Statistical Science
2003, Vol. 18, No. 1, 33–70

Statistical Challenges in Functional Genomics

Paola Sebastiani, Emanuela Gussoni, Isaac S. Kohane and Marco F. Ramoni

Abstract. On February 12, 2001 the Human Genome Project announced the completion of a draft physical map of the human genome—the genetic blueprint for a human being. Now the challenge is to annotate this map by understanding the functions of genes and their interplay with proteins and the environment to create complex, dynamic living systems. This is the goal of *functional genomics*. Recent technological advances enable biomedical investigators to observe the genome of entire organisms in action by simultaneously measuring the level of activation of thousands of genes under the same experimental conditions. This technology, known as *microarrays*, today provides unparalleled discovery opportunities and is reshaping biomedical sciences. One of the main aspects of this revolution is the introduction of computationally intensive data analysis methods in biomedical research. This article reviews the foundations of this technology and describes the statistical challenges posed by the analysis of microarray data.

Key words and phrases: Bioinformatics, classification, clustering, differential analysis, gene expression, functional genomics, microarray.

CONTENTS

Paola Sebastiani is Assistant Professor, Department of Mathematics and Statistics, University of Massachusetts, Amherst, Massachusetts 01003 (e-mail: sebas@math.umass.edu). Emanuela Gussoni is Assistant Professor of Pediatrics, Department of Medicine, Children's Hospital, Harvard Medical School, Boston, Massachusetts 02115 (e-mail: gussoni@rascal.med.harvard.edu). Isaac S. Kohane is Associate Professor of Pediatrics and Director, Children's Hospital Informatics Program, Harvard Medical School, Boston, Massachusetts 02115 (e-mail: isaac_kohane@harvard.edu). Marco F. Ramoni is Assistant Professor of Pediatrics, Children's Hospital Informatics Program, Harvard Medical School, Boston, Massachusetts 02115 (e-mail: marco_ramoni@harvard.edu).

1. THE HUMAN GENOME PROJECT

The Human Genome Project (HGP) is a multiyear effort, coordinated by the Department of Energy and

the National Institutes of Health, to create a reference sequence of the entire DNA and to identify the estimated 30,000–40,000 genes of the human genome. Officially started in 1990, the HGP is expected to render its final results in 2005, but the staggering technological advances of the past few years will probably allow completion of the project by April 2003. By then, the total cost of the project will be in excess of $3 billion, making the HGP one of the most funded single scientific endeavors in history, in the same league as the Manhattan Project and the Apollo Space Program. The rationale behind such a herculean effort is that a panoramic view of the human genome will dramatically accelerate advances in biomedical sciences and develop new ways to treat, cure or even prevent the thousands of diseases that afflict humankind. The HGP is also delivering a wealth of commercial opportunities: sales of DNA-based products and technologies are projected to exceed $45 billion by 2009 in the United States alone.

In June 2000, Craig Venter of Celera Genomics, U.S. President Clinton and the leaders of HGP consortium announced the completion of a "working draft" DNA sequence of the human genome, the details of which were published in February 2001 in dedicated issues of *Nature* and *Science* (volume 409 of *Nature*, published February 15, 2001 and available at http://www.nature.com/genomics/human/, reports the findings of the publicly sponsored HGP, while volume 291 of *Science*, published February 16, 2001 and available at http://www.sciencemag.org/content/vol291/issue5507, focuses on the findings of the draft sequence reported by the privately funded company Celera Genomics). The result of these efforts is a map of the human genes. This map consists of about 30,000–40,000 protein-coding genes (International Human Genome Sequencing Consortium, 2001), only twice the number of protein-coding genes in a worm or a fly. Because about 50% of these discovered genes have known functions, the challenge now is to annotate this map by understanding the functions of genes and their interplay with proteins and the environment to create complex, dynamic living systems. This is the goal of *functional genomics*.

Several projects around the world are currently under way to discover gene functions and to characterize the regulatory mechanisms of gene activation. One avenue of research focuses on gene expression level and exploits the recent technology of microarrays (Duggan et al., 1999; Lipshutz, Fodor, Gingeras and Lockhart, 1999; Lockhart and Winzeler, 2000; Lockhart et al., 1996) to obtain a panoramic view of the activity of the genome of entire organisms. Microarray technology is reshaping traditional molecular biology by shifting its paradigm from a hypothesis driven to a data driven approach (Lander, 1999). Traditional methods in molecular biology generally work on a "one gene in one experiment" basis, making the whole picture of gene functions hard to obtain. Microarray technology makes it possible to simultaneously observe thousands of genes in action and to dissect the functions, the regulatory mechanisms and the interaction pathways of an entire genome.

A fundamental component of functional genomics is the development of computational methods able to integrate and understand the data generated by microarray experiments. Typical experimental questions investigated with microarray experiments are the detection of genes differentially expressed in an abnormal/tumor cell compared to a normal cell, the identification of groups of genes that characterize a particular class of tumors and the recognition, at the molecular level, of novel subclasses of tumors and the detection of gene regulatory mechanisms. Although the avalanche of genome data produced with microarrays grows daily, no consensus exists about the best quantitative methods to analyze them. Many methods lack appropriate measures of uncertainty, make dubious distributional assumptions and are hardly portable across experimental platforms. Furthermore, little is known about how to design informative experiments, how to assess whether an experiment has been successful, how to measure the quality of information conveyed by an experiment and, therefore, the reliability of the results obtained. The specific character of gene expression data opens unique statistical problems.

The aim of this article is to offer an overview of these problems and the main approaches proposed to tackle them. To make the article self-contained, the next section will review essential biological notions and we refer to Griffiths et al. (2000) for more technical details. Section 3 describes the two most used microarray platforms: cDNA and synthetic oligonucleotide microarrays. Experimental design issues are described in Section 4, and Section 5 focuses on data quality issues. Section 6 describes techniques used for the analysis of gene expression data measured in comparative experiments, while Section 7 focuses on the supervised and unsupervised methods used to analyze gene expression data from experiments that compare several conditions. Section 8 lists some of the critical open problems and the challenges they pose to the statistical community.

2. THE BIOLOGY OF GENE EXPRESSION

Cells are the fundamental working units of every living system. The nucleus of each cell contains the chromosomes that carry the instructions needed to direct the cell activities in the production of proteins via the DNA (deoxyribonucleic acid). The structural arrangement of DNA looks like a ladder twisted into a helix, where the sides of the ladder are formed by molecules of sugar and phosphate, and the rungs consist of pairs of nucleotide bases A (adenine), T (thymine), C (cytosine) and G (guanine) joined by hydrogen bonds. In base pairing, A pairs with T and G pairs with C. Each strand of the double helix consists of a sequence of nucleotides that is made of one of the four bases A, T, G or C, a molecule of sugar and one molecule of phosphate. The particular order of the bases arranged along the sugar–phosphate backbone is called the DNA sequence. The *genome* is an organism's complete DNA and encodes the *genetic code* required to create a particular organism with its own unique traits. The nucleotide bases A, T, C and G are the letters that spell out these genetic instructions by producing a three-letter word code, where each specific sequence of three DNA bases (codons) encodes an amino acid. Amino acids are the basic units of proteins, which perform most life functions.

With few exceptions, all human cells contain the same DNA, but despite carrying the same set of instructions, cells are actually different. These differences are due to the fact that, stimulated by cell regulatory mechanisms or environmental factors, segments of DNA express the genetic code and provide instructions to the cells on when and in what quantity to produce specific proteins. These segments of DNA are the *genes* and the process by which they become active is called their *expression*.

The modern concept of gene expression dates back to 1961, when the theory of genetic regulation of protein synthesis was first described by Jacob and Monod (1961). The fundamental discovery was that differential gene expression, that is when and in what quantities a gene is expressed, determines differential protein abundance, thus inducing different cell functions. The *gene expression level* is an integer valued or continuous measure that provides a quantitative description of the gene expression by measuring the number of intermediary molecules produced during this process. These molecules are the mRNA (messenger ribonucleic acid) and the tRNA (transfer ribonucleic acid),

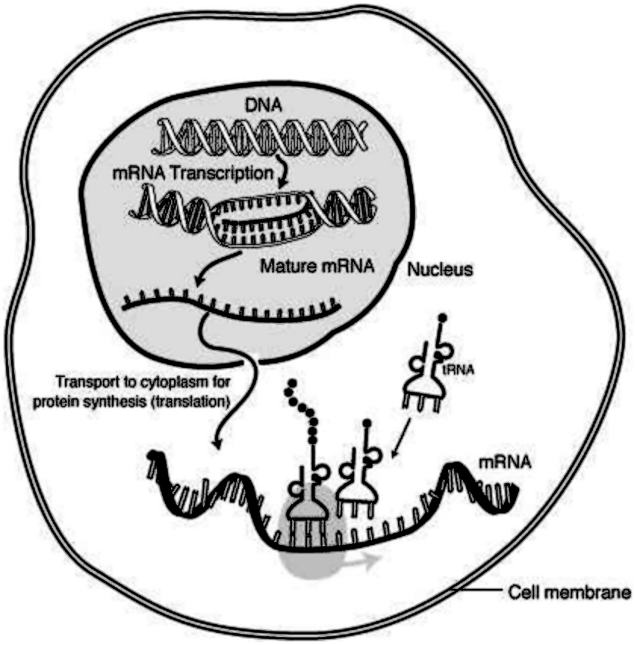

FIG. 1. *During the expression process, a complementary copy of a gene code is transcribed into the mRNA. An appropriately modified copy migrates from the nucleus to the cytoplasm where it serves as a template for the protein synthesis. Picture taken from National Human Genome Research Institute* (2001).

and they are produced during the two steps of *transcription* and *translation* that lead to the synthesis of a protein. This two-step representation of the protein-synthesis process is depicted in Figure 1 and constitutes the *central dogma of molecular biology* (Crick, 1970):

Transcription. The first step of a gene expression is the creation of a complementary copy of the gene sequence stored in one of the two DNA complementary strands. The complementary copy of the gene DNA code transcribes U (uracil) for A, A for T, G for C and C for G into the mRNA.

Translation. The mRNA transcript is moved from the nucleus to the cellular cytoplasm, where it serves as a template on which tRNA molecules, which carry amino acids, are lined up. The amino acids are then linked together to form a protein chain.

Because gene expression consists of copying DNA code into mRNA molecules, a measure of the gene expression level is the abundance of mRNA produced during this process (Schena, Shalon, Davis and Brown, 1995). This is the main intuition behind the large scale measurement of gene expression levels in microarrays that is described in the next section.

3. LARGE SCALE MEASUREMENT OF GENE EXPRESSION LEVELS

Quantitative methods to measure gene expression levels have been available to biologists for more than 20 years. Northern and Southern blots (see Alwine, Kemp and Stark, 1977; White, 1995) are techniques used to identify and locate mRNA and DNA sequences that are complementary to a segment of DNA. While these techniques are limited to examining a small number of genes at a time, a more recent technique, called serial analysis of gene expression (SAGE; Velculescu, Zhang, Vogelstein and Kinzler, 1995), is able to measure the global gene expression from entire cells. SAGE technology was introduced in 1995 by a team of cancer researchers at Johns Hopkins to rapidly identify differences between cancer and normal cells. The main intuition behind this technology was that short but specific stretches of DNA are sufficient to uniquely identify the genes expressed in a particular cell. SAGE uses these *short sequence tags* to mark the transcripts of a gene and to identify the number of transcripts generated by each gene, thus providing a measure of the gene expression. This technology is useful to detect and quantify the absolute expression level of both known and unknown genes, but it is time-consuming because it involves multiple steps and extensive sequencing to identify the appropriate tags (Lockhart and Barlow, 2001). Microarray technology has rendered efficient this process by measuring, simultaneously, the relative expression level of a large number of genes and, in so doing, is reshaping the epistemological and methodological vision of molecular biology and biomedical sciences.

3.1 Microarray Technology

The basic idea behind microarray technology is to simultaneously measure the relative expression level of thousands of genes within a particular cell population or tissue. Two key technical concepts behind this measurement process are *reverse transcription* and *hybridization*.

Reverse transcription. The mRNA transcript of a gene can be experimentally isolated from a cell, and reverse-transcribed into a complementary DNA copy called cDNA. A collection of cDNAs transcribed from cellular mRNA constitutes the cDNA library of a cell. Similarly, double-stranded cDNA can be reverse-transcribed into a complementary copy called cRNA. Technical details are described in Griffiths et al. (1999, Chapter 12).

Hybridization. Hybridization is the process of base pairing two single strands of DNA or RNA (Lennon and Lehrach, 1991). DNA molecules are double-stranded and these two strands melt apart at a characteristic melting temperature, usually above 65°C. As the temperature is reduced and held below the melting temperature, single-stranded molecules bind back to their counterparts. The process of binding back is based again on the principle of base pairing, so that only two complementary strands can hybridize. In the same way, a mRNA molecule can hybridize to a melted cDNA molecule when the mRNA contains the complementary code of the cDNA strands. When hybridization occurs, a single-stranded DNA binds strongly to complementary RNA in a way that prevents the DNA strands from reassociating with each other (Southern, Mir and Shchepinov, 1999).

Microarray technology is used to measure the relative level of expression of genes in a particular cell or tissue by hybridizing a labeled cDNA representation of the cellular mRNA to cDNA sequences (cDNA microarrays) or by hybridizing a labeled cRNA representation of the cellular mRNA to short specific segments known as *synthetic oligonucleotides* (synthetic oligonucleotide microarrays; Duggan et al., 1999; Lipshutz, Fodor, Gingeras and Lockhart, 1999). Synthetic oligonucleotides—also referred to as *oligos* in the biomolecular jargon—are short sequences of single-stranded cDNA that bind readily to their complements. The tethered cDNA sequences or oligos are called *probes*, while the cDNA or cRNA representation of cellular mRNA extracted from the cell is called the *target* (this is the suggested common terminology of Phimister, 1999). In both cases, the probes represent either genes of known identity or segments of functional DNA, also known as ESTs (expressed sequence tags). The target is labeled with fluorescent dye and hybridized to the probes. The higher the amount of cDNA or cRNA hybridized to a probe, the more intense the fluorescent dye signal will be on that probe. The relative mRNA abundance of a gene in a particular cell or tissue is therefore measured by the emission intensity of the probes. Synthetic oligonucleotide and cDNA microarrays are the two most popular microarray technologies and are described in the next two sections.

3.2 cDNA Microarrays

The cDNA technology (see Figure 2) was developed at Stanford University (Schena et al., 1995), although

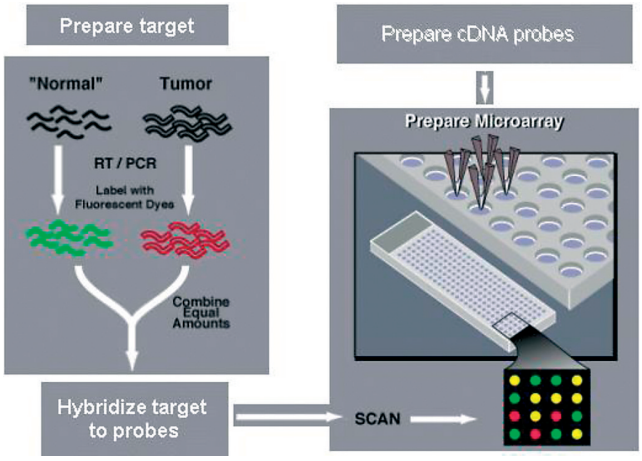

FIG. 2. *A sketch of cDNA microarray technology. Selected probes are amplified by PCR and the PCR product is printed to a glass slide using a high-speed robot. The targets are labeled representations of cellular mRNA obtained by reverse transcriptions of total RNA extracted from the test and reference cells, and the pooled target is allowed to hybridize with the cDNA spotted on the slides. Once the hybridization is completed, the slides are washed and scanned with a scanning laser microscope able to measure the brightness of each fluorescent spot; brightness reveals how much of a specific DNA fragment is present in the target.*

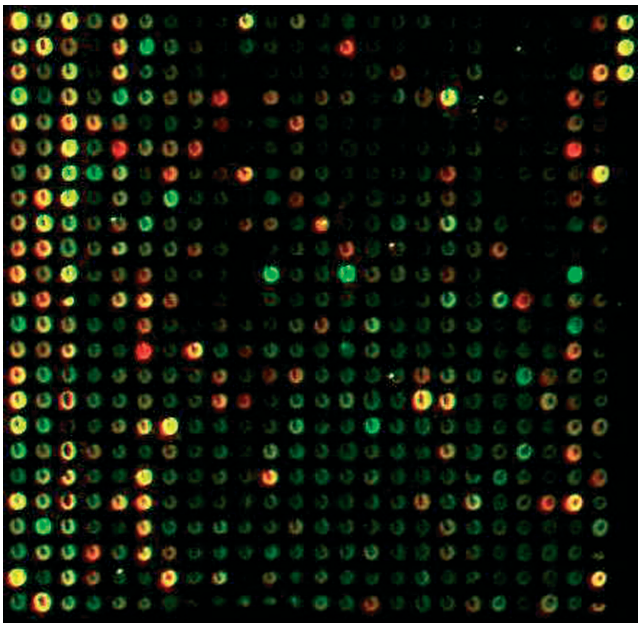

FIG. 3. *A scanned image produced from a cDNA microarray experiment. Each spot represents a gene. Grey spots denote genes that were expressed in neither type of cell; colored spots identify genes that were expressed in one of the two cells or both. The color of the spot discloses the relative expression of the gene in the two cells.*

similar concepts can be traced back as far as the mid 1980s (Ekins and Chu, 1999). The first step in the production of the microarray is selection of the probes to be placed on the microarray and amplification of the corresponding cDNA clones by a technique known as polymerase chain reaction (PCR). The PCR allows multiple rounds of amplification of a minimal amount of DNA to produce sufficient quantities of a sample. The cDNA microarrays are produced by spotting PCR samples of cDNA strands in approximately equal amounts on a glass slide using a high-speed robot. Each strand of cDNA identifies uniquely with its code, a gene or an EST, so that each spot in the microarray corresponds to a gene or an EST.

To prepare the target, investigators extract total RNA or mRNA produced from two types of cells, for example, healthy and tumor cells or test and reference cells. Then, by using a single round of reverse transcription, the mRNA from the two samples is fluorescently labeled with Cy3 (green) and Cy5 (red), and the target mixture is hybridized to the probes on the glass slides. During the hybridization, if segments of the mRNA representation in the target find their complementary portion among the samples of cDNA on the glass slide, they will bind together. When the hybridization is complete, the glass slide is washed and laser excitement of the glass slide is used to yield a luminous emission that is then measured by a scanning microscope. Fluorescence measurements are made with a microscope that illuminates each spot and measures fluorescence for each dye separately, thus providing a measure of the relative mRNA abundance for each gene in the two cells. The intensity of the green spot measures the relative mRNA abundance of the gene in the cell that had reverse-transcribed mRNA labeled with Cy3, while the intensity of the red spot measures the relative mRNA abundance of the gene in the cell that had reverse-transcribed mRNA labeled with Cy5. Grey spots denote genes that were expressed in neither cell type.

These measurements provide information about the relative level of expression of each gene in the two cells. The monochrome images can be pseudocolored to provide a quantitative measure of the relative expression of each gene in the two cells. This measure is adjusted to account for background noise caused by high salt and detergent concentrations during the hybridization or contamination of the target. Further details are discussed in Section 3.4. Figure 3 shows one of these images in which spots are colored in red, green, yellow and grey. Each spot corresponds to a gene, and the color of the spot discloses whether the gene is expressed (colored) or not and the relative level of expres-

sion in the two targets. Usually a measurement scale is provided to associate each color tone with a ratio between expression level in the two cells (Brown and Botstein, 1999; Schena et al., 1995).

Two limitations of cDNA technology are the risk of cross-hybridization and the large amount of total RNA (50–200 μg) required to prepare the target (Duggan et al., 1999). Cross-hybridization occurs when fragments of the reverse-transcribed mRNA in the target hybridize to similar complementary probes, thus producing false detections. The large amount of mRNA for target preparation has implications on the range of detection, so that genes expressed at low level—less than 1 transcript per 100,000—may fail to be detected. Several schemes to increase detection specificity are under development and a discussion can be found in Duggan et al. (1999).

3.3 Synthetic Oligonucleotide Microarrays

High-density synthetic oligonucleotide microarrays are fabricated by placing short cDNA sequences (*oligonucleotides*) on a small silicon chip by means of the same photolithographic techniques used in computer microprocessor fabrication. This proprietary technology, developed and commercialized by Affymetrix under the trademark GeneChip®, allows the production of highly ordered matrices that contain between 17,000 genes in the Affymetrix Murine Genome U74 set and 33,000 genes in the Affymetrix Human Genome U133 set.

The rationale behind this technology is based on the concept of probe redundancy: a *set* of well-chosen small segments of cDNA is not only sufficient to uniquely identify a specific gene, but also reduces the chances that fragments of the target will randomly hybridize to the probes, thus reducing the chances of cross-hybridization. Therefore, synthetic oligonucleotide microarrays represent each gene not by its cDNA, but by a set of fixed-length independent segments unique to the DNA of the gene as shown in Figure 4. On the GeneChip® platform, each oligonucleotide (probe) is 25 bases long and each gene is represented by a number of *probe pairs* ranging from 11 in the new Human Genome U133 set to 16 in the Murine Genome U74v2 set and the Human Genome U95v2. A probe pair consists of a perfect match (PM) probe and a mismatch (MM) probe. Each PM probe is chosen on the basis of uniqueness criteria and proprietary, empirical rules designed to improve the odds that probes will hybridize with high specificity. The MM probe is identical to the corresponding PM probe except for

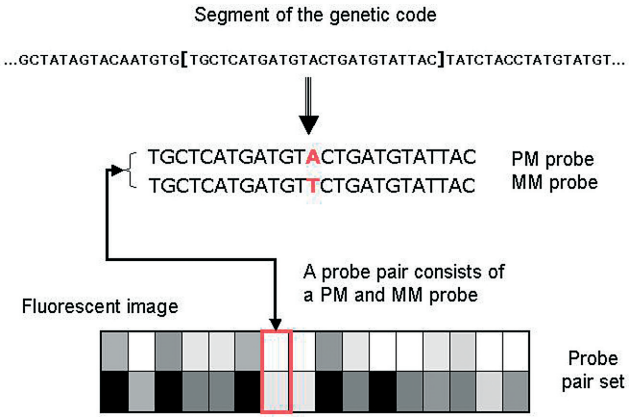

FIG. 4. *An oligonucleotide microarray associates a gene with a set of probe pairs, in this case* 20*. Each probe pair consists of a perfect match probe* (*PM*) *and a mismatch probe* (*MM*)*. Each PM probe is* 25 *bases long and is paired with the MM probe, in which the central base of the oligonucleotide is inverted. After hybridization of the target to the probes, the microarray is read with a laser scanner to produce an image and the intensity of the MM probes is used to correct the intensity of the PM probes. Image adjusted from Lipshutz, Fodor, Gingeras and Lockhart,* (1999).

the base in the central position, which is replaced with its complementary base as shown in Figure 4. The inversion of the central base makes the MM probe a further specificity control because, by design, hybridization of the MM probe can be attributed to either cross-hybridization or background signal caused by the hybridization of cell debris and salts to the probes (Lipshutz, Fodor, Gingeras and Lockhart, 1999; Lockhart et al., 1996). Each cell of an Affymetrix oligonucleotide microarray consists of millions of samples of a PM or MM probe, and probes that tag the same gene are scattered across the microarray to avoid systematic bias.

To prepare the target, investigators extract total RNA from a cell or tissue. The mRNA is reverse-transcribed into cDNA, which is made double-stranded and then converted into cRNA using a transcription reaction that fluorescently labels the target. Once hybridization has occurred, the microarray is washed and scanned with a standard laser scanner. The scanner generates an image of the microarray that is gridded to identify the cells that contain each probe and analyzed to extract the signal intensity of each probe cell.

Although less flexible than cDNA microarrays because the experimenter cannot select the probes, synthetic oligonucleotide microarrays offer several advantages. Besides the decreased chance of cross-hybridization, synthetic oligonucleotide microarrays require a smaller amount of total RNA to prepare the target (5 μg), have a wider dynamic range [the

hybridization signal is linearly related to up to 500-fold mRNA abundance (Lipshutz, Fodor, Gingeras and Lockhart, 1999) compared to 10-fold in cDNA microarrays (Relogio et al., 2002)] and a high detection specificity [mRNA transcript representations present in the target at a relative abundance of less than 1 in 10^6 can be detected (Lipshutz, Fodor, Gingeras and Lockhart, 1999)].

3.4 From Images to Data

In both cDNA and oligonucleotide microarrays, hybridization of the target to the probes determines a chemical reaction that is captured into a digital image by a scanning laser device. The next step is to translate the intensity of each hybridization signal into a table with numerical measures. The quality of the image analysis process is crucial for accurate interpretation of the data, and a variety of algorithms and software tools tailored to the different aspects of cDNA and oligonucleotide microarray images have been developed; see Bowtell (1999) and Kohane, Kho and Butte (2002).

The main steps of cDNA microarray image analysis are gridding, segmentation and intensity extraction, which have been reviewed in Smyth, Yang and Speed (2003). The gridding step recovers the position of the printed spots that correspond to the probes into the image. Because the position of the spots in the microarray is known, gridding is relatively straightforward although a series of parameters have to be estimated to account, for example, for shifts or rotations of the microarray into the image or small translations of the spots. The segmentation consists of a classification of the image pixels into foreground and background, where foreground pixels correspond to spots of interest in the microarray and background pixels correspond to noise resulting from high salt and detergent concentrations during the hybridization, or contamination of the target. Several segmentation methods have been proposed, which differ by the geometry of the spot they produce. For example, the method implemented in ScanAlyze (available at http://rana.lbl.gov/EisenSoftware.htm) fixes a circle with constant diameter to all spots in the image, whereas the method implemented in the Axon software GenePix (available at http://www.axon.com/GN_GenePixSoftware.html) estimates the diameter for each spot separately. The method developed by Chen, Dougherty and Bittner (1997) and implemented in QuantArray (available at http://www.packardbioscience.com/products/521.asp) uses repeatedly the Mann–Whitney test to label groups of eight pixels at a time as background or foreground. The package Spot (available at http: // experimental.act.cmis.csiro.au/Spot/index.php) for the R software implements an adaptive shape segmentation developed by Yang et al. (2001).

The intensity extraction step calculates the intensity of the red and green fluorescence of each spot, the background intensity and some quality measures. The background intensities are used to correct the foreground intensities and, hence, the red and green intensities that become the primary data for the subsequent analysis. Background correction is motivated by the fact that intensity measured for each fluorescent channel includes a contribution that is not due to the hybridization of the target to the probes. Most packages calculate the foreground intensity as the mean or the median pixel values. To correct the intensity of the two channels, an estimate of the background intensity is usually subtracted from the foreground intensity. For example, ScanAlyze calculates the corrected intensity by the average number of foreground pixels for each channel minus the median number of background pixels. Corrected intensity values are calculated as the difference between median foreground pixels and background pixels in QuantArray. Spot computes the background intensity by a nonlinear filter called morphological opening, which works by creating a background image for the whole microarray and by sampling this background image at the nominal centers of the spots. Further details and empirical comparisons of different segmentation and background correction methods can be found in Yang et al. (2001). Note that background correction introduces negative values when the foreground intensity is lower than the background intensity. Because background intensity larger than foreground intensity is considered an error, spots with negative corrected intensities are either disregarded or replaced by an arbitrary small positive number.

The analysis of oligonucleotide microarray images exploits the fact that the image produced by the scanning laser device describes the probes by squares of an approximately known number of pixels organized in a lattice. Furthermore, the image contains some alignment features that are recognizable as the checkerboard patterns at the corners of the image in Figure 5. Because the approximate physical dimension of each probe in the image is known, once the positions of the alignment features are determined, a basic grid is created to determine the pixels that describe each probe

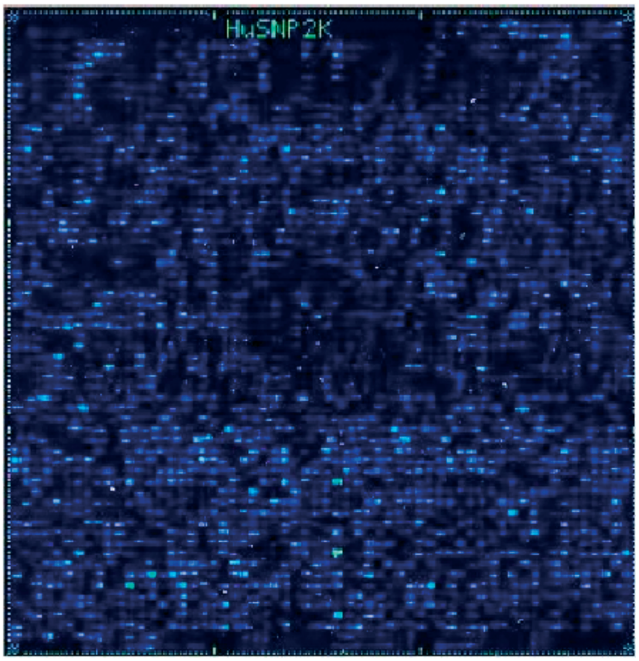

FIG. 5. *Scanned image of a synthetic oligonucleotide microarray. Grid cells represent probes and the intensity of each matrix cell measures the quantity of hybridized oligonucleotides in a probe. The checkerboard patterns at the corners of the image are the alignment features used to grid the image. Image courtesy of Affymetrix.*

cell by using some form of linear interpolation. To extract the intensity of each probe, the original proprietary algorithm employed by Affymetrix software used the 75th percentile of the pixel intensities, after removing the boundary pixels where intensity could be distorted.

Awareness of potential misalignment of the basic gridding algorithm, with consequent failure to extract the correct signal intensity, has led researchers to develop adaptive pixel selection algorithms; see Schadt, Li, Su and Wong (2000) and Zuzan et al. (2001). The adaptive pixel selection algorithm of Schadt et al. (2000) begins by removing pixels of extreme intensity and then iteratively adjusting the edges by removing those pixels that contribute most significantly to the coefficient of variation. Some constraints are also imposed to avoid bias of boundary pixels. The algorithm in Zuzan et al. (2001) corrects for misalignment, resulting in an improved selection of pixels attributed to individual probe cells and a substantial reduction in the variance of pixel intensities. The main motivation of this method is the fact that probe cells are often not equally spaced, so that gridding by linear interpolation can cause misalignment by as many as three pixels. To accommodate this deformation, the algorithm uses an iterative procedure that translates the initial location of probe cells by maintaining the lattice structure of neighbor cells. The most recent Affymetrix software for image analysis has a new algorithm to compute the background intensity that accounts for potential spatial effects. Essentially, the image is split into k (default value 16) rectangular zones and background intensity is computed as the lowest 2% intensity of the cells in each rectangular zone. The background intensity for each cell is calculated as a weighted average of the background intensities in each rectangular zone, with weights that account for the distance of the cell from the centers of each rectangular zone. This estimate of the background intensity is then subtracted to the probe cell intensity. Negative values resulting from background adjustments are set equal to a user defined value (the default value is 0.5).

Because the relative mRNA abundance is represented by the intensity of a probe pair set that consists of a number of probe pairs, the intensities of the probe cells are summarized to yield a relative measure of the gene expression level. The latest statistical algorithm produced by Affymetrix (MAS 5.0) generates, for each probe set, three measures: a detection call, a detection p-value and a signal value. The detection calls assess the quality of the hybridization, whereas the detection p-values represent the confidence in this assessment. The signal is a proxy for the relative expression level of the gene represented by the probe set. Full details are described in Affymetrix, Inc. (2002) and we summarize them briefly.

Detection calls and p-values are generated by first calculating a discriminant score R_i for each probe pair PM and MM given by

$$R_i = \frac{I(\mathrm{PM}_i) - I(\mathrm{MM}_i)}{I(\mathrm{PM}_i) + I(\mathrm{MM}_i)},$$

where $I(\mathrm{PM}_i)$ and $I(\mathrm{MM}_i)$ are the extracted intensities for the ith perfect match probe and mismatch probe. The score R_i is bounded above by 1 and measures the ability of the ith probe pair to detect its intended target. A positive value implies that the perfect match intensity $I(\mathrm{PM}_i)$ is larger than $I(\mathrm{MM}_i)$, and the strength of detection ability of the ith probe pair increases with R_i. A negative value implies that the mismatch intensity $I(\mathrm{MM}_i)$ exceeds $I(\mathrm{PM}_i)$ and highlights a poor detection ability of the ith probe pair. To avoid bias, saturated cells (defined as mismatch probe cells with intensity above a fixed threshold) as well as probe cells in which $I(\mathrm{PM}_i) \leq I(\mathrm{MM}_i) + 0.015$ are disregarded.

To determine the detection p-value, the scores R_i computed for the probe pairs in a probe set are compared with a user defined threshold τ (typically $\tau = 0.015$), and the null hypothesis of no difference between the median discrimination score and τ is tested by the one-sided Wilcoxon signed rank test. The detection p-value is simply the p-value computed by assuming an asymptotic normal distribution for the Wilcoxon signed rank statistic when more than 12 probe pairs are used, whereas exact calculations are carried out when the retained number of probe pairs is less than 12. Detection calls describe whether the hybridization of a probe set has occurred (P for present), has not occurred (A for absent) or has been only marginal (M), and are assigned on the basis of a significance range for the detection p-value. Suggested settings are to call the hybridization present if the detection p-value is smaller than 0.04, marginal if the detection p-value is between 0.04 and 0.06, and absent otherwise. Last, the signal that measures the relative expression level of a probe set is computed by the one-step Tukey biweight estimate, which essentially produces a robust average of the differences between $I(\mathrm{PM}_i)$ and $I(\mathrm{MM}_i)$, with weights that take into account the distance between $I(\mathrm{PM}_i) - I(\mathrm{MM}_i)$ and the median intensity difference.

The rationale behind the use of paired PM and MM probes is that the specific hybridization represented by the intensity of the PM probes should be stronger than the nonspecific hybridization represented by the intensity of the MM probes, and such a consistent pattern across the probe set is unlikely to occur by chance. Several studies support this claim, for example, Kane et al. (2000) and Lockhart et al. (1996). However, mismatch values $I(\mathrm{MM}_i)$ can be higher than perfect match values $I(\mathrm{PM}_i)$ for a number of reasons, such as cross-hybridization occurring when the probe sequence has high homology with another unknown sequence or errors in the probe sequences that cause low specificity. Therefore, a weighted average of the difference between $I(\mathrm{PM}_i)$ and $I(\mathrm{MM}_i)$ can produce negative intensity values. In fact, the previous Affymetrix software MAS 4.0 used to return probe set intensity values called *average difference* that could be negative. A series of rules is employed by the latest Affymetrix software MAS 5.0 to avoid the calculation of negative signal values. Particularly if the mismatch value $I(\mathrm{MM}_i)$ is higher than the perfect match $I(\mathrm{PM}_i)$, then the mismatch is assumed to provide no additional information about the estimate of the signal and it is replaced by an imputed value called idealized mismatch (IM). This idealized mismatch is either a value smaller than $I(\mathrm{PM}_i)$ or an estimate based on the average ratio between perfect match and mismatch values.

4. EXPERIMENTAL QUESTIONS AND EXPERIMENTAL DESIGN

Both cDNA microarrays and oligonucleotide microarrays provide a panoramic view of the activity of genes under particular experimental conditions, and are nowadays used to answer the same broad classes of questions. In the following discussion, we will term the set of expression levels measured for a gene across different conditions its *expression profile*, whereas we will use the term *sample molecular profile* to denote the expression level of the genes measured in a sample in a particular condition.

4.1 Experimental Questions

By providing a measure of expression of a gene in terms of its mRNA abundance, microarray technology lets the experimenters observe the molecular profile of a cell, or cell line—distinct families of cells grown in culture—in a particular condition. The simplest experiment we can devise using this technology is a *comparative* experiment, illustrated in Figure 6, to identify the genes differentially expressed in two conditions. An example of this experimental setting is the comparison of metastatic versus nonmetastatic derivatives of a tumor cell line (Lander, 1999), in which samples of cells from the two conditions are extracted from several patients. The experimental conditions can be specific levels of controllable environmental factors, such as extreme temperatures or starvation, or the modification (*knock-in*) or the removal (*knock-out*) of a specific portion of the genome.

More complex experimental questions involve molecular profiling of several conditions at a time to characterize, for example, the genomic fingerprint of different

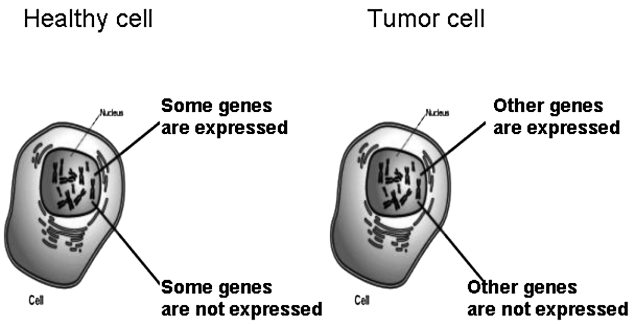

FIG. 6. *Microarray technology enables investigators to detect the genes that are differentially expressed in two samples.*

types of cancer (Alizadeh et al., 2000) or the effect of changing several experimental factors simultaneously (Churchill and Oliver, 2001). In both cases, each sample consists of the gene expression levels measured in cells grown or observed in a particular condition, and different samples can be assumed to be stochastically independent. A different class of experimental questions involves the study of the temporal evolution of gene expression profiles, so that different samples may be stochastically dependent. Studies in this class try to understand, for instance, the process that turns a locally growing tumor into a metastatic killer (Clark, Golub, Lander and Hynes, 2000), the yeast sporulation cycle (Spellman et al., 1998) or the response of human fibroblasts to serum (Iyer et al., 1999). Although the dependency structure among samples requires a different analysis, the common feature of these experiments is to compare the molecular profiles of cells in different conditions.

Advanced experiments investigate the regulatory mechanism of cells observed in different experimental conditions. When functioning normally, regulatory pathways in cells modulate the level and duration of gene expression, thus ensuring that cells respond to physiological and extracellular stimuli in an appropriate manner. However, a broad range of diseases can result when the activation of a gene regulation pathway triggers either an under- or overproduction of certain proteins. Fundamental problems are the discovery of new gene regulatory pathways and of causal dependencies among gene expression (Pilpel, Sudarsanam and Church, 2001).

4.2 Experimental Design

The design of microarray experiments is a critical, albeit still neglected, issue of modern functional genomics. One important difference between cDNA microarrays and synthetic oligonucleotide microarrays is that the former are designed for so-called competitive hybridization: two targets can be simultaneously hybridized to the probes on one microarray. The first key issue in designing a comparative cDNA experiment is to choose between direct and indirect comparisons. In the first case, the target mixture is hybridized to the same microarray, whereas, in the second case, the mRNA representations from the two treated cells are mixed with the reverse-transcribed mRNA from a reference cell and the two mixtures are hybridized to two different microarrays. A discussion of the pros and cons of direct versus indirect comparisons is given in Yang and Speed (2002).

Besides technical issues of probe/microarray choice and design, the most fundamental design issue common to both cDNA and synthetic oligonucleotide microarrays is the choice of the number of replications required to stake a statistically sound claim. Although microarray technology has rendered gene expression measurement blazingly fast, the cost of a single experiment—up to $1200 for a single high resolution synthetic oligonucleotide microarray—is still a significant factor in the experimental choices of biomedical investigators. Comparative experiments reported in mainstream biomedical journals were originally limited to one replication of an experiment (DeRisi, Iyer and Brown, 1997). Arguments have been made to show that a single replication of a comparative experiment is not sufficient to achieve reproducible results (Lee, Kuo, Whitmorei and Sklar, 2000), but despite the increasing awareness that data generated by even the most accurate microarray are very noisy, many discoveries reported in mainstream journals are often based on experiments with three replications (Wyrick et al., 1999).

The main problem of this experimental design aspect is caused by the parallel nature of experiments conducted with microarrays: the number of replicates necessary to obtain an accurate measure of the expression level of a gene g may not be the same number needed for a different gene. Furthermore, responses to the microenvironment conditions, such as the time of day or washing conditions, appear to have a significant impact on gene expression. Lander (1999), leader of one of the largest genomic centers in the world, reports that "It is well known among *aficionados* that comparison of the same experiment performed a few weeks apart reveals considerably wider variation than seen when a single sample is tested by repeated hybridization." Therefore, while replicated experiments should increase the amount of information needed to carry out a statistical analysis, they may also increase variability among replicates.

An additional experimental design issue arises from the common problem of mRNA paucity. It is often the case that a single cell is unable to produce detectable mRNA in the desired condition. In this situation, common practice is either to *pool* together the mRNA extracted from different samples or to amplify the cellular RNA. While obvious reasons of variability control suggest using the same pooled sample for each experimental condition, determination of the number of units to pool together is still an open issue. An

interesting discussion of this experimental design issue is given in Yang and Speed (2002).

When the objective of the experiment is the study of the temporal evolution of a biological system, the researchers need also to choose the time points to sample. These experiments are usually performed by sampling the gene expression profile using a microarray at predefined temporal intervals and then mounting these snapshots of the genome activity into "movies" that capture the dynamics of the process. The specificity of each gene becomes, here, even more important: the optimal sample points to observe the evolution of a gene during a process may not be the same for another gene on the same microarray.

In more complex experiments conducted to study the effect of different experimental factors, the choice of the number of replications is paired with the choice of the experimental treatments to test. Some recent research has addressed the issue of the experimental design for microarray data (Churchill and Oliver, 2001; Kerr and Churchill, 2001b, c; Pan, Lin and Le, 2002; Yang and Speed, 2002) by proposing classical factorial experimental designs, but we believe the choice of the experimental design is very much an open problem. The theory of statistical experimental design seeks experimental plans that allow a specific statistical analysis to be carried out to test particular hypotheses (Cox and Reid, 2000). Because to date no agreement exists about the appropriate statistical analysis of gene expression data produced with microarrays and because many experiments with microarrays are conducted to generate rather than test hypotheses, critical experimental design issues are still far from being solved.

5. DATA PREPROCESSING

To answer the experimental questions, the quantitative measurements of gene expression data produced by microarray experiments are analyzed using statistical and machine learning methods. A common strategy to reduce data variability and dimensionality is to perform two preprocessing operations known as *normalization* and *filtering* on either the raw or transformed data, before undertaking any data analysis. The goal of the normalization operation is to remove systematic distortions across microarrays to render comparable the experiments conducted under different conditions. The aim of the filtering operation is twofold: to reduce variability by removing those genes that have measurements that are not sufficiently accurate and to reduce the dimensionality of the data by removing genes that are not sufficiently differentiated. The transformation of the raw data, advocated by several authors, should even out the intensity values that are usually extremely skewed.

5.1 To Log or Not To Log Transform?

Suppose the microarray experiment was conducted to compare the expression level of G genes in two cells. For each gene g, denote by (y_{g1}, y_{g2}) the pair of relative expression levels measured in the two conditions. If the experiment is conducted by a direct comparison with one cDNA microarray, (y_{g1}, y_{g2}) will denote the corrected intensity values for the red and green channels in the spot that corresponds to the gene g. When the experiment is conducted with two synthetic oligonucleotide microarrays, (y_{g1}, y_{g2}) will denote the signal values for the probe set that describes the gene g.

Because the corrected intensity values are highly skewed, log-transforming the raw data (y_{g1}, y_{g2}) produced by cDNA microarray experiments is strongly recommended by several authors to even out intensity values; see, for example, Yang et al., 2001. A fairly common assumption is that the logarithmic transformation produces normally distributed data (Nadon and Shoemaker, 2002). As an example, the histogram in the top plot of Figure 7 describes the distribution of corrected intensity values for the red channel in an experiment conducted to compare gene expression levels in normal and malignant lymphocytes (Alizadeh et al., 2000). The background correction was done by subtracting the median background intensity from the average foreground intensity for each spot, and the foreground and background intensities were computed using ScanAlyze. Note that the background correction introduces a small proportion of negative values and, typically, spots with negative corrected intensities are either disregarded or the negative intensity is replaced by an arbitrary small number. The distribution of positive intensity values is extremely skewed and the histogram in the bottom plot of Figure 7 describes the distribution of the log-transformed intensities after removal of the negative values. The logarithmic transformation removes most of the original asymmetry, but some right skewness is still visible. Other examples are reported in the Speed group microarray page (http://www.stat.berkeley.edu/users/terry/zarray/Html/log.html). This residual lack of symmetry after the logarithmic transformation is typical of Gamma distributed data (McCullagh and Nelder, 1989), so that rather than the logarithmic transformation, some power or

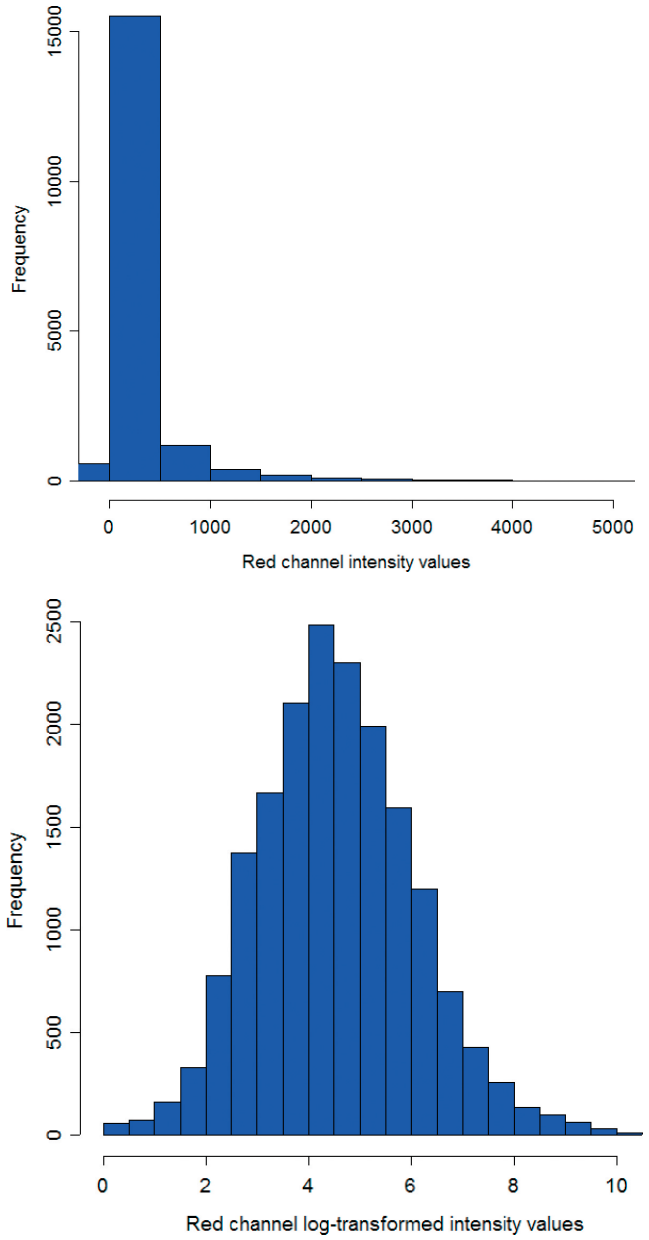

FIG. 7. *Histograms of the corrected intensity values for the red channels* (*top*) *and the corrected intensity values after negative intensities were removed and positive intensities were log-transformed* (*bottom*).

variance transformations would be more suitable. Examples of such transformations are discussed in Rocke and Durbin (2001).

Although the choice for the best data transformation of the red and green corrected intensities is still an open problem, it is generally acknowledged that the corrected intensity values measured with competitive hybridization on cDNA microarrays should be transformed. No such similar consensus exists when the data (y_{g1}, y_{g2}) are produced by synthetic oligonucleotide microarrays, possibly because the original Affymetrix statistical software MAS 4.0 used to return a fairly large proportion of negative intensity values, often 25% of the data. Authors have suggested using the cubic root transformation (Tusher, Tibshirani and Chu, 2000) or the log transformation of appropriately truncated data (Ibrahim, Chen and Gray, 2002), but well accepted data analysis protocols (Golub et al., 1999) do not use any data transformations. Empirical evidence that the bulk of positive values produced by synthetic oligonucleotide microarrays analyzed with MAS 4.0 follow a log-normal distribution was provided by Hoyle, Rattray, Jupp and Brass (2002), whereas questions about the correct transformation to use, if any, of intensity values calculated by the Affymetrix statistical software MAS 5.0 are still open.

5.2 Normalization of Microarray Data

A well known problem with cDNA technology is the consistent imbalance of the fluorescent intensities of the two dyes Cy3 (green) and Cy5 (red)—Cy3 is systematically less intense than Cy5 (Quackenbush, 2001; Yang et al., 2001). Although a simple dye-swap experiment in which each hybridization is repeated twice with reversed dye assignment to the two targets would be the best way to remove this systematic bias (Yang and Speed, 2002), normalization techniques are commonly used to render the gene expression levels measured by the two different dyes comparable (Duggan et al., 1999). Synthetic oligonucleotide microarrays do not suffer from a known systematic distortion similar to the dye fluorescence imbalance of cDNA microarrays, but a comparative experiment conducted on this platform requires hybridization of each target to a different microarray. A variety of experimental errors, including variations of the amount of mRNA used to create the target hybridized to each microarray or the quantity of dye used to fluorescently label each target, may introduce errors. Normalization techniques are therefore used in an attempt to "remove" the experimental errors.

Assuming that the amount or type of dye used to label the two targets as well as variations of the quantity of cellular mRNA used in the two targets induces contaminations, the observed expression level y_{g2} masks the correct expression level $\tilde{y}_{g2}$ one would observe if the second experiment were conducted in exactly the same conditions of the first experiment. Formally, we can write

$$y_{g2} = f(\tilde{y}_{g2})$$

and normalization techniques consist of estimating the function $f(\cdot)$ to recover $\tilde{y}_{g2} = f^{-1}(y_{g2})$. *Total*

intensity normalization approximates $f(\cdot)$ with the zero-intercept regression line $y_{g2} = \beta \tilde{y}_{g2}$ and estimates β by $(\sum_g y_{g1})/(\sum_g y_{g2})$ (Quackenbush, 2001). The rationale behind this choice is that the total quantity of mRNA representation hybridizing from each target should be the same. When β is estimated by the ratio $(\bar{y}_1/\bar{y}_2)$, where $\bar{y}_1$ and $\bar{y}_2$ are the average expression levels in the two targets, the technique is also called *total mean normalization*. Variants of this procedure estimate β by the ratio of the medians or by the ratio of the trimmed means.

An alternative technique, known as *normalization with calibration*, relies on the assumption that only a very small proportion of genes in a microarray should have substantially different levels of expression across the two cells. Following this principle, the function $f(\cdot)$ is approximated with the regression line $y_{g2} = (\tilde{y}_{g2} - \alpha)/\beta$, and the parameters α and β are estimated from the data (y_{g1}, y_{g2}) by fitting the linear regression $y_1 = \alpha + \beta y_2$. In so doing, the regression line for y_1 versus $\tilde{y}_2$ will have zero intercept—thus removing systematic deviations—and unitary slope—thus capturing the intuition that the majority of gene expression levels across the two experimental conditions should remain unchanged. Normalization with calibration can be adjusted to account for specific nonlinear effects, and nonparametric regression techniques, such as lowess regression, have been proposed to handle possibly nonlinear transformations (Bhattacharjee et al., 2001; Yang et al., 2001) or spatial effects (Irizarry et al., 2001; Yang et al., 2002).

All these normalization techniques can be used either globally or locally. Global normalization uses all genes in the microarray to identify a transformation of the expression data to calibrate the measures in the two samples. Local normalization uses only those genes known to remain constantly expressed across the two particular experimental conditions or *housekeeping genes*, a library of genes believed to have nearly constant expression levels in a variety of experimental conditions. Well accepted protocols (Bhattacharjee et al., 2001; Coller et al., 2000; Golub et al., 1999) use the subset of genes detected as hybridized by the Affymetrix software.

One problem of normalization with calibration applied to intensity data is that when $\alpha > 0$, small values of the systematically larger intensity are replaced by negative numbers. To avoid this bias, other normalization techniques try to calibrate the ratios y_{g2}/y_{g1} (Chen, Dougherty and Bittner, 1997) or the log ratios $\log(y_{g2}/y_{g1})$ (Yang et al., 2001, 2002). Clearly, these techniques are applicable to microarray data that can be paired, as for example data generated by direct comparisons with cDNA microarrays.

Extending normalization techniques to repeated experiments is not straightforward. Yang et al. (2001) provided a comprehensive overview of normalization techniques for repeated experiments with cDNA microarrays. For oligonucleotide microarrays, a common approach to normalization of multiple experiments is to choose one replication as a baseline and to apply normalization with calibration or total intensity normalization to the other replications (Golub et al., 1999). Because the results will differ according to the chosen baseline, authors have suggested computing the baseline as the average expression profile across all microarray samples (Tusher, Tibshirani and Chu, 2001). An open question remains whether normalization of replicated experiments with oligonucleotide microarrays is needed at all. In replicated experiments in which more than one microarray is hybridized to a replication of the same target, changes in the amount of cellular mRNA used to prepare the target or changes in the amount of fluorescent dye should be considered part of the experimental error. If no systematic errors are introduced, one can assume that the measurement observed for gene g in the replicate k of the experimental condition i is

$$y_{gik} = \mu_{gi} + \varepsilon_{gik},$$

where ε_{gik} is the error in replicate k and μ_{gi} is the correct expression level of gene g in condition i. The assumption that the experiment is reproducible would require that, on average, the experimental errors compensate, so that normalization is not necessary. This is, for example, the approach adopted by Olshen and Jain (2002). The error variance can be modeled to account for the different sources of variability. An approach along this line is presented in Jin et al. (2001) and Wolfinger et al. (2001) for the analysis of repeated cDNA-based expression levels transformed in log scale.

The issue of normalization of repeated comparative experiments differs from the normalization needed when more than two experimental conditions—either different targets or the same target tested at different time steps—are analyzed. For example, when the objective of the whole experiment is to examine the temporal behavior of a genomic system during a cell cycle, it is common practice to take only one replication of the gene expression data at each time point (Eisen, Spellman, Brown and Botstein, 1998; Iyer et al., 1999;

Spellman et al., 1998) and use standard normalization techniques to make the expression levels measured at different time points comparable. Although a preferable solution would be to take a few replicates of each measurement, cost constraints often make this solution impractical.

5.3 Filtering

Several techniques are available to reduce data dimensionality and variability by removing some gene measurements. It is surprising to realize that ad hoc rules are commonly used and that the choice of the genes to be removed can differ substantially according to the microarray platform and the technique chosen to analyze the data.

For expression data measured with a cDNA microarray, it is common practice to disregard those genes with negative or small expression levels (before or after normalization). Typically, all those spots in which the foreground intensity does not exceed the background intensity by more than 1.4-fold are disregarded or replaced by an arbitrary small number. The Affymetrix statistical software MAS 5.0 assigns a detection call to each probe set to assess the amount and quality of hybridization, and it is suggested to discard all genes that have expression levels labeled as A (absent) or M (marginal) in all samples. This procedure is justified by the empirical evidence that expression levels smaller than 10 are actually measurement errors (Affymetrix, 2002). However, a large proportion of genes would often be discarded by this procedure, and investigators tend to adopt less stringent criteria to select a subset of the genes to be further analyzed. A common strategy is to retain only those genes that have minimum fold-change that exceeds a particular threshold d in a preset number of experiments c, for example, $d = 3$ and $c = 1$ in Causton et al. (2001). The choice $c = 2$ was originally suggested by DeRisi et al. (1997) to analyze expression levels measured with cDNA microarrays, and an insightful analysis of the empirical success of this rule is described in Sabatti, Karsten and Geschwind (2001). Golub et al. (1999) suggested to further score genes by their standard deviation, so as to limit the analysis to those genes that vary most across experiments; a similar approach was proposed in Efron, Tibshirani, Storey and Tusher (2001). Other authors remove "spiked" genes, that is, those genes with one abnormally large or abnormally small measurement (Thomas, Olsen, Tapscott and Zhao, 2001). The recent book by Kohane, Kho and Butte (2002) contains a comprehensive description of other filtering techniques most commonly used.

All these filters depend on arbitrary thresholds used to decide when a value is abnormally large or small, or when the variability of the measurements is too high. The impact of normalization and filtering strategies is unclear and few systematic studies are available to provide investigators with a description of the properties of these preprocessing techniques and guidance on choosing the one most appropriate for their particular problem.

6. ANALYSIS OF COMPARATIVE EXPERIMENTS

This section describes the most popular techniques for the analysis of gene expression data in repeated comparative experiments. The objective of the analysis is to identify the genes with a significant expression change across two conditions. The approaches to this problem can be classified in two broad categories. Methods in the first category, known as fold analysis, estimate the ratio between the expression levels of each gene in the two conditions, whereas methods in the second category use the data to estimate the expected difference in expression of each gene in the two conditions.

6.1 Fold Analysis

Early comparative experiments based on cDNA microarray technology measured differences of gene expression across two conditions in terms of the fold-change: the ratio of the expression levels (DeRisi, Iyer and Brown, 1997; Schena et al., 1995, 1996). Particularly genes that showed a negative or positive fold-change of at least 2 were deemed to be differentially expressed across the two conditions. The need to choose a threshold to identify significant differentially expressed genes in two conditions is the motivation of a series of articles focused on statistical fold analysis.

We let $\rho_g = \mu_{g1}/\mu_{g2}$ denote the unobservable "true" fold-change for gene g in the two conditions. When $\rho_g = 1$, the expression level of the gene g has not changed, while $\rho_g < 1$ and $\rho_g > 1$ indicate differential expression of the gene g in the two conditions. Particularly, $\rho_g < 1$ means that the gene is *down-regulated* by condition 1, whereas $\rho_g > 1$ means that the gene is *up-regulated* by condition 1. Statistical approaches to ratio-based differential analysis estimate the ratio ρ_g with some statistic r_g and decide whether deviations of the estimate r_g from 1 can be attributed to a real difference of the gene expressions in the two conditions, rather than sampling variability. In the first published work that followed this approach, Chen, Dougherty

and Bittner (1997) used the naive ratio estimator $r_g = y_{g1}/y_{g2}$. Assuming that the measurements from the two different channels (corresponding to the Cy3 and Cy5 fluorescent dyes) are independent and normally distributed, and that they have constant coefficient of variation for all genes in both conditions, they derived an approximate distribution of the ratio statistic r_g that can be used to find a $(1-\alpha)\%$ confidence interval for the ratio ρ_g. The assumption of a constant coefficient of variation c in the two conditions lets the distribution of r_g depend on c, which is estimated by maximum likelihood. They also proposed an iterative procedure to simultaneously estimate c and the normalization factor to render comparable the measurements from the two channels.

As noted in Newton et al. (2001), this approach disregards ancillary information during the computation of the distribution of the ratio statistic, because the product $y_{g1} \times y_{g2}$ contains information about the variability of r_g. Furthermore, despite the fact that expression levels should be positive numbers, the measurements of the two channels are assumed to follow normal distributions. This inappropriate distributional assumption is corrected in Newton et al. (2001) by assuming that the measurements of the two channels follow Gamma distributions, and a Bayesian method is proposed to estimate the fold-change of each gene to account for the "between microarrays" variability. Although this second approach is based on sounder distributional assumptions about gene expression measurements, it relies on the unconventional assumption that the experimental error across microarrays also follows a Gamma distribution.

Distributional assumptions aside, both approaches treat the pair of measurements of each gene in the cDNA microarray as independent, but this choice does not seem to be always correct. In direct comparisons, the same spot of cDNA in the microarray is simultaneously hybridized to the pool of mRNA representation in the target mixture. In other words, the two targets compete for hybridization to the probes so that, by design, each pair of measurements should be treated as a matched pair. Alternative approaches that directly model the ratio $r_g = y_{g1}/y_{g2}$ or its logarithm $l_g = \log(r_g)$ overcome this difficulty. The method introduced by Lee et al. (2000) uses a mixture model to describe the joint distribution of the log ratio of the measurements from the two channels as follows:

$$f(l_g) = pf_E(l_g) + (1-p)f_U(l_g).$$

The parameter p is the unknown proportion of genes that are differentially expressed, $f_E(l_g)$ is the density function of l_g when the gene g is differentially expressed and $f_U(l_g)$ is the density function of l_g when the gene g is not differentially expressed. By assuming a normal distribution for l_g, for each g, the mixture components are estimated by using the expectation-maximization (EM) algorithm (Dempster, Laird and Rubin, 1977). The estimates are then used to compute the posterior probability

$$pf_E(l_g)/f(l_g)$$

that each gene g is differentially expressed in the two experiments. When more than one replication is available, this procedure is applied to a "polished" summary of the original expression ratios that is computed as follows. By taking into account the sources of variability of each gene measurement, the log ratio $l_{gk} = \log(y_{g1k}/y_{g2k})$ of the paired measurements for each gene g is modeled by

$$l_{gk} = \mu + \alpha_g + \beta_k + (\alpha\beta)_{gk} + \varepsilon_{gk}, \quad g = 1, \ldots, G,\ k = 1, \ldots, n, \tag{1}$$

where G is the total number of genes in the microarray and n is the total number of replicates of the experiment. The parameters α_g represents the "gene effect," described as the correct fold-change of gene g across all replications of the experiment. The parameter β_k captures the "microarray effect," due, for example, to between microarray differences in the fluorescent dye or the amount of mRNA used to prepare the target. The interaction term $(\alpha\beta)_{gk}$ accounts for possible variations of each gene fold-change in each replication of the experiment. The errors ε_{gk} are assumed to have zero mean. Although the authors acknowledge that all the effects in model (1) should be treated as random effects, they propose to estimate the parameters α_g using the standard two-way analysis of variance estimator

$$\hat{\alpha}_g = \frac{1}{n}\sum_k l_{gk} - \frac{1}{nG}\sum_{gk} l_{gk}, \quad g = 1, \ldots, G,$$

which does not require any assumptions about the error distribution. Each estimate $\hat{\alpha}_g$ is then used as a proxy of l_g to estimate the posterior probability that the gene g is differentially expressed. Note that, in the absence of pure replications, model (1) is overparameterized because the gene-array interaction $(\alpha\beta)_{gk}$ and the random error ε_{gk} are not distinguishable. In fact, pure replications are rarely conducted, and the microarray effect should be treated as a random block effect. Several authors have modified this approach by relaxing the parametric assumption on the mixture model (Efron, Storey and Tibshirani, 2001; Pan, Lin and Le,

2001), by using a larger number of fixed effects to model dye and spot effects (Kerr and Churchill, 2001c) or by using random effects (Wolfinger et al., 2001).

The scope of this stream of work is limited to direct comparisons of gene expression data with cDNA microarrays, where two targets are hybridized to the probes on the same microarray. In this case, the expression measurements from each microarray are paired by design. When cDNA is used for indirect comparisons or the expression data are measured with synthetic oligonucleotide microarrays, there is no unique pairing of the data. To conduct the fold analysis on repeated experiments, researchers compute the average of the normalized expression levels in the two experimental conditions and impose an arbitrary threshold on the ratio (or log ratio) of the two averages. Unfortunately, no consensus exists about this threshold, even across different studies on the same organism by the same investigators (Holstege et al., 1998; Wyrick et al., 1999). Typically, this threshold varies between 2 and 3 (Glynne et al., 2000; Jackson-Grusby et al., 2001; Ly, Lockhart, Lerner and Schultz, 2000; Roberts et al., 2000), but it can be as low as 1.7 (Lee, Weindruch and Prolla, 2000). No published work addresses the problem of the extent of false positive and false negative rates produced by this "naive" fold analysis. A very elegant solution is presented in Ibrahim, Chen and Gray (2002), which describes a Bayesian method for fold analysis under the assumption that appropriately truncated gene expression data follow a log-normal distribution.

6.2 Differential Analysis

We now describe the hypothesis that a gene g is not differentially expressed in two experimental conditions by $H_0 : \mu_{g1} = \mu_{g2}$, while differential expression occurs under the alternative hypothesis $H_a : \mu_{g1} \neq \mu_{g2}$. To identify the set of genes that are differentially expressed, one needs to test the null hypothesis for each gene and then select the set of genes for which the null hypothesis is rejected. We continue to denote by y_{gik_i}, $g = 1, \ldots, G$, $i = 1, 2$ and $k_i = 1, \ldots, n_i$, the expression level data generated by a comparative experiment. When the expression levels are measured with cDNA microarrays by direct comparisons, the replications of each condition are equal, say $n_1 = n_2 = n$, while there is no need to impose this restriction for data measured with oligonucleotide microarrays or indirect comparisons with cDNA microarrays. The standard statistic used to test the null hypothesis is the t-statistic

$$t = \frac{|\bar{y}_{g1} - \bar{y}_{g2}|}{\sqrt{s_g^2}},$$

where $\bar{y}_{g1}$ and $\bar{y}_{g2}$ are the average expression levels of gene g in the two conditions and s_g^2 is an estimate of the variance σ_g^2 of the sample mean difference. Large values of the t-statistic would offer evidence in favor of differential expression. The two main problems are choice of the estimate s_g^2 and identification of a threshold to reject the null hypothesis.

When the two samples are not independent—as for data collected with cDNA microarrays in direct comparisons—an appropriate estimate of σ_g^2 appears to be

$$\begin{aligned} s_{Dg}^2 &= \frac{\sum_k [(y_{g1k} - y_{g2k}) - (\bar{y}_{g1} - \bar{y}_{g2})]^2}{n(n-1)} \\ &\equiv \frac{s_{g1}^2}{n} + \frac{s_{g2}^2}{n} - 2\frac{s_{g12}}{n}, \end{aligned} \tag{2}$$

where the term $s_{g12} = \sum_k (y_{g1k} - \bar{y}_{g1})(y_{g2k} - \bar{y}_{g1})/(n-1)$ is an estimate of the covariance of the two sample means. When the two samples are independent—for example, when data are collected with oligonucleotide microarrays or indirect comparisons with cDNA microarrays—σ_g^2 can be estimated by

$$\begin{aligned} s_{Ig1}^2 &= \frac{\sum_{k_1} (y_{g1k_1} - \bar{y}_{g1})^2}{n_1(n_1-1)} + \frac{\sum_{k_2} (y_{g2k_2} - \bar{y}_{g2})^2}{n_2(n_2-1)} \\ &:= \frac{s_{g1}^2}{n_1} + \frac{s_{g2}^2}{n_2} \end{aligned} \tag{3}$$

or by

$$\begin{aligned} s_{Ig2}^2 &= \frac{\sum_i \sum_{k_i} (y_{gik_i} - \bar{y}_g)^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \\ &:= s_{gp}^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right), \end{aligned} \tag{4}$$

where $\bar{y}_g$ is the average expression across the two experiments. The estimate in Equation (3) is appropriate when the variances of the gene expression data in the two conditions are different; its use was suggested in Dudoit, Yang, Callow and Speed (2002) and Lönnstedt and Speed (2002). The estimate in Equation (4) uses the typical pooled estimate of the common variance and it is used less often; see, for example, Olshen and Jain (2002). Because of the large variability of gene expression data measured with microarrays, some forms of penalization for the denominator of the t-statistic have been suggested. For example, Golub et al. (1999) suggested estimating σ_g by the quantity

$$s_{S2Ng} = \frac{s_{g1}}{\sqrt{n_1}} + \frac{s_{g2}}{\sqrt{n_2}}$$

and referred to the ratio $|\bar{y}_{g1} - \bar{y}_{g2}|/s_{S2Ng}$ as the *signal-to-noise ratio*. Because $s_{S2Ng} > s_{Ig1}$ unless either $s_{g1} = 0$ or $s_{g2} = 0$, the signal-to-noise ratio statistic penalizes those genes that have large variances in both conditions compared to those genes that have a large variance in one class and a low variance in the other. The justification for this choice is that when a gene is differentially expressed in the two conditions, it is biologically reasonable to expect expression data to be distributed with very different variances. One objection to this justification is that one is interested in the distribution of the t-statistic under the null hypothesis of no differential expression.

Other forms of penalization are justified by the fact that because of the wide range of measurements, the estimate s_{Ig1} may be very small for some gene g and may produce an inflated value of the t-statistic. Therefore, authors have suggested estimating σ_g by $a + s_{Ig1}$ and choosing the constant a to minimize the coefficient of variation of the t-statistic (Tusher, Tibshirani and Chu, 2000), while Efron, Tibshirani, Storey and Tusher (2001) suggested replacing a by the 90th percentile of the standard error of all the genes.

The most popular approach to choosing a threshold is distribution-free. The main idea is to compute the value of the t-statistic from the data in which the sample labels that represent the experimental conditions are randomly reshuffled. By repeating this process several times, it is possible to construct the empirical distribution of the t-statistic under the null hypothesis of no differential expression. From the empirical distribution function, one can select a gene-specific threshold to reject the null hypothesis with a particular significance. This method is implemented in the popular program GeneCluster 2.0b (available at http://www-genome.wi.mit.edu/cancer/software/genecluster2/gc2.html) that conducts the differential analysis based on the signal-to-noise ratio statistic or the standard t-statistic in which $s_g = s_{Ig1}$. The program Sam (available at http://www-stat.stanford.edu/~tibs/SAM/index.html) implements a distribution-free differential analysis using the t-statistic with denominator $a + s_{Ig1}$. Because of the large number of genes, algorithms also have been developed for multiple comparison adjusted p-values; see, for example, Dudoit, Yang, Callow and Speed (2002).

Distribution-free methods tend to be widely used in practice, although few authors have suggested making distribution assumptions on the gene expression data. For example, Baldi and Long (2001) introduced a Bayesian parametric version of the analysis based on the t-statistic, in which expression data transformed in logarithmic scale are assumed to follow a normal distribution. Usually, the t-statistic is applied to data that are normalized using one of the methods described in Section 5.2. A model-based approach to simultaneously normalize and estimate the difference of gene expression between two experimental conditions was presented in Thomas, Olson, Tapscott and Zhao (2001). Although their integrated modeling approach is appealing, a limitation is the large sample approximate distribution for the t-statistic.

7. ANALYSIS OF MULTIPLE CONDITIONS

Some of the most interesting applications of microarray technology are based on data collected under multiple experimental conditions. These conditions can be, for example, different known classes of the same tumor—such as acute leukemia (Golub et al., 1999) or non-Hodgkin's lymphoma (Alizadeh et al., 2000)—or controlled experimental factors such as sex and age (Jin et al., 2001). The different experimental conditions can also be time points when the experimenter wishes to analyze the evolution of a physiological response (Iyer et al., 1999), identify genomic features of a cell cycle (Pilpel, Sudarsanam and Church, 2001) or track down the genetic mechanisms that switch a locally growing tumor into a metastatic killer (Clark, Golub, Lander and Hynes, 2000). These different experiments are designed to answer different questions and they require different data analysis tools.

7.1 Main Objectives

Data are typically collected in a $G \times n$ array Y, where G is the number of genes that have expression levels measured in each of the n samples. Each row $y_g = (y_{g1}, \ldots, y_{gn})$ collects the expression level y_{gj} for gene g measured in the n samples, while each column $e_j = (y_{1j}, \ldots, y_{Gj})$ collects the expression level of the G genes in sample j. The expression levels can be either absolute or relative with respect to a common reference sample. The n samples are typically collected from $c \leq n$ conditions. We will continue to denote by n_i the number of samples taken in each condition i, so that $n = \sum_{i=1}^{c} n_i$. The main experimental goals of multiple microarray experiments fall neatly into two broad classes:

Class prediction. The experimenter chooses c conditions and measures repeatedly the expression level of

the same set of genes in each condition. Each condition is regarded as a class label and the goal of the analysis is to detect the genes that are differentially expressed in at least two conditions or that are good predictors of the class. The analysis described in Section 6 is a particular example of this type of analysis, although its goal is mainly to "describe" the molecular differences of two conditions. In cancer genomic experiments, for example, the goal may be the development of new diagnostic tools based on the molecular profiles of tumor cells. To do this, the experimenter may collect samples from patients known to be affected by different types of the same tumor class—such as different types of leukemia (Golub et al., 1999) or breast cancer (West et al., 2001)—and use each patient sample as an instance of the molecular profile of the specific type of tumor. The goal of the analysis would be to determine the molecular profile of each type of tumor to make possible a molecular-based diagnosis of a specific tumor (Lakhani and Ashworth, 2001).

Class discovery. Multiple microarray experiments can also be used to help investigators create new classifications by discovering new classes characterized by a specific molecular profile. There is little doubt that the current taxonomy of cancer lumps together molecularly distinct diseases with distinct clinical phenotypes, with the consequence that patients who receive the same diagnosis can have different clinical courses and treatment responses (Alizadeh et al., 2000). For example, in the analysis of gene expression data collected from tissues of breast cancer patients, the goal may be the identification of new molecular taxonomies of breast cancers characterized by particular profiles. Again, the advantage of such discovery could be to aid the diagnosis, as well as to tailor treatments to more specific diagnoses. Sometimes, the distinction among different classes is observable only through the dissection of the dynamics of the genomic system. In these cases, the different conditions are represented by time points and the goal is to identify groups of genes that behave in a similar way.

The solution to class prediction problems requires the development of classification rules able to label the molecular profile of a sample, whereas the goal of class discovery studies is to create new classes from the available data. Formally, the distinction between the two tasks is that the former relies on a labeled data set, while the latter relies on an unlabeled data set. Supervised and unsupervised machine learning methods are currently used to tackle both tasks.

7.2 Supervised Classification

Supervised classification techniques are used to learn a classification rule from a set of labeled cases (called the *training set*) to classify new unlabeled cases in a *test set*. Each condition i is regarded as a class label, and the columns of the data matrix Y are the labeled cases used to learn mappings of molecular profiles to class labels. This mapping can be constructed in two ways. One approach models the dependency of the class labels on the gene expression, and this dependency is used to compute the probability of each class label given its molecular profile. The classification can be based on a decision rule that selects a class by minimizing the expected loss. We call this approach model-based in contrast to a model-free approach that partitions the space of gene expression data so that each element of the partition corresponds to one and only one class label. Well known model-based classification methods are multinomial logistic or probit regression (McCullagh and Nelder, 1989) and naive Bayes classifiers (Hand, 1997). In multinomial logistic/probit regression, the probability distribution of the class labels $p(i|y_1, \ldots, y_G)$, $i = 1, \ldots, c$, is modeled as

$$p(i|y_1, \ldots, y_G) = F^{-1}\left(\beta_0 + \sum_g \beta_g y_g\right),$$

where F is the cumulative distribution function of the logistic distribution or of the standard normal distribution and β_g are regression parameters. The probabilities are estimated directly from the training set and to classify a case with known gene expression data, say $y_1, \ldots, y_G$, it is sufficient to compute the probability $p(i|y_1, \ldots, y_G)$ for all i and select the class with maximum probability. The classification rule can be adjusted to account for misclassification costs. A difficulty with this approach, known as the "small n large p" problem, is the typical sparseness of microarray data, which often consist of thousands of genes (large p) and few observations for each gene (small n). A Bayesian method for fitting probit regression and tackling the "small n large p" problem was proposed by West et al. (2001) for the classification of different types of breast cancers.

Naive Bayes classifiers rely on the assumption that expression measurements within a microarray are conditionally independent given the class membership, so that the stochastic dependency between class labels and gene expression values can be modeled as

$$p(i, y_1, \ldots, y_G) = p(i) \prod_g p(y_g|i),$$

where $p(y_g|i)$ is the density function of the expression level of gene g in class i and $p(i)$ is the marginal probability of the ith class. Once the terms $p(i)$ and $p(y_g|i)$ are estimated from the training data, it is possible to predict the class of a new unlabeled case by computing the posterior distribution of the class labels given the gene expression values observed in the new case. The conditional independence assumption of the classifier simplifies the dependency structure of the class labels on the gene expression data, and the classification rule can be learned efficiently and accurately, despite the small number of observations available for each gene (Keller, Schummer, Hood and Ruzzo, 2000).

The classification accuracy of both regression and naive Bayes classifiers can be improved by selecting the subset of genes with the highest predictive accuracy. In logistic regression, for example, the selection of genes can be done by using standard large sample model selection techniques, which are reliable when the number of observations for each pair (y_g, i) is at least 25 (McCullagh and Nelder, 1989). Similar feature selection methods are available for the naive Bayes classifier (Mitchell, 1997). However, the staggering cardinality of the model space requires the adoption of heuristic search strategies. For example, if one limits attention to the set of all additive logistic regression models, the cardinality of the model space would be 2^G, where G can be as large as 12,625 as in the case of experiments carried out with the Affymetrix Human Genome U95A chip.

Although model-based approaches provide a quantification of the uncertainty of the predictive model and a principled way to select a subset of the most predictive genes, model-free approaches are currently the most popular. Examples of model-free approaches to classification are methods for discriminant analysis such as Fisher linear discriminant analysis, nearest neighbor classification trees (Hand, 1997) and support vector machines (Vapnik, 1998). A comprehensive review of classical statistical methods for discriminant analysis applied to gene expression-based tumor discrimination was presented by Dudoit, Fridlyand and Speed (2002) and gives a critical assessment of the pros and cons of each method. The selection of genes with predictive properties is often based on heuristic rules, such as filtering out genes with a fold-change that does not exceed a particular threshold (Tamayo et al., 1999) or selecting genes that are highly correlated with a dummy pattern of 0's and 1's that mirrors the class partition (Golub et al., 1999).

Support vector machines are a supervised classification technique that is increasingly popular and that uses the training data Y in which genes known to belong to the same functional class are assigned the same class label, say $i = 1$, and genes known not to be members of that class are assigned the same different class label, say $i = -1$. The two-labeled data constitute the training set for the support vector machine that is used to learn to distinguish between members and nonmembers of the functional class on the basis of their expression data. Formally, a support vector machine maps the binary labeled training data $y_1, \ldots, y_G$ into a high-dimensional feature space F, where $f_g = \phi(y_g)$. In the feature space F, the two classes of data are separated by a hyperplane (w, b) with maximum margin γ. The optimal solution is known to be $w = \sum_g \alpha_g i_g \phi(y_g)$, where i_g is the label assigned to the gene g and the parameters α_g are positive real numbers chosen to maximize the function

$$\sum_g \alpha_g - \sum_{gh} \alpha_g \alpha_h i_g i_h \langle \phi(y_g), \phi(y_h) \rangle$$

$$\text{subject to} \quad \sum_g \alpha_g i_g = 0,$$

where $\langle \phi(y_g), \phi(y_h) \rangle$ is the dot product in the feature space. The real number b is found by maximizing the hyperplane margin:

$$\gamma = \min_g i_g \{ \langle w, \phi(y_g) \rangle - b \}.$$

Having learned the expression features of the two classes, the support vector machine can be used to recognize and classify the genes in the data set on the basis of their expression (Brown et al., 2000). The classification is based on the decision function

$$\begin{aligned} d(y) &= \text{sign}(\langle w, \phi(y) \rangle - b) \\ &= \text{sign}\left(\sum_g \alpha_g i_g \langle \phi(y_g), \phi(y) \rangle - b \right), \end{aligned}$$

so that if the decision function for the new gene with expression profile y is $d(y) > 0$, the gene is assigned to the same functional class of the genes labeled by $i = 1$ in the training set. Note that the parameter α_g associated with the profile y_g expresses the weight that this point has on the decision function. Particularly, only a subset of the initial training point will have nonzero weights α_g. These points are called the support vectors. Because both the learning algorithm and the decision function depend on the dot product $\langle \phi(y_g), \phi(y_h) \rangle$, the specification of the map $\phi(\cdot)$ can be done indirectly

via the kernel function $K(x, y) = \langle \phi(x), \phi(y) \rangle$. Typical kernel functions are the dot product, when $\phi(\cdot)$ is the identity, and some power or exponential function of the dot product.

7.3 Unsupervised Classification and Clustering

Unsupervised classification techniques, such as clustering or multidimensional scaling, can be used to group either genes with a similar expression profile or samples (e.g., patients) with a similar molecular profile, or both. The average-linkage hierarchical clustering proposed by Eisen, Spellman, Brown and Botstein (1998) is today one of the most popular analytical methods to cluster gene expression data. Given a set of n expression values measured for G genes, this approach recursively clusters genes, or samples, according to some similarity measure of their measurements. When applied to gene expression profiles, the method treats each row of the $G \times n$ data matrix Y as an n-dimensional vector and iteratively merges genes into a single cluster. Relationships among the genes are represented by a tree (*dendrogram*), the branch lengths of which reflect the degree of similarity between the genes. The similarity measure commonly used is the correlation between pairs of gene expression data, but other measures have been used, such as Euclidean distance or information-theoretic metrics. The resulting tree sorts the genes in the original data array Y, so that genes or groups of genes with similar expression patterns will be adjacent. The ordered table can be displayed graphically, together with the dendrogram, for the investigators' visual inspection. Figure 8 provides an example of such a graphical display, which is known as an Eisen plot. Software for the cluster analysis and visualization is available from the Eisen Lab web page (http://rana.lbl.gov/EisenSoftware.htm).

The same approach can be applied to the columns of the data matrix to identify samples with a similar molecular profile. Hierarchical clustering applied to the rows and columns of the data array Y will return a sorted image of the original data. The image of the sorted data is typically used to support the operation of partitioning genes or samples into separated groups with common patterns. This operation is done by visual inspection, by searching for large contiguous patches of color that represent groups of genes that share similar expression patterns or groups of samples that share similar molecular profiles. Identification of these patches allows the extraction of subgroups of genes to be used to recluster the samples and, conversely, the extraction of subgroups of experiments to be used

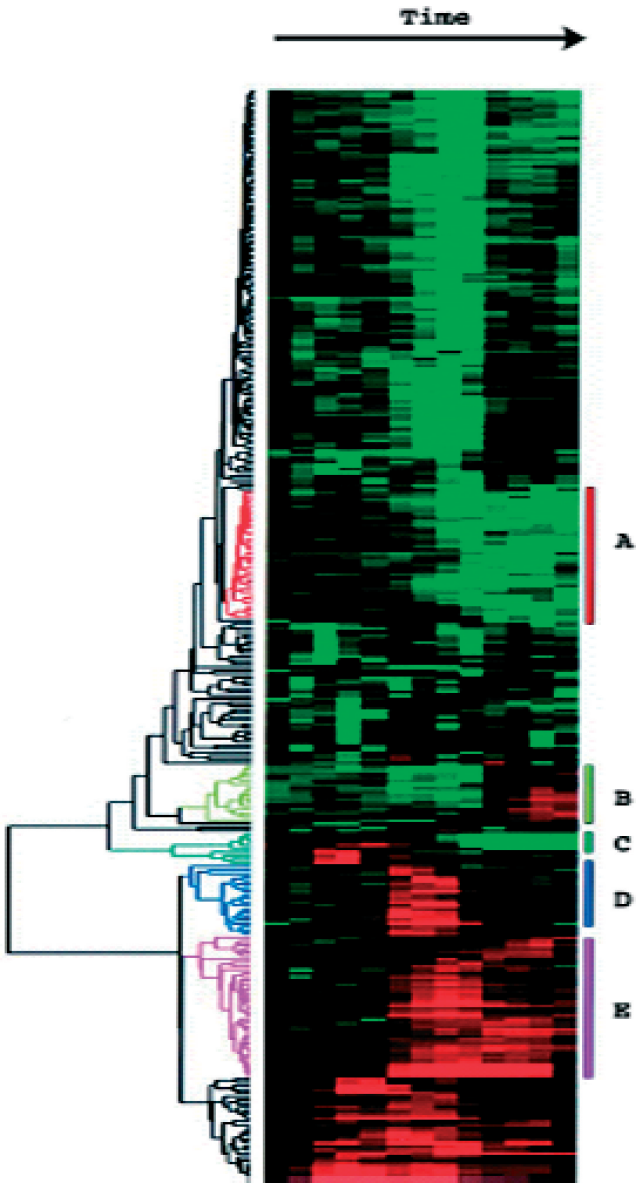

FIG. 8. *Example of an Eisen plot applied to* 517 *gene expression data measured in* 13 *experiments displaced along time. The image is a graphical display of the data array* Y *with rows sorted using the average-linkage hierarchical clustering procedure. Each row of the image represents a gene and each column represents an experiment. Each cell* (g, j) *of the image represents the fold-change of gene* g*, relative to the first time point expression value, in logarithmic scale. Cells with log fold-change equal to* 0 *are colored black, increasingly positive log fold-changes are reds of increasing intensity and increasingly negative log fold-changes are greens of increasing intensity. A representation of the dendrogram is appended to the image. Contiguous patches of color, labeled by the investigators with the letters A, B, C, D and E, are taken to indicate groups of genes that share similar expression patterns. The image is reproduced from Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. U.S.A.* **95** 14863–14868, *with permission. Copyright (1998) National Academy of Sciences, U.S.A.*

to recluster gene expression patterns. Although the choice of subsets is arbitrary and the final result heavily depends on the genes or samples selected at each step of the procedure, this method has been successfully applied to identify, for example, new molecular classes of non-Hodgkin's lymphoma (Alizadeh et al., 2000), cutaneous malignant melanoma (Bittner et al., 2000), breast cancer (Sorlie et al., 2001) and lung cancer (Bhattacharjee et al., 2001).

Notwithstanding these interesting results, this approach is not without problems. The subjective nature of partitioning by visual inspection may lead one to disregard some important information or to include irrelevant information. Decades of cognitive science research have shown that the human eye tends to overfit observations, selectively discount variance and "see" patterns in randomness (Tversky and Kahneman, 1974; Gilovich, Vallone and Tversky, 1985). Permutation tests are sometimes used to validate the partitions found by this procedure (Eisen et al., 1998) and a bootstrap-based validation technique is presented in Kerr and Churchill (2001a). The gap statistics of Tibshirani, Walther and Hastie (2001) can also be used to find the optimal number of groups in the data. A second problem of this approach is the dilution of distance measures in average-linkage hierarchical clustering. When genes are assigned to the same subtree, the similarity measure between subtrees or between single genes and subtrees is computed by using a subtree profile calculated as the average of the subtree member profiles. As the subtree grows, this average profile becomes a less adequate representation of the subtree members. A solution to this problem can be the adoption of single-linkage clustering or complete-linkage clustering (Quackenbush, 2001).

Relevance networks (Butte et al., 2000) are a non-hierarchical clustering method which does not suffer from this dilution problem. For each pair of genes, the method computes a similarity between their expression measures, such as correlation or mutual information on appropriately discretized expression measures, and assigns genes that have a similarity measure above a preset threshold to the same cluster. This method can be regarded as a graphical representation of the matrix of all pairwise distances between gene expression profiles, since genes assigned to the same cluster are linked by an edge that has a thickness proportional to the similarity between the two elements. Although this method does not rely on visual inspection, the division into clusters is entrusted to an arbitrary threshold.

When some prior knowledge about the number of groups in the data is available, k-means clustering can be used as an alternative to hierarchical clustering to provide an optimal grouping of rows and/or columns of the data array Y into a preset number of clusters. The k-means clustering starts with a random assignment of the rows (columns) of the data matrix into k disjoint groups, and the rows (columns) are iteratively moved among the clusters until a partition with optimal properties is found. Typically, the criterion to find the optimal partition is to minimize the within-cluster variability while maximizing the between-cluster variability. The within-cluster variability is measured by the average distance between cluster members and the cluster profile, while the between-cluster variability is a measure of the distance of each cluster member from the other cluster profiles. Tavazoie et al. (1999) used k-means clustering to identify groups of genes with similar patterns across different experimental conditions. Similar to k-means clustering are the self-organizing maps of Kohonen (1997). A self-organizing map uses a two- or three-dimensional projection of each cluster profile and provides a straightforward graphical representation of the result. Self-organizing maps have been used to identify classes of genes with similar functions in the yeast cell cycle (Tamayo et al., 1999) and have been combined with the nearest neighbor classification method to discriminate between two types of acute leukemia (Golub et al., 1999). An implementation of the method is GeneCluster 2.0b.

One potential danger of searching an optimal sorting of the data array Y by independently looking for an optimal arrangement of rows and columns is to overlook the association between gene expression data and samples. Clustering methods that address the issue of simultaneously sorting rows and columns of the matrix Y have been proposed, such as gene shaving (Hastie et al., 2000), biclustering (Cheng and Church, 2000), coupled two-way clustering (Getz, Levine and Domany, 2000) and the plaid model (Lazzeroni and Owen, 2002). Gene shaving is a block clustering technique for clustering genes and samples simultaneously. The algorithm uses an iterative procedure to identify subsets of highly correlated genes that vary greatly between samples. Biclustering is a method for simultaneously clustering genes and samples by using a similarity measure of genes and samples. The idea of coupled two way clustering is to cluster pairs of small subsets of genes and samples. The rationale of this approach is that only a small subset of the genes is expected to participate in any cellular processes, which

by themselves are supposed to take place only in a subset of the samples. Therefore, the algorithm looks for pairs of a relatively small subset of genes and samples that yield stable and significant partitions. The plaid model is a block clustering technique that produces overlapping clusters.

All these clustering methods are model-free: they do not rely on any assumptions about the distribution of genes or samples. In contrast, model-based procedures (Banfield and Raftery, 1993; Cheeseman and Stutz, 1996) regard clustering as the task of merging the observations generated by the same probability distribution. Cast in this framework, the simultaneous clustering of genes and samples can be regarded as the task of identifying a hidden variable that labels the cells of the array Y. In this way, the problem of simultaneously grouping rows and columns could be solved by estimating the hidden variable and, subsequently, by finding the genes and the samples that share the same label. If we let H be the hidden variable that assigns the same label (r, c) to similar cells of Y, then the likelihood function of the data matrix Y, conditional on a known labeling h of rows and columns, can be represented as

$$p(Y|h, \theta) = \prod_r \prod_c \prod_{g(r)} \prod_{j(c)} p(y_{g(r)j(c)}|\theta_{r,c}),$$

where $\theta = \{\theta_{r,c}\}$. The index $g(r)$ specifies the genes assigned the same label r, whereas the index $j(c)$ specifies the samples assigned the same label c, and $p(y_{g(r)j(c)}|\theta_{r,c})$ is the density function of the genes and samples assigned the same label pair (r, c). The overall likelihood can then be written as $\sum p(Y|h, \theta)p(h|\eta)$, where $p(h|\eta)$ is the probability that $H = h$ that depends on parameters η. The EM algorithm can be used to estimate the unknown parameters for a specification of the density function $p(y_{g(r)j(c)}|\theta_{r,c})$ and the probabilities $p(h|\eta)$. Alternatively, if some initial labeling of the experiments is available, an agglomerative clustering procedure can be used to iteratively relabel rows and columns. Some relevant work in this area was presented in Yeung et al. (2001) for one-dimensional clustering and in Bhattacharjee et al. (2001). Although model-based clustering relies on distributional assumptions of gene expression profiles and samples, the validity of these assumptions can be assessed using statistical validation techniques. One of the main advantages of a model-based approach is the possibility of using sound statistical methods to assess the significance of the similarity between genes or samples and to identify the best number of clusters consistent with the data (Fraley and Raftery, 2002).

7.4 Time Series Analysis

Several applications of genomewide clustering methods focus on the temporal profiling of gene expression. The intuition behind this analytical approach is that genes that show a similar expression profile over time are acting together, because they belong to the same or, at least similar, functional categories. Temporal profiling offers the possibility of observing the regulatory mechanisms in action and tries to break down the genome into sets of genes that are involved in the same, or at least related, processes. However, the clustering methods described in the previous section rest on the assumption that the set of observations for each gene is exchangeable over time: pairwise similarity measures, such as correlation or Euclidean distance, are invariant with respect to the order of the observations and if the temporal order of a pair of series is permuted, these distance measures will not change. While this assumption holds when expression measures are taken from independent biological samples, it may no longer be valid when the observations are a time series.

Although the functional genomic literature is becoming increasingly aware of the specificity of temporal profiles of gene expression data, as well as of their fundamental importance in unravelling the functional relationships between genes (Clark, Golub, Lander and Hynes 2000; Coller et al., 2000; International Human Genome Sequencing Consortium, 2001), traditional clustering methods are still used to group genes on the basis of their similarity. For example, Holter et al. (2001) described a method for characterizing the time evolution of gene expression levels by using a time translational matrix to predict future expression levels of genes based on their expression levels at some initial time, thus capturing the inherent dependency of observations in time series. This approach relies on the clustering model obtained using a timeless method, such as singular value decomposition (Alter, Brown and Botstein, 2000), and then infers a linear time translational matrix for the characteristic modes of these clusters. The advantage of this approach is that it provides, via the translational matrix, a stochastic characterization of a clustering model that takes into account the dynamic nature of temporal gene expression profiles. However, the clustering model which this method relies upon is still obtained by disregarding the dynamic nature of the observations, while we expect that different assumptions on the correlation between temporal observations will affect the way in which gene profiles are clustered together.

When the goal is to cluster gene expression patterns measured at different time points, the observations for each gene are serially correlated and clustering methods should take into account this dependency. The method of Ramoni, Sebastiani and Kohane (2002) is a Bayesian model-based approach to cluster temporal gene expression patterns that accounts for the temporal dependencies using autoregressive models. The method represents gene expression dynamics as autoregressive equations and uses an agglomerative procedure to search for the most probable set of clusters, conditional on the available data. Features of this method are the ability to take into account the dynamic nature of gene expression time series during clustering and a principled way to identify the number of distinct clusters. As the number of possible clustering models grows exponentially with the number of observed time series, a distance-based heuristic search procedure is used to render the search process feasible. In this way, the method retains the important visualization capability of hierarchical clustering but acquires an independent measure to decide when two series are different enough to belong to different clusters. Furthermore, the reliance of this method on an explicit statistical model of gene expression dynamics makes it possible to use standard statistical techniques to assess the goodness of fit of the resulting model and validate the underlying assumptions. When the autoregressive order is equal to zero, this method subsumes, as a special case, model-based clustering of atemporal (i.e., independent) observations. The method is implemented in the program Caged (available at http://www.genomethods.org/caged) described in Sebastiani, Ramoni and Kohane (2003).

8. OPEN CHALLENGES

Microarray technology makes possible the simultaneous execution of thousands of experiments to measure gene expression levels in a variety of conditions. This article has reviewed the biology of gene expression, the technology of microarrays and several statistical issues involved in the analysis of gene expression data, including experimental design, data quality, data analysis and validation. Although a massive effort is under way to improve methods and technology, several issues are still open and are particularly relevant to the statistical community.

Experimental design. The design of a microarray experiment is an unprecedented challenge. The main character of microarray technology is to make possible the parallel execution of thousands of experiments that are not independent of each other. For example, measurements of the gene expression data are subjected to common experimental errors, such as those due to the amount of fluorescent dye used to label the target in each experimental replicate or the amount of mRNA in each sample target. The challenge is the design of parallel and dependent experiments that can exploit the full power of this technology. Because no agreement exists about the appropriate statistical analysis of gene expression data produced with microarrays and because many experiments with microarrays are conducted to generate rather than test hypotheses, critical experimental design questions are still far from being answered.

Quality assessment and normalization. A very important issue in analyzing gene expression data is the ability to assess whether the execution of an experiment was successful, that is, to evaluate the quality of the experimental data. By this we mean the ability to decide whether the effects of random components, such as variations in the amount of dye or variations of the mRNA samples, are not large enough to irremediably mask the signal in the data. The normalization and gene filtering techniques discussed in Section 5 seem to be ad hoc bias-correction procedures, but their effect is unclear and their use is questionable in many applications. Some initial efforts in this direction were presented by Hoyle, Rattray, Jupp and Brass (2002). They investigated whether probability distributions such as the Benford law of the first significant digit or the Zipf law can provide reference distributions to be used as the gold standard in data quality assessment. Although their results are very preliminary, they are suggestive and open the way to a general probabilistic means to measure the reliability and quality of microarray data.

Differential analysis. The last two years have witnessed an increasing number of research articles that propose methods for the differential analysis of gene expression data measured in comparative experiments. Many of these methods use one of the t-statistics described in Section 6 with an ad hoc chosen denominator, and the most disconcerting fact is the lack of empirical and theoretical studies to help choose the best method. The consequence seems to be that the choice of the differential analysis method is driven by the availability of software rather than the quality and appropriateness of the method. Furthermore, many of the original problems associated with gene expression data measured by the Affymetrix software MAS 4.0

have been overcome by the latest statistical software MAS 5.0 with a consequent change in research priorities. In particular, the improved quality of the data produced by the new software opens the way to the development of full parametric methods.

Survival analysis. While extensive work has been conducted to develop methods for the differential analysis of gene expression data measured in two conditions, very little is known about the analysis of gene expression data in which the training signal is a continuous variable. Particularly important to cancer genomics applications is the development of methods for the selection of genes that are predictive of the survival time of patients treated with a particular therapy. Some preliminary work is this area can be found in Nguyen and Rocke (2002) and Park, Tian and Kohane (2002).

Metric selection. An open issue in the analysis of gene expression data is the selection of the metrics most suitable to answer specific biological questions. As an example, popular clustering methods use correlation, Euclidean distance, Kullback–Liebler information distance and, typically, different distances to sort gene expression profiles in different ways. Similarly, the classification induced by support vector machines depends on the specification of the kernel function. An important contribution would be the development of a formal way to determine which metrics are most relevant, or robust, for different questions.

Does clustering provide the right answer? Clustering techniques are extremely popular tools for the comparative analysis of gene expression data collected in a variety of conditions. The main reason for using clustering methods is the intuition that co-regulated genes have similar patterns, or similar levels of expression (Eisen, Spellman, Brown and Botstein, 1998). However, clustering techniques by themselves cannot discover the dependency structure between genes. Popular knowledge representation formalisms such as Bayesian networks (Cowell, Dawid, Lauritzen and Spiegelhalter, 1999) and dynamic Bayesian networks seem to be the ideal modeling tool for capturing the dependency structure among genes. The big challenge is whether the data structure available—a large number of parameters for few observations—makes Bayesian networks induced from gene expression data reliable. The wealth of genomic information grows daily and one may imagine that full Bayesian methods could be used to integrate the data with prior knowledge in a coherent way. Some initial attempts are discussed in Friedman, Linial, Nachman and Pe'er (2000), Segal et al. (2001) and Yoo, Thorsson and Cooper (2002).

Validation. Validation of cluster analysis is a very important issue that deserves further attention. Because clusters of similar genes/experiments are often identified by visual inspection or by imposing arbitrary thresholds, an independent quantitative validation of the results is required to assess whether the clusters are indeed capturing the signal in the data. Permutation tests as in Bhattacharjee et al. (2001) or bootstrapping the results (Kerr and Churchill, 2001a), are often used to show that clustering applied to data in which the signal has been removed does not identify meaningful groups of genes/experiments. However, it is important to stress that these tests do not prove the functional validity of the groups identified in the data. An increasing number of studies use an independent biological validation of the identified groups (Alizadeh et al., 2000; Golub et al., 1999), but on such a small number of cases (e.g., 40 patients in Alizadeh et al., 2000), this validation does not seem to provide much support. Some authors have shown the validity of their results by using different clustering techniques (Bhattacharjee et al., 2001; Bittner et al., 2000). The development of sound validation tests ranks among the top priorities in the field.

Lander (1999) wrote that developing experimental designs able to take advantage of the full power of microarray technology is the challenge for biologists of this century, but he also acknowledged that the greatest challenges are fundamentally analytical. The newly born functional genomic community is in great need of tools for data analysis and visual display of the results, and the statistical community could offer an invaluable contribution toward efficient collection and use of functional genomic data.

ACKNOWLEDGMENTS

This research was supported by NSF Grant ECS-01-20309, by NHLBI Grant HL-99-024 and by the Genomics Core of Beth Israel Deaconess Medical Center, Boston, MA. The authors are grateful to Stefano Monti, Whitehead Institute, and Joseph Jerry, Department of Veterinary and Animal Sciences, University of Massachusetts, for their insightful comments, and to the referees and the Editor for their invaluable suggestions.

REFERENCES

AFFYMETRIX, INC. (2002). Statistical Algorithms Description Document. Available from http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf.

ALIZADEH, A. A., EISEN, M. B., DAVIS, R. E., MA, C., LOSSOS, I. S., ROSENWALD, A., BOLDRICK, J. C., SABET, H., TRAN, T., YU, X., POWELL, J. I., YANG, L., MARTI, G. E., MOORE, T., HUDSON, JR., J., LU, L., LEWIS, D. B., TIBSHIRANI, R., SHERLOCK, G., CHAN, W. C., GREINER, T. C., WEISENBURGER, D. D., ARMITAGE, J. O., WARNKE, R., LEVY, R., WILSON, W., GREVER, M. R., BYRD, J. C., BOTSTEIN, D., BROWN, P. O. and STAUDT, L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403** 503–511.

ALTER, O., BROWN, P. O. and BOTSTEIN, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. U.S.A.* **97** 10101–10106.

ALWINE, J. C., KEMP, D. J. and STARK, G. R. (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.

BALDI, P. and LONG, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes. *Bioinformatics* **17** 509–519.

BANFIELD, J. D. and RAFTERY, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49** 803–821.

BHATTACHARJEE, A., RICHARDS, W. G., STAUNTON, J., LI, C., MONTI, S., VASA, P., LADD, C., BEHESHTI, J., BUENO, R., GILLETTE, M., LODA, M., WEBER, G., MARK, E. J., LANDER, E. S., WONG, W., JOHNSON, B. E., GOLUB, T. R., SUGARBAKER, D. J. and MEYERSON, M. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. U.S.A.* **98** 13790–13795.

BITTNER, M., MELTZE, P., CHEN, Y., JIANG, Y., SEFTOR, E., HENDRIX, M., RADMACHER, M., SIMON, R., YAKHINI, Z., BEN-DOR, A., SAMPAS, N., DOUGHERTY, E., WANG, E., MARINCOLA, F., GOODEN, C., LUEDERS, J., GLATFELTER, A., POLLOCK, P., CARPTEN, J., GILLANDERS, E., LEJA, D., DIETRICH, K., BEAUDRY, C., BERENS, M., ALBERTS, D., SONDAK, V., HAYWARD, N. and TRENT, J. (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **406** 536–540.

BOWTELL, D. D. L. (1999). Options available—from start to finish—for obtaining expression data by microarray. *Nature Genetics* **21** 25–32.

BROWN, M. P. S., GRUNDY, W. N., LIN, D., CRISTIANINI, N., SUGNET, C. W., FUREY, T. S., ARES, JR., M. and HAUSSLER, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. U.S.A.* **97** 262–267.

BROWN, P. O. and BOTSTEIN, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nature Genetics* **21** 33–37.

BUTTE, A. J., TAMAYO, P., SLONIM, D., GOLUB, T. R. and KOHANE, I. S. (2000). Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl. Acad. Sci. U.S.A.* **97** 12182–12186.

CAUSTON, H. C., REN, B., KOH, S. S., HARBISON, C. T., KANIN, E., JENNINGS, E. G., LEE, T. I., TRUE, H. L., LANDER, E. S. and YOUNG, R. A. (2001). Remodeling of yeast genome expression in response to environmental changes. *Molecular Biology of the Cell* **12** 323–337.

CHEESEMAN, P. and STUTZ, J. (1996). Bayesian classification (AutoClass): Theory and results. In *Advances in Knowledge Discovery and Data Mining* (V. M. Fayyad et al., eds.) 153–180. MIT Press, Cambridge, MA.

CHEN, Y., DOUGHERTY, E. R. and BITTNER, M. L. (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics* **2** 364–374.

CHENG, Y. and CHURCH, G. M. (2000). Biclustering of expression data. In *Proc. 8th International Conference on Intelligent Systems for Molecular Biology* 93–103. AAAI Press, Menlo Park, CA.

CHURCHILL, G. A. and OLIVER, B. (2001). Sex, flies and microarrays. *Nature Genetics* **29** 355–356.

CLARK, E. A., GOLUB, T. R., LANDER, E. S. and HYNES, R. O. (2000). Genomic analysis of metastasis reveals an essential role for RhoC. *Nature* **406** 532–535.

COLLER, H. A., GRANDORI, C., TAMAYO, P., COLBERT, T., LANDER, E. S., EISENMAN, R. N. and GOLUB, T. R. (2000). Expression analysis with oligonucleotide microarrays reveals that MYC regulates genes involved in growth, cell cycle, signaling, and adhesion. *Proc. Natl. Acad. Sci. U.S.A.* **97** 3260–3265.

COWELL, R. G., DAWID, A. P., LAURITZEN, S. L. and SPIEGELHALTER, D. J. (1999). *Probabilistic Networks and Expert Systems*. Springer, New York.

COX, D. R. and REID, N. (2000). *The Theory of the Design of Experiments*. CRC Press, Boca Raton, FL.

CRICK, F. H. C. (1970). Central dogma of molecular biology. *Nature* **227** 561–563.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 1–38.

DERISI, J. L., IYER, V. R. and BROWN, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278** 680–686.

DUDOIT, S., FRIDLYAND, J. and SPEED, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Statist. Assoc.* **97** 77–87.

DUDOIT, S., YANG, Y. H., CALLOW, M. J. and SPEED, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarrays experiments. *Statist. Sinica* **12** 111–139.

DUGGAN, D. J., BITTNER, M., CHEN, Y., MELTZER, P. and TRENT, J. M. (1999). Expression profiling using cDNA microarrays. *Nature Genetics* **21** 10–14.

EFRON, B., STOREY, J. D. and TIBSHIRANI, R. (2001). Microarrays, empirical Bayes methods, and false discovery rate. Technical report, Dept. Statistics, Stanford Univ.

EFRON, B., TIBSHIRANI, R., STOREY, J. D. and TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151–1160.

EISEN, M. B., SPELLMAN, P. T., BROWN, P. O. and BOTSTEIN, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.* **95** 14863–14868.

EKINS, R. and CHU, F. W. (1999). Microarrays: Their origins and applications. *Trends in Biotechnology* **17** 217–218.

FRALEY, C. and RAFTERY, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.* **97** 611–631.

FRIEDMAN, N., LINIAL, M., NACHMAN, I. and PE'ER, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology* **7** 601–620.

GETZ, G., LEVINE, E. and DOMANY, E. (2000). Coupled two-way clustering analysis of gene microarray data. *Proc. Natl. Acad. Sci. U.S.A.* **97** 12079–12084.

GILOVICH, T., VALLONE, R. and TVERSKY, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology* **17** 295–314.

GLYNNE, R., AKKARAJU, S., HEALY, J. I., RAYNER, J., GOODNOW, C. C. and MACK, D. H. (2000). How self-tolerance and the immunosuppressive drug FK506 prevent B-cell mitogenesis. *Nature* **403** 672–676.

GOLUB, T. R., SLONIM, D. K., TAMAYO, P., HUARD, C., GAASENBEEK, M., MESIROV, J. P., COLLER, H., LOH, M., DOWNING, J. R., CALIGIURI, M. A., BLOOMFIELD, C. D. and LANDER, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286** 531–537.

GRIFFITHS, A. J. F., MILLER, J. H., SUZUKI, D. T., LEWONTIN, R. C. and GELBART, W. M. (2000). *An Introduction to Genetic Analysis*, 7th ed. Freeman, New York (Available from http://www.ncbi.nlm.nih.gov/books/.)

HAND, D. J. (1997). *Construction and Assessment of Classification Rules*. Wiley, New York.

HASTIE, T., TIBSHIRANI, R., EISEN, M. B., ALIZADEH, A. A., LEVY, R., STAUDT, L., CHAN, W. C., BOTSTEIN, D. and BROWN, P. O. (2000). "Gene shaving" as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology* **1**(2) research0003.1-3.21.

HOLSTEGE, F. C. P., JENNINGS, E. G., WYRICK, J. J., LEE, T. I., HENGARTNER, C. J., GREEN, M. R., GOLUB, T. R., LANDER, E. S. and YOUNG, R. A. (1998). Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95** 717–728.

HOLTER, N. S., MARITAN, A., CIEPLAK, M., FEDOROFF, N. V. and BANAVAR, J. (2001). Dynamic modeling of gene expression data. *Proc. Natl. Acad. Sci. U.S.A.* **98** 1693–1698.

HOYLE, D. C., RATTRAY, M., JUPP, R. and BRASS, A. (2002). Making sense of microarray data distributions. *Bioinformatics* **18** 576–584.

IBRAHIM, J. G., CHEN, M. H. and GRAY, R. J. (2002). Bayesian models for gene expression with DNA microarray data. *J. Amer. Statist. Assoc.* **97** 88–99.

INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM (2001). Initial sequencing and analysis of the human genome. *Nature* **409** 860–921.

IRIZARRY, R. A., PARMIGIANI, G., GUO, M., DRACHEVA, T. and JEN, J. (2001). A statistical analysis of radiolabeled gene expression data. In *Proc. 33rd Symposium on the Interface: Computing Science and Statistics*. Interface Foundation of North America, Fairfax Station, VA.

IYER, V. R., EISEN, M. B., ROSS, D. T., SCHULER, G., MOORE, T., LEE, J. C. F., TRENT, J. M., STAUDT, L. M., HUDSON, JR., J., BOGUSKI, M. S., LASHKARI, D., SHALON, D., BOTSTEIN, D. and BROWN, P. O. (1999). The transcriptional program in the response of human fibroblasts to serum. *Science* **283** 83–87.

JACKSON-GRUSBY, L., BEARD, C., POSSEMATO, R., TUDOR, M., FAMBROUGH, D., CSANKOVSZKI, G., DAUSMAN, J., LEE, P., WILSON, C., LANDER, E. S. and JAENISCH, R. (2001). Loss of genomic methylation causes p53-dependent apoptosis and epigenetic deregulation. *Nature Genetics* **27** 31–39.

JACOB, F. and MONOD, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology* **3** 318–356.

JIN, W., RILEY, R. M., WOLFINGER, R. D., WHITE, K. P., PASSADOR-GURGEL, G. and GIBSON, G. (2001). The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nature Genetics* **29** 389–395.

KANE, M. D., JATKOE, T. A., STUMPF, C. R., LU, J., THOMAS, J. D. and MADORE, S. J. (2000). Assessment of the sensitivity and specificity of oligonucleotide (50 mer) microarrays. *Nucleic Acids Research* **28** 4552–4557.

KELLER, A. D., SCHUMMER, M., HOOD, L. and RUZZO, W. L. (2000). Bayesian classification of DNA array expression data. Technical Report UW-CSE-2000-08-01, Dept. Computer Science and Engineering, Univ. Washington, Seattle.

KERR, M. K. and CHURCHILL, G. A. (2001a). Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *Proc. Natl. Acad. Sci. U.S.A.* **98** 8961–8965.

KERR, M. K. and CHURCHILL, G. A. (2001b). Experimental design for gene expression microarrays. *Biostatistics* **2** 183–201.

KERR, K. M. and CHURCHILL, G. A. (2001c). Statistical design and the analysis of gene expression microarray data. *Genetical Research* **77** 123–128.

KOHANE, I. S., KHO, A. T. and BUTTE, A. J. (2002). *Microarrays for an Integrative Genomics*. MIT Press, Cambridge, MA.

KOHONEN, T. (1997). *Self Organizing Maps*, 2nd ed. Springer, Berlin.

LAKHANI, S. R. and ASHWORTH, A. (2001). Microarray and histopathological analysis of tumours: The future and the past? *Nature Reviews Cancer* **1** 151–157.

LANDER, E. S. (1999). Array of hope. *Nature Genetics* **21** 3–4.

LAZZERONI, L. and OWEN, A. B. (2002). Plaid models for gene expression data. *Statist. Sinica* **12** 61–86.

LEE, C. K., WEINDRUCH, R. and PROLLA, T. A. (2000). Gene-expression profile of the ageing brain in mice. *Nature Genetics* **25** 294–297.

LEE, M. T., KUO, F. C., WHITMORE, G. A. and SKLAR, J. (2000). Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl. Acad. Sci. U.S.A.* **18** 9834–9839.

LENNON, G. G. and LEHRACH, H. (1991). Hybridization analyses of arrayed cDNA libraries. *Trends in Genetics* **7** 314–317.

LIPSHUTZ, R. J., FODOR, S. P. A., GINGERAS, T. R. and LOCKHART, D. J. (1999). High density synthetic oligonucleotide arrays. *Nature Genetics* **21** 20–24.

LOCKHART, D. J. and BARLOW, C. (2001). Expressing what's on your mind: DNA arrays and the brain. *Nature Reviews Neuroscience* **2** 63–68.

LOCKHART, D. J., DONG, H., BYRNE, M. C., FOLLETTIE, M. T., GALLO, M. V., CHEE, M. S., MITTMANN, M., WANG, C., KOBAYASHI, M., HORTON, H. and BROWN, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology* **14** 1675–1680.

LOCKHART, D. J. and WINZELER, E. A. (2000). Genomics, gene expression and DNA arrays. *Nature* **405** 827–836.

LÖNNSTEDT, I. and SPEED, T. P. (2002). Replicated microarray data. *Statist. Sinica* **12** 31–46.

LY, D. H., LOCKHART, D. J., LERNER, R. A. and SCHULTZ, P. G. (2000). Mitotic misregulation and human aging. *Science* **287** 2486–2492.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.

MITCHELL, T. (1997). *Machine Learning*. McGraw–Hill, New York.

NADON, R. and SHOEMAKER, J. (2002). Statistical issues with microarrays: processing and analysis. *Trends in Genetics* **18** 265–271.

NATIONAL HUMAN GENOME RESEARCH INSTITUTE. (2001). Talking glossary of genetic terms. Available from http://www.genome.gov/glossary.cfm.

NEWTON, M. A., KENDZIORSKI, C. M., RICHMOND, C. S., BLATTNER, F. R. and TSUI, K. W. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8** 37–52.

NGUYEN, D. V. and ROCKE, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* **18** 39–50.

OLSHEN, A. B. and JAIN, A. N. (2002). Deriving quantitative conclusions from microarray expression data. *Bioinformatics* **18** 961–970.

PAN, W., LIN, J. and LE, C. T. (2001). A mixture model approach to detect differentially expressed genes with microarray data. Technical report, Division of Biostatistics, School of Public Health, Univ. Minnesota.

PAN, W., LIN, J. and LE, C. T. (2002). How many replicates of arrays are required to detect gene expression changes in microarrays experiments? A mixture model approach. *Genome Biology* **3**(5) research 0022.1-22.10.

PARK, P. J., TIAN, L. and KOHANE, I. S. (2002). Linking gene expression data with patient survival times using partial least squares. *Bioinformatics* **18** S120–S127.

PHIMISTER, B. (1999). Going global. *Nature Genetics* **21** 1.

PILPEL, Y., SUDARSANAM, P. and CHURCH, G. M. (2001). Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genetics* **29** 153–159.

QUACKENBUSH, J. (2001). Computational analysis of microarray data. *Nature Reviews Genetics* **2** 418–427.

RAMONI, M., SEBASTIANI, P. and KOHANE, I. S. (2002). Cluster analysis of gene expression dynamics. *Proc. Natl. Acad. Sci. U.S.A.* **99** 9121–9126.

RELOGIO, A., SCHWAGER, C., RICHTER, A., ANSORGE, W. and VALCARCEL, J. (2002). Optimization of oligonucleotide-based DNA microarrays. *Nucleic Acids Research* **30**(11) e51.

ROBERTS, C. J., NELSON, B., MARTON, M. J., STOUGHTON, R., MEYER, M. R., BENNETT, H. A., HE, Y. D., DAI, H., WALKER, W. L., HUGHES, T. R., TYERS, M., BOONE, C. and FRIEND, S. H. (2000). Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* **287** 873–880.

ROCKE, D. M. and DURBIN, B. (2001). A model for measurement error for gene expression arrays. *Journal of Computational Biology* **8** 557–569.

SABATTI, C., KARSTEN, S. L. and GESCHWIND, D. (2001). Thresholding rules for recovering a sparse signal from microarray experiments. *Math. Biosci.* **176** 17–34.

SCHADT, E. E., LI, C., SU, C. and WONG, W. H. (2000). Analyzing high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry* **80** 192–202.

SCHENA, M., SHALON, D., DAVIS, R. W. and BROWN, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270** 467–470.

SCHENA, M., SHALON, D., HELLER, R., CHAI, A., BROWN, P. O. and DAVIS, R. W. (1996). Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci. U.S.A.* **93** 10614–10619.

SEBASTIANI, P., RAMONI, M. and KOHANE, I. (2003). Bayesian model-based clustering of gene expression dynamics. In *The Analysis of Microarray Data: Methods and Software* (G. Parmigiani, R. Irizarry and S. L. Zeger, eds.). Springer, New York.

SEGAL, E., TASKAR, B., GASCH, A., FRIEDMAN, N. and KOLLER, D. (2001). Rich probabilistic models for gene expression. *Bioinformatics* **17** S243–252.

SMYTH, G. K., YANG, Y. H. and SPEED, T. P. (2003). Statistical issues in cDNA microarray data analysis. In *Functional Genomics: Methods and Protocols* (M. J. Brownstein and A. B. Khodursky, eds.). Humana, Totowa, NJ.

SORLIE, T., PEROU, C. M., TIBSHIRANI, R., AAS, T., GEISLER, S., JOHNSEN, H., HASTIE, T., EISEN, M. B., VAN DE RIJN, M., JEFFREY, S. S., THORSEN, T., QUIST, H., MATESE, J. C., BROWN, P. O., BOTSTEIN, D., EYSTEIN LØNNING, P. and BORRESEN-DALE, A. L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U.S.A.* **98** 10869–10874.

SOUTHERN, E., MIR, K. and SHCHEPINOV, M. (1999). Molecular interactions on microarrays. *Nature Genetics* **21** 5–9.

SPELLMAN, P. T., SHERLOCK, G., ZHANG, M. Q., IYER, V. R., ANDERS, K., EISEN, M. B., BROWN, P. O., BOTSTEIN, D. and FUTCHER, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell* **9** 3273–3297.

TAMAYO, P., SLONIM, D., MESIROV, J., ZHU, Q., KITAREEWAN, S., DMITROVSKY, E., LANDER, E. S. and GOLUB, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. U.S.A.* **96** 2907–2912.

TAVAZOIE, S., HUGHES, J. D., CAMPBELL, M. J., CHO, R. J. and CHURCH, G. M. (1999). Systematic determination of genetic network architecture. *Nature Genetics* **22** 281–285.

THOMAS, J. G., OLSON, J. M., TAPSCOTT, S. J. and ZHAO, L. P. (2001). An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Research* **11** 1227–1236.

TIBSHIRANI, R., WALTHER, G. and HASTIE, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *J. Roy. Stat. Soc. Ser. B Stat. Methodol.* **63** 411–423.

TUSHER, V. G., TIBSHIRANI, R. and CHU, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U.S.A.* **98** 5116–5121.

TVERSKY, A. and KAHNEMAN, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science* **185** 1124–1131.

VAPNIK, V. (1998). *Statistical Learning Theory*. Wiley, New York.

VELCULESCU, V. E., ZHANG, L., VOGELSTEIN, B. and KINZLER, K. W. (1995). Serial analysis of gene expression. *Science* **270** 484–487.

WEST, M., BLANCHETTE, C., DRESSMAN, H., HUANG, E., ISHIDA, S., SPANG, R., ZUZAN, H., OLSON, JR., J. A., MARKS, J. R. and NEVINS, J. R. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **98** 11462–11467.

WHITE, B. (1995). Southerns, Northerns, Westerns, and Cloning: "molecular searching" techniques. *MIT Biology Hypertextbook*. MIT Press. Available at http://web.mit.edu/esgbio/www/rdna/rdna.html.

WOLFINGER, R. D., GIBSON, G., WOLFINGER, E. D., BENNETT, L., HAMADEH, H., BUSHEL, P., AFSHARI, C. and PAULES, R. S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology* **8** 625–637.

WYRICK, J. J., HOLSTEGE, F. C. P., JENNINGS, E. G., CAUSTON, H. C., SHORE, D., GRUNSTEIN, M., LANDER, E. S. and YOUNG, R. A. (1999). Chromosomal landscape of nucleosome-dependent gene expression and silencing in yeast. *Nature* **402** 418–421.

YANG, Y. H., DUDOIT, S., LUU, P., LIN, D. M., PENG, V., NGAI, J. and SPEED, T. P. (2002). Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* **30**(4) e15.

YANG, Y. H., DUDOIT, S., LUU, P. and SPEED, T. P. (2001). Normalization for cDNA microarray data. In *Microarrays: Optical Technologies and Informatics* (M. L. Bittner, Y. Chen, A. N. Dorsel and E. R. Dougherty, eds.) 141–152. SPIE, Bellingham, WA.

YANG, Y. H. and SPEED, T. P. (2002). Design issues for cDNA microarray experiments. *Nature Reviews Genetics* **3** 579–588.

YEUNG, K. Y., FRALEY, C., MURUA, A., RAFTERY, A. E. and RUZZO, W. L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics* **17** 977–987.

YOO, C., THORSSON, V. and COOPER, G. F. (2002). Discovery of causal relationships in a gene-regulation pathway from a mixture of experimental and observational DNA microarray data. In *Proc. Pacific Symposium on Biocomputing* **7** 498–509. Available from http://psb.stanford.edu.

ZUZAN, H., BLANCHETTE, C., DRESSMAN, H., HUANG, E., ISHIDA, S., MARKS, J. R., NEVINS, J. R., SPANG, R., WEST, M. and JOHNSON, V. E. (2001). Estimation of probe cell locations in high-density synthetic-oligonucleotide DNA microarrays. Technical report, Institute of Statistics and Decision Sciences, Duke Univ., Durham, NC.

Comment

Henry V. Baker

The following monograph will attempt to highlight several issues raised by Sebastiani, Gussoni, Kohane and Ramoni regarding the use and analysis of DNA microarray experiments and will focus on noise associated with uncontrolled experimental variables, methods to reduce this noise, methods to reduce experimental error, and the use of unsupervised and supervised analysis methods.

Henry V. Baker is Professor, Department of Molecular Genetics and Microbiology, University of Florida, Gainesville, Florida 32610-0266 (e-mail: hvbaker@ufl.edu).

1. DNA MICROARRAYS—ONE OF THE GREAT UNINTENDED CONSEQUENCES OF THE HUMAN GENOME PROJECT

In the late 1980s as the Human Genome Project was being debated, its detractors argued that it would result in little more than a DNA sequence made up of A's, T's, G's and C's, if even that were possible. On the other hand, the proponents argued that the genomic sequence would provide valuable insights into biology that would have untold ramifications on human health. DNA microarray technology spawned from the Human Genome Project provides a window to the dynamic genome as it functions within cells to allow them to respond to their environment. DNA microarray technology has been compared to both the

telescope and the microscope. Both of those earlier technologies changed the way the world we live in was viewed. DNA microarray technology as a tool is changing the way the living world is viewed.

Early investigations utilizing microarray technology demonstrated the promise and potential of this new technology as a major research tool in the biological sciences. Unfortunately some early studies also served to illustrate potential pitfalls associated with improper experimental methodologies and inadequate or faulty computational analysis.

2. USE OF MEASUREMENTS OF HYBRIDIZATION SIGNAL INTENSITIES TO INFER GENE EXPRESSION INVOLVES MANY STEPS, SOME OF WHICH ARE POORLY UNDERSTOOD

Key to the design, analysis and interpretation of microarray experiments is the understanding that the parameter being measured, signal intensity of indirectly labeled probes, is many steps removed from the parameter being inferred, gene expression. A typical microarray experiment represents a large-scale physiological study in which cells are isolated, RNA is harvested, and labeled representations of the harvested RNA are prepared and then used in hybridization experiments to indirectly label the nucleic acid probes that constitute the array. The signal intensity of the label at each probe on the array is taken as a measure of gene expression for the genes specified by the probes on the array. It is also important to realize that the inferred gene expression is not that of a single cell, but rather that of a population of cells. In some cases the population of cells under investigation is composed of many different cell types, each of which may have varied expression profiles unique to itself.

Microarray experiments are sensitive, albeit indirect, assays capable of measuring the genomic response to subtle changes in the environment that occur during the RNA harvesting process. Uncontrolled experimental variables may be introduced at any step in the wet laboratory workup of microarray experiments and may add to observed variances from array to array. In designing microarray experiments, it is important to recognize areas where uncontrolled experimental variables may be introduced so that they may be guarded against, although in some cases they are unavoidable. In these instances variations in the experimental protocol should be documented.

Potential sources of uncontrolled experimental variables vary with individual applications. In clinical studies involving patient volunteers, the greatest potential for uncontrolled experimental variables exists. For instance, in clinical studies aimed at identifying gene expression differences between various tumors and normal tissue, uncontrolled experimental variables may include age of subject, diet, diurnal variations in gene expression, type of anesthesia used, length of ischemia prior to tissue removal, time from tissue removal to RNA stabilization and method of RNA isolation. Studies involving laboratory animals are potentially affected by similar sources of uncontrolled experimental variables as encountered in clinical studies except that it is usually within the investigator's power to control many of the variables such as animal demographics, diet, light–dark cycles and time of day at which material is harvested. In considering experiments with tissue or cell cultures, passage number also needs to be added to the list of potential sources of variation mentioned above for material harvested directly from animals. Important variables to consider in experiments aimed at measuring changes in gene expression in response to exposure to a drug or other stimulus include concentration and potency, length of exposure, time between stimulus administration and RNA harvest, and stabilization.

In addition to apparent gene expression differences associated with uncontrolled experimental variables, biases and artifacts may be introduced by virtue of the methods used at each step of the procedure, including cellular and tissue harvest, RNA isolation and labeling methods. Differential recovery of specific cell types from tissue may bias the gene expression profile observed for a particular tissue type. Likewise, RNA isolation protocols may introduce bias if they differentially recover membrane bound RNA versus soluble RNA. In one large study involving gene expression profiling of human leukocytes before and after *Staphylococcus aureus* enterotoxin B (SEB) treatment, the largest response variable was method of RNA isolation and not SEB stimulation, although with both RNA isolation protocols, gene expression differences due to SEB stimulation were readily apparent. In the spotted array realm, differential incorporation of nucleotide analogs that contain Cy3 and Cy5 during labeling reactions is well known, necessitating the need of dye swap experimental designs. Labeling reactions that involve limited amplifications of the target material, such as those used with Affymetrix GeneChips and some spotted array protocols, can result in skewering if unequal amplification occurs during the in vitro transcription reactions.

The last step in the indirect labeling of the array is the hybridization of the targets to the probes. The hybridization reaction is governed in large part by the specific sequences of the individual targets and probes, and is affected by the ability of the target to form secondary structures with itself and other molecules that may be present in the hybridization mixture. Target molecules that form extensive secondary structure with themselves tend to produce dimmer signals than targets that are devoid of secondary structure. Some hybridization protocols employ a target fragmentation step in an attempt to circumvent secondary structure problems. Other factors that may affect hybridization from experiment to experiment, and hence hybridization signal intensity, are temperature and duration of hybridization, both of which are important experimental parameters that should be highly controlled.

Microarray experiments by their nature are very complex experiments that indirectly provide a measure of gene expression. The steps between gene expression at the level of mRNA expression and measurement of the signal intensities of the probes arrayed are numerous and some are poorly understood. Yet with care it is possible to use microarrays as a tool to begin to discern the dynamic changes that occur within cells as they respond to their environment, but great precautions must be taken to avoid contaminating the data set with noise that results from uncontrolled experimental variables.

3. SPOTTED ARRAYS—ISSUES WITH SPOT MORPHOLOGY, CONCENTRATION AND QUALITY OF PROBES

Spotted array technology and quality has improved since spotted arrays were first introduced. Early experiments utilized cDNA products as probes, which presented a number of technical challenges, not the least of which was the production of the PCR products to be arrayed. A number of factors contribute to variances with spotted arrays. Use of cDNAs of gene coding sequences usually identified by the presences of open reading frames (orfs) resulted in probes of different lengths from one gene to the next. The size of genes can vary by more than an order of magnitude. For example, an orf for one gene may be 300 nucleotides in length and for another gene, 3000 nucleotides. If, during the labeling reactions, 1 labeled molecule were incorporated per 100 nucleotides, the orf of 300 nucleotides would incorporate 3 labeled molecules on average, while the orf of 3000 nucleotides would incorporate 30 labeled molecules. Longer probes are capable of hybridizing to more labels and hence, for a given amount of gene expression, longer genes would tend to have greater hybridization signal intensities than shorter genes that express mRNA transcripts at the same molar amounts. Spotting PCR fragments of cDNAs is giving way to the use of oligonucleotides of constant length, typically on the order of 60 to 70 mers, thereby eliminating differences in hybridization signal intensities due to differences in probe length.

Spot morphology and consistency are major quality control issues in the fabrication of spotted arrays. Initially, spotted arrays were prepared by dipping metal pins into DNA solutions of the various probes and repeatedly touching the pins to a solid support, be it glass slides or nylon membranes. The DNA solutions were deposited onto the surface of the solid support by virtue of the process of capillary attraction. This process tends to be imprecise and variable during the course of the arraying process. Size and shape of the spots are also affected by relative humidity at the time of spotting, and the method and duration of spot rehydration. Use of ink jet technology in the spotting process has served to reduce some of the variability inherent in the spotting process.

Competitive hybridization reactions with mixed labels (Cy3/Cy5) served to circumvent some of the problems associated with inconsistencies of spot size and morphology from one array to the next, since the fluorescence intensity of each label is measured at each probe on a single array. The resulting data typically are reported as a ratio or a log ratio of one dye labeled target to the other and not in terms of absolute fluorescence, thereby minimizing the effect of differences in spot size and morphology. The use of ratios or log ratios provides a difference measure, not an absolute measure of gene expression. In the case of arrays prepared on nylon membranes, where competitive hybridizations with fluorescent dyes cannot be used due to autofluorescence of the membrane, spot-to-spot variation from array to array can be circumvented to some extent by stripping and reusing arrays. The process of stripping hybridized target from the probe is harsh and can remove bound probe, thus leading to variances in the resulting data set due to loss of probe affixed to the array from one hybridization to the next.

4. NUMBER OF REPLICATES AND THE LEVEL AT WHICH TO REPLICATE

Microarray experiments are no different from any other experiment; for meaningful results, experiments must be replicated. The question that confronts most

serious investigators is not whether to replicate, but the number of replicates to perform and the level at which to replicate. Differences in gene expression due to uncontrolled experimental differences tend to dampen, while differences due to the controlled response variable tend to reinforce with replication. The number of replicates to perform is dependent, in part, on the noise associated with the system under study. For the simplest of experiments, such as those aimed at identifying differences between two cell lines, a minimum of three replicates per condition should be budgeted. Four replicates per condition are better and more appropriate if one is considering cross-validation methods such as leave-one-out cross-validation or other statistical validation measures.

The level of replication is dictated in part by the question being addressed. If one were interested in determining the variances in hybridization signal intensities associated with length of hybridization time, then technical replicates would be appropriate. One would want to have hybridization time as the sole variable and one would design an experiment where one preparation of labeled target was prepared and repeatedly hybridized to different arrays for various amounts of time. If, however, one wishes to determine the gene expression differences between different tumor types, say glioblastoma multiforme and anaplastic oligodendroglioma, then the level of replication should be at the biological level of the tumors. Different tumors should be assayed from different donors, holding all other variables constant, so that an inference can be made about gene expression patterns in a particular type of tumor. With clinical specimens more replicates are usually required than for laboratory studies utilizing cell lines or isogenic strains due to the higher coefficients of variation in hybridization signal intensities usually encountered with clinical material.

5. SUPERVISED VERSUS UNSUPERVISED ANALYSIS METHODS

The development of analytical methods for use on data sets derived from microarray expression studies is a rapidly changing and progressing field. Most investigators utilize a combination of supervised and unsupervised methods in their analysis, and the individual methods of analysis used are somewhat of an art form that varies from investigator to investigator.

The first level of microarray data analysis is usually supervised. One simply seeks to determine which genes are most affected by a particular condition or treatment protocol. In this line of endeavor the investigator makes use of the class labels of the samples, for example, wild type versus mutant, to determine the probes that display differential signal intensities. As Sebastiani et al. noted, early studies tended to rely on fold-change differences and not the use of statistics. In several studies published early in the microarray era it was not even clear that replicates were performed. Among reports that utilize p-values or estimates of error based on permutations of the data set in setting significance levels, the cutoff levels used remained largely arbitrary. In some cases p-values as low as $p = 0.05$ for arrays with greater than 12,000 probes have been used. With such a modest threshold, 600 probes would be expected to exceed the threshold by chance alone. Clearly for larger arrays a Bonferroni correction should be applied to the traditional p-value of 0.05, or a more stringent p-value such as $p = 0.001$ should be used or an estimate of the false discovery rate based on permutations of the data set should be included in setting significance cutoffs.

In many cases supervised analyses are used for the purpose of identifying probes that can be used for class prediction, for example, to diagnose and differentiate diseased from normal tissue. In this case the goal of the investigator is to identify probes that are predictors of the class labels, which then can be used in future studies to identify the nature of the specimen as normal or diseased using one or more of several prediction models. Investigators, however, should be aware that microarray experiments exemplify the "small n large p" trap. The number of probes on a typical microarray, tens of thousands, vastly exceeds the number of categories into which the arrays can be classified. Thus, by chance alone it is likely that many probes can be identified out of a typical data set that can distinguish between the small numbers of class labels in a typical study. Cross-validation studies and Monte Carlo simulations should be employed to gauge the significance of the probes identified as predictors.

Supervised analyses are only as good as the supervision applied. In cases where the class labels of the specimens are definitively known—as in comparing gene expression of a wild-type tissue versus tissue of a knockout organism, where the genotypes of the wild type and knockout are precisely known—supervised analyses can be very powerful. However, when phenotypic distinctions are subtle and class labels are known with less certainty—as is the case often encountered in the clinical setting, where highly skilled pathologists

may disagree over the diagnosis of a tumor as a particular cancer or grade—supervised analysis methods are hampered by misclassification errors at the supervision stage.

Unsupervised analysis methods, including hierarchical clustering, k-means clustering and self-organizing maps, can be used as tools for class discovery in situations where standard methods of assigning class labels are incomplete or inadequate. In situations where class labels can be assigned with impunity, as in studies designed to identify gene expression differences between wild-type and knockout animals, unsupervised cluster analysis can be used as an assessment of overall reproducibility of measurements between replicates. Experimental replicates should cluster together according to the controlled response variable in this case; that is, wild type with wild type and knockout with knockout. If replicates do not cluster together or if clustering occurs according to some other identifiable variable, such as date of tissue harvest or date of labeled target preparation, then responses to uncontrolled experimental variables are likely contaminating the data set, obscuring gene expression changes resulting from the controlled experimental variable.

6. SUMMARY

DNA microarray technology has established itself as a major new research tool for the analysis of gene expression. The inference of gene expression is indirect and involves many steps, each of which can become a source of noise if left uncontrolled. The nature and characteristics of microarray experiments present a number of challenges that must be overcome so as to obtain high quality data sets with low noise and high informational content. With appropriate experimental design, execution, and proper analytical and statistical methods, DNA microarray technology will likely take its place with the microscope as an invaluable tool in the biological sciences, providing a window with which to view the genome as it dynamically responds to changes in its intracellular and extracellular environment.

Comment

Gary A. Churchill

1. INTRODUCTION

The field of molecular biology has made tremendous advances in the half century since the discovery of the structure of the DNA molecule. The functions and mechanisms of nucleic acids have been unraveled through a series of singular, often elegant, experiments that had conclusions that did not require statistical interpretation. Indeed, it has been a commonly held view among molecular biologists that experiments that required statistical interpretation were not done well. Statisticians are equally guilty of ignoring developments in molecular biology and the two fields have gone their separate ways for all this time. Until now.

It seems that nothing has caught the attention of the statistical community like microarrays. The promise of large and highly structured data sets waiting to be mined for valuable information has caught on like a gold rush. The presence of more than 60 microarray talks at the most recent Joint Statistical Meeting is just one indication of the amount of statistical attention that this technology is receiving. Likewise, many biologists, who for years prided themselves on never having computed a t-test, suddenly find themselves in need of statistical advice on the interpretation of these large and complex data sets.

The intertwined histories of genetics and statistics go back to the very roots of both fields. Indeed, the grandfather of statistics, R. A. Fisher, is known to many primarily for his contributions in genetics. (Upon learning of the origin of the F-statistic, a surprised colleague once asked, "Do you mean to say that Fisher was a *statistician* too?") The deep connection of the past makes the present rift all the more important to bridge, but the divergence of goals, language and concepts presents challenges. As statistical concepts are adapted to applications in molecular biology, the terminology is often misused and, as a result, fundamental concepts are misconstrued; the problem is not one-sided. Statisticians are too often willing to accept the most superficial

Gary A. Churchill is a Staff Scientist affiliated with The Jackson Laboratory, Bar Harbor, Maine 04609 (e-mail: garyc@jax.org).

understanding of biological concepts which they spin into elaborate statistical models that have little connection with the data or the reality of biological processes. There is much to relearn on both sides.

I wish to express my admiration for the authors of this article for taking on the task of bridge building. Their review serves to connect two divergent fields of investigation. The task is daunting. Even in the relatively narrow context of gene expression microarray technology, which is the primary focus of the review, there are thousands of publications that express diverse and often conflicting opinions on complex issues, the background of which requires two Ph.D. degrees to grasp. My commentary may focus on some points of disagreement, but it is mostly an opportunity to voice some of my own opinions on the issues raised here. On the whole, Sebastiani, Gussoni, Kohane and Ramoni have done a remarkable job. A lot of ground is covered and one is appropriately left with the impression that we have a long way still to go. After all, there is 50 years of catching up to do between old friends.

2. GENE EXPRESSION TECHNOLOGIES

There are a number of new (and some not so new) methods to assay relative quantities of mRNA species in a sample. Most of these methods rely on the miraculous property of specific hybridization between complementary nucleic acid molecules. However, none of the hybridization-based methods is capable of absolute quantitation of mRNA species and none is accepted as a gold standard, even for relative measurements. Alternative approaches to mRNA quantitation, so-called tag count methods (Velculescu, Zhang, Volgelstein and Kinzler, 1995; Brenner et al., 2000), are based on sequencing of short signature tags which results in counts of mRNA molecules. Thus they promise to yield direct measurement of mRNA abundance with (apparently) fewer biases and greater depth. Tag count methods are not constrained by the availability of probe sequences and are enhanced by, but do not require, complete genome sequence data. They are time consuming and involve multiple steps (potential sources of variation and bias), but the same is true of microarray experiments. We should not dismiss these methods lightly and, for the statistician, they offer abundant opportunities to explore new methods of analysis.

The discussion of Affymetrix technology by Sebastiani et al. is too uncritical. The studies cited that claim superior performance of this system were either conducted by Affymetrix employees, using conditions (i.e., large sets of probe pairs) that are not used in practice, or by groups studying the properties of longer oligos in contexts also quite different from actual practice. The characterization of MAS 5.0 as a "well accepted data analysis protocol" suggests complacency in an area where skepticism should reign. Someone needs to question the fundamental principles on which these analyses are based. The role of the MM probe is one point of concern. A careful examination of the logic that justifies the MM probes is needed. What is being measured by the probes (PM and MM) and how should this information be combined? Some efforts to study the properties of oligonucleotide probe sets have been made (Li and Wong, 2001; Irizarry et al., 2003), but it appears that further critical investigation of this measurement system is needed.

Two-color cDNA microarrays present their own set of challenges, both in the wet lab and in their analysis. An underappreciated advantage of cDNA arrays is that they are inherently comparative. Pairing is a well established and powerful approach to controlling variation when experimental materials are heterogeneous (Fisher, 1951). Direct comparison of RNA samples on the same slide effectively eliminates an important source of variation—fluctuations in size and quality of the printed spots—which might otherwise contribute to noise in the measurements. Single color systems, such as Affymetrix arrays, must rely on tightly controlled production to minimize the between array component of variance. When there are more than two samples of interest in a cDNA microarray experiment, some comparisons must necessarily be indirect and the resulting experimental design has an incomplete blocking structure. This restriction of the two-color system has led to a practice of making all of the comparisons in a microarray experiment to a common reference sample, which can lead to inefficient experiments and potential biases (Kerr and Churchill, 2001b, c). Additional concerns arise because it is apparent that the relationship between mRNA concentration and signal intensity is not identical for the two most commonly used dyes. However, this problem is easily corrected by repeating hybridizations with dye labels reversed. Affymetrix arrays almost certainly suffer from the same nonlinearity dye effects; however, they are not as simple to detect or to correct.

Possibilities for new gene expression technologies simply await the next imaginative technique for manipulating nucleic acids. One thing is certain, the technology will change. If statisticians are going to play a role in these developments it will be best not to wait

for things to change, but rather to keep abreast and to anticipate. By the time we will have worked out all of the nuances of any one of the existing gene expression technologies, it will have become obsolete and everyone will be using the next greatest thing. The real challenge here is to stay one step ahead.

3. DESIGN OF EXPERIMENTS

Despite the trend toward "discovery driven" experimentation, it is my belief that the very best experiments are motivated by specific scientific questions (hypotheses). Such questions may involve interrogating thousands of genes, but the focus of a hypothesis helps one to avoid the situation of having assembled an uninterpretable hodgepodge of data. The goal of an experiment drives the choice of experimental units and the treatments or conditions under which they will be examined. A good experiment should include replication of units at appropriate levels in the design to ensure that valid estimates of error are available for assessing the significance of results. Many microarray experiments are carried out without appropriate replication. Quoting Fisher (1951), "perhaps these should not be called *experiments* at all, but be added merely to the body of *experience* on which, for lack of anything better, we may have to base our opinions" (emphasis added). The temptation to overinterpret microarray data must be kept in check and it should be the responsibility of the statistical community to set a good example in this regard.

What is appropriate replication? In most cases this will involve independent sampling of biological units to assure that inferences apply in a broad sense. In some situations this is not feasible or desired and technical replication, achieved by repeated measurement of the same biological material, can be used to provide a narrow sense inference that refers only to the samples in hand. The generalization of narrow sense inference to a wider biological context is not necessarily wrong, but it is subject to an unassessable degree of error. Technical replication plays an important role in microarray experiments by providing increased precision in the measurement of individual samples. Use of multiple arrays is the most effective method and it provides an appropriate variance estimate for narrow sense inferences. Repeated spotting of the same clone on a single microarray is rarely, if ever, appropriate for error estimation, but it can be effective for quality control and increased precision of measurements.

Although counterintuitive, increased variability among replicates can be a desirable property of an experiment. If, by our attempts to control variability, we introduce correlations into the data that are not properly accounted for, we run the risk of replicating biases (Rosenbaum, 2001). The recommendation by Sebastiani et al. to use the same pooled sample for experimental conditions is dangerously misleading. By repeatedly measuring the same pooled samples, we will amplify the apparent significance of both real and chance differences among the conditions. It would be better to use several independent pools and thereby obtain an estimate of the between pools variance.

Decisions regarding the allocation of resources, such as the relative balance of biological and technical replication in an experiment, can be informed by knowledge of the magnitude of variance components. One such analysis is summarized in Figure 1. This particular analysis illustrates the importance of technical replication (the largest variance components are between arrays) and suggests that pooling, at least of inbred mice, is not an effective variance reduction strategy. Last, certainly not first, one must make the choice of direct versus indirect comparisons. Direct comparisons are more efficient, but there can also be practical considerations that motivate indirect comparisons via a common reference sample (Yang and Speed, 2002; Churchill, 2002).

As statisticians we should not be fooled by the apparent novelty of microarray experiments. The classical principles of experimental design are based on sound arguments that apply to the new situation. Factorial designs are still the most efficient means for investigating the simultaneous effects of multiple experimental factors. Combinatorial strategies will be essential if we ever hope to tackle the enormous task of understanding the functions of all genes in a genome (Jansen, 2003). Left to the biologists, this task will be taken on by "knocking out" one gene at a time, an inefficient and inadequate strategy.

4. DATA TRANSFORMATIONS

The most important reason to log transform microarray data is to obtain a scale on which the sources of variation in the experiment are (roughly) additive. Logarithm is the obvious choice because of the assumed proportionality of effects in microarray experiments. Twice as much RNA should produce twice as much signal over a wide range of absolute quantities. A secondary goal of transformation is to obtain constant variance across the full intensity range of the signal. Logarithm tends to overcorrect variance heterogeneity,

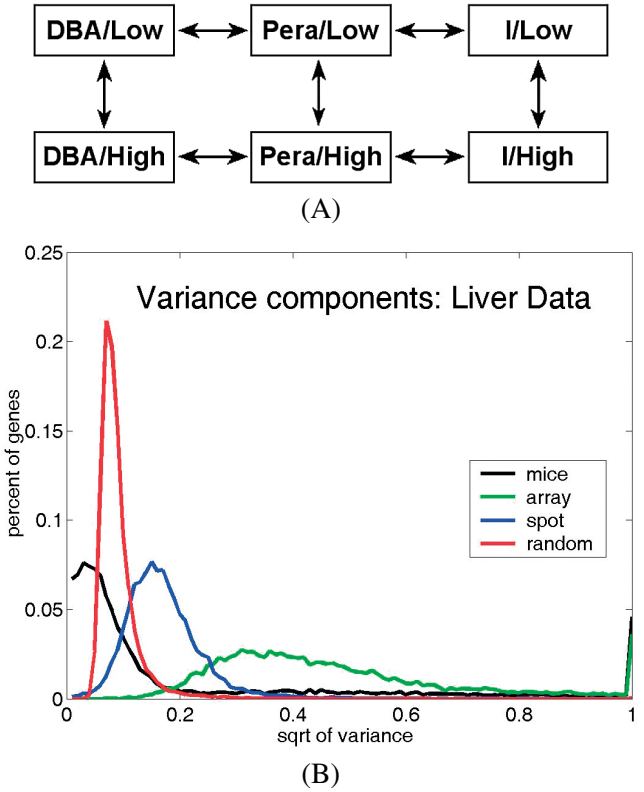

FIG. 1. *A factorial experiment was carried out to measure gene expression in liver tissue for all combinations of three inbred mouse strains on two different diets.* (A) *Direct dye swap comparisons were made using cDNA microarrays. Arrays were printed with duplicate clones and the entire experiment was repeated two times using independent biological samples. A mixed model analysis of variance was carried out on a gene-by-gene basis and restricted maximum likelihood (REML) estimates of the variance components corresponding to measurement error, spots, arrays and mice were computed for each gene on the array.* (B) *Smoothed histograms of the variance component estimates.*

resulting in higher variances at low intensities. A third function of data transformation is to correct for nonlinear effects associated with physical properties of the dyes. Although the latter purpose predominates in discussions of data transformation for microarrays, most solutions to this problem are variations on the logarithmic transform. A more detailed discussion of these topics can be found in Cui, Kerr and Churchill (2003).

The presence of negative values in background corrected data does not provide a valid argument against using the logarithmic transform. Rather it is an indication that the background correction method is flawed. Converting an inherently positive value (essentially a count of photons that hit a photomultiplier tube) into a negative value violates what should be a fundamental principle of modeling, the Hippocratic oath of data transformations: first, do no harm.

Filtering based on data values is also a common and dubious practice in cDNA microarray experiments. It results in a missing data pattern that does not satisfy the standard *missing at random* (MAR) condition and thus could lead to biases. Correcting these effects would require a model of the missingness mechanism. Consider the case of a spot which is near the background intensity on one microarray and is discarded. On another array the same spot appears bright, but since the corresponding dim spot has been culled, a potentially interesting result is not noted.

Among the many small problems that have yet to be addressed in microarray analysis, missing data methods stand out in my mind as one of the more pressing. Despite my objections to filtering, it is quite common that scratches or debris obscure data points and these *should* be discarded from further analysis (and unless the scratch is targeted, they should be MAR). A few missing points may be acceptable loss in a large experiment, but the cumulative effects of missing data can sometimes lead to elimination of the majority of genes. Robust imputation methods seem most promising.

5. THE ANALYSIS OF VARIANCE

Not all of the problems presented by microarray analysis are novel. With the current emphasis on novelty in scientific research it can be easy to overlook connections between new problems and older ones. This can lead to a failure to see obvious solutions. Rather than spend the effort to reinvent classical techniques, we should take advantage. When a giant offers his shoulders, stand on them—and enjoy the view.

The recognition that cDNA microarray experiments have split plot features and an incomplete blocking structure opens the possibility to apply sophisticated analysis techniques that have been developed for similar experiments in other contexts. Analysis of variance (ANOVA) models for microarray data were introduced by Kerr, Martin and Churchill (2000) and have been extended to allow for random variation at multiple levels in the experiment (Wolfinger et al., 2001). Although a discussion of mixed model analysis of variance falls outside the scope of this commentary, a key feature of mixed model analysis is worth keeping in mind. Mixed models can be constructed to admit general variance–covariance structures among the observations. There are dramatic correlations present in most microarray data sets and failure to account for these can lead to overly optimistic inferences.

Do we believe that the ANOVA model is true? Certainly not. It is, however, a flexible model that can provide a reasonable approximation to the truth; furthermore, it leads to powerful and robust analysis methods. The strongest modeling assumption is additivity of effects. Intensity data on the raw scale have both additive and multiplicative components (Rocke and Durbin, 2001), but for the bulk of observations, multiplicative effects are dominant. Variations on the logarithm transform are available to obtain approximate additivity over the full intensity range of the data (Cui, Kerr and Churchill, 2003). Distributional assumptions in testing can be overcome by permutation analysis, provided that there is sufficient replication in the experiment. We currently prefer approaches that permute sample labels over residual permutations. Permutation can also be used to compute multiple *testing* adjustments (both Family-wise error rate and False discovery rate control), but it has the obvious disadvantage of computational burden. In practice we find that permutation analysis results agree well with tabulated statistics, but we remain skeptical. The least squares fitting procedures could be robustified and, last, the distributional assumptions on the random components of the mixed model might be relaxed somewhat by considering normal mixtures. There are plenty of opportunities to improve microarray analysis within the ANOVA framework.

6. OPEN CHALLENGES

Sebastiani et al. have identified a number of interesting open problems in microarray analysis. However, their vision is perhaps too narrow in light of the title of their article. The real challenges in functional genomics will come from attempts to integrate multiple diverse data types, including, but not limited to, gene expression data, data on quantities and modification states of proteins, metabolite flux and the allelic states of genes. All of these must be combined to achieve a detailed understanding of the physiological state of an organism.

In addition to the human genome, a number of other complete genomic sequences are available or soon will be. These include mouse, rat, fruitfly, nematode and yeast. These organisms present opportunities for experimental manipulation of the genome and the environment in which it exists. Unraveling the developmental program of gene expression will be possible only in experimental organisms. The same is true of gene expression response to environmental challenges. The problem of relating gene expression variation to genetic polymorphisms is another wide open challenge, where controlled crosses are likely to yield valuable insights (Brem, Yvert, Clinton and Kruglyak, 2002). An interesting twofold multiple testing problem arises in this context and is in need of immediate attention, because many groups are already running experiments of this type. It will also be of interest to study natural populations. We will want to examine the role of gene expression variation in adaptation and evolution (Oleksiak, Churchill and Crawford, 2002). In human populations there are complex sampling issues that have been largely ignored in gene expression studies, but are a cause for concern in an epidemiological context.

Tissue samples are mixtures of cell types. They range in complexity from the fairly homogeneous liver to the brain that has thousands of cell types. Tumor samples are notoriously heterogeneous. In a mammary tumor study, it may be helpful to subtract out the signal contributed by the adipose and normal epithelial components so as to focus on expression in the cancerous cells themselves. Whenever we assay gene expression in a tissue, we are observing a mixture of cell types. Is it possible to deconvolve this mixture? Perhaps we can, but it looks like a challenging problem to me.

There is plenty of room for novel statistical developments in functional genomics. Hierarchical cluster methods for the analysis of gene expression data caught on like the hoola hoop. I, for one, will be glad to see them fade. However, what will fill the void? Sebastiani et al. mentioned the promise of Bayesian network models. These are often elaborate models that can specify complex conditional dependence relationships among many variables acting simultaneously. There is a concern that we cannot adequately infer the parameters (or the structure) of a Bayesian network model with currently available data which sample only a small number of states of the system. Here is a challenge in design. How can we construct a set of observations that can be obtained within limits of available resources and still yield sufficient information to estimate such complex models?

The real challenges for the statistician are to stay connected to the biology even (especially) when things get messy, to stay in touch with the realities of the data and to keep the (biological) goals of an investigation foremost in mind. This will mean that one should develop and apply methods of data analysis that are effective and robust even if they are not mathematically elegant. At the same time we should not forget

our roots or ignore the lessons of the past. We should not allow the flashy appeal of computationally intensive visualization to take precedence over sound design and proper inference techniques. I would like to close by, again, quoting from Fisher (1951): "Statistical procedure and experimental design are only two different aspects of the same whole, and the whole comprises all the logical requirements of the complete process of adding to natural knowledge by experimentation." It is a great tradition to uphold. We should not lose sight of that.

Rejoinder

Paola Sebastiani, Emanuela Gussoni, Isaac S. Kohane and Marco F. Ramoni

We would like to thank the discussants for their contributions that have significantly enhanced the value of the article by reinforcing some of our points and drawing attention to other challenging problems.

Microarray technology offers incredible opportunities for scientific discoveries, but there are still serious limitations at technology and data analysis level. We agree with Curchill's comment that our description of the Affymetrix technology is uncritical, but the main objective of our contribution was to describe rather than evaluate microarray technology in the context of data analysis for the benefit of the statistical community. We have noted elsewhere (Kuo et al., 2002) that there are cross-platforms and cross-generation reproducibility issues related to the measurement of gene expression level from synthetic oligonucleotide microarrays and we have addressed elsewhere more technical details of this technology (Kohane, Kho and Butte, 2002). Still, we must stress that this technology is considered more reliable than cDNA microarrays (Kuo et al., 2002). The high risk of cross-hybridization of cDNA technology is well documented, as pointed out by Kothapalli, Yoder, Mane and Loughran (2002) and Li, Gu, Mohan and Baylink (2002), just to mention two recent publications that quantify this risk in controlled experiments, and the technology of synthetic oligonucleotide microarrays based on competitive hybridization attempts to minimize this risk through a careful oligonucleotide selection. The probe set selection also normalizes for GC content [the A and T bases are known to be less stable than the G and C bases, so that sequences with larger contents of A and T nucleotide bases do not have the same chances of hybridization as sequences with larger contents of G and C nucleotide bases (Kohane, Kho and Butte, 2002)]. There is no doubt, however, that both the probe set selection and the preprocessing of perfect match and mismatch intensity values implemented in MAS 5.0 could be improved, as shown by recent efforts by Antipova, Tamayo and Golub (2002), Irizarry et al. (2003) and Kasif et al. (2002).

Both discussants emphasized the importance of experimental design of microarray experiments. We fully agree with Baker that adding replications improves the results of any analytical methods. However, we hesitate to adhere to his suggestion to budget a minimum of three replicates. The costs of microarray experiments still impose serious sample size limitations and the designer of the experiment needs to trade off the number of independent samples with the number of replications. The best solution depends, of course, on the objective of the analysis: if the interest is to have an accurate estimate of the error variance, then an experiment with a large number of replications and a small number of independent samples will be preferable to an experiment with one replication of each independent sample. However, in experiments in which the variability between sample units is expected to be large, such as clinical samples, it is better to invest in independent samples rather than replications. This dilemma in the design of the experiments and the lack of an "out-of-the-box" answer shows the need to research this area further. We agree with Churchill's observation that pooling samples can be misleading and, in fact, we do not recommend this strategy in the article, but simply describe it as a strategy used because of mRNA paucity. Although the continuous improvement of amplification techniques will render the problem of mRNA paucity less serious, further research is needed to assess the effect of amplification of the mRNA in the preparation of the target.

The analysis of microarray data is full of challenges and there is wide disagreement about the best preprocessing steps to conduct. We do not share Churchill's view about the need to log transform the data. Although it is generally acknowledged that

the corrected intensity values measured with competitive hybridization on cDNA microarrays should be transformed, the choice of the best transformation to use is still an open issue. The log transformation is just an example of a wider family of power transformations that could be used, such as the cubic root transformation (Tusher, Tibshirani and Chu, 2001), which, incidentally, is more appropriate when data follow a Gamma rather than a log-normal distribution. We would suggest that the decision as to whether and how to transform the data should not be made independently of the data analysis to conduct. Furthermore, careful modeling of gene expression data could eliminate the need for arbitrary data transformation and normalization (Sebastiani and Ramoni, 2002). After all, appropriate modeling rather than "ad hoc" data transformation was the motivation behind the introduction of the family of generalized linear models (McCullagh and Nelder, 1989).

The discussion contributions have emphasized the fact that gene expression analysis is an important and challenging area for statisticians and data analysts in general. Particularly, we fully endorse Churchill's conclusions about the need for close collaborations between statisticians and biologists: interdisciplinary collaborations are crucial for the full exploitation and understanding of functional genomic data.

ADDITIONAL REFERENCES

ANTIPOVA, A. A., TAMAYO, P. and GOLUB, T. R. (2002). A strategy for oligonucleotide microarray probe reduction. *Genome Biology* **3**(12) research0073.1-0073.4.

BREM, R. B., YVERT, G., CLINTON, R. and KRUGLYAK, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science* **296** 752–755.

BRENNER, S., JOHNSON, M., BRIDGHAM, J., GOLDA, G., LLOYD, D. H., JOHNSON, D., LUO, S., MCCURDY, S., FOY, M., EWAN, M., ROTH, R., GEORGE, D., ELETR, S., ALBRECHT, G., VERMAAS, E., WILLIAMS, S. R., MOON, K., BURCHAM, T., PALLAS, M., DUBRIDGE, R. B., KIRCHNER, J., FEARON, K., MAO, J. and CORCORAN, K. (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature Biotechnology* **18** 630–634.

CHURCHILL, G. A. (2002). Fundamentals of experimental design for cDNA microarrays. *Nature Genetics* **32** 490–495.

CUI, X., KERR, M. K. and CHURCHILL, G. A. (2003). Data transformations for cDNA microarray data. In *Statistical Applications in Genetics and Molecular Biology*. To appear.

FISHER, R. A. (1951). *The Design of Experiments*, 6th ed. Oliver and Boyd, Edinburgh.

IRIZARRY, R. A., BOLSTAD, B. M., COLLIN, F., COPE, L. M., HOBBS, B. and SPEED, T. P. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research* **31**(4) e15.

JANSEN, R. C. (2003). Studying complex biological systems using multifactorial perturbation. *Nature Reviews Genetics* **4** 145–151.

KASIF, S., WENG, Z., DERTI, A., BEIGEL, R. and DELISI, C. (2002). A computational framework for optimal masking in the synthesis of oligonucleotide microarrays. *Nucleic Acids Research* **30**(20) e106.

KERR, M. K., MARTIN, M. and CHURCHILL, G. A. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology* **7** 819–837.

KOTHAPALLI, R., YODER, S. J., MANE, S. and LOUGHRAN, JR., T. P. (2002). Microarray results: How accurate are they. *BMC Bioinformatics* **3**: 22.

KUO, W. P., JENSSEN, T. K., BUTTE, A. J., OHNO-MACHADO, L. and KOHANE, I. S. (2002). Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* **18** 405–412.

LI, X., GU, W., MOHAN, S. and BAYLINK, D. J. (2002). DNA microarrays: Their use and misuse. *Microcirculation* **9** 13–22.

LI, C. and WONG, W. H. (2001). Model based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci. U.S.A.* **98** 31–36.

OLEKSIAK, M. F., CHURCHILL, G. A. and CRAWFORD, D. L. (2002). Variation in gene expression within and among natural populations. *Nature Genetics* **32** 261–266.

ROSENBAUM, P. R. (2001). Replicating effects and biases. *Amer. Statist.* **55** 223–227.

SEBASTIANI, P. and RAMONI, M. (2002). Bayesian differential analysis of gene expression data. In *Proc. Joint Statistical Meeting. Section on Bayesian Statistical Sciences*. Amer. Statist. Assoc., Washington, 3146–3148.