Experimental Nephrology

Exp Nephrol 2002;10:64–74

Global Analysis of Gene Expression: Methods, Interpretation, and Pitfalls

Ryan M. Fryer[a] Jeffrey Randall[a] Takumi Yoshida[a] Li-Li Hsiao[a]
Joshua Blumenstock[a] Katharine E. Jensen[a] Tudor Dimofte[a]
Roderick V. Jensen[b] Steven R. Gullans[a]

[a]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Mass., and
[b]Department of Physics, Wesleyan University, Middletown, Conn., USA

Key Words
Gene expression · HuGE index · Microarray · Bioinformatics · Standard gene set

Abstract
Over the past 15 years, global analysis of mRNA expression has emerged as a powerful strategy for biological discovery. Using the power of parallel processing, robotics, and computer-based informatics, a number of high-throughput methods have been devised. These include DNA microarrays, serial analysis of gene expression, quantitative RT-PCR, differential-display RT-PCR, and massively parallel signature sequencing. Each of these methods has inherent advantages and disadvantages, often related to expense, technical difficulty, specificity, and reliability. Further, the ability to generate large data sets of gene expression has led to new challenges in bioinformatics. Nonetheless, this technological revolution is transforming disease classification, gene discovery, and our understanding of regulatory gene networks.

Introduction

Global analysis of gene expression has emerged as a major advance in biomedical research. Traditional methods of studying gene regulation led investigators to focus on one gene at a time in any particular biological context. As a result, many important biological changes were either missed or uncovered in a serendipitous manner. Over the past 10 years, numerous methods were developed to allow investigators to examine mRNA expression levels of thousands of genes in a single experiment. These large-scale approaches are opening a new vista with regard to gene discovery, disease subclassification, and ultimately characterization of coordinately regulated gene networks. Herein, we discuss utility, strategies, and caveats of global analysis of mRNA expression. We will not discuss the methods used to isolate differentially expressed genes such as subtraction hybridization, as these are not typically used to quantify transcript levels.

Utility of Global Expression Analysis

Global analysis of RNA expression has three fundamental but related uses: disease subclassification, identification of key genes, and elucidation of biological path-

KARGER
Fax + 41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

1018–7782/02/0102–0064$18.50/0

Accessible online at:
www.karger.com/journals/exn

Steven R. Gullans
Harvard Institutes of Medicine
77 Avenue Louis Pasteur
Boston, MA 02115 (USA)
Tel. +1 617 525 5712, Fax +1 617 525 5711, E-Mail sgullans@rics.bwh.harvard.edu

ways (table 1). With regard to disease subclassification, by analyzing large-scale patterns of gene expression, investigators can deduce similarities within and among patient populations and thereby more accurately classify them. Several studies showed the power of this subclassification process, particularly in studies of cancer [1–4] and in studies on the cellular response to chemotherapeutic susceptibility [5]. To identify key genes involved in a biological process, one typically is seeking those genes that are induced or repressed which are then prioritized for further study. This typically involves going to the literature to acquire knowledge about known genes and often define potential new roles for known genes. In addition, investigators often seek to perform a functional characterization of novel, orphan genes. Biological pathway discovery is a less appreciated use of gene expression analysis, but remains the most tantalizing with regard to understanding the complexities of large networks of interacting genes and their encoded proteins. Work in yeast, particularly from the Brown laboratory [6–9], pioneered the pathway discovery modality by showing that time course studies using many different conditions and stimuli can identify coordinately regulated genes and their related upstream regulatory gene promoter regions [10, 11].

Table 1. Utility of global gene expression

Disease subclassification
Identification of key genes
Elucidation of biological pathways

Table 2. Methods of gene expression analysis

Comparative EST
Differential-display RT-PCR
Macroarrays
MPSS
Microarrays
Northern blot analysis
RT-PCR
SAGE

Methods of Performing Global Expression Analysis

There are a number of technologies available for analyzing mRNA expression levels or differential mRNA expression (table 2). These methods include Northern blots, RT-PCR, macroarrays, microarrays, differential-display RT-PCR, serial analysis of gene expression (SAGE), comparative expressed sequence tag (EST) analysis, and massively parallel signature sequencing (MPSS). It should be noted that all of these systems have advantages and disadvantages (table 3). Furthermore, none is considered the 'best' at this time. In general, the sequence-based methods such as SAGE, MPSS, and comparative EST analysis, in which an automated sequencer is used to identify the transcript, provide the greatest specificity with regard to gene identity. On the other hand, the alternative, hybridization-based methods are typically less expensive per gene analyzed and can process more samples.

Table 3. Advantages and disadvantages of the methods described

	Micro-arrays	Macro-arrays	MPSS	Northern blot	SAGE	RT-PCR	Differential-display RT-PCR	Comparative EST
Technical difficulty	●●	●	●●●	●	●●●	●	●	●●
Setup expense	●●●	●	●●●	●	●●	●●	●●	●●●
Cost of analysis	●●	●●	●●●	●	●●●	●	●●	●●●
Bioinformatics needs	●●●	●●	●●●	●	●●●	●	●●	●●●
Number of genes	●●●	●●	●●●	●	●●●	●	●●	●●●
Number of samples	●●	●●	●	●	●	●●●	●●	●
Flexibility	●●	●●●	●●	●●●	●●	●●	●●●	●●
Gene specificity	●●	●●	●●●	●●●	●●●	●●	●●	●●●

● = Low; ●● = medium; ●●● = high.

Of note, sequence-based methods such as SAGE, comparative EST sequencing, and MPSS are the most accurate with regard to transcript identification but, for academic scientists, these approaches are technically challenging and relatively labor-intensive and expensive when analyzing many samples (e.g., in large number of patients or in a time course study). In contrast, the hybridization-based formats are relatively easy to use, but can have ambiguities with regard to nonspecific signal detection as well as difficulty in monitoring low copy number transcripts.

The traditional *Northern blot* [12] analyzes a single gene at a time and is considered by most investigators to be the 'gold standard' in terms of quantifying mRNA expression levels. The inherent limitation of a Northern blot is that typically only one gene at a time is analyzed. In addition, there are variations from blot to blot that undermine large-scale quantitative comparisons. Analyzing more than 10–20 RNA samples requires multiple blots, and the methods consume considerable amounts of RNA and other reagents that makes large-scale analysis costly and time-consuming.

In the 1980s and early 1990s, strategies for large-scale analysis of gene expression emerged. Initially dot-blots were developed as 'reverse-Northern blots', in which each gene being analyzed was blotted onto a nylon filter membrane and an RNA sample was labeled and hybridized to the blot. Expression levels were quantified using radioactivity and an appropriate scanner. These *macroarrays* have proven very reliable for quantifying gene expression levels and continue to be used. There are numerous commercial sources providing macroarrays. In general, these 'low-tech' arrays are relatively inexpensive, work extremely well, and can be easily scaled to allow analysis of thousands of genes. Furthermore, they have the great advantage that any laboratory can create its own macroarrays using an in-house or a commercially available library of genes (i.e., oligonucleotides or ESTs) to be spotted. The major drawbacks of filter-based macroarrays are that (1) they can be difficult to process in a high-throughput manner due to membrane stretching and other anomalies; (2) there can be problems quantitatively comparing multiple blots because of inconsistencies in labeling, and (3) the blots typically have a high background signal, making it difficult to analyze low-abundance transcripts.

In the early 1990s, *RT-PCR* [13, 14] analysis of mRNA expression became prominent for analysis of mRNA expression. This method's ability to be used in a 96-well format, as well as its requirement for only small amounts of RNA, led to its development as a high-throughput, large-scale technology. However, early recognition that RT-PCR was only semiquantitative undermined its use as a simple tool. The use of competing templates to provide calibration standards allowed quantitation to be implemented but slowed its utilization as a large-scale platform for many genes. In recent years, however, commercial enterprises have developed real-time PCR instruments that have led to truly quantitative RT-PCR. Today, with appropriate instrumentation, an investigator can use a 96-well format to measure mRNA levels in hundreds of samples each day. Nonetheless, this approach essentially remains a serial analysis when analyzing many genes, as an investigator must test one sample and one gene in each reaction tube. Thus, it is relatively expensive, if one wishes to analyze hundreds or thousands of different transcripts in many samples. Thus, quantitative real-time RT-PCR is typically considered a validation method, suitable for confirming expression levels of small numbers of genes in many biological samples.

Differential-display RT-PCR and similar PCR-based techniques were introduced in the early 1990s [15, 16] as a way of identifying mRNA expression differences in two or more samples. These techniques use short oligonucleotide primers (e.g., 10mers) to amplify arbitrary subsets of 25–100 genes in an RNA sample. These amplified products are visualized using polyacrylamide gel electrophoresis, and differences in the amount of a PCR product are seen from the intensity of individual bands in the gel. Genes of interest are then cut out of the gel, isolated, and sequenced. This approach has been largely used as a screening tool for identifying differentially expressed genes. However, with some technical modifications and use of more sophisticated instruments and appropriate informatics, investigators have shown that differential-display RT-PCR methods can generate quantitative information regarding mRNA expression levels, and individual gene identities can be determined [17]. The great advantage of this approach is that it requires no a priori knowledge that a gene exists or what its sequence is. This make it very useful for studies of organisms for which there is little genomics infrastructure. Moreover, because this approach involves PCR amplification, it often generates new genes that do not exist in public databases. Finally, it is very flexible. The major drawback is that it would be difficult for academic laboratories to automate, and there is always some uncertainty regarding the identity of each gene being analyzed.

The advent of automated, high-throughput sequencing technology led to the *comparative EST sequence analysis.* Pioneered by Venter and colleagues [18, 19] in the early

1990s, comparative EST analysis provided the first large-scale analysis of gene expression. The strategy involves creating cDNA libraries representing all expressed mRNAs in a cell or tissue. Then, by sequencing thousands of arbitrarily chosen cDNAs, a database is created that identifies and counts all the genes that are expressed which are termed ESTs. The method has the inherent advantages that it does not require prior knowledge of the existence of a transcript to measure its level of expression. In addition, being a sequence-based method, this strategy is highly reliable with regard to transcript identification. Nonetheless, this method has fallen into relative disfavor because of significant drawbacks. In particular, it is very expensive to sequence enough ESTs to generate a full profile of gene expression. This issue is compounded by the fact that most transcripts are of relatively low abundance, so an investigator must sequence tens or hundreds of thousands of ESTs to generate a statistical sampling of a pool of RNA to identify differentially expressed genes. Nonetheless, EST sequencing will remain an important tool for studies of organisms the genomes of which are not yet sequenced, as other approaches (e.g., microarray, SAGE) typically require a database and/or repository of genes to be useful. Furthermore, this approach can uncover novel transcripts, particularly splice variants that are likely to be missed by other methods.

Developed in the laboratories of Vogelstein and Kinzler [20, 21], *SAGE* is a very clever analytical method that has been effectively used in studies of cancer [22, 23]. SAGE uses a series of biochemical reactions to create a library of short DNA fragments (13 base pairs), each one being derived from a single mRNA molecule in a biological sample. This population of DNA fragments, known as SAGE tags, is concatenated together to form longer DNA molecules that are then sequenced, as done with comparative EST sequencing. However, a single molecule contains DNA fragments representing >30 different RNA molecules, so a single sequencing reaction is sampling the expression of >30 transcripts. A database of SAGE tags is created, and specialty software distinguishes each SAGE tag, bins and counts identical tags, and thus measures the abundance and distribution of each transcript in the RNA pool being studied. Among a population of an estimated 40,000 mammalian genes, a 13-nucleotide sequence is a relatively unique identity for a gene (1 per 4^{13} or ~70 million). Thus, additional software is available to identify the gene that corresponds to each SAGE tag.

Performing a comparison of two RNA samples typically involves analyzing 30,000–100,000 SAGE tags in each sample to generate statistically valid comparisons. The advantage of this strategy is that it provides a highly reliable identification of each gene at less cost than EST library sequencing. The major disadvantages of this approach are that it is technically challenging to make SAGE tag libraries, that it is relatively slow to perform a study, that it is costly to perform a single experiment, and, most importantly, that its power can only be fully appreciated when there is an EST or genome database available to allow identification of each SAGE tag. Thus, its application has been predominant in human and mouse studies. Of note, a public database of SAGE tags is available at http:/www.ncbi.nlm.nih.gov/SAGE/ with links to the Cancer Genome Anatomy Project established by the National Cancer Institute to generate the information and technological tools needed to decipher the molecular anatomy of the cancer cell.

MPSS is a novel method that, like SAGE, captures the power of sequencing for gene identification and combines it with the clout of a parallel-processing system to provide ultrahigh-throughput analysis [24, 25]. Briefly, MPSS involves construction of a cDNA library which is captured on millions of microbeads and assembled into a planner array. Individual beads are then individually imaged during a series of complex multistep sequencing reactions, and the DNA sequence of hundreds of thousands of transcripts is determined at once. By sequencing 16–20 bases, unique identities are deduced for each molecule, and a database of all transcripts is created. Thus, sequence analysis provides gene identification, and a simple count of the number of copies of each sequence reveals transcript abundance or expression level. This approach was shown to be effective in yeast and human cells for measuring gene expression levels. There are several major advantages of this strategy. It provides sequence level gene identification, and it can identify transcripts not previously known to exist. Furthermore, the miniature, parallel-processing system allows interrogation of hundreds of thousands of transcripts at once, providing unprecedented sensitivity. MPSS, like SAGE, is most valuable when used in conjunction with an existing genome or EST database. However, at this time it is unclear how widespread this relatively elaborate and complex analytical system will become.

DNA microarrays, or 'gene chips', have become the most popular platform among scientists for performing global gene expression analysis. Microarrays provide a relatively rapid, reliable, reproducible, and quantitative approach for simultaneously monitoring expression levels of thousands of genes [7, 26–29]. Basically, the approach is to create a spotted array of thousands of different DNA

molecules (i.e., oligonucleotides or cDNAs) corresponding to thousands of different genes. Then, starting with an RNA sample, a series of biochemical reactions generates a fluorescently labeled cRNA or ss-cDNA probe – note that the term 'probe' has been used differently within the community. This probe is then hybridized to the microarray and scanned with a laser scanner. The expression levels are measured by the fluorescence intensity of bound probe to each spot. Multiple microarray platforms now exist from a variety of commercial entities. In addition, as pioneered at Stanford, many institutions have robotic equipment enabling them to create their own custom microarrays.

The predominant microarray platforms, to date, are Affymetrix oligonucleotide microarrays and glass slide cDNA or oligonucleotide microarrays [30–32]. Other innovative microarray technologies include 'flow-through' microarrays and fiberoptic bead arrays [33]. Though important differences exist among the platforms, they share many of the same advantages and disadvantages. In particular, they allow simultaneous analysis of thousands of genes in a single sample. In addition, the identity of each gene is known a priori. It is possible to process many different RNA samples quickly and efficiently, making possible studies of large numbers of patient samples or detailed time course studies. Finally, the data are quantitative and can be compared among laboratories and across different experiments. The disadvantages begin with the fact that microarrays are generally considered to be relatively expensive, though the prices have decreased as much as ten-fold over the past 4 years. In addition, there is always some level of uncertainty regarding binding specificity, so that the measured expression level of one gene may be corrupted by 'nonspecific' binding of another gene with similar sequence. A number of strategies seek to reduce this problem, such as the 'mismatch' analysis provided by Affymetrix or using informatics to design 'gene-specific' oligos for the chip. In addition, the level of sensitivity of microarrays is less than that of sequence-based methods, though improvements in signal detection are emerging all the time [34–36]. Finally, the technical reproducibility has been a major concern in the microarray field and led to the perception that the method can detect only twofold or greater changes.

Pitfalls of DNA Microarrays

Because microarrays are becoming a ubiquitous approach, it is worthwhile to consider in some detail the problems that can arise in using them. These problems range from technical issues associated with the chip, labeling, and scanning to bioinformatics issues such as image analysis and gene identification. Because many institutions are setting up their own robotic systems with more variables to deal with, the following discussion is directed primarily at custom-made glass slide microarrays, though similar concerns can arise with any platform.

The first step is to fabricate a microarray using a glass slide, a spotting robot, and DNA for spotting. The quality of the microarray slide is essential. A high-quality slide is composed of an evenly planed glass slide to reduce optical noise. This slide is generally coated with a positively charged material (e.g., polylysine) that binds to and immobilizes the spotted material. Covalent attachment chemistries are available, though there is dispute regarding whether they are necessary. A poor surface integrity can lead to washing away of immobilized material, ultimately reducing sensitivity and signal of the spotted product.

The source of DNA for spotting is a major consideration, as one can use cDNAs and ESTs or synthesized oligos. For cDNAs and ESTs, the major problem has been correct tracking of the DNA in each well to be sure that it is the correct gene [37]. The recent availability of oligonucleotides from commercial vendors likely obviates this problem, though they are more expensive and require greater care with regard to designing gene-specific oligos. Oligos range in length from 25mers (Affymetrix) to 70mers from various vendors. To date, there is no clear consensus regarding the best material to spot on microarrays.

There are many good commercial arraying instruments that all appear to work well. They differ with regard to cost, throughput, spot density, and ease of use. The buffer used to spot a microarray determines size and morphology of the spots. A good spotting buffer will help produce identical spot size and solid round spots, without wasting DNA by excessive prespotting to remove excess material from the surface of the pin. A variety of buffers are known to work. The DNA needs to be covalently linked to the slide using either high-energy ultraviolet rays or exposure to high temperatures (80 °C). The concentration of the immobilized material should ideally be around 50 μM to provide superfluous binding sites for the labeled probe. Indeed, limiting the immobilized target directly decreases the sensitivity of the system.

Most microarray protocols use total RNA and not poly(A) RNA which can be contaminated with oligo(dT) in the isolation process. There are a variety of procedures for fluorescently labeling RNA, including a simple reverse transcription with fluorescently modified oligos or an in

vitro transcription process [31], and recently a dendrimer tagging procedure has been reported [34]. Labeling procedures are fraught with potential problems making it difficult to evenly and reproducibly label all transcripts from a given RNA sample. Among the problems that exist are difficult protocols, weak signal, high background fluorescence, quenching, high cost, inconsistent labeling, and signal compression. At this point, it is fair to say that this is not routine procedure for most academic researchers and that the benefit of commercial labeling kits and dedicated technicians is the optimal strategy.

The hybridization of the labeled probe to the immobilized target is critically dependent upon efficient mixing. Proper mixing can significantly reduce the hybridization time and increase the reproducibility of data from chip to chip. Many laboratories still hybridize under a coverslip which we feel is unacceptable. Dehydration, air bubbles, dust, and leaking of sample over the edge of the slide are some of the problems inherent to the use of coverslips. Indeed, due to lack of mixing, very long incubations are required which dramatically increase the likelihood of the aforementioned problems. Hybridization chambers or automated hybridization systems are recommended, though one needs to be cautious. Of note, bright signals are not necessarily the optimal result, as they may represent nonspecific binding.

Most confocal laser scanners on the market today will produce a high-quality digital image of the fluorescent signals from the chips. The major differences are the ability to change laser power and photomultiplier tube settings as well as the number of slides the scanner can hold and scan at one time. In our system, increasing the laser power by 10% results in a twofold increase in signal intensity; however, this is not true for photomultiplier tube settings which are very nonlinear in their behavior. In general, scanners should be adjusted to maximize the dynamic range and to reduce the problem of signal saturation.

The final step for transforming microarray scanned images into a database of expression values is image processing. This process requires specialty software customized to identify individual spots, to determine signal and background, and to exclude artifactual signals. Many approaches have been advocated [38], and, in general, improvements in all aspects of the technology, particularly reducing background signal and enhancing true signal, have facilitated routine image processing to generate reliable data. Nonetheless, it should be noted that errant data points will always haunt the field, and, when feasible, an investigator should examine and confirm the primary image data when anomalous biological findings arise. From a pragmatic perspective, many technical problems are recognized and obviated by using duplicate spotting on different regions of a chip and performing replicate experiments with different chips.

Bioinformatics

Bioinformatics utilizes statistics and computer algorithms to effectively classify, or cluster, genes or biological samples (e.g., patients) into distinct groups based on gene expression by comparing sets of gene expression data throughout the course of disease or following the application of one or numerous pharmacological agents. These gene expression profiles may be obtained from biopsy specimens of normal or diseased tissue or through the monitoring of the effect of drugs or other biological stimuli on gene expression profiles in animals or isolated cell systems. In the case of biopsy specimens, these profiles may then be utilized to form a prognosis and diagnosis and, therefore, provide a means for rational drug therapy.

Proper interpretation of the large data sets generated by any method for global analysis of gene expression requires tools for effective mining of the data. Researchers must often correct for the high levels of noise inherent in microarray experiments. In other words, it is important to measure instrument error in every experiment. This requires performing multiple replicates to measure a gene-by-gene reproducibility in the data. In addition, biological noise can be measured by analyzing multiple RNA samples representing 'identical' biological conditions. Hughes et al. [39], using 63 'identical' yeast cultures, showed a number of genes to be fluctuating in expression level, many by severalfold, under conditions researchers would traditionally consider to be unchanged. Thus, biological noise can be significant and needs to be considered. Unfortunately, obtaining a precise measurement of biological noise can be very expensive. Thus, we prefer to simply perform replicates of individual samples and apply stringent criteria for identifying gene changes.

Analysis of replicates using microarrays (and reverse labeling for two-color microarrays) has shown us that there can be significant variation in technical error among platforms, among investigators, among lots of reagents and chips, etc. Thus, it is inappropriate to apply a uniform filter, such as two- or fourfold changes. Rather it is important to measure instrument error. In the case of Affymetrix GeneChips, the ability to resolve differences in expres-

sion level is a function of expression level, termed 'average difference'; abundant transcripts have small errors (<2-fold), and low-abundance transcripts have large errors (>8-fold). Once instrument error is measured, data can be corrected or filtered and interpreted correctly. One consequence of this approach is that in most microarray studies, most data are discarded prior to in-depth data mining; in some cases >90% of the data are not reliable. At the end, only reproducible data are meaningful for further analysis.

Statistical Analysis of Differential Gene Expression

A scatter plot of the gene expression levels of one control sample measured on one microarray versus a replicate of the same sample on a second microarray can provide a simple (and often sobering) assessment of the instrumental noise in the data. In particular, a superposition of the scatter plot of control versus control (using log-log scales) over the usual experiment versus control scatter plot provides a simple 'eyeball' test for the genes with significant changes in expression. Briefly, the experimental data points that lie well outside the 'noise cloud' defined by the same versus the same scatter plot are most likely to be significantly changed between the control and the experimental conditions. Even though hundreds or thousands of genes may exhibit twofold or greater changes on the microarrays, we often find that only a few tens of genes meet this 'eyeball' criterion. Although expensive, additional replicates of the microarrays for both the control samples and the experimental samples are essential for more rigorous statistical tests such as t tests or Anova to determine significant changes in gene expression on a gene-by-gene basis.

Patterns of Gene Expression

DNA and oligonucleotide microarray technology has made possible the analysis of expression of thousands of genes simultaneously, and various statistical techniques have been developed to interpret these data efficiently and effectively. Typically, the goal is to identify potential target genes for further analysis or to cluster samples according to global similarity. These techniques can be classified as either *supervised* or *unsupervised,* although the majority of experiments will make use of both. Supervised methods require the direction of a scientist with a priori knowledge of the data, such as disease class, gene function, transcriptional regulation, or tissue type. An unsupervised technique, on the other hand, lets the gene expression patterns direct the analysis, regardless of preexisting expectations [40]. Our experience has shown that data should initially be analyzed using an unsupervised approach, as this can reveal unappreciated, systematic data anomalies or identify more subclasses than anticipated, such as three subclasses of cancer when only two are thought to exist.

Hierarchical clustering is a type of exploratory data analysis aimed at classifying and grouping data into meaningful subsets. In bioinformatics, hierarchical clustering algorithms are generally used to assess the similarities among tissue samples or other biological samples. Various hierarchical algorithms exist [41], and most provide a numerical indicator of cluster quality as well as an intuitive tree plot of the similarity between samples as represented by the length of the branch connecting them. Hsiao et al. [42] recently utilized hierarchical clustering algorithms to effectively cluster 19 nondiseased human tissue types utilizing 451 housekeeping genes, differentially expressed in all 19 tissues. Other studies have shown the utility of hierarchical clustering for identifying disease subtypes [2, 4]. Since many parameters can be varied in the standard hierarchical clustering methods, such as the choice of metric (e.g., Euclidean, Manhattan, or some special weighted measure of distance), it is important to explore both multiple clustering algorithms and settings to develop hypotheses about the groupings of samples or genes. In particular, groupings that are robust to changes in methods and settings are most likely to be real rather than artifacts of a particular clustering algorithm.

Self-organizing maps (SOM) are a type of unsupervised algorithm well suited for the clustering of genes into functionally meaningful groups. Tamayo et al. [43] recently applied a SOM system, implemented using their publicly available Genecluster software, to interpret gene expression patterns, with application to hematopoietic differentiation. Through the iterated adjustment of representative clusters, the SOM groups genes that behave similarly. SOMs are capable of handling large data sets, and the implementation by Tamayo et al. [43] provides easy visualization and interpretation of the system.

To effectively demonstrate the utility of SOMs, a myeloid leukemia cell line, HL-60, was analyzed upon stimulation with a phorbol ester to initiate macrophage differentiation. The temporal changes in gene expression were monitored using Affymetrix oligonucleotide arrays containing over 5,000 human genes and over 1,000 ESTs. Upon configuration of the SOM algorithm, they identified numerous genes previously thought to be important during differentiation as well as linking other, previously unrecognized genes to this phenomenon. These same methods were later successfully applied to a more com-

plex analysis of four different cell lines, including HL-60, U937, Jurkat, and NB4 cells, and could easily be applied to nephrology to predict novel genes important in normal renal function and in pathophysiologies. Additionally, others [44] have used SOM in combination with other mapping algorithms to more clearly visualize relationships between individual gene sets. Indeed, these methods of clustering reliably predicted functional similarities among genes.

Bioinformatics Pitfalls

Although considerable progress has been made in analyzing global gene expression, many questions remain. In addition to those methods described above, many analytical approaches are being tested. Many of these are available in standard statistical software (e.g., S-Plus, SyStat, SAS) and appear to be very useful (such as principal-component analysis). Today, however, no universally accepted protocol exists regarding how to conduct a bioinformatic analysis. As a result, everyone has a personal favorite arsenal of statistical tools. Commonly used methods, such as SOM and hierarchical clustering, carry inherent assumptions and biases. In addition, various strategies of normalization exist. As a result, two researchers using the same data could reach different conclusions because of differing approaches and selection of parameters. Undoubtedly, as larger and more robust data sets arise, the community will begin to settle on standard approaches that should be implemented and also pursue more advanced strategies that can utilize the perceived power of large public databases. Thus, the best approach is to use several tools and to view the results as hypotheses that need to be validated.

Towards the Development of Standards

A major goal of the gene expression research community is to build large-scale public databases using data from all expression-profiling platforms [45]. These data repositories could then be used to generate or substantiate new hypotheses, using information compiled from multiple biological systems. A major obstacle to this endeavor is developing standards that allow cross-comparisons and cross-validation. Considerable effort is being devoted to developing standards (e.g., MIAME, MAML) for annotating expression experiments and facilitating data sharing (see www.mged.org). In addition, however, biological standards and reference data sets are likely to be required. No standard RNA samples have been uniformly adopted, though suitable commercial products are emerging. To facilitate cross-validation of data sets and methods, we recently conceived and implemented the idea of a set of genes that should be available on all platforms, termed the Standard Gene Set (SGS).

The SGS currently consists of a set of 96 ubiquitously expressed housekeeping genes. To assist the research community, we identified the genes corresponding to the human SGS in many commercial microarrays, oligonucleotide sets, and cDNA sets, including those available from Affymetrix, Operon, Research Genetics, Compugen, Motorola, and Perkin-Elmer. These genes were obtained as part of our Human Gene Expression (HuGE) Index Project which evaluated expression profiles of 19 normal human tissues [42]. It should be noted that these 96 housekeeping genes are ubiquitously expressed, however, at different levels in different tissues and can vary following experimental stimulation. The SGS will allow comparison of DNA microarray data between chips in the same laboratory as well as cross-laboratory comparison of gene expression analysis. The SGS is designed to assist in the standardization of microarray type studies. To date, the SGS can be used as a quality control measure in human DNA microarray expression analysis and will likely be adapted to include mouse, rat, and other species. These genes will ensure a higher level of *data quality,* and the development of a larger SGS (e.g., 384 genes) could further enhance the quality.

The genes included in the SGS, as well as accession numbers and links to associated gene databases, can be found as a database at the HuGE Index (http://www.hugeindex.org). The ultimate objective of the HuGE Index is to characterize the mRNA expression levels in all human tissues. The publicly available mRNA expression levels of thousands of genes are obtained using high-density oligonucleotide array technology and aim at providing a comprehensive database to understand the expression of human genes in normal human tissues [42]. Expression profile data can be downloaded and include many major organ systems such as brain, kidney, liver, lung, muscle, duodenum, colon, prostate, and spleen.

Fingerprinting Human Diseases

Biological stimuli, disease pathologies, and pharmacological manipulations induce a distinct biological phenotype or 'molecular fingerprint' that can be characterized using RNA expression analysis. This fingerprint is effectively created by the gene expression profile of the cell or tissue and is distinct for that biological state or condition. It needs to be appreciated that advanced informatics tools often yield fingerprints or diagnostics that are a nonintuitive transformation of the expression levels of many genes

rather than a simple list of genes that are increased and/or decreased in a disease sample.

When creating a diagnostic fingerprint, an investigator must use a subset of microarray data as a training set to establish an algorithm or analytical test and then use this test to analyze new samples, de novo. This approach was used to distinguish different human acute leukemias. Golub et al. [1] accurately distinguished acute myeloid leukemia (AML) from acute lymphoblastic leukemia (ALL) without previous identification or knowledge of the clinical diagnosis. Using bone marrow samples from 38 patients, these authors first identified 1,100 genes that were highly correlated with an AML-ALL diagnosis based on gene expression data generated from high-density oligonucleotide microarrays. Additionally, they were able to identify a subset of 50 of the 1,100 genes that enabled distinction between AML or ALL. This 50-gene set was then applied to 34 independent samples where they were able to assign 29 samples to either AML or ALL with accuracy. In fact, using only 10 of the 50 genes gave the same results. Notably, 10 of the 34 samples were collected from peripheral blood rather than bone marrow, suggesting the power of this methodology to differentiate between disease types using dissimilar tissues. To further extend the effectiveness of the molecular diagnosis of ALL, Golub et al. [1] also applied gene expression analysis algorithms to predict whether ALL cases derived from either A or B lineage. Therefore, gene expression analysis can be used clinically to assist in the diagnosis of specific pathologies and may be extended from leukemias to other pathologies and organ systems. In fact, current technology is sufficient to develop effective diagnostic tests for most diseases using global expression profiling. It should be realized that the clinical application may not involve a microarray analysis, as simpler and cheaper methods are likely to be sufficient.

Global Gene Expression and Nephrology

Increasing amounts of data linking gene expression with kidney biology are being generated, including studies of development, human pathology, and animal models of disease [46 ,47]. A particularly intriguing and clinically relevant study performed by Moch et al. [29] showed the value of using a combination of DNA microarrays for gene discovery and *tissue arrays* for validation of pathologically important genes. Microarray analysis of the renal cell carcinoma (RCC) cell line CRL-1933 versus nondiseased renal samples was used to identify 89 differentially expressed genes in the cancer cell line. Then using a renal tissue array containing 532 tumors, these authors confirmed the importance of one of these transcripts, vimentin, in RCC. A 37-month follow-up survival analysis confirmed that histological assessment of the vimentin expression on the tumor array in clear-cell RCC was positively associated with shorter survival time versus those without vimentin-positive tissues. These data suggest that microarrays can be used clinically to predict effective prognostic biomarkers for disease, including RCC.

DNA microarrays can generate important information concerning the genes and gene groups important during the progression of renal diseases as well as play an important role in the establishment of specific gene function in the kidney. Monti et al. [48] recently utilized microarray profiling to assess the genes responsible for physiological blood pressure compensation in bradykinin B_2 knockout mice, providing clues as to the molecular mechanism responsible for this phenomenon. Since the B_2 receptor is an important component of cardiovascular homeostasis such as the regulation of vasodilation and natriuresis-diuresis, compensatory gene expression changes were monitored in 12,000 mouse genes and ESTs. These authors identified 20 candidate genes that were upregulated in the transgenic mice and 59 genes that were downregulated in B_2-receptor-deficient mice. Grouping these genes into functionally related classes identified two gene families likely to impact cardiovascular function: serine proteases and aquaporins. The serine protease genes were upregulated and are known to be essential for the conversion of high-molecular-weight kininogen to bradykinin. The aquaporin gene family is important in the transport of water across the membrane of the proximal tubule of the kidney. Monti et al. [48] found that only specific subtypes of the aquaporin gene family, AQP4 and AQP1, were downregulated in B_2-deficient mice. Therefore, they utilized microarrays in the elucidation of the molecular mechanism of blood pressure compensation in mice lacking functional B_2 receptors and several differentially expressed genes in the kidney that are likely to be of functional importance.

Expression profiling can identify key genes in animal models of renal pathologies. Recently, Nagasawa et al. [28] applied microarray analysis to monitor gene expression changes in a mouse model of massive proteinuria due to intraperitoneal bovine serum albumin injections, resulting in intrinsic renal toxicity, fibrosis, and deterioration of the renal function. Utilizing a hierarchical clustering algorithm, these authors demonstrated temporal gene changes associated with proteinuria when analyzed on days 0, 7, and 21 after protein overload. Nagasawa et al. [28] identified numerous genes upregulated on days 7 and

21 versus day 0, including osteopontin which has been implicated in various renal diseases. Numerous genes were also found to be downregulated. They concluded that over 10% of gene transcripts in the kidney are altered under conditions of excessive protein loading. Of note, they used a variety of approaches to validate their observations. The use of animal models, such as used by Nagasawa et al. [28], in the development of these profiles may be important, since many genes are retained in alternative species such as mouse, rat, and nonhuman primates and likely are translated into proteins with similar in vivo function as their human counterpart. This work demonstrates that these techniques are useful tools in the analysis of gene expression in the era of functional genomics.

Conclusions

Global gene expression is a very powerful strategy for uncovering new and important biology in nephrology. Numerous methods are available, but investigators must be cautious in using them, as many technical problems and experimental biases can influence the results. As the field matures, these strategies will lead to new diagnostic tests, to delineation of key genes involved in renal physiology and pathology, and potentially to an understanding of the complex gene regulation network that underlies kidney function. Furthermore, these discoveries will certainly guide future human disease treatment.

References

1 Golub T, Slonim D, Tamayo P, Huard C, Gaasenbeek M, Mesirov J, Coller H, Loh M, Downing M, Caligiuri M, Bloomfield C, Lander E: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 1999;286:531–537.

2 Kahn J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 2001;7:673–679.

3 Takahashi M, Rhodes D, Furge K, Kanayama H, Kagawa S, Haab B, Teh BT: Gene expression profiling of clear-cell renal cell carcinoma: Gene identification and prognostic classification. Proc Natl Acad Sci USA 2001;98:9754–9759.

4 Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, Sampas N, Dougherty E, Wang E, Marincola F, Gooden C, Lueders J, Glatfelter A, Pollock P, Carpten J, Gillanders E, Leja D, Dietrich K, Beaudry C, Berens M, Alberts D, Sondak V: Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 2000;406:536–540.

5 Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS: Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. Proc Natl Acad Sci USA 2000;97:12182–12186.

6 Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO, Herskowitz I: The transcriptional progam of sporulation in budding yeast. Science 1998;282:699–705.

7 Lashkari DA, DeRisi JL, McCusker JH, Namath AF, Gentile C, Hwang SY, Brown PO, Davis RW: Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc Natl Acad Sci USA 1997;94:13057–13062.

8 Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO: Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 2000;11:4241–4257.

9 Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: Comprehensive identification of cell cycle-regulated genes of the yest *Saccharomyces cerevisiae* by microarray hybridization. Mol Biol Cell 1998;9:3273–3297.

10 Cohen BA, Mitra RD, Hughes JD, Church GM: A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nat Genet 2000;26:183–186.

11 Roth FP, Hughes JD, Estep PW, Church GM: Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotechnol 1998;16:939–945.

12 Sabelli PA, Shewry PR: Northern analysis and nucleic acid probes. Methods Mol Biol 1995;49:213–228.

13 Cale JM, Shaw CE, Bird IM: Optimization of a reverse transcription-polymerase chain reaction (RT-PCR) mass assay for low-abundance mRNA. Methods Mol Biol 1998;105:351–371.

14 Freeman WM, Walker SJ, Vrana KE: Quantitative RT-PCR: Pitfalls and potential. Biotechniques 1999;26:112–122, 124–125.

15 Liang P, Pardee AB: Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science 1992;257:967–971.

16 McClelland M, Mathieu-Daude F, Welsh J: RNA fingerprinting and differential display using arbitrarily primed PCR. Trends Genet 1995;11:242–246.

17 Shimkets RA, Lowe DG, Tai JT, Sehl P, Jin H, Yang R, Predki PF, Rothberg BE, Murtha MT, Roth ME, Shenoy SG, Windemuth A, Simpson JW, Simons JF, Daley MP, Gold SA, McKenna MP, Hillan K, Went GT, Rothberg JM: Gene expression analysis by transcript profiling coupled to a gene database query. Nat Biotechnol 1999;17:798–803.

18 Adams MD, Soares MB, Kerlavage AR, Fields C, Venter JC: Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. Nat Genet 1993;4:373–380.

19 Lee NH, Weinstock KG, Kirkness EF, et al: Comparative expressed-sequence-tag analysis of differential gene expression profiles in PC-12 cells before and after nerve growth factor treatment. Proc Natl Acad Sci USA 1995;92:8303–8307.

20 Velculescu VE, Vogelstein B, Kinzler KW: Analysis uncharted transcriptomes with SAGE. Trends Genet 2000;16:423–425.

21 Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: Serial analysis of gene expression. Science 1995;270:484–487.

22 Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, Vogelstein B, Kinzler KW: Gene expression profiles in normal and cancer cells. Science 1997;276:1268–1272.

23 Polyak K, Xia Y, Zweier JL, Kinzler KW, Vogelstein B: A model for p53-induced apoptosis. Nature 1997;389:300–305.

24 Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, Roth R, George D, Eletr S, Albrecht G, Vermaas E, Williams SR, Moon K, Burcham T, Pallas M, DuBridge RB, Kirchner J, Fearon K, Mao J, Corcoran K: Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol 2000;18:630–634.

25 Tyagi S: Taking a census of mRNA populations with microbeads. Nat Biotechnol 2000;18: 597–598.
26 Hsiao L, Stears R, Hong R, Gullans S: Prospective use of DNA microarrays for evaluating renal function and disease. Curr Opin Nephrol Hypertens 2000;9:253–258.
27 Kurella M, Hsiao L, Yoshida T, Randall J, Chow G, Sarang S, Jensen R, Gullans S: DNA microarray analysis of complex biologic processes. J Am Soc Nephrol 2001;12:1072–1078.
28 Nagasawa Y, Takenaka M, Kaimori J, Matsuoka Y, Akagi Y, Tsujie M, Imai E, Hori M: Rapid and diverse changes in gene expression in the kidneys of protein-overloaded proteinuria mice detected by microarray analysis. Nephrol Dial Transplant 2001;16:923–931.
29 Moch H, Schraml P, Bubendorf L, Mirlacher M, Kononen J, Gasser T, Mihatsch M, Kallioniemi O, Sauter G: High-throughput tissue microarray analysis to evaluate genes uncovered by cDNA microarray screening in renal cell carcinoma. Am J Pathol 1999;154:981–986.
30 Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ: High-density synthetic oligonucleotide arrays. Nat Genet 1999;21:20–24.
31 Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL: Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 1996;14:1675–1680.
32 Jain KK: Applications of biochip and microarray systems in pharmacogenomics. Pharmacogenomics 2000;1:289–307.
33 Walt DR: Techview: Molecular biology. Bead-based fiberoptic arrays. Science 2000;287:451–452.
34 Stears R, Getts R, Gullans S: A novel, sensitive detection system for high-density microarrays using dendrimer technology. Physiol Genomics 2000;3:93–99.
35 Schweitzer B, Kingsmore S: Combining nucleic acid amplification and detection. Curr Opin Biotechnol 2001;12:21–27.
36 Zhang Y, Price BD, Tetradis S, Chakrabarti S, Maulik G, Makrigiorgos GM: Reproducible and inexpensive probe preparation for oligonucleotide arrrays. Nucleic Acids Res 2001;29: E66–E66.
37 Knight J: When the chips are down. Nature 2001;410:860–861.
38 Yang MC, Ruan QG, Yang JJ, Eckenrode S, Wu S, McIndoe RA, She JX: A statistical method for flagging weak spots improves normalization and ratio estimates in microarrays. Physiol Genomics 2001;7:45–53.
39 Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J, Bard M, Friend SH: Functional discovery via a compendium of expression profiles. Cell 2000;102:109–126.
40 Raychaudhuri S, Sutphin PD, Chang JT, Altman RB: Basic microarray analysis: Grouping and feature reduction. Trends Biotechnol 2001;19:189–193.
41 Kaufman L, Rousseeuw PJ: Finding groups in data: An introduction to cluster analysis, 1990. New York, John Wiley & Sons, 1990.
42 Hsiao LL, Dangond F, Yoshida T, Hong R, Jensen RV, Misra J, Haverty P, Weng Z, Mutter GL, Frosch MP, MacDonald ME, Milford EL, Crum CP, Bueno R, Pratt RE, Mahadevappa M, Warrington JA, Stephanopoulos G, Stephanopoulos G, Gullans SR: A compendium of gene expression in normal human tissues reveals tissue-selective genes and distinct expression patterns of housekeeping genes. Physiol Genomics 2001;7:97–104.
43 Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR: Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA 1999;96:2907–2912.
44 Toronen P, Kolehmainen M, Wong G, Castren E: Analysis of gene expression data using self-organizing maps. FEBS Lett 1999;451:142–146.
45 Gardiner-Garden M, Littlejohn TG: A comparison of microarray databases. Brief Bioinform 2001;2:143–158.
46 Stuart RO, Bush KT, Nigam SK: Changes in global gene expression patterns during development and maturation of the rat kidney. Proc Natl Acad Sci USA 2001;98:5649–5654.
47 Bard J: A bioinformatics approach to investigating developmental pathways in the kidney and other tissues. Int J Dev Biol 1999;43:397–403.
48 Monti J, Gross V, Luft F, Franca A, Schulz H, Rainer D, Sharma A, Hubner N: Expression analysis using oligonucleotide microarrays in mice lacking bradykinin type 2 receptors. Hypertension 2001;38:E1–E3.