High-throughput Biology in the Postgenomic Era

Albert Hsiao, PhD, and Michael D. Kuo, MD

J Vasc Interv Radiol 2006; 17:1077–1085

Abbreviations: cDNA = complementary DNA, CGH = comparative genome hybridization, SNP = single nucleotide polymorphism

HIGH-THROUGHPUT biological methods, namely, methods that perform thousands of simultaneous measurements of biological molecules, have rapidly transformed the landscape of biomedical research during the past decade. Perhaps most central to this transformation has been the sequencing of the human genome and subsequent free release of genomic information, emphasized as a critical goal in the Bermuda Statement by participants of the Human Genome Project (1). With this massive undertaking largely completed, investigators are faced with new tools and unique challenges in this postgenomic era. The large volume of information generated by the Human Genome Project has facilitated the development of many novel platforms for profiling each stage in the flow of biological information: from DNA to RNA to protein to the myriad of protein interactions, inspiring advancement of the burgeoning new areas of genomics, transcriptomics, proteomics, and interactomics, respectively.

Collectively, these core aspects of modern high-throughput biology each aim to provide a cross-sectional snapshot of fundamental biology to simultaneously assess the direct or downstream influence of thousands of genes. Ultimately, this may allow us to then identify and characterize the entire space of biomolecules that constitute the composite catalogue of possible therapeutic targets and their roles in disease, and to thereby affect disease at a molecular level through improved rational design of new classes of micro- and nanoscale molecular therapeutic agents and bioactive medical devices. Improved understanding of individual targets promises to be useful in the assessment of risk, diagnosis, prognosis, and therapy of human disease, lending to the possibilities of personalized medicine.

As these high-throughput biological tools have the potential to introduce significant changes that could affect the practice of clinical medicine, it is critical that interventional radiologists understand them so they can critically evaluate and integrate them directly into their own research and clinical efforts.

Herein we review the implications of high-throughput biology in the postgenomic era for biomedical research and clinical practice. In the first section, we discuss the basic principles behind high-throughput tools. We principally focus on array-based high-throughput biological methods, a key high-throughput biological tool, and describe how they are being used to uncover different aspects of biology. In the second half, we discuss interpretation of high-throughput data and the challenges high-throughput biological techniques present for data analysis. These discussions provide an understanding of the high-throughput technologies currently being applied in basic research, the biomedical questions they can be used to address, and a foundation for the interpretation of novel results.

ARRAY-BASED HIGH-THROUGHPUT PLATFORMS

Array-based high-throughput methods have rapidly been adopted by the biomedical research community. Since the initial description of microarray devices more than a decade ago, they have grown to encompass many aspects of molecular and cell biology (2,3). Originally devised as a method for rapid analysis of DNA or RNA samples, they have since been extended to assay protein/DNA, protein/protein, and cell-level interactions. These tools allow simultaneous reporting on the behavior of thousands of genes, transcripts, and their products, facilitating a transition from qualitative analysis of a few genes to quantitative analysis of gene networks and cell physiology (**Fig 1**). Whereas traditional approaches generally focus on the response of a few genes to a biological perturbation, high-throughput experimental tools now allow one to investigate many genes without prior knowledge of what genes are im-

From the Department of Radiology (A.H., M.D.K.) and Center for Translational Medical Systems (M.D.K.), University of California San Diego Medical Center, San Diego, California 92103. Received February 22, 2006; revision requested April 10; final revision received and accepted May 1. **Address correspondence to** M.D.K.; E-mail: mkuo@ucsd.edu

Neither of the authors have identified a conflict of interest.

DOI: 10.1097/01.RVI.0000228840.12260.FF

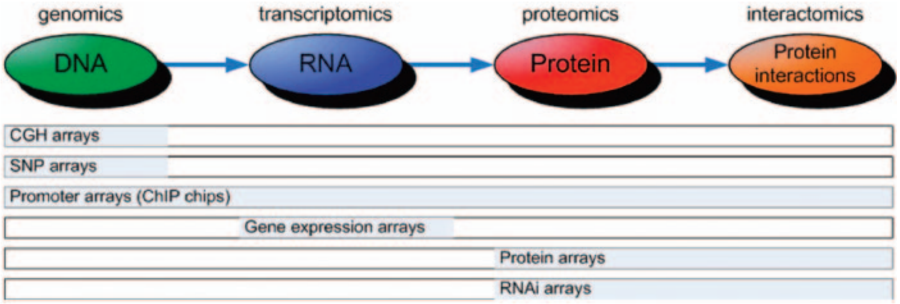

Figure 1. Overview of array platforms and the dimensions of molecular biology that they can help us understand. CGH and SNP arrays explore DNA content and polymorphic sequences, respectively. Promoter arrays span many dimensions of molecular biology, depending on the specific protocol and experimental design used. Protein and RNA interference arrays address the content and activity of individual proteins.

portant, heralding a paradigm shift in scientific inquiry to discovery-driven science. Because many of the genetic and epigenetic mechanisms underlying human disease are not yet known, high-throughput biology has the potential to revolutionize the way that biomedical research is conducted and to accelerate the discovery of new therapeutic agents to address human disease.

Gene Expression Microarrays

The most common and well-developed array platform is the gene expression microarray. Several variations of this tool are commonly used, and each comes with its own caveats, but the general principle behind these arrays is the same. Each gene expression array is manufactured through the highly ordered spotting, printing, or in situ synthesis of thousands of molecular probes on a glass slide or nylon membrane. These probes consist of short oligonucleotides or single-stranded complementary DNA (cDNA) sequences fixed to the array surface. Each probe is designed to be specific for a particular RNA species. RNA from a sample of interest—whether from cell culture, harvested human or animal tissue, or biopsy—is extracted, purified, copied, and labeled with a fluorescent dye. When the fluorescent-labeled RNA is hybridized onto the array, each probe hybridizes to its specific and complementary sequence on the array via Watson-Crick base pairing. After excess sample is washed off, the remaining fluorescent-labeled RNA can be quantitated as a function of its fluorescence signal intensity with the use of confocal microscopy, thereby providing a measure of the proportion of a given RNA species in the original sample (**Fig 2**). The end result is the ability to perform massively parallel gene expression hybridization experiments that allow one to simultaneously "visualize" and quantify the expression of every gene on the array at once. Microarray core facilities at many research institutions are already capable of manufacturing these arrays. Alternatively, they can be purchased through large commercial manufacturers such as Affymetrix (Santa Clara, CA) and Agilent (Palo Alto, CA).

Affymetrix gene expression arrays are commercial arrays that consist of hundreds of thousands of oligonucleotide probes, each precisely synthesized on the surface of a glass slide by computer-aided photolithography. Affymetrix has taken several unique steps to improve the reliability of gene expression measurements from their arrays. Instead of reading the brightness of each spot directly as a measure of gene expression, Affymetrix arrays include several perfect match and mismatch probes for each gene of interest. Combined, these probes are known as a probe set. Signals for each probe in a probe set are normalized and averaged during postprocessing and integrated to obtain probe set gene expression scores. Several methods for this level of data processing are now available and have been reviewed elsewhere (4,5).

In contrast to Affymetrix arrays, most other gene expression platforms use robotic spotters to "print" thousands of cDNA or oligonucleotide probes in precise locations on a glass slide or membrane. These probes are then chemically fixed on the surface of the microarray. Although this method is generally less precise for controlling the amount of probe at each location on the array, robotic spotting is relatively cheap and widely available. RNA samples can be hybridized to each array in a manner similar to that described earlier. Because the amount of probe present at each spot on the array is variable, the signal measured from chip to chip will also vary, even if the hybridized RNA sample is exactly the same. To overcome inherent chip-to-chip variability, many investigators have additionally used differential labeling of two RNA samples. One RNA sample is labeled with a green fluorescent dye (Cy3), and the other is labeled with a red fluorescent dye (Cy5). Cohybridization of two samples at the same time allows the relative difference in hybridization intensity to be used as a finer measure of gene expression, which can account for the amount of probe spotted on the chip.

Two common experimental designs for two-channel microarrays include the reference design and the paired experiment (**Fig 3**). In a reference design experiment, an RNA sample of interest is cohybridized to the array along with a reference RNA sample. The same reference RNA is used for all microarrays in the study. For example, to examine the gene expression patterns of hepatocellular carcinomas, Cheung et al (6) extracted RNA from resected tumors and cohybridized these with a standard reference pooled from several different cell lines. The standard liver reference should theoretically have the same gene expression across all arrays and can therefore be used to normalize for variability in probe spotting. In comparison, a paired microarray experiment cohybridizes samples that are experimentally related without a common reference. One publicly available example from the Alliance for Cellular Signaling involves a time course of endotoxin-treated macrophages (7). In this experiment, investigators took cultured macrophages and treated them with endotoxin or vehicle for six different time intervals (2, 4, 6, 8, 24, and 48 hours). At each time point, an endotoxin treatment and a vehicle treatment were differentially labeled and

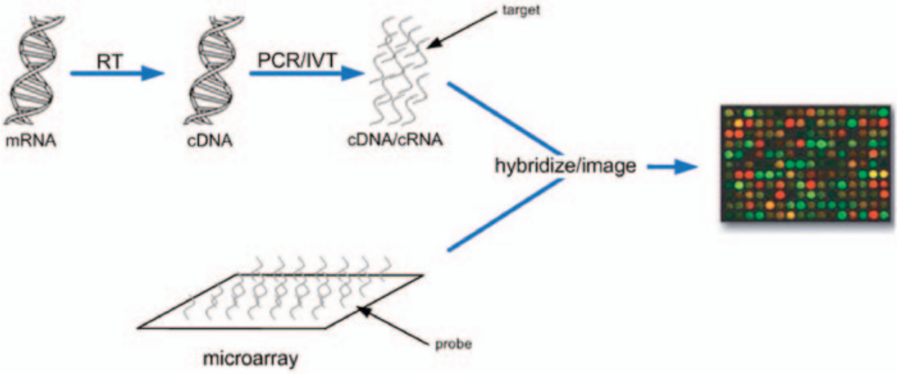

Figure 2. Overview of gene expression profiling. Messenger RNA is isolated from tissues or cells and copied, labeled, and hybridized onto microarrays, which are subsequently scanned by a confocal microscope. Computational methods are subsequently used to interpret the resulting image. (RT-PCR = reverse transcriptase polymerase chain reaction; IVT = in vitro transcription.)

cohybridized on an Agilent oligonucleotide array. The investigators also took the additional care of "dye-swapping" to control for any bias introduced by labeling with either dye. Because each dye-swap pair was performed in triplicate, this data set contained a total of six arrays for each time point. This kind of experimental design simultaneously addresses the chip-to-chip variability and the inherent variability of vehicle-treated cells. Although the reference design and paired design are similar, each has characteristic advantages and disadvantages. The choice of a particular design may be limited by the resources available (eg, biological material, financial resources, labor constraints) and by the quality of the arrays. This choice may ultimately limit the analytic methods that can be applied and therefore the quality and interpretation of the results. Each of these factors should be considered before a particular microarray platform or experimental design is decided on.

Comparative Genomic Hybridization Arrays

Other types of microarrays are designed with similar concepts in mind (**Table**). Rather than targeting RNA, comparative genome hybridization (CGH) arrays bind DNA sequences with high specificity. These arrays potentially allow rapid and high-resolution profiling of DNA copy number changes across an entire genome. Because it was recognized that DNA copy number changes are characteristic of many different cancers, these arrays were devised to rapidly assess which chromosomal segments are highly replicated or lost relative to background DNA (8). CGH arrays are often designed with cDNA probes periodically spaced across the genome while avoiding highly repetitive chromosomal regions. When a fragmented DNA sample is hybridized on its surface, the array can be used to assess the relative proportions of each chromosomal segment. This allows for much greater spatial resolution (approximately 1 Mb) than traditional comparative genomic hybridization methods (approximately 20 Mb) and can also be performed on a genome-wide basis, as opposed to fluorescence in situ hybridization, which generally operates on a gene-by-gene basis. In normal human tissue, copy numbers should be fairly consistent across all chromosomal segments. In pathologic tissue such as cancer, which is typified by genomic instability, DNA copy numbers may increase or decrease for specific chromosomal segments, revealing regions of amplification or deletion. This information may also uncover oncogenes or tumor-suppressor genes (9–11). Further, localized changes in copy number are diagnostic of some cancers and have even been used to classify certain tumors (12). Although these arrays have been successfully used to characterize copy number defects of some cancers, researchers are currently working on improving the copy number resolution of these arrays. Although large copy number changes are readily detected by current renditions of this platform, accurate assessment of single-copy number changes are still being worked out.

SNP Arrays

Single nucleotide polymorphism (SNP) arrays are also constructed with probes that bind specific DNA sequences. However, unlike CGH arrays, SNP array probes are generally designed around specific loci known to be polymorphic among individuals of a given species. Such arrays potentially allow rapid genotyping of individuals at polymorphic sites. It is estimated that there are several million sites in the human genome. Because this is currently beyond the scope of modern array technology, many SNP arrays have been designed around polymorphic sites of known biomedical significance. Researchers have only recently begun to use these platforms to associate genetic loci with increased susceptibility to human disease or response to certain drugs (13). For example, investigators have used SNP arrays to demonstrate genome-wide SNPs that result in loss of heterozygosity events in several cancers such as small-cell lung cancer and bladder cancer (14–16). More recently, Garraway et al (17) used a novel approach integrating SNP array and gene expression array data to uncover a melanoma-specific oncogene, MITF. Using SNP arrays, they were first able to isolate a chromosomal region of DNA copy gain in a melanoma cell line. Then, by integrating gene expression array data tightly associated with this region of amplification, they were able to isolate and identify the novel lineage-specific oncogene MITF as the primary driver. Interestingly, they noted that MITF is a master regulator for melanocyte differentiation and was amplified in melanomas but not in premalignant lesions such as melanocytic nevi. In addition, patients who had melanomas with this amplification had worse survival and a greater tendency to experience metastasis. This novel lineage-specific oncogene

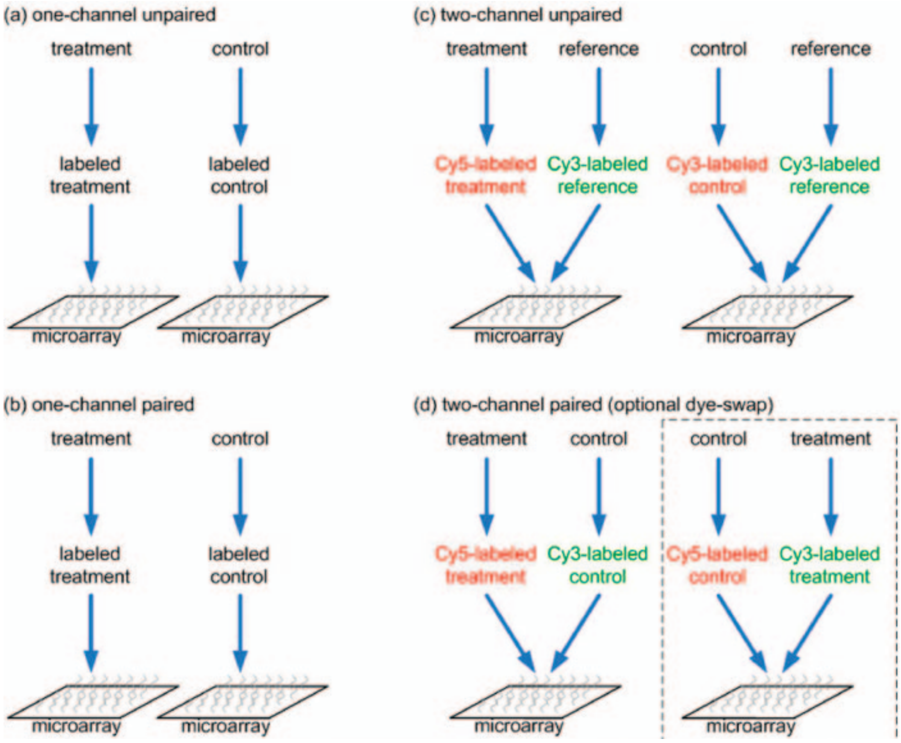

Figure 3. Common experimental designs used on one-channel and two-channel microarrays. In an example experiment, an investigator may want to measure RNA content with and without treatment with a new therapeutic agent. The investigator may choose from several possible experimental designs. **(a)** One-channel microarrays measure targets from a single sample per array. Treatments and controls are measured on separate arrays. **(b)** One-channel microarrays may also be used to measure paired controls and treatments. **(c)** Two-channel microarrays may involve comeasurement of standardized reference samples to control for variability between arrays. **(d)** Two-channel arrays may also be used without reference samples, recognizing the inherent paired nature of measuring both samples from a single array. When the paired two-channel design is used, experimental samples may be dye-reversed across multiple arrays to overcome dye bias.

may serve as a novel therapeutic target for the treatment of melanoma.

Although SNP arrays offer the advantage over CGH arrays of greater resolution at any given locus by being able to measure changes in DNA copy number and detect loss of heterozygosity, they cannot yet cover the entire genome. The number of predicted SNPs present in the human genome is estimated at more than 10 million; current SNP arrays are able to handle only a fraction of this number. Nevertheless, these results are very promising from a clinical standpoint because they demonstrate that these tools may help us stratify the risk of development of disease and better target preventive therapies to those most likely to benefit from them. They may help us identify patients who are more likely to show a response to a particular drug and identify those who are more likely to experience an adverse reaction.

Promoter Arrays

At a more basic level, one of the most exciting new array formats for furthering our understanding of gene regulation is the promoter array. Promoter arrays also use oligonucleotide or cDNA probes to bind DNA sequences, but unlike SNP and CGH arrays, promoter array probes are targeted to promoter regions upstream of genes of interest. Let us take, for example, an investigator interested in determining the DNA localization of peroxisome proliferator-activated receptor–γ coactivator 1α across the human genome in normal and pathologic liver tissue. To use these arrays, investigators first perform a chromatin immunoprecipitation (ChIP) experiment with an antibody against a particular transcriptional regulator. In this case, the investigator may use an antibody against peroxisome proliferator-activated receptor–γ coactivator 1α. The immunoprecipitated DNA/protein complex is fixed with formalin, which crosslinks the proteins with each other and with adjacent DNA. After fragmentation of this DNA, the crosslinks are chemically reversed. Ligation-mediated polymerase chain reaction can then be used to amplify and label the remaining DNA fragments. The resulting labeled DNA is relatively enriched for promoter segments in the vicinity of the immunoprecipitated transcriptional regulator. The investigator can then compare the resulting array images to determine which DNA-binding sites are accessible to peroxisome proliferator−activated receptor–γ coactivator 1α in normal tissue but not in pathologic tissue. Such arrays are therefore useful for identifying the transcriptional targets of individual transcription factors. Because many current drugs have direct effects on the transcriptional apparatus, this relatively new tool will be increasingly valuable for biomedical research.

This technology has already been used to help uncover the regulatory transcriptional networks of several important transcription factors such as C-Myc in Burkitt lymphoma and members of the hepatocyte nuclear factor family in liver and pancreatic islets (18–20). The insights gained from these studies could allow us to better understand the global role of these transcription factors, the genes they regulate, and how their dysregulation may directly or indirectly contribute to disease. In addition, they may ultimately facilitate the development of better-targeted therapeutic agents. These ChIP chip experiments have the potential to more precisely elucidate the direct mechanisms of drug action. They may be used not only to identify the specific genes that are targeted by these drugs but also to identify other transcriptional regulators that also play a role.

RNA Interference Arrays

Cell-based microarrays represent the next phase in array-based high-

Current Array-based Methods for High-throughput Biology

Method	Probe	Target
SNP	cDNA, oligonucleotide	Fragmented DNA
CGH	cDNA, oligonucleotide	Fragmented DNA
Promoter (ChIP chip)	cDNA	Fragmented DNA "enriched" by ChIP
Gene expression	cDNA, oligonucleotide	Messenger RNA
Protein	Antibodies, proteins, substrates	Protein/protein function

Note.—Multiple platforms exist for probing many dimensions of cellular and molecular biology. Each array consists of probes fixed to the array surface. These probes are each designed to quantify the relative amount of a molecular target. ChIP = chromatin immunoprecipitation.

throughput biology. With these arrays, it is possible to probe a spectrum of cell types or experimental conditions to assess whole-cell responses to various stimuli. One example of this approach is the RNA interference microarray (21). RNA interference is an experimental method that takes advantage of one of the cell's natural defense mechanisms. On a small scale, RNA interference allows investigators to specifically inhibit the translation of individual genes. Array-based RNA interference facilitates high-throughput analysis of hundreds of inhibitory RNAs simultaneously. With this platform, inhibitory RNA sequences are generally spotted and fixed on the array surface. The arrays are incubated with adherent cells of interest, which facilitates uptake of the inhibitory RNA and subsequent knockdown of cellular genes. At a fundamental level, this powerful technology allows one to uncover the large-scale roles of particular genes and their genome-wide influence. Several proof-of-concept studies have been published about this novel platform (22,23). For example, Berns et al (22) constructed a library of expression vectors for short hairpin RNAs, which were designed to individually silence 7,914 different genes. The investigators then devised a system to screen for genes that, when "knocked down," could escape cell cycle arrest (22). This system allowed the investigators to identify not only p53 as a critical regulator of their cell cycle arrest model, but five new genes as well. Because of its high-throughput nature and its ability to generate new leads in understanding human disease, this technology may be invaluable for many other fields of biomedical research. Until now, one of the limitations to the broad use of RNA interference arrays has been the absence of complete and effective RNA interference libraries. Several academic and commercial entities are devising such libraries, making the use of this platform more feasible (23). Because this new high-throughput tool is potentially so powerful, it is likely to become an essential tool in the repertoire of high-throughput biology.

Protein Arrays

Drug discovery may benefit not only from nucleotide-based arrays but also from the new arrival of protein microarrays. These devices are being actively developed to investigate the many dimensions of protein biology, including cellular protein content, protein/protein interactions, and enzymatic activity. Several groups have applied this technology to characterize cancer (24–26), autoimmune disease (27,28), and experimental models of type I diabetes mellitus (29). To identify novel protein/protein interactions, others have arrayed thousands of proteins and characterized the extent of antibody, calmodulin, lipid, and integrin $\alpha_{IIb}\beta_3$ binding (30,31). Even more recently, Ramachandran et al (32) demonstrated that protein microarrays could be "self-assembled" from cDNA by use of a cell-free mammalian reticulocyte lysate. Because protein arrays have been relatively difficult to manufacture, this discovery has the potential to make protein arrays more broadly accessible by coopting widely available tools for the printing of cDNA arrays. With such arrays more accessible, investigators will be able to rapidly assay protein/protein interactions of any protein of interest. Array-based methods for the assessment of protein enzymatic activities are also beginning to be developed. By arraying more than 1,000 protein kinases in microwells of a flexible silicone sheet, it is also possible to assess the ability of each of these kinases in phosphorylating a series of potential substrates (33). As each of these protein array modalities are further developed, investigators will be able to more easily explore the highly uncharacterized dimension of protein interaction.

As can be seen by these technologies, with the sequencing of the human genome and advances in fabrication methods, high-throughput arrays can be manufactured and used for many biological substrates. Although gene expression microarrays continue to be the most widely used array platform, other arrays designed to silence individual genes, assess DNA content, and assess protein interactions are also being developed. In parallel to these advances, statistical methods for the interpretation of these high-throughput data have also been rapidly developed. Such methods are essential for answering specific biological questions, for visualizing the sheer volume of data produced, and for providing biological interpretations that best fit the pattern of data observed.

INTERPRETATION OF ARRAY DATA

Background

Because of its high-throughput nature, array-based data methods produce information orders of magnitude greater than conventional experiments and introduce statistical challenges not encountered in conventional biology. Although the principles discussed herein are applicable to other array-based platforms, we will focus our attention on gene expression arrays to simplify discussion. In many cases, the magnitudes of array signals are not dramatically different from the level of background noise. Many technical improvements, including the use of multiple probes for single genes and cohybridization of multiple samples, have gradually improved array platforms, but this continues to be an important issue because biological samples themselves have considerable variability. Although this is also a challenge for conventional biological methods, this problem is magnified because of the sheer number of simultaneous experiments being performed on a microarray.

Suppose one would like to identify genes that are altered in expression between hepatocellular carcinoma and adjacent liver tissue. For this purpose, several paired samples can be obtained from a variety of patients. With a conventional biological approach, we may hypothesize that a particular gene should be expressed more greatly in the tumor sample than in normal liver because it enhances cell proliferation. We can perform the corresponding assay, isolating RNA from each biopsy specimen, and compare the expression of the gene between tumor and normal tissue. A Student *t* test and a conventional threshold for statistical significance (α value of 0.05) may then be used to assess whether gene expression is significantly different between tumor and normal tissue. By doing so, we can effectively assess the reproducibility of this experiment, given the same tumor and normal tissue.

In a microarray experiment, in which no specific genes are isolated, we simultaneously perform this experiment on thousands of genes. Total cellular RNA may be extracted from each of these samples and hybridized on corresponding arrays, and gene expression scores for every measured gene can be obtained from the confocal image. Because so many RNA species are being measured, we must account for the thousands of parallel statistical tests being performed or risk identifying false-positive findings based on the sheer number of experiments attempted. With use of the same statistical tests and significance thresholds, we would expect one of every 20 tests to have a false-positive result by chance alone. Therefore, for an array of 10,000 probes, approximately 500 false-positive findings would be expected. If we do not adequately address these issues, many of the genes we identify may be spurious and may not reliably be found again in future studies.

Several approaches can be used to address this problem by changing the statistical test used to determine statistical significance and/or changing the threshold for statistical significance. One such approach is to use Bonferroni correction. The Bonferroni correction corrects for the number of parallel tests by dividing the significance threshold by the number of statistical tests being performed ($\alpha_{\mathrm{Bonf}} = \alpha/\mathrm{n}$). Some investigators refer to this threshold as an array-wide false-positive rate (34). In other words, for an α_{Bonf} value of 0.05, there is a 5% chance that at least one false-positive result will be observed in the list of "significant" differences in gene expression. With an n on the order of 10^5, this threshold for significance becomes very small, and changes in gene expression must be extremely dramatic to be detected. As previously mentioned, array signals are rarely dramatically different from background noise. Therefore, for most statistical tests that are commonly used, Bonferroni thresholds are too strict, and few if any genes can be found that achieve this level of statistical significance. Another commonly used statistical threshold is the false-discovery rate. Instead of strictly relating the threshold to the number of statistical tests performed, the false-discovery rate is based on the number of "significant" genes identified. One way of performing this is by ranking the *P* values for hundreds of genes and choosing a threshold for significance through an iterative process. Suppose a false-discovery rate of 0.05 is desired. By choosing this false-discovery rate, one expects that approximately one of every 20 genes identified will be a false-positive finding. The *P* value threshold is iteratively changed until the expected number of false-positive findings equals 0.05. Generally less strict than the Bonferroni threshold, false-discovery rate can be a viable alternative and has been integrated into several statistical tools for microarray analysis (5,35).

The choice of a statistical test to interpret gene expression data is highly dependent on the experimental design and on the assumptions that the investigator is willing to accept. Statistical assumptions about the underlying data can be powerful if valid and can increase the sensitivity for real changes in gene expression. Therefore, a wide spectrum of statistical tools have been developed, each addressing different sets of initial assumptions. They range from nonparametric tests (35) to the conventional parametric tests (eg, *t* test, analysis of variance), to the novel parametric tests (eg, Cyber-T, VAMPIRE) (34,36). Although not all these methods have been thoroughly compared in the literature, there is increasing evidence of improved sensitivity and specificity in highly parameterized statistical tests that account for the relationship between gene expression and noise (4).

Higher-order analyses are also possible with array data. Clustering methods are commonly used to group data by similarities in signal patterns. Suppose we again examine the pattern of expression of human hepatocellular carcinomas. We may use clustering tools to group genes that are most similar in expression pattern across all samples. This form of clustering is theoretically useful for bringing together genes that share common regulation (37). Alternatively, we may also cluster tumor samples to group tumors that are most related based on their global expression profile. This form of clustering can bring together tumors that are most similar and can separate tumors that are most different in their gene expression profile. Finally, we may cluster samples on both "axes" in such a manner that groups of genes that are similar are grouped along with samples that share similar global expression profiles (**Fig 4**). When no previous categorizations are available for the data, this can be useful for the stratification of different cancer types. Several investigators have used this type of approach to discover new tumor classes, some of which have distinct differences in clinical outcome based on these new molecular classifications (38).

More recent algorithms are beginning to integrate the knowledge contained in biological databases into the interpretation of array data. Instead of relying solely on patterns in expression, several authors (39–41) have devised tools that integrate annotation databases such as Gene Ontology, Kyoto Encyclopedia of Genes and Genomes, and the TRANSFAC transcription factor database to facilitate interpretation. These databases provide a highly structured framework for defining specific biological terms and mapping these terms to individual genes. Specifically, the Gene Ontology database stores information about spe-

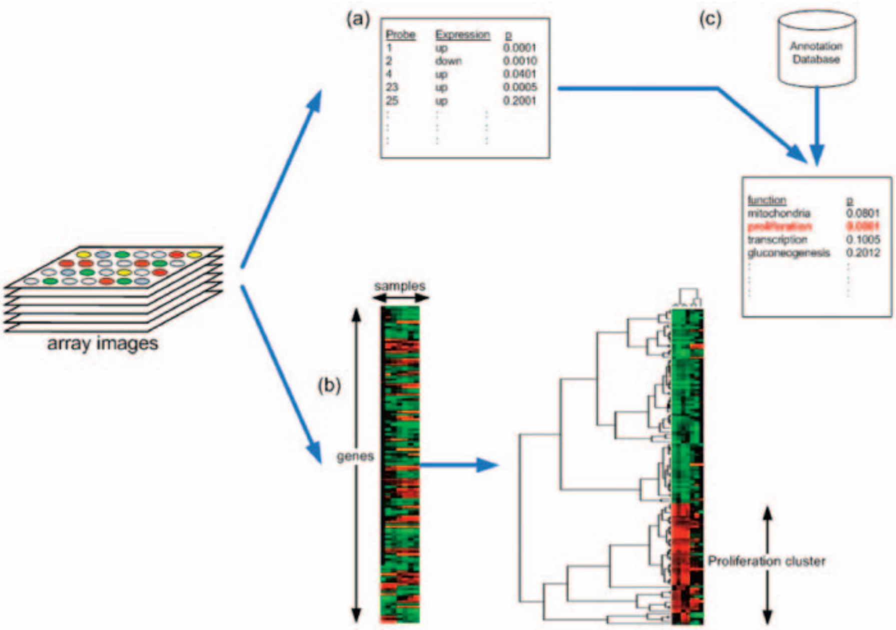

Figure 4. Typical analyses of array data. **(a)** The simplest and most basic analysis involves the identification of probes associated with significant differences in intensity between sets of samples. **(b)** A set of probes may also be used in clustering. Clustering methods can be used to identify probes and samples that are most similar. In gene expression profiling, clustered genes often represent genes that are functionally related by gene regulation or functional role in the cell. Samples that cluster together share similar measurements across many genes. These clusters can be used to define new functional relationships (eg, cell proliferation cluster). **(c)** Functional enrichment analysis is used to identify previously defined groups of genes that are unusually enriched. An analysis algorithm is used to identify a subset of genes of interest, the annotations associated with these genes are explored in annotation databases, and these are compared to the genes in the background. In this example, a set of probes is significantly enriched for proliferation (highlighted in red) at a *P* value of .0001.

cific molecular functions, biological processes, and cellular locations (39). The Kyoto Encyclopedia of Genes and Genomes database provides a resource to associate genes with signaling networks and metabolic maps (40). The TRANSFAC database describes the location and nature of upstream transcription factor binding sites (41). Tools that interface these databases range from those that solely provide database access to tools that integrate sophisticated statistical calculations to identify the most robust biological patterns in a set of gene expression data (5,42–44).

The development of high-throughput array platforms has led to an explosion of biological data and has stimulated the creation of several databases for providing public access. Among the most widely used public repositories are the Gene Expression Omnibus provided by National Center for Biotechnology Information, the Stanford Microarray Database, and ArrayExpress at the European Bioinformatics Institute (45–47). Because of the high dimensionality of these data, it is likely that much of the knowledge contained in these easily accessible data sets has yet to be uncovered. In addition, as novel approaches for interpreting these data are devised, individual investigators may find it increasingly valuable to reinterpret previously published data to guide scientific discovery and to identify new avenues of scientific research.

CONCLUSION

High-throughput methods have already had a considerable impact on biomedical research. When adopting high-throughput technology, new investigators should identify the most effective platform and experimental design that best fits their biological question, given their specific resource limitations. Unlike traditional biological experiments, these experiments are generally processing intensive, and considerable care during the analysis is required to obtain reproducible results and interpretations. In cancer, for example, these new tools have pointed us to a better understanding of the cell cycle–regulatory pathways that lead to tumorigenesis. In type II diabetes mellitus, they suggest that mitochondrial function may be a limiting factor to metabolic function in insulin resistance (48). In rheumatologic disease, they may be valuable for characterizing the pattern of autoantibodies associated with each clinical syndrome (28). As new commercial products arrive to take advantage of improved diagnostic and prognostic accuracy and advances in miniaturization, high-throughput technology may also exert a similar influence on clinical practice. These devices may shape the decisions that physicians and patients make regarding the form and extent of chemotherapy to pursue, for example. Medical therapeutic agents will also likely evolve as these tools enhance decision making. In the future, we may improve targeted therapy for even finer classifications of disease that are currently indistinguishable by histopathologic examination. Although they are not addressed in this report, information systems for the storage and interface of this biomedical knowledge must also adapt to allow physicians to provide optimal care to their patients. Medical practice has and will continue to undergo great transformations, and it is likely that high-throughput biological tools will play an important role (**Appendix**).

References

1. Marshall E. Bermuda rules: community spirit, with teeth. Science 2001; 291:1192.
2. Fodor SPA, Rava RP, Huang XC, et al. Multiplexed biochemical assays with biological chips. Nature 1993; 364:555–556.
3. Schena M, Shalon D, Davis RW, et al. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995; 270:467–470.
4. Choe S, Boutros M, Michelson A, et al. Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. Genome Biol 2005; 6:R16.
5. Hsiao A, Ideker T, Olefsky JM, et al. VAMPIRE microarray suite: a web-

based platform for the interpretation of gene expression data. Nucleic Acids Res 2005; 33(Web Server issue):W627–W632.

6. Cheung ST, Chen X, Guan XY, et al. Identify metastasis-associated genes in hepatocellular carcinoma through clonality delineation for multinodular tumor. Cancer Res 2002; 62:4711–4721.
7. Gilman AG, Simon MI, Bourne HR, et al. Overview of the Alliance for Cellular Signaling. Nature 2002; 420:703–706.
8. Pollack JR, Perou CM, Alizadeh AA, et al. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat Genet 1999; 23:41–46.
9. Albertson DG, Ylstra B, Segraves R et al. Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene. Nat Genet 2000; 25:144–146.
10. Wilhelm M, Veltman JA, Olshen AB, et al. Array-based comparative genomic hybridization for the differential diagnosis of renal cell cancer. Cancer Res 2002; 62:957–960.
11. Cheng KW, Lahad JP, Kuo WL et al. The RAB25 small GTPase determines aggressiveness of ovarian and breast cancers. Nat Med 2004; 10:1251–1256.
12. Bredel M, Bredel C, Juric D, et al. High-resolution genome-wide mapping of genetic alterations in human glial brain tumors. Cancer Res 2005; 65:4088–4096.
13. Syvanen AC. Toward genome-wide SNP genotyping. Nat Genet 2005; 37(suppl):S5–10.
14. Lindblad-Toh K, Tanenbaum DM, Daly MJ, et al. Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. Nat Biotech 2000; 18:1001–1005.
15. Hoque MO, Lee J, Begum S, et al. High-throughput molecular analysis of urine sediment for the detection of bladder cancer by high-density single-nucleotide polymorphism array. Cancer Res 2003; 63:5723–5726.
16. Veltman JA, Fridlyand J, Pejavar S, et al. Array-based comparative genomic hybridization for genome-wide screening of DNA copy number in bladder tumors. Cancer Res 2003; 63:2872–2880.
17. Garraway LA, Widlund HR, Rubin MA, et al. Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. Nature 2005; 436:117–122.
18. Ren B, Robert F, Wyrick JJ, et al. Genome-wide location and function of DNA binding proteins. Science 2000; 290:2306–2309.
19. Li Z, Van Calcar S, Qu C, et al. A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. Proc Natl Acad Sci U S A 2003; 100: 8164–8169.
20. Odom DT, Zizlsperger N, Gordon DB, et al. Control of pancreas and liver gene expression by HNF transcription factors. Science 2004; 303:1378–1381.
21. Wheeler DB, Carpenter AE, Sabatini DM. Cell microarrays and RNA interference chip away at gene function. Nat Genet 2005; 37(suppl):S25–30.
22. Berns K, Hijmans EM, Mullenders J, et al. A large-scale RNAi screen in human cells identifies new components of the p53 pathway. Nature 2004; 428: 431–437.
23. Paddison PJ, Silva JM, Conklin DS, et al. A resource for large-scale RNA-interference-based screens in mammals. Nature 2004; 428:427–431.
24. Hudelist G, Pacher-Zavisin M, Singer C, et al. Use of high-throughput protein array for profiling of differentially expressed proteins in normal and malignant breast tissue. Breast Cancer Res Treat 2004; 86:281–291.
25. Wang X, Yu J, Sreekumar A, et al. Autoantibody signatures in prostate cancer. N Engl J Med 2005; 353:1224–1235.
26. Belov L, de la Vega O, dos Remedios CG, et al. Immunophenotyping of leukemias using a cluster of differentiation antibody microarray. Cancer Res 2001; 61:4483–4489.
27. Robinson WH, Fontoura P, Lee BJ, et al. Protein microarrays guide tolerizing DNA vaccine treatment of autoimmune encephalomyelitis. Nat Biotech 2003; 21:1033–1039.
28. Robinson WH, DiGennaro C, Hueber W, et al. Autoantigen microarrays for multiplex characterization of autoantibody responses. Nat Med 2002; 8:295–301.
29. Quintana FJ, Hagedorn PH, Elizur G, et al. Functional immunomics: microarray analysis of IgG autoantibody repertoires predicts the future response of mice to induced diabetes. Proc Natl Acad Sci U S A 2004; 101(suppl 2):14615–14621.
30. Michaud GA, Salcius M, Zhou F, et al. Analyzing antibody specificity with whole proteome microarrays. Nat Biotechnol 2003; 21:1509–1512.
31. Zhu H, Bilgin M, Bangham R, et al. Global analysis of protein activities using proteome chips. Science 2001; 293: 2101–2105.
32. Ramachandran N, Hainsworth E, Bhullar B, et al. Self-assembling protein microarrays. Science 2004; 305:86–90.
33. Zhu H, Klemic JF, Chang S, et al. Analysis of yeast protein kinases using protein chips. Nat Genet 2000; 26:283–289.
34. Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatics 2001; 17:509–519.
35. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2001; 98:5116–5121.
36. Hsiao A, Worrall DS, Olefsky JM, et al. Variance-modeled posterior inference of microarray data: detecting gene-expression changes in 3T3-L1 adipocytes. Bioinformatics 2004; 20:3108–3127.
37. Eisen MB, Spellman PT, Brown PO, et al. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998; 95:14863–14868.
38. Sorlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A 2001; 98:10869–10874.
39. Gene Ontology Consortium. Creating the Gene Ontology Resource: design and implementation. Genome Res 2001; 11:1425–1433.
40. Kanehisa M, Goto S, Kawashima S, et al. The KEGG resource for deciphering the genome. Nucleic Acids Res 2004; 32:D277–D280.
41. Wingender E, Chen X, Fricke E, et al. The TRANSFAC system on gene expression regulation. Nucleic Acids Res 2001; 29:281–283.
42. Liu G, Loraine AE, Shigeta R, et al. NetAffx: Affymetrix probesets and annotations. Nucleic Acids Res 2003; 31: 82–86.
43. Beissbarth T, Speed TP. GOstat: find statistically overrepresented gene ontologies within a group of genes. Bioinformatics 2004; 20:1464–1465.
44. Zhang B, Schmoyer D, Kirov S, et al. GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using gene ontology hierarchies. BMC Bioinformatics 2004;5:16.
45. Sherlock G, Hernandez-Boussard T, Kasarskis A, et al. The Stanford Microarray Database. Nucleic Acids Res 2001; 29:152–155.
46. Parkinson H, Sarkans U, Shojatalab M, et al. ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic Acids Res 2005; 33(suppl 1):D553–555.
47. Barrett T, Suzek TO, Troup DB, et al. NCBI GEO: mining millions of expression profiles— database and tools. Nucleic Acids Res. 2005; 33(suppl 1): D562–566.
48. Mootha VK, Lindgren CM, Eriksson KF, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 2003; 34:267–273.

APPENDIX:
GLOSSARY

Term	Definition
CGH	Comparative genomic hybridization, a method to determine the relative number of chromosomal segments present in a particular sample
α	Threshold for type I error, the likelihood of a false-positive result
Amplification	Abnormal increase in the number of copies of a particular segment of DNA; can be associated with oncogenes,
Bonferroni correction	Statistical correction that accounts for the number of statistical tests performed, often performed by dividing the significance threshold α by the number of tests ($\alpha_{bonf} = \alpha/n$).
cDNA	Complementary DNA, usually obtained through reverse-transcription of mRNA
"ChIP chip"	Chromatin immunoprecipitation combined with microarray analysis; generally used to assess the location of transcription factors and other proteins across the genome
cRNA	Complementary RNA, usually obtained from reverse-transcription of mRNA, followed by in vitro transcription
Deletion	Abnormal loss in the number of copies of a particular segment of DNA; can be associated with tumor suppressor genes
FDR	False-discovery rate. A statistical correction that accounts for the number of statistical tests by setting an *a priori* rate of acceptable false-positives.
Fluorescence in situ hybridization	A physical genomic mapping technique that utilizes fluorescently labeled DNA probes which bind to chromosomes or chromatin and which can then be detected by fluorescence microscopy. It is commonly used to identify the chromosomal location of a particular genomic sequence or to detect a chromosomal abnormality such as a translocation
Genomics	The comprehensive study of the structure and function of the entire set of genes in a cell or organism
Human Genome Project	A large international research project coordinated by the NIH and DOE to map and sequence the DNA in the entire human genome
Interactomics	The study of the complement of biomolecular interactions
LM-PCR	Ligation-mediated polymerase chain reaction, a method used to prime all nucleotide species for replication, followed by PCR-based replication
mRNA	Messenger RNA, the transcribed RNA sequences responsible for protein expression
Oligonucleotide	A short string of nucleotides, typically consisting less than 25 bases
Oncogene	A gene capable of transforming normal cells into cancer cells and are typically involved in cell growth or differentiation. Examples include myc, ras, and HER-2/*neu*
Probe	A chemical (eg, oligonucleotides, cDNA, antibodies, substrates) fixed to the array surface, which is specifically bound or modified by a target of interest; typically thousands of individual probes are placed at specific locations across the array; in the case of a gene expression array, probes are oligonucleotides or cDNA specific for a particular mRNA
Probe set	A collection of probes that cumulatively measure signal for a particular nucleotide species; devised by Affymetrix to provide a more robust measurements from their gene expression microarrays
Proteomics	The systematic study of the protein complement of the genome
RNAi	RNA interference. The method of performing gene silencing by using small inhibitory RNA sequences.
RNAi array	RNA interference array; designed to simultaneously assess the effect of RNAi on thousands of individual genes
Single nucleotide polymorphism (SNP)	Common but small variations in genomic DNA sequences that occur at a frequency of approximately 1 in 1,000 bases and in ≥1% of the population
siRNA	Small inhibitory RNA, used to inhibit translation of specific gene transcripts
Target	Biomolecules being measured by the array; in the case of a gene expression array, targets are the extracted mRNA from a biopsy or other biological specimen
Transcriptomics	The study of the gene expression (mRNA) levels of the set of all genes in a given population in a given condition
Tumor suppressor genes	Genes which generally inhibit the uncontrolled growth of cells. Examples include p53, and BRCA1
Two-channel microarray	An array that measures signals from two samples simultaneously
Watson-Crick base pairing	Complementary nitrogenous base pairing that connects complementary strands of DNA or RNA and consist of hydrogen bond associations between purines and pyrimidines