot due to factors such as unevenness in slide surface properties, dust contamination, and incomplete washing, leading to high levels of signal variability for low-abundance mRNA species even in 2-color systems.

The high levels of variability of microarray data also mean that subtle changes in experimental conditions may significantly alter the results, making it difficult for separate laboratories to compare experimental data. In addition, the lack of standard controls, the predominant use of relative signals (ratios), and the adoption of incompatible data formats contribute to poor comparability between studies.

Despite the hard-wired variability introduced by chip manufacturing conditions, most of the published studies to date using microarray-based expression analysis include only limited numbers of replicates. In fact, many studies conduct the experiment only once. Considering the potential sources of assay variation, the need for sufficiently replicated studies is underscored.

Microarray Data Normalization

Because of variability of microarray data for single sample arrays and for further analysis of 2-color system arrays, each must be brought into the same scale to compare 2 or more arrays. How to perform this normalization of gene expression levels across multiple arrays, thus removing systematic variation between the arrays and rendering different experiments comparable, remains an issue that is not yet fully resolved. Many of the early microarray studies in the literature simply ignored this issue. A more statistically rigorous approach is needed.

One difficulty has been that leading microarray manufacturers have not published statistical error models for their products. Thus, users are unclear how much to adjust data for variations in spot intensity, hybridization patterns, and intensity measurement sensitivity. Software does exist to allow for array-to-array comparisons by using a scaling factor to normalize gene expression patterns across arrays. However, in general, these algorithms assume that intensity differences between arrays are linearly related with a zero y-intercept. This assumption allows software to trim the tails off distributions of expression from different arrays at statistical cutoffs and then simply move the distributions along an axis to a common level to provide comparisons. However, this linear relationship often does not hold true. When the average expression level of 1 array is higher than that of a second array, a longer tail will be trimmed from the second array. Thus, a greater number of genes from the first array will be counted as being expressed because their expression level is above the statistical cutoff point. In this case, the 2 arrays cannot be considered comparable.

Although bioinformatic software has recently been developed that offers more statistically robust normalization, the cost of these commercially available programs (combined with the already expensive microarrays) has been prohibitive for many research-
Standardization of these processes awaits the development of improved methods of normalization leading to valid statistical models widely available to all researchers.

To this end, Schadt et al\(^5\) have developed a standard nonlinear curve technique for normalizing the data in arrays that do not demonstrate a linear relationship between data sets. This model performs well when the 2 samples being compared demonstrate a low number of differentially expressed genes. However, when expression profiles of 2 samples vary to a greater extent, Schadt et al\(^5\) recommend a rank-selection method. Using this method, genes expressed on an array are ranked from highest to lowest level of expression. Then, for the array expressing a greater number of genes, the genes with the lowest expression levels are removed from the list until the 2 arrays list a comparable number of expressed genes. This type of rank-selection method has gained support from other groups, but it too has limitations.\(^5\) Removing low-expression level data points restricts the study to the more extreme and easily detected entities, a technique that blunts the genomic-scale potential of microarray technology.

Efforts continue to improve comparability between arrays. Jones\(^5\) recently applied a statistical model to normalize spotted cDNA array data that takes into account not only the differences in numbers of genes expressed between arrays, but also the interarray variations in fluorescent dye intensity and mechanical error occurring in the printing process. Nevertheless, the issue of how to properly normalize array data has not been settled. Researchers must continue to demand statistical rigor in their comparisons before they can believe the mathematical results of their data.

LOGISTICAL LOGJAM

Microarrays deliver massive amounts of data on tens of thousands of genes. The result is an immense quantity of biological information that must be analyzed, presented, and archived in a meaningful way.

Data Analysis

In human studies, the number of hybridizations that can be performed for any set of experimental conditions is often restricted by the limited number of obtainable tissue samples and by the expense of arrays. Restricted numbers of hybridizations for each experiment hamper the ability to assess the biological significance of variation within or between given sets of conditions. Thus, for the assessment of thousands of genes in a setting of limited hybridizations, the importance of reliable and sophisticated algorithms for data analysis becomes amplified.\(^5\)

A logical beginning is to examine the extremes, that is, genes with significant differential expression in individual samples. For example, a comparison of 2 samples can be visualized in the form of a simple bivariate scatterplot in which the expression profile of 1 sample (x-axis) is plotted against that of the second sample (y-axis). The distribution pattern generally demonstrates that the expression ratios cluster around the line in which x is equal to y (indicating comparable levels), with individual genes falling varying distances from this line. Additional lines can be placed on the scatterplot to represent various fold changes of expression. Data points that fall above or below these lines represent genes exhibiting expression ratios greater or less than the specified fold change. Thus, one can begin by examining those genes that demonstrate a 10-fold or greater change in expression level. To expand the number of genes under investigation, one can examine genes that demonstrate a 5-fold or greater change, or a 3-fold or greater change, and so forth. Many studies define a 2-fold or greater change in expression level to represent significant differential expression. The 2-fold threshold, however, as noted above, has been shown to be statistically invalid.\(^6\)

Although this simple technique can be efficient and effective for focusing on expanding sets of differentially expressed sequences, again, such an analysis does not take advantage of the full potential of genome-scale experiments to enhance our comprehension of cellular biology that would be provided by an inclusive analysis of the entire repertoire of transcripts in a cell as it goes through a biological process.\(^3\) A more holistic approach, which allows the deciphering of patterns from the entire data set, is needed.

Data Organization and Presentation

Statistical algorithms can be applied to detect and extract patterns within profiling data. It is a basic assumption of many gene expression studies that knowledge of where and when a gene is expressed provides information about the function of the gene. Therefore, an important beginning is to organize genes on the basis of similarities in their expression profiles.\(^3\) However, even this basic tenet deserves critical consideration. Similarity of gene expression profile does not mandate similarity of function or mechanistic pathway, and it may be purely coincidental. Nevertheless, the idea of clustering genes on the basis of their expression patterns is well established and cluster analysis has become the most widely used statistical technique applied to large-scale gene expression data.\(^2\)

Although various cluster methods can usefully organize tables of gene expression measurements, the resulting ordered but still massive collection of numbers remains difficult to assimilate. Thus, another important component of genome-wide expression data exploration is the development of powerful data visualization methods and tools. Approaches have been developed that present clustering results in simple graphical displays such as dendrograms, which represent relationships among genes by a tree whose branch lengths reflect the degree of similarity in expression between the genes. Similarity is mathematically defined.\(^5\) The computed trees can be used to order genes in the original data table such that genes or groups of genes with similar expression level patterns are

©2001 American Medical Association. All rights reserved.
the original experimental observations. Visual assimilation is then more intuitive.

**Data Archiving and Mining**

Ultimately, successful interpretation of gene profiling studies is likely to be dependent on the integration of experimental data with external information resources. As multiple experiments involving multiple cell types and tissues from multiple laboratory groups accumulate, data archiving may well become the watershed issue. Ideally, all data, in a suitably standardized form, would be freely accessible in the public domain. Even assuming a willingness to share the data, such utopian goals would require a user-friendly and powerful database system and standardization of correction and normalization procedures such that data points from various projects become comparable. The National Center for Biotechnology Information Entrez system (http://www.ncbi.nlm.nih.gov/Entrez/) does provide useful data in this regard, but current databases may be limited in scope or computability. A major focus of infrastructure development to support genomic-scale expression studies will need to be in the area of electronic biological pathway databases and resources.

**CONFIRMATIONAL STUDIES**

The development of more sophisticated analytical algorithms and databases will help lend credence to the biological significance of differential gene expression determined by microarray analysis. In the meantime, several studies have begun to examine the sensitivity and specificity of microarray-based experiments. Sensitivity, defined as the minimum reproducible signal detected by a given array scanning system, has been reported for microarrays to be approximately 10 mRNA copies per cell, which is slightly inferior to the sensitivity of Northern blot analyses. Specificity studies showed that for a given probe any nontarget transcripts with more than 75% sequence similarity may show cross-hybridization. The problem of clone misidentification and the need for clone confirmation have also been addressed. One study found that of 1189 bacterial stock cultures, only 62.2% were uncontaminated and contained cDNA inserts that had significant sequence identity with published data for the ordered clones. Thus, the use of sequence-verified clones for cDNA microarray construction is warranted.

**Methods at the RNA Level**

Reverse transcriptase polymerase chain reaction (RT-PCR) is a method often used to verify microarray data. Although RT-PCR is not well suited to quantitation, the relative technological ease of this assay and the ability to rapidly monitor multiple samples make it a useful technology. Hybridization data can be verified and multiple putative markers can be screened in a short period.

Several other studies have used real-time quantitative RT-PCR (TaqMan, PE Applied Biosystems, Foster City, Calif) Real-time PCR is a technique that increases the quantitative ability of RT-PCR by providing accurate and reproducible information on RNA copy number (Figure 3). In this method, a fluorogenic probe (labeled at the 5' end with a reporter fluorochrome and at the 3' end with a quencher fluorochrome) is annealed to 1 strand of the target cDNA sequence between the forward and reverse PCR primers. As Taq polymerase extends the forward primer, its nuclease activity displaces the reporter fluorochrome from the quencher label and allowing the detection of a fluorescent signal.
signal that is proportional to the amount of PCR product generated in each cycle. For mRNA expressed at moderate-to-high levels, and for which cDNA probes are available, Northern blot analysis works well, but it is not well suited for low-copy mRNA. Furthermore, only a small number of genes can be analyzed with this conventional method.

Methods at the Protein Level

DNA microarray technology is limited to the study of gene expression at the mRNA level. However, it has been established that mRNA levels do not necessarily correlate with protein levels. Moreover, the level of expression or even presence of a protein is not tightly linked to physiological consequences. An investigation conducted by Winzeler et al, for example, provides a cautionary tale. Their study demonstrated that genes upregulated in yeast growing in minimal medium did not prove to be more important for growth than genes that were not upregulated. They found only 2 of 8 genes required for yeast growth in minimal medium to be induced. The lesson to be learned is that genes that are not differentially expressed may be of equal functional importance in disease states compared with those that are.

Furthermore, the regulation of some genes may be at the translational rather than the transcriptional level, which would preclude detection by DNA microarrays. Posttranslational modification of proteins is also an important mode of regulation that cannot be detected by DNA microarrays. Protein activity, particularly receptor activity, is heavily dependent on phosphorylation, for example. DNA and mRNA reveal nothing about whether a given protein is active, and can be deceptive when used to speculate about quantities of proteins. It has been demonstrated that the correlation between mRNA and protein abundance is less than 0.5, emphasizing that ideally, mRNA expression studies should be accompanied by analyses at the protein level. Radioimmunoassay and immunohistochemistry have been used in a number of studies. These techniques, however, are not well suited to detecting low levels of expression, and they require the availability of an antibody specific for the protein to be studied.

The field of proteomics, the large-scale parallel analysis of the proteins that are present in a cell, is developing rapidly, but has problems of its own. Proteins vary in abundance by many orders of magnitude within a given cell, and there is no PCR equivalent for the amplification of proteins. Moreover, proteins fold in many known (and unknown) ways that affect their function. The feasibility of the microarray analysis of proteins has begun to be explored. Antibodies attached to microarrays can be used to bind to and quantitatively detect proteins that have been tagged with fluorescent dyes. Skeptics doubt the plausibility of identifying thousands of unknown proteins in this manner. The diverse chemistry of various proteins poses serious difficulties, and it will be challenging to find antibodies for every protein. Thus, although it is important to incorporate protein analyses into expression profiling studies, current platforms are technically limiting.

Functional Studies

Confirming the role of a gene initially identified in a microarray experiment in animal models with transgene or knockout studies provides a particularly powerful alternative platform. Transcript function, rather than mere presence, is addressed. However, this approach is ill-suited for high-throughput conditions. It may be ideal for an in-depth investigation of 1 or 2 genes of interest, but it is not practical for confirming large quantities of profiling data.

Confirmational studies are useful to corroborate the biological significance of differential gene expression determined by microarray analysis. While improved databases and more reliable statistical models will help to lend greater authority to array data, alternative platforms can be used to assess the relevance of genes first identified by array comparisons. It should be realized, however, that the alternative technologies are not intended for large-scale analyses. Realistically, only selected sequences from the array data can be confirmed with other platforms in the short-term, a retreat from the initial purpose of the genome-scale investigation by microarray.

CONCLUSIONS

Microarrays can be expected to prove extremely valuable as tools for the study of the genetic basis of complex diseases. The ability to measure expression profiles across entire genomes provides a level of information not previously attainable. Although complicated issues must be resolved, the potential payoff is big. Microarrays make it possible to investigate differential gene expression in normal vs diseased tissue, in treated vs nontreated tissue, and in different stages during the natural course of a disease, all on a genomic scale. Gene expression profiles may help to unlock the molecular basis of phenotype, response to treatment, and heterogeneity of disease. They may also help to define patterns of expression that will aid in diagnosis as well as define susceptibility loci that may lead to the identification of individuals at risk. Finally, as specific genes are identified and their functional roles in the development and course of disease are characterized, new targets for therapy should be identified.

Despite the problems of defining normal, understanding tissue heterogeneity, making arrays comparable, analyzing and archiving massive quantities of data, and performing confirmational studies in alternative platforms, expression profiling with microarrays stands as a truly revolutionary technology. As we continue to delve into the possibilities, we will surely progress in our understanding of current issues and complications. No doubt the ride on the high-throughput highway will be exhilarating.

Author Contribution: Ms King and Dr Sinha both contributed to the conceptualization and writing of this article.
REFERENCES


