BIOCHEMISTRY AND MOLECULAR BIOLOGY EDUCATION
Vol. 32, No. 4, pp. 217–227, 2004

Articles

The Microarray Revolution

PERSPECTIVES FROM EDUCATORS

Received for publication, September 4, 2003, and in revised form, January 23, 2004

Jay L. Brewster‡§¶, K. Beth Beason§‖, Todd T. Eckdahl§, and Irene M. Evans§‡‡**

*From the ‡Natural Science Division, Pepperdine University, Malibu, CA 90263, §The Genome Consortium for Active Teaching (GCAT), Department of Biology, Davidson College, Davidson, NC 28035, ‖Department of Biochemistry and Cell Biology, Rice University, Houston, TX 77251, **Department of Biology, Missouri Western State College, Saint Joseph, MO 64507, and ‡‡Department of Biological Sciences, Rochester Institute of Technology, Rochester, NY 14623*

In recent years, microarray analysis has become a key experimental tool, enabling the analysis of genome-wide patterns of gene expression. This review approaches the microarray revolution with a focus upon four topics: 1) the early development of this technology and its application to cancer diagnostics; 2) a primer of microarray research, designed to guide the beginner; 3) a highlight of the Genome Consortium for Active Teaching (GCAT), a worldwide consortium of faculty who are integrating microarrays into the undergraduate teaching laboratory; and 4) the use of microarrays in the biotechnology industry with a look forward to future applications. A central theme within this review is the profound relevance of new, bioinformatics-based, technologies to undergraduate students within the biosciences.

Keywords: Microarray technology, undergraduate education, Genome Consortium for Active Teaching (GCAT), bioinformatics, gene expression.

One of the most powerful new technologies to emerge from the age of genome sequencing comes from the tiny microarray slide, carrying the capacity to comparatively scan genome-wide patterns of gene expression for any organism with a sequenced genome. First developed in research laboratories examining model organisms (yeast, mustard), microarrays are now being used worldwide to study everything from cancer biology and drug development to the evolutionary biology of microbes. Already, the basics of array technology are being adapted to characterize more than gene expression, to include the diagnosis of disease predisposition in humans, the rapid identification of specific viruses in infected humans, and protein analysis through the burgeoning field of proteomics. In this review, we consider the history of microarray development, fundamentals of the technology involved, and applications for the medical and pharmaceutical industries. In addition, we offer an introduction to the Genome Consortium for Active Teaching (GCAT), a collection of faculty committed to the inclusion of microarrays in the undergraduate teaching laboratory.

THE BEGINNING OF MICROARRAYS

Although the concept of using microarrays can be traced back 25 years to the introduction of the Southern blot [1], modern microarray analysis was introduced in 1995 by a Stanford University research team led by Pat Brown and Ron Davis. Their seminal publication was titled "Quantitative monitoring of gene expression patterns with a complementary DNA microarray" and has since been cited over 1,500 times [2]. The authors described the use of a robotics system to spot DNA oligonucleotides onto glass slides in ordered arrays, generating microarray slides or "gene chips." Labeled cDNAs made from varied samples of *Arabidopsis thaliana* (mustard plant) mRNA were hybridized to DNA on the chips. Only 45 oligonucleotide sequences were spotted in this first experiment, representing a tiny fraction of the total number of the genes present in *Arabidopsis*, but the work inspired many new experiments and soon whole genomes of species such as yeast, bacteria, mice, and humans were being spotted onto glass slides. Some early studies asked questions about genome size and diversity in different yeast strains or changes in gene expression as yeast experienced varied growth conditions. As was to be expected from such a revolutionary new technology, there were many surprises. Lashkari *et al.* [3] showed that laboratory yeast strains sometimes discard DNA fragments, encoding whole sets of genes. Researchers realized that culturing cells in common rich broths selects for the fastest growing cells, offering an advantage for cells that have discarded genes that are unnecessary for rapid growth. DeRisi *et al.* [4] showed the concerted induction and repression of numerous genes and pathways as yeast responded to environmental changes such as the depletion of glucose. By looking for upstream regulatory elements and transcription

¶ To whom correspondence should be addressed: Natural Science Division, Pepperdine University, 24255 Pacific Coast Highway, Malibu, CA 90263. Tel.: 301-506-4321; Fax: 310-506-4785; E-mail: jay.brewster@pepperdine.edu.

factors shared by genes regulated in concert, they were able to use the microarray data set to characterize gene regulatory pathways.

These early articles clearly identified the power of microarray analysis, and many laboratories decided to add this technology to their repertoire. In one of the footnotes to the DeRisi article [4], the authors suggest it would take a well-organized laboratory only 6 months to set up yeast array experiments. This would include everything from building an array printer, to generating oligonucleotide probes for all 6,400 yeast genes, to printing and utilizing the arrays. DeRisi and Brown developed and posted a "how to" manual for building an array printer from scratch, known as the Mguide [5], and laboratories throughout the world began to build. Today, commercial array printers and scanners are widely available, as are commercial prespotted slides. As a result, the use of microarrays in basic and applied research is growing at an extraordinary rate.

Microarrays Illuminate Many Areas of Biological Science

An examination of the published literature from 1997 to 2004 demonstrates that microarray technology has provided a powerful method for analysis of biological problems. Developmental biologists have measured changes in gene expression in organisms at different developmental stages. Neuroscientists have studied patterns of gene expression in varied areas of the brain before and after specific tasks are performed,and compared transcriptional patterns in pathological (*e.g.* Alzheimer's) *versus* nonpathological brains. Molecular biologists have looked at changes in gene expression when specific mutations or gene knockouts were present in an organism. Tissue-specific gene expression patterns in normal kidney and heart have been compared with those found in abnormal pathological conditions such as kidney failure and heart dysfunction. From the study of micropathogenesis to evolution, microarray expression analysis has identified gene candidates and signaling pathways for investigation. Thus, microarray analysis provides interesting leads in almost all fields of biology, offers a genome-wide glimpse into genetic "terra incognita," and challenges scientists to explore this unknown world.

Microarrays and Cancer

Some of the most dramatic breakthroughs have been in cancer diagnosis and pathology where microarrays are being used to identify and classify tumors based on their gene expression patterns. Golub *et al.* did a proof-of-principle study designed to distinguish acute myeloid leukemia from acute lymphoblastic leukemia using patterns of gene expression in patients' bone marrow samples [6]. The study showed that tumor gene profiling correctly identified the cancer type in 36 out of 38 patients, with the remaining two identified as "uncertain." These scientists suggest that a battery of tumor "class predictor" genes can be used for diagnostic confirmation or clarification of unusual cases. This point was dramatically illustrated by using the tumor predictors in an actual case in which a boy had classic symptoms of acute leukemia, but his tumor cells had atypical morphology for this disease. Microarray analysis of the boy's tumor cells suggested that he did not have leukemia because there was low expression of leukemia class predictor genes. Instead, the genes expressed suggested muscle cancer. The boy was eventually diagnosed with rhabdomyosarcoma and his treatment changed accordingly [6]. Because these two types of cancer have quite different treatment modalities, microarray analysis may have saved this boy's life.

Pat Brown, Ash Alizadeh, and David Botstein along with National Cancer Institute (NCI)[1] researcher Louis M. Stoudt, have used microarrays to characterize diffuse large B-cell lymphomas and divide them into at least two distinct diseases with significant differences in survival rates [7]. The subgrouping of tumors according to expression patterns has led to optimal treatment modalities with associated lifespan extension. An example of using microarray analyses for the hierarchical clustering of cancer cell types is shown in Fig. 1. The 60 tumor cell lines used by the NCI to screen anti-cancer compounds were classified solely by gene expression patterns. The results reveal a correlation between expression pattern and the cell type from which the tumors originated [8]. Microarray analysis of tumors can be expected to yield significant gains in the future, improving accuracy of disease diagnosis and ensuring the most effective treatment regimen is prescribed for each patient.

Microarrays and SARS

Microarray analysis attracted public attention recently when it was used to identify the virus that causes severe acute respiratory syndrome (SARS), a highly contagious disease that has become a worldwide health concern. In 2003, the U.S. Centers for Disease Control and Prevention released the latest test for the SARS virus. The test was developed in the DeRisi laboratory (University of California, San Francisco, CA), using a microarray device to quickly identify the virus. The SARS epidemic has highlighted the power of microarray technology for broad application in research and medical diagnostics.

A PRIMER OF MICROARRAY METHODOLOGY

The traditional method for quantification of gene expression uses a single, gene-specific DNA/RNA probe to screen RNA samples that have been immobilized on a nylon matrix, and is called an RNA or Northern blot [9]. Northern analysis identifies quantitative differences of expression between samples, but only for the gene selected. This method is cumbersome when large numbers of genes are being examined and would be impossible to use on a genome-wide scale. Microarray technology is based on a similar process of hybridizing complementary probe and target strands of nucleic acids. Gene chips are produced containing 30,000 or more spots on a slide ($\sim$100 μm spacing), each spot containing DNA oligonucleotides or cDNA clones specific for a known gene. These gene chips

[1] The abbreviations used are: NCI, National Cancer Institute; SARS, severe acute respiratory syndrome; NSF, National Science Foundation; GMO, genetically modified organism; SNP, single nucleotide polymorphism.

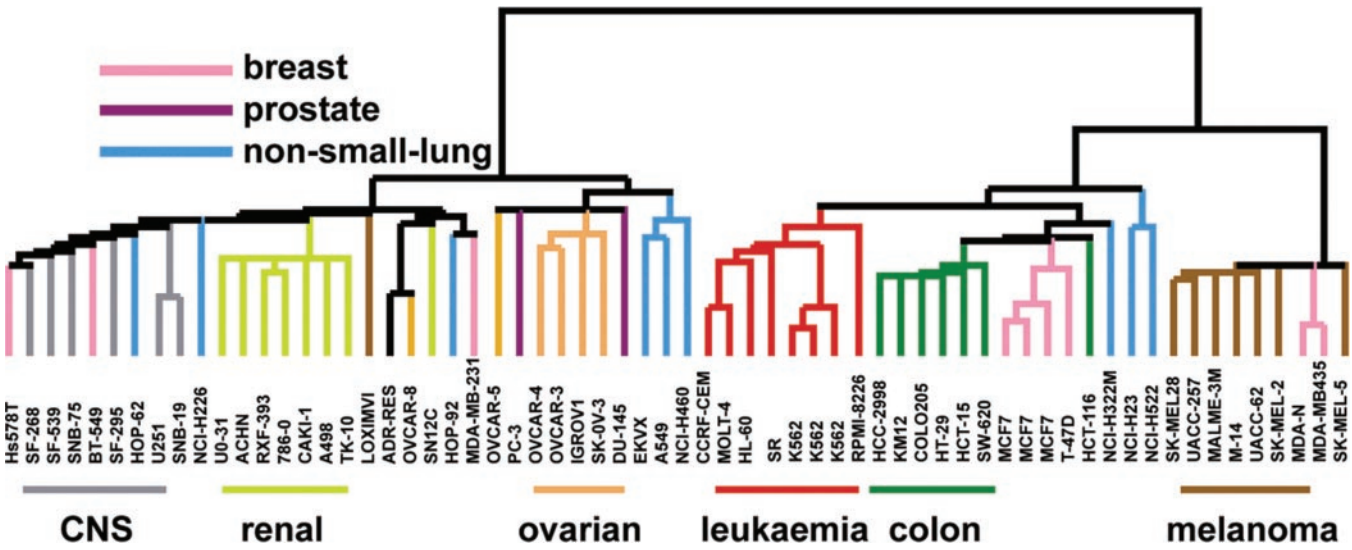

FIG. 1. **A dendrogram summarizing the hierarchical clustering of 60 cancer cell lines used by the NCI to screen cancer drugs.** Ross *et al.* [8] analyzed each cell line by microarray analysis and revealed that cell lines originating from similar cell types (breast, prostate, lung) were grouped together. Adapted from Ref. 16, reprinted by permission of Pearson Education.

are probed with fluorescently labeled mRNA or cDNA, and comparisons of gene induction and repression are made using alternate-colored labels for distinct RNA samples [2]. The hybridizations performed on a single gene chip are equivalent to performing tens of thousands of comparative Northern blots in 1 day (Fig. 2). Microarrays can simultaneously compare the expression of all known genes in each paired sample, offering a powerful tool for the analysis of gene expression patterning.

Microarray analysis consists of three major components: array fabrication, target preparation and hybridization, and data collection and analysis. Each of these components is described below, and some of the leading manufacturers of the necessary reagents are provided.

Array Fabrication

The basic iterative step in microarray production is performed by a robot and involves spotting a small volume of DNA solution from a microtiter plate onto a glass microscope slide coated with poly-L-lysine or aminosilane [2, 5, 10–12]. The spotted DNAs are typically oligonucleotides or PCR-amplified cDNA clones. This process is repeated until as many as 30,000 DNA spots are applied to precise locations on the slide. After printing, the DNA is covalently cross-linked to the glass slide. Arrays manufactured by Affymetrix (Santa Clara, CA) use photolithography and solid-phase chemistry to synthesize small oligomers on the surface of a glass slide. For each of the genes being examined, a complementary oligonucleotides is synthesized. In addition, corresponding sets of oligomers are synthesized with known mismatches for the gene sequence. These sets of oligomers offer a measure of binding specificity during hybridization. Unlike some of the array methodologies listed below, Affymetrix arrays compare control and experimental gene expression profiles using two separate arrays, which are scanned separately. The signal from a reference or control is then compared with that of an experimental sample using appropriate software.

The terminology of array fabrication has become a point of confusion in recent years. The synthesized oligonucleotides in Affymetrix microarrays are referred to as probes, and the labeled mRNA or cDNA that are hybridized to them are called targets. cDNA arrays are referenced differently, with immobilized cDNA targets, and labeled probes. Differences in nomenclature stem from Southern blotting terminology, where the probe is a known sequence that has been labeled and is hybridized to DNA fragments immobilized on a membrane. For the purpose of this review, and in order to incite the least amount of confusion in our readers, we will refer to the labeled, nonimmobilized mRNA or cDNA as a probe.

Probe Preparation and Hybridization

The first and most critical step in probe preparation is isolation of total or poly(A)$^+$ RNA from control and experimental sources. The purified RNA must always be visualized by denaturing gel electrophoresis to verify the integrity of the ribosomal RNA bands. If the RNA is degraded, it will not be useful for labeling. Using supplies and reagents that are certified RNase-free promotes successful RNA isolation. Ambion (Austin, TX) offers numerous products for working with RNA, from pipette tips to buffers and RNase inhibitors. Qiagen (Valencia, CA) offers several easy-to-use kits for isolating RNA from bacterial, yeast, plant, and animal cells. Other protocols for RNA isolation are found on Pat Brown's laboratory website (cmgm.stanford.edu/pbrown/mguide/index.html) and the protocols page for the Genome Consortium for Active Teaching (GCAT) (www.bio.davidson.edu/projects/GCAT/GCATprotocols.html).

Once extracted from the two populations, the RNA samples are typically labeled with fluorescent dyes in order to generate probes. The commercial cyanine dyes Cy3 and Cy5 are commonly used in labeling reactions. Other dyes, such as Alexa Fluor® 546 and Alexa Fluor® 647 (Molecular Probes, Eugene, OR), are becoming more popular because the cyanine dyes, especially Cy5, are unstable and

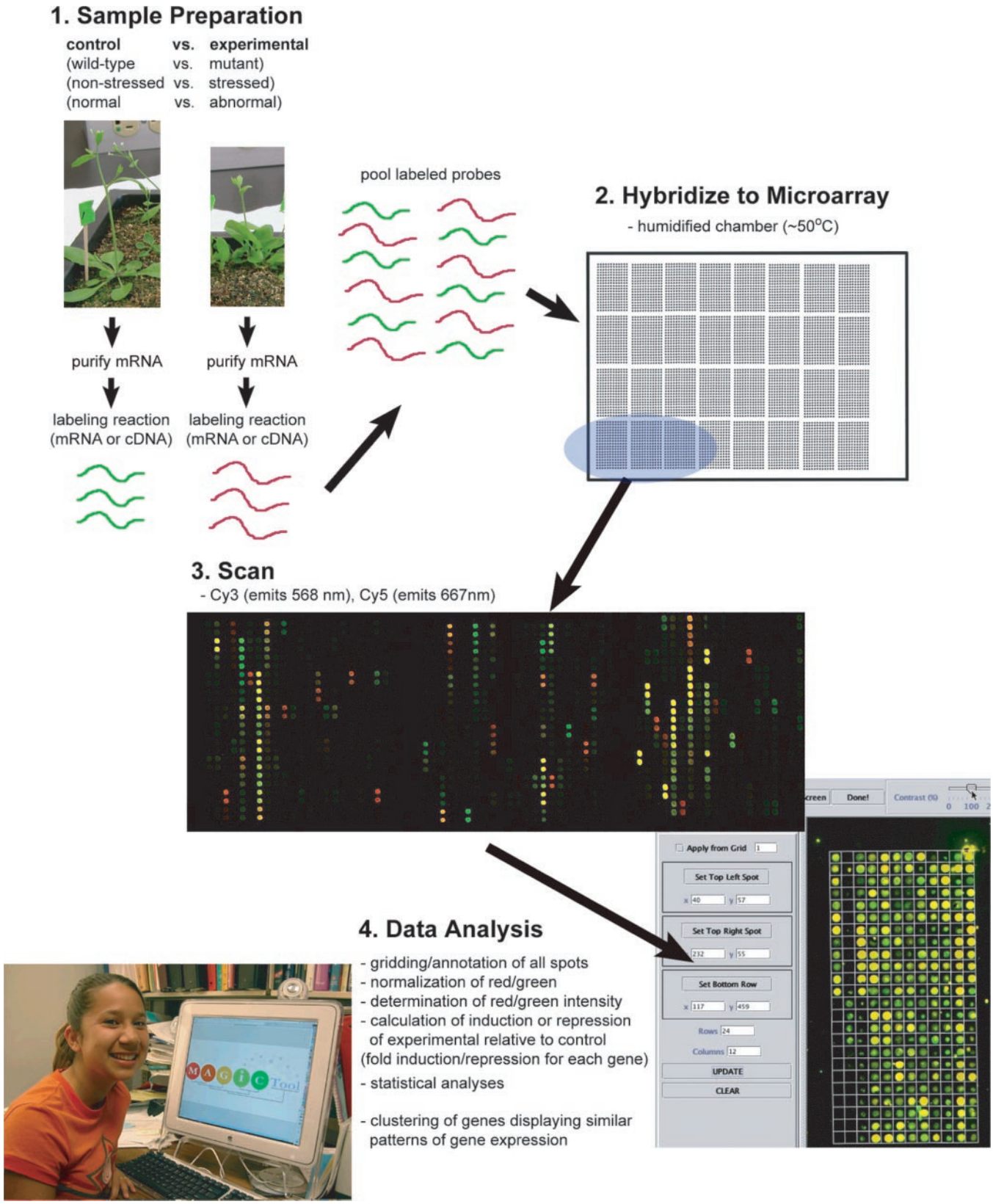

FIG. 2. **An overview of microarray analysis.** *1*, mRNA is purified from experimental and control samples, and fluorescent label is incorporated either during or after cDNA synthesis or through direct labeling of the mRNA to generate probe. Labeled probes from experimental and control are then pooled. *2*, The pool of labeled probes are hybridized to a microarray slide containing thousands of spotted oligonucleotides or cDNAs. *3*, Slides are scanned for fluorescence emissions at selected wavelengths, and a high-resolution image file is generated. *4*, Using microarray analysis software, the image file is annotated to identify each spot, and then fluorescence intensities are quantified and analyzed as described in the text. The *A. thaliana* photographs were contributed by Steve Davis (Pepperdine University, Malibu, CA). The microarray image was contributed by GCAT faculty member Laura Hoopes and undergraduate research student Allen Kuo (Pomona College, Claremont, CA). The data analysis images are from MAGIC, software generated by GCAT faculty member Laurie Heyer with a team of undergraduate students (Davidson College, Davidson, NC).

susceptible to degradation by ozone, light, and the lasers used in slide scanning. Genisphere Inc. (Hatfield, PA) has recently introduced a new stabilizer specifically to address Cy5 instability. Fluorescently labeled probes can be prepared by several different methods including direct or indirect cDNA labeling, cDNA labeling with fluorescent dendrimers, direct mRNA labeling, and direct or indirect labeling of amplified RNA [5, 10, 13–15]. RNA amplification may be the method of choice when isolating RNA from limiting samples, such as those from tissue biopsies. For use in microarray protocols requiring cDNA synthesis, SuperScript II reverse transcriptase from Invitrogen Life Technologies (Carlsbad, CA) is recommended because it generates high cDNA yields. An alternative enzyme is ImProm-II reverse transcriptase from Promega Corporation (Madison, WI). Promega offers a training support program for educators using molecular biology techniques, and many of their products are offered to educators at a discount (contact Diana Long; Fax: 608-277-2601).

In the direct cDNA labeling method, fluorescently modified deoxynucleotides are incorporated during the first-strand cDNA synthesis from an RNA template using reverse transcriptase [5, 10, 14]. Although this procedure is relatively straightforward, fluorescently modified nucleotides are bulky and incorporate less efficiently than unmodified nucleotides. In the indirect cDNA labeling method, aminoallyl-modified nucleotides are incorporated during the reverse transcription reaction, and fluorescent dyes are subsequently coupled to the reactive amino groups in the cDNA. Because the amine-modified nucleotides resemble unmodified nucleotides more than the fluorescently labeled nucleotides used for direct labeling, the reverse transcription reaction is more efficient. One disadvantage of the indirect labeling method is that the procedure takes more time to perform. A newer method uses fluorescent dendrimer complexes to label cDNA [13, 14]. After cDNA synthesis, a fluorescent dendrimer with hundreds of dye molecules per complex is hybridized to the cDNA. Genisphere offers discount pricing and other promotions of 3DNA (dendrimer) sample kits with Cy3/Cy5 or Alexa 546/647 dyes to GCAT members, and protocols using 3DNA products are on the GCAT website (www.bio.davidson.edu/projects/GCAT/GCATprotocols.html).

The labeled probes prepared from the two RNA sources are co-hybridized to the same DNA chip. The conditions during this step must be optimized to promote specific binding of labeled probe to its target and reduce background. Important parameters include hybridization temperature, length of hybridization, concentration of salts, pH of the solution, and the presence or absence of denaturants such as formaldehyde in the hybridization buffer. Chips are often prehybridized with a solution containing bovine serum albumin to block nonspecific binding of labeled probe to the surface. Hybridization and wash solutions must be evenly distributed over the chip to maximize interactions between probe and target sequences and minimize background fluorescence. During hybridization the chips are stored in a humidified, temperature-controlled, darkened environment. Small, affordable, aluminum chambers that house one or two chips work well, and can simply be placed in a standard incubator or water bath during the hybridization steps. These chambers are available from Monterey Industries (Richmond, CA).

Data Collection and Analysis

The hybridized array is typically scanned with a system that uses lasers as a source of excitation light and photomultiplier tubes as detectors [13]. This system is capable of differentiating the fluorescently labeled probes. Most commercially available array scanners scan sequentially, meaning the scanner acquires one image at a time and then builds the ratio image after acquiring images at both fluorescence excitation wavelengths. Other scanners use simultaneous dual laser scanning to acquire both images at the same time, reducing scan times and eliminating potential errors associated with aligning two separately generated images.

After scanning, a grid must be placed on the image and the spots representing the arrayed genes must be identified. The background fluorescence is calculated locally for each spot and is subtracted from the hybridization intensities. Differentially expressed genes are identified by comparing the fluorescence intensity of control and experimental probes hybridized to each spot [11–13, 16, 17]. Typically, the experimental target sequences are labeled with Cy5, which fluoresces red light (667 nm), and control targets are labeled with Cy3, which fluoresces green light (568 nm). The ratio of red to green signal can then be used as a measure of the effect of the experimental treatment on the expression of each gene. A ratio of 1 (yellow spot) indicates no change in the expression level between experimental and control samples, while a ratio greater than 1 (red spot) indicates increased transcription in the experimental sample, and a ratio less than 1 (green spot) indicates decreased transcription in the experimental sample. A scatter plot is a very useful representation of the expression data; the signal intensities of the experimental and control samples are plotted along the *x*- and *y*-axes, and the ratio values are plotted as a distance from the diagonal [13]. The diagonal separates spots with higher activity than the control sample from spots with lower activity than the control. The scatter plot provides a visualization of the fluorescence ratios obtained from the experimental and control samples. One can then easily choose points that represent a severalfold increase or decrease in gene expression and focus additional analyses on these genes.

With just one experimental condition and a control, the data analysis is limited to a list of regulated genes ranked by the fold-change or by the significance of the change determined in a *t* test. Normalization of data must be performed to compare separate arrays. With multiple experimental conditions (*e.g.* time-points or drug doses), the genes are often grouped into clusters that behave similarly under the different conditions. Complex computational methods such as hierarchical clustering or *k*-means are used to analyze the massive amounts of data generated by these experiments. Gene clusters are visualized with trees or color-coded matrices by placing genes with similar patterns of expression into a clustered group (Fig. 3). Image processing and analysis software is commercially available, and several packages are available as freeware

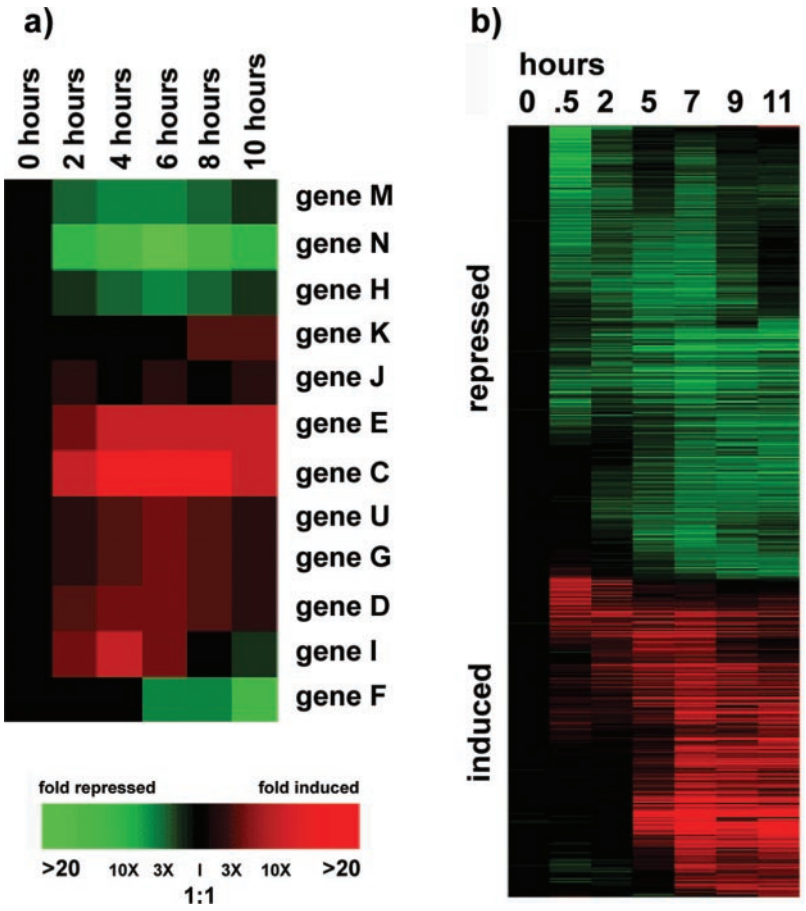

FIG. 3. **Clustering of gene expression patterns.** *a*, the ratio of gene expression in control relative to experimental for individual genes is displayed using a color scale. *Black* indicates no change in expression, while an increase in the experimental relative to the control is shown as *red*, and a decrease in the experimental relative to the control is shown as *green*. Genes displaying similar patterns of induction or repression are clustered together. *b*, clustering of thousands of genes by patterns of gene induction or repression following a treatment. Adapted from Ref. 16, reprinted by permission of Pearson Education.

(www.bio.davidson.edu/projects/GCAT/GCATprotocols.html, www.tigr.org/softlab/, and www.nhgri.nih.gov/DIR/LCG/15K/HTML/img_analysis.html). Laurie Heyer (Davidson College, Davidson, NC) and a group of undergraduate students have written MAGIC (Microarray Genome Imaging and Clustering) tool, a free program for microarray data analysis that is designed with the undergraduate student in mind (www.bio.davidson.edu/projects/magic/magic.html). MAGIC is an interface to the free microarray software developed by Michael Eisen (University of California, Berkeley, CA), including ScanAlyze, Cluster, and Treeview (rana.lbl.gov). MAGIC simplifies use of these programs and offers careful tutorials for each step of data analysis, from annotation and normalization to gene clustering.

Limitations of Expression Analysis and Confirmation of Results

Microarray analysis of gene expression does have limitations that researchers must consider. In gene expression, the correlation between induced mRNA and induced levels of protein are not always well aligned. Translational and post-translational regulatory mechanisms that impact the activity of various cellular proteins are not examined by DNA microarrays, though the emerging field of proteomics is beginning to address this issue. Other limitations of microarray analysis include the impact of alternative splicing during transcript processing and the limited detectability of unstable mRNAs.

Differential gene expression results must be confirmed through direct examination of selected genes. These analyses are typically at the level of RNA blot or quantitative RT-PCR to examine transcripts of a specific gene [9], and/or detection of protein concentration using immunoblots. Additional studies often include alteration of gene function with targeted mutations, antisense technology, or protein inhibition.

MICROARRAYS IN THE TEACHING LABORATORY: GCAT

In December 1998, Pat Brown presented his work on DNA microarrays at the annual meeting of the American Society for Cell Biology. In the audience were A. Malcolm Campbell of Davidson College and Mary Lee Ledbetter of the College of the Holy Cross, who were inspired by the power and simplicity of this technology. They began to develop the concept of a national effort to include microarrays in the undergraduate curriculum. The following year, they founded the Genome Consortium for Active Teaching (GCAT) as a nonprofit educational consortium to bring functional genomics methods into undergraduate courses and independent student research. Consortium members work to make microarray experiments affordable through cost sharing, to provide a clearinghouse of information, raw data, and analyzed results for use in teaching genomics, and to develop a network of teachers using functional genomics. According to GCAT founder Malcolm Campbell, "Biology is being transformed with genomics research, and we need to join the party."

GCAT faculty recognized the high costs of faculty effort, institutional commitments, and funding support needed to

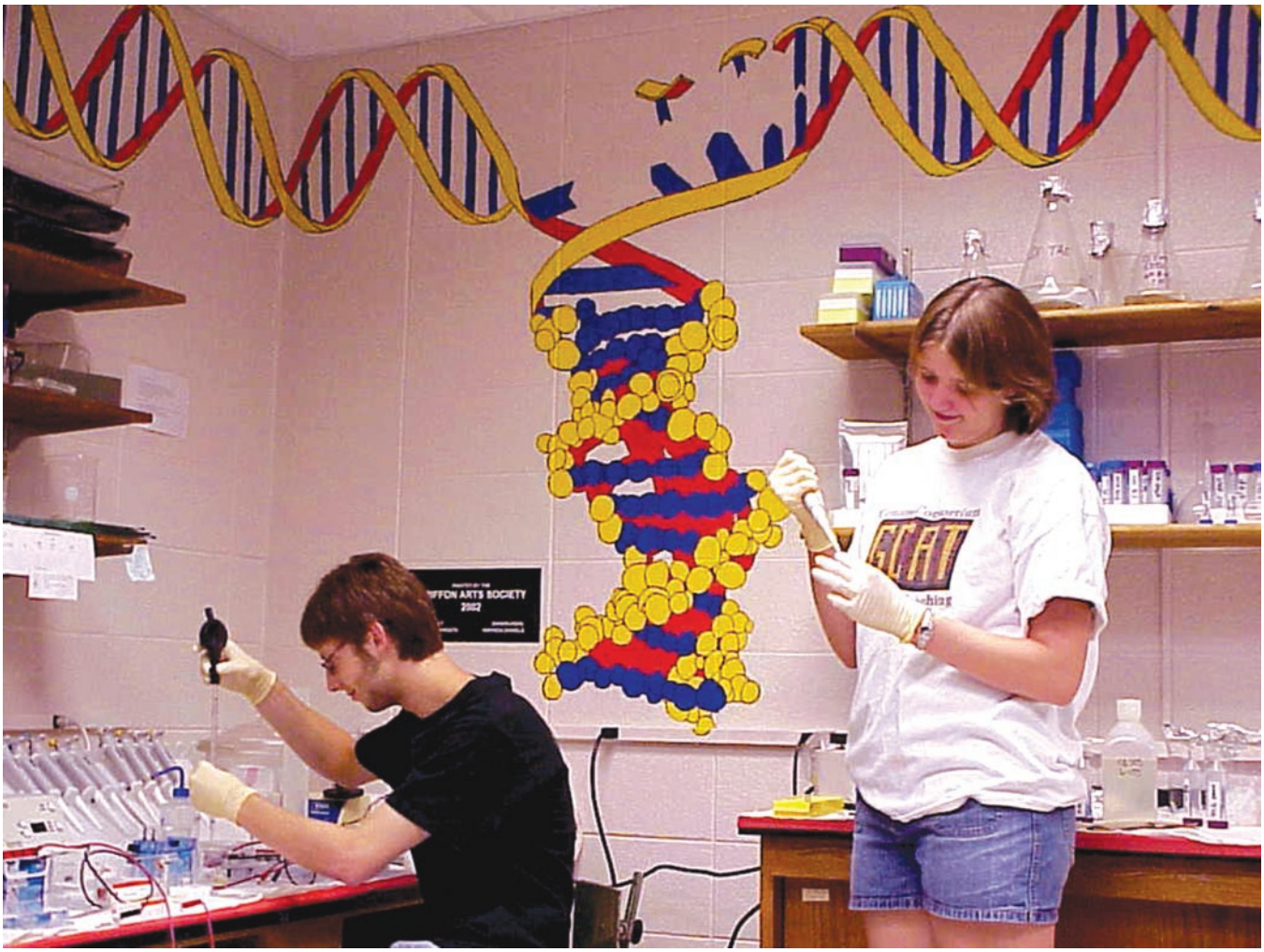

FIG. 4. **Student researchers Bart Phillips and Sara Freel perform microarray analysis on yeast cells treated with anti-cancer drugs.** These students are from Missouri Western State College (Saint Joseph, MO) and are mentored by GCAT faculty member Todd Eckdahl.

incorporate new genomics-based research strategies into the classroom. But they also shared the conviction that incorporation of arrays into the undergraduate curriculum would be valuable for the following reasons. First, microarray projects represent an excellent example of the use of molecular biology and strengthen students' conceptual understanding and problem-solving skills associated with the molecular biology laboratory. Second, proper experimental designs demand advanced thinking and planning if one is to generate reliable and reproducible data. Microarray experiments teach students to critically assess their experiments for reliability, reproducibility, and the inclusion of proper controls. Third, data analysis allows students to experience first-hand the data-rich environment of the genomics era through the use of data sets that are freely available in the public domain. Students analyze data sets that might include the entire yeast (or another organism's) genome and use microarray analysis software to cluster groups of genes displaying similar patterns of induction/repression. This intensive introduction to genomics cannot be replaced by any amount of lecturing or demonstration. GCAT faculty are committed to providing their undergraduate students opportunities to engage in modern genomics in a meaningful way, and although the inclusion of microarrays in the curriculum can be challenging to students and faculty alike, consortium members enthusiastically believe the program is working.

By the fall semester of 2000, GCAT consisted of 23 faculty from the United States and Canada engaging their students in functional genomics investigations (Fig. 4) and an additional 50 people on the GCAT listserv (www.bio.davidson.edu/Biology/GCAT/GCAT-L.html). Pat Brown graciously donated 135 microarray slides printed with the complete set of about 6,400 yeast open reading frames for use by GCAT members. Richard Bookman of the University of Miami also donated yeast mini chips with 96 yeast genes spotted 10 times on each microarray slide for student practice. Genisphere, Inc. provided mini labeling kits using their patented 3DNA dendrimer technology at a reduced price for GCAT members. Lee Hood provided access to the Institute for Systems Biology scanners for GCAT members, and student microarray results were posted on the Stanford Microarray Database. GCAT was off to an exciting start, but integration of microarrays into the undergraduate curriculum proved to be challenging for GCAT faculty.

Microarrays for Undergraduates: Removing the Obstacles

The key challenges to incorporating array analysis in the undergraduate classroom related to faculty time and training, internal support, access to a slide scanner, and the identification of user-friendly array analysis software. Many faculty have little experience with the relatively new array-based technologies and are hesitant to retrain themselves. In addition, there are financial limitations. An array laser scanner can cost up to $100,000, and each array slide can cost $200-$450. From its inception, GCAT has attempted to address these challenges.

In July 2001, GCAT received a Research in Undergraduate Institutions (RUI) grant from the National Science Foundation for the purchase of an array scanner to be based at Davidson College. The grant was authored by collaborative group of GCAT faculty including Malcolm Campbell and Laurie Heyer at Davidson College, Laura Hoopes (Pomona College, Pomona, CA), and Todd Eckdahl (Missouri Western State College, Saint Joseph, MO). The grant provides support for undergraduate research and teaching activities of GCAT faculty, who now have direct access to the DNA microarray reader at Davidson College. Consortium members can send their hybridized microarrays to be scanned and can retrieve their image files by ftp from the Institute for Systems Biology server for data analysis.

GCAT has experienced significant growth since its inception. In 2002–03, 40 GCAT members worked with 396 chips from *Escherichia coli*, yeast, human, *Arabidopsis*, and mouse. Chips donated by major research laboratories included yeast chips from Lee Hood (Institute for Systems Biology, Seattle, WA), *E. coli* chips from Frederick Blattner (Univ. of Wisconsin), *Arabidopsis* chips from Ellen Wisman (Michigan State University), and human chips from Richard Bookman (Univ. of Miami). Silicon Genetics (Redwood City, CA) provided free access to their GeneSpring software package for microarray data analysis to GCAT members. Faculty within the program are beginning to publish teaching modules and research results generated by undergraduates using GCAT arrays and equipment [18].[2] GCAT faculty at several institutions engage undergraduate students in microarray-based research. For example, Liz Vallen and her students at Swarthmore College are studying DNA replication mutants in yeast, while Dennis Revie at California Lutheran University guides students in an investigation of the effects of hepatitis C virus on cultured human cells. At Mount Saint Mary's College, Myra Derbyshire and her undergraduates are studying genes that modify yeast chromatin structure. Also, GCAT members Laura Hoopes of Pomona College and Todd Eckdahl of Missouri Western State College have been awarded National Institutes of Health AREA grants for microarray-based undergraduate research projects.

In addition to the distribution of microarray slides and the use of shared equipment, the consortium has provided a supportive network for faculty who want to take on the challenge of learning this new technology. In 2002, members of GCAT and faculty from around the world gathered to discuss educational genomics at the American Society for Microbiology conference. Presentations by GCAT faculty heralded the success of the approach and described goals for continued improvement of the program (www.bio.davidson.edu/people/macampbell/ASM/ASM.html) [19]. After the symposium, many faculty members requested workshops for learning how to use array technology in teaching and doing research with undergraduates. A clear need for such training was recognized, and a second GCAT grant submission to the National Science Foundation (NSF) was developed.

GCAT received a NSF award for a workshop, which was held in summer 2003, at the Institute for Systems Biology in Seattle. A group of GCAT faculty met with leading microarray researchers and developers to examine new techniques in cDNA labeling and data analysis and to redesign GCAT assessment efforts. Results from the workshop have been posted at the GCAT website (www.bio.davidson.edu/Biology/GCAT/workshop.html). The group received NSF support for a second workshop, held at Georgetown University in July, 2004. The workshop was intended for new users and provided background information on microarrays, hands-on experience with microarray data analysis, and microarray hybridization procedures. A major outcome of the project is the establishment of a group of knowledgeable and confident undergraduate faculty. This core group will serve as a valuable resource for many other faculty who wish to incorporate microarray technology into their undergraduate teaching and research activity. A second major outcome of the workshop is a downloadable laboratory module for using yeast microarrays. Interested faculty should check the GCAT website for information.

Participation in GCAT

In order to become a GCAT member, a faculty member must agree to the following terms: 1) All work must be performed by faculty and their undergraduate students. 2) All data obtained will be public domain. 3) Faculty must arrange for payment for chips, currently at the modest rate of no more than $50 for the first chip and $20 for each additional chip. Faculty are also responsible for any additional costs associated with making probes, growing cells, and sending the chips for scanning. 4) Faculty must be willing to help other faculty by answering questions that come from the GCAT-L listserv. 5) Faculty must be willing to take a risk and try something very new, knowing that it may not work out the first time. 6) Faculty must be willing to participate in the assessment component of GCAT, coordinated by Mark Salata (www.bio.davidson.edu/Biology/GCAT/assessment/assess.html). A preliminary GCAT assessment report is available online (www.bio.davidson.edu/Biology/GCAT/assessment/01_02/assess01_02.html).

A FUTURE FOR MICROARRAYS

Genome sequencing has fundamentally changed the biological sciences, including basic and applied research,

[2] J. J. Campanella, C. Du, Q. Vega, O. Gomes, W. Graff (2003) Microarray analysis in Brassicaceae: Hybridization of *Arabidopsis thaliana* cDNA arrays with cDNA from a radiation-induced plant tumor and normal plant tissues, submitted for publication.

medicine, and the pharmaceutical industries. Undergraduates training for any career in the biological sciences will soon require at least an introduction to bioinformatics if they are to compete for jobs in these exciting career fields. A recent report by the National Research Council stressed the need for improved integration of mathematics and computer science in undergraduate biology courses, citing the impact of bioinformatics on the biological sciences [20]. Here we highlight the role of bioinformatics within applied biological research, focusing on the work of a leading biotechnology company, and upon a new application of microarrays in medicine, genotype determination.

Microarrays in Biotechnology

Ceres Inc. (Malibu, CA) is a biotechnology company that uses plant genomics and bioinformatics to develop improved crop plants through genetic engineering. Their focus is upon *Arabidopsis thaliana,* which offers a short life cycle, the relatively simple generation of transgenic plants, and a sequenced genome [21]. Information gained from their studies is applied to crop plants through collaborative efforts with Monsanto (St. Louis, MO). The scientists at Ceres generate full-length cDNA libraries, characterizing each gene sequence represented in the library through nucleotide sequencing, database searches, protein product analysis, and expression analysis using microarray analysis. They generate transgenic plants that express each of these cDNAs at high levels and analyze these plants through a comprehensive phenotypic screening program. The whole process is designed to enable a very large number of genes to be evaluated simultaneously. This type of multifaceted characterization of genes generates a database of integrated information about each gene [22]. Researchers then utilize this information to design gene modifications that may offer improved crop performance in areas that include stress-resistance (drought, heat, cold), resistance to insects/viruses, or even plant structure (seed/leaf size, fruit characteristics).

Microarrays are a fundamental platform for transcript profiling, offering quantitative expression data for the transcriptome in each tissue, stage of development, and under each condition. Ceres employs expression analysis in the following ways.

Gene Annotation—Using microarrays and expressed sequence tags (collections of short cDNA fragments, known as ESTs), researchers scan patterns of gene expression in different tissues and under varied conditions to identify clusters of genes displaying similar patterns and possibly sharing functional relevance for the organism [23]. This is especially helpful when some members of the annotated group have been well characterized. Literature searches yield information regarding the role these genes and their protein products play in cell biology, and that information can then be utilized to hypothesize the function of poorly characterized genes placed in the same expression group. For Ceres, this process aids in the identification of genes that might be manipulated to impact specific plant traits.

Identifying Promoter Elements—The careful control of transgene expression (when, where, and how much) is a vital consideration for the production of genetically modified organisms (GMOs). Carefully annotated promoters offer pinpoint precision in control of transgene expression. Microarray analyses offer quantitative data on the expression of every known gene in the organism being studied. Researchers can select genes displaying the desired pattern of expression and use the *A. thaliana* genome information to extract the DNA sequence of the promoter sequence for that gene [24, 25]. That promoter can then be used for the generation of transgenic plants, offering rigid control of the transgene [26]. By comparing the promoters of genes displaying a similar pattern of expression, scientists can identify shared regulatory sequences and use them to custom design a promoter for the chosen application. The information available from a reliable database of gene expression patterns from varied tissues and conditions offers a valuable tool for promoter discovery.

Characterization of GMOs—The addition of a transgene to a complex, multicellular organism can have unanticipated secondary influences upon phenotype, and a critical concern in GMO production is to assess the transgenic organism for any of these changes. Microarray analysis of transgenic plants is a very sensitive method of screening for any alterations in plant biology. Similar applications of microarrays are being used in the pharmaceutical industry to determine the influence of a particular drug on cellular function, using the microarray as a critical component in assessing drug toxicity or secondary effect upon cellular function [27]. More traditional analyses focus upon cell survival, or specific physiological characteristics, but these are limited in their ability to detect subtle influences of a drug. Thus, the microarray can be used to compare treated cells (genetically altered, drug treated) with untreated cells and to quantify differences effectively.

Genotype Analysis—Detection of SNPs

There are a growing number of applications for chip technology beyond gene expression analysis. One highly anticipated application has been the rapid determination of genotype using oligonucleotides arrays. Individuals in any population display differences in phenotype (traits), and currently it is very difficult to identify the specific genetic makeup (or genotype) that determines any given phenotype. Ultimately scientists need to follow the segregation of each gene as it passed from one generation to the next, and establish a correlation between traits and the alleles of every gene. Traditional strategies for genotype determination have been laborious and limited, scanning hundreds or a few thousand genetics markers to crudely examine the genotype of each individual at relatively low resolution. The markers being used in these newer genotyping strategies are at the level of single nucleotide polymorphisms (SNPs), which occur at high frequency in the genome, about every 1,000 base pairs [28]. If all SNPs for each individual in a pedigree could be determined, researchers could follow genetic information at high resolution as it is passed from generation to generation. But determining over a million SNPs for each sample is a daunting task. To offer high-throughput determination of SNPs, oligonucleotide microarrays have been developed

for the rapid and accurate analysis of genotype [29–32]. Perlegen Sciences (Mountain View, CA) and Affymetrix have collaborated to develop microarrays for the detection of SNPs in humans. Introduced in 2001, the first generation of these tests (GeneChip® HuSNP) examines 1,500 SNPs for each DNA sample. Using a manufacturing process that has been adapted from the semiconductor industry, Perlegen Sciences is now developing a protocol that will utilize tens of millions of probes on a glass wafer to characterize ~1.5 million SNPs for each individual sample (www.perlegen.com). Rapid SNP-based genotyping will offer applications in preventive medicine, diagnosis of disease, characterization of complex traits, forensic science, and even the development of effective pharmaceutical treatment of disease [33–35].

CONCLUSIONS

It is clear that the development of array-based technologies represents a fundamental shift in the way scientists study living organisms. More that just a new experimental technique, microarrays employ the massive data generated by genome sequencing and use that data to comparatively examine genome-wide expression patterns. The data sets produced offer a high-resolution map of the genes being regulated to mediate cellular differentiation, adaptation, division, and evolution. This technology is a first glimpse into how researchers will utilize genome data in coming years. Educators are beginning to teach genomics, proteomics, and bioinformatics in undergraduate biology and computer science courses and a few, such as those in the GCAT group, are beginning to include these technologies in the teaching laboratory. Far from being just a lesson in new technology, these courses are introducing the new age of biology to the next generation of scientists. It is an exciting, data-rich age in which the entire landscape of experimental possibility has changed. Every graduate pursuing a career in the life sciences will be impacted by these changes, and it is critical that educators begin to develop curriculum accordingly. Biology is indeed being transformed by genomic research, and undergraduate faculty and students need to join the party.

Acknowledgments—We thank Ken Feldman, Richard Flavell and Yiwen Fang at Ceres, Inc. for generously offering their time, expertise, and scientific vision. We thank Doug Swartzendruber, Tom Vandergon, and Lee Kats for their careful proofreading of the manuscript.

REFERENCES

[1] E. M. Southern (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis, *J. Mol. Biol.* **98**, 503–517.
[2] M. Schena, D. Shalon, R. W. Davis, P. O. Brown (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science* **270**, 467–470.
[3] D. A. Lashkari, J. L. DeRisi, J. H. McCusker, A. F. Namath, C. Gentile, S. Y. Hwang, P. O. Brown, R. W. Davis (1997) Yeast microarrays for genome wide parallel genetic and gene expression analysis, *Proc. Natl. Acad. Sci. U. S. A.* **94**, 13057–13062.
[4] J. L. DeRisi, V. R. Iyer, P. O. Brown (1997) Exploring metabolic and genetic control of gene expression on a genomic scale, *Science* **278**, 680–686.
[5] The Mguide. Version 2.0 (1999) The Brown Lab's complete guide to microarraying for the molecular biologist: cmgm.stanford.edu/pbrown/mguide/index.html.
[6] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, E. S. Lander (1999) Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science* **286**, 531–537.
[7] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Bostein, P. O. Brown, L. M. Stoudt (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature* **403**, 503–11.
[8] D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M Van de Rijn, M. Waltham, A. Pergamenschikov, J. C. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein, P. O. Brown (2000) Systematic variation in gene expression patterns in human cancer cell lines, *Nat. Genet.* **24**, 227–238.
[9] J. C. Alwine, D. J. Kemp, G. R. Stark (1977) Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes, *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5350–5354.
[10] P. Hegde, R. Qi, K. Abernathy, C. Gay, S. Dharap, R. Gaspard, J. E. Hughes, E. Snesrud, N. Lee, J. Quackenbush (2000) A concise guide to microarray analysis, *BioTechniques* **29**, 548–562.
[11] W. M. Freeman, D. J. Robertson, K. E. Vrana (2000) Fundamentals of DNA hybridization arrays for gene expression analysis, *BioTechniques* **29**, 1042–1055.
[12] D. L. Bowtell (1999) Options available—From start to finish—for obtaining expression data by microarray, *Nat. Genet. Suppl. 21*, 25–32.
[13] M. Schena (2003) *Microarray Analysis*, Wiley-Liss, Hoboken, NJ.
[14] A. Richter, C. Schwager, S. Hentze, W. Ansorge, M. W. Hentze, M. Muckenthaler (2002) Comparison of fluorescent tag DNA labeling methods used for expression analysis by DNA microarrays, *BioTechniques* **33**, 620–630.
[15] R. N. Van Gelder, M. E. von Zastrow, A. Yool, W. C. Dement, J. D. Barchas, J. H. Eberwine (1990) Amplified RNA synthesized from limited quantities of heterogenous cDNA, *Proc. Natl. Acad. Sci. U. S. A.* **87**, 1663–1667.
[16] A. M. Campbell, L. J. Heyer (2003) *Discovering Genomics, Proteomics & Bioinformatics*, CSHL Press and Benjamin Cummings, San Francisco, CA.
[17] S. Knudsen (2002) *A Biologist's Guide to Analysis of DNA Microarray Data*, Wiley-Liss, New York.
[18] D. Wallack (2001) Genomics and microarrays in an undergraduate research class. *Cur. Q.* **21**, 126–129.
[19] A. M. Campbell (2002) Meeting report: Genomics in the undergraduate curriculum—Rocket science or basic science? *Cell Biol. Educ.* **1**, 70–72.
[20] National Research Council (2003) *BIOL2010: Transforming Undergraduate Education for Future Research Biologists*, The National Academies Press, Washington, DC.
[21] The Arabidopsis Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*, *Nature* **408**, 796–815.
[22] J. Donson, Y. Fang, G. Espiritu-Santo, X. Weimei, A. Salazar, S. Miyaomoto, V. Armendarez, W. Volkmuth (2002) Comprehensive gene expression analysis by transcript profiling, *Plant Mol. Biol.* **48**, 75–79.
[23] W. Zhu, S. S. Schlueter, V. Brendel (2003) Refined annotation of the *Arabidopsis* genome by complete expressed sequence tag mapping, *Plant Physiol.* **132**, 469–484.
[24] K. Maleck, A. Levine, T. Eulgem, A. Morgan, J. Schmid, K. A. Lawton, J. L. Dangl, R. A. Dietrich (2000) The transcriptome of *Arabidopsis thaliana* during systemic acquired resistance, *Nat. Genet.* **26**, 403–410.
[25] R. V. Davuluri, I. Grosse, M. Q. Zhang (2001) Computational identification of promoters and first exons in the human genome, *Nat. Genet.* **29**, 412–417.
[26] B. A. Krizek, V. Prost, R. M. Joshi, T. Stoming, T. C. Glenn (2003) Developing transgenic *Arabidopsis* plants to be metal-specific bioindicators, *Environ. Toxicol. Chem.* **22**, 175–181.
[27] F. Boess, M. Kamber, S. Romer, R. Gasser, D. Muller, S. Albertini, L. Suter (2003) Gene expression in two hepatic cell lines, cultured primary hepatocytes, and liver slices compared to the in vivo liver gene expression in rats: Possible implications for toxicogenomics use of in vitro systems, *Toxicol. Sci.* **73**, 386–402.
[28] B. A. Salisbury, M. Pungliya, J. Y. Choi, R. Jiang, X. J. Sun, J. C.

Stephens (2003) SNP and haplotype variation in the human genome, *Mut. Res./Fund. Mol. Mech. Mutagen.* **526**, 53–61.

[29] J. G. Hacia, J. B. Fan, O. Ryder, L. Jin, K. Edgemon, G. Ghandour, R. A. Mayer, B. Sun, L. Hsie, C. M. Robbins, L. C. Brody, D Wang, E. S. Lander, R. Lipshutz, S. P. Fodor, F. S. Collins (1999) Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays, *Nat. Genet.* **22**, 164–167.

[30] N. Patil, A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi, C. R. Hacker, C. R. Kautzer, D. H. Lee, C. Marjoribanks, D. P. McDonough, B. T. Nguyen, M. C. Norris, J. B. Sheehan, N. Shen, D. Stern, R. P. Stokowski, D. J. Thomas, M. O. Trulson, K. R. Vyas, K. A. Frazer, S. P. Fodor, S. P. Cox (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21, *Science* **294**, 1669–1670.

[31] D. J. Cutler, M. E. Zwick, M. M. Carrasquillo, C. T. Yohn, K. P. Tobin, C. Kashuk, D. J. Mathews, N. A. Shah, E. E. Eichler, J. A. Warrington, A. Chakravarti (2001) High-throughput variation detection and genotyping using microarrays, *Genome Res.* **11**, 1913–1925.

[32] K. Lindroos, S. Sigurdsson, K. Johansson, L. Ronnblom, A. Syvanen (2002) Multiplex SNP genotyping in pooled DNA samples by a four-color microarray system, *Nucleic Acids Res.* **30**, e70.

[33] C. Debouck, P. N. Goodfellow (1999) DNA microarrays in drug discovery and development, *Nat. Genet.* **21**, 48–50.

[34] H. Primdahl, F. P. Wikman, H. von der Maase, X. G. Zhou, H. Wolf, T. F. Orntoft (2002) Allelic imbalances in human bladder cancer: genome-wide detection with high-density single-nucleotide polymorphism arrays, *J. Natl. Cancer Inst.* **94**, 216–223.

[35] M. O. Hoque, C. C. Lee, P. Cairns, M. Schoenberg, D. Sidransky (2003) Genome-wide genetic characterization of bladder cancer: A comparison of high-density single-nucleotide polymorphism arrays and PCR-based microsatellite analysis, *Cancer Res.* **63**, 2216–2222.