sion profiles are DNA microarrays and serial analysis of gene expression (SAGE). Because of the large amount of data that is generated from these experiments, special computational tools are required for obtaining, storing, and analyzing the results.

## DNA Microarray Technology

A DNA microarray (DNA chip, or gene chip) experiment consists of hybridizing a nucleic acid sample (target) derived from the messenger RNAs (mRNAs) of a cell or tissue to single-stranded DNA sequences (probes) that are bound in an ordered arrangement to a solid platform. One type of DNA microarray is constructed by spotting polymerase chain reaction (PCR)-amplified cDNA sequences from the mRNAs of a single cell or all or specific sets of the coding sequences of an organism onto a glass slide or nylon membrane. In this case, about 10,000 different probes can be arrayed in a 1-$cm^2$ area.

An alternative microarray system utilizes sets of oligonucleotides as probes, usually representing thousands of genes. For one commonly used platform, the probes are synthesized directly (in situ) on a solid surface (quartz wafer) by a light-directed process known as photolithography. Thousands of copies of an oligonucleotide with the same specific nucleotide sequence are synthesized in a predefined position (probe cell or feature) on the array surface. For this type of microarray, the probes are typically 10 to 40 nucleotides, and several probes with different sequences for each gene will be synthesized on the microarray. Longer oligonucleotides up to 100 nucleotides can also be used. A complete whole-genome oligonucleotide array may contain more than 500,000 probes representing as many as 30,000 genes.

Generally, the design of the probes (probe set) for a microarray depends on the objective of the experiment and the degree of resolution that is required. Computer programs determine probe sequences that are specific for their target sequences, are least likely to hybridize with nontarget sequences (cross-hybridize), have no secondary structure (foldback) that would prevent hybridization with the target sequence, and have similar melting (annealing) temperatures, so that all target sequences can bind to their complementary probe sequences under the same conditions. Oligonucleotide microarrays may consist of probes that represent an entire genome, a single chromosome, selected genomic regions, or selected coding regions from one or several different organisms. Repetitive sequences are not included in genomic DNA microarrays.

Typically, for most gene expression profiling experiments that utilize microarrays, mRNA is extracted from cells or tissues and purified, and cDNA is synthesized using reverse transcriptase and the extracted mRNA as a template. Usually, mRNA is extracted from two or more sources whose expression profiles are compared; for example, from diseased versus normal tissue or from cells grown under different conditions. The cDNA from each source is labeled with a different fluorophore by incorporating fluorescently labeled nucleotides during cDNA synthesis. For example, a green-emitting fluorescent dye (Cy3) is used for the normal (reference) sample and a red-emitting fluorescent dye (Cy5) for the test sample (Fig. 5.4). After being labeled, the cDNA samples are mixed and hybridized to

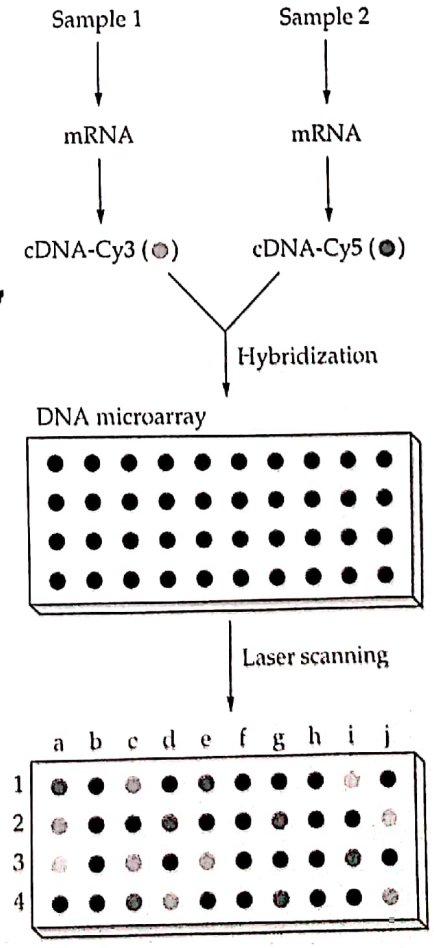



FIGURE 5.4 Gene expression profiling with a DNA microarray. mRNA is extracted from two samples (sample 1 and sample 2), and during reverse transcription, the first cDNA strands are labeled with the fluorescent dyes Cy3 and Cy5, respectively. The cDNA samples are mixed and hybridized to an ordered array of either gene sequences or gene-specific oligonucleotides. After the hybridization reaction, each probe cell is scanned for both fluorescent dyes and the separate emissions are recorded. Probe cells that produce only a green or red emission represent genes that are transcribed only in samples 1 and 2, respectively; yellow emissions denote genes that are active in both samples; and no emissions (black) represent genes that are not transcribed in either sample.

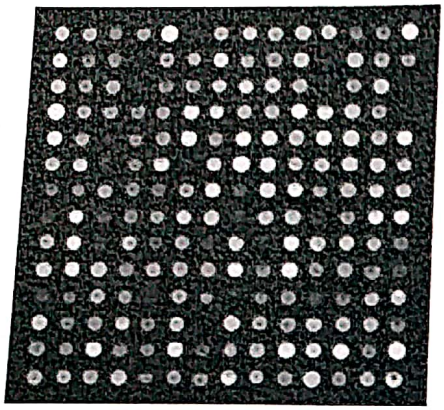


**FIGURE 5.5** Fluorescence image of a DNA microarray hybridized with Cy3- and Cy5-labeled cDNA. Reproduced from http://biotech.biology.arizona.edu/Resources/DNA_analysis.html. Courtesy of N. Anderson, University of Arizona.

the same microarray. A laser scanner determines the intensities of Cy5 and Cy3 for each probe cell. The probe cells have different colors depending on the amounts of Cy3 and Cy5 that are present, and the ratio of red (Cy5) to green (Cy3) fluorescence intensity of a probe cell indicates the relative expression levels of the represented gene in the two samples (Fig. 5.5). To avoid variation due to inherent and sequence-specific differences in labeling efficiencies between Cy3 and Cy5, reference and test samples are often reverse labeled and hybridized to another microarray. In the above example, reverse labeling (dye swapping) would entail labeling the reference sample with Cy5 and the test sample with Cy3. Alternatively, for some microarray platforms, the target sequences from each source are labeled with the same fluorescent dye, and reference and test samples are hybridized to different microarrays.

In an alternative strategy, mRNA is purified with an oligo(dT) sequence that binds to the poly(A) tail of eukaryotic mRNA and has a short extension (T7 primer) containing the sequence for the bacteriophage T7 RNA polymerase promoter (Fig. 5.6). The oligo(dT)–T7 sequence primes the synthesis of cDNA from mRNA using reverse transcriptase. Synthesis of the second DNA strand using DNA polymerase results in double-stranded cDNA that contains the T7 RNA polymerase promoter. Next, T7 RNA polymerase is used to produce RNA copies (complementary RNA [cRNA] or antisense RNA) from the second cDNA strand as a template in the presence of biotin-labeled ribonucleotides. This reaction results in linear amplification and biotinylation of cRNA, which is then fragmented into pieces from 50 to 100 nucleotides in length that are optimal for hybridization. After hybridization, the microarray is treated with streptavidin that is bound to the fluorescent protein phycoerythrin. Streptavidin binds specifically to the biotin residues of hybridized cRNA, and hybridization can be detected by emissions from phycoerythrin that are elicited during laser scanning.

Because of the vast amount of data generated by microarray experiments, specialized software has been developed to maximize the output of information. Analyses of two-dye and one-dye hybridized microarrays are similar. A common method by which information is extracted from two-dye hybridized microarrays is summarized here. Each probe cell of a two-dye hybridized microarray is scanned using a confocal scanning microscope. Following laser excitation of the dye, fluorescence emitted from each probe cell, detected at both 532 and 635 nm for Cy3 and Cy5, respectively, is collected through the microscope's objective lens and converted to an electrical signal via a photomultiplier tube. The intensities of fluorescence emitted by both dyes for each probe cell, along with background readings for the microarray, are recorded and stored. Background fluorescence is determined by measuring the fluorescence from blank areas where probe cells have not been spotted and is subtracted from the fluorescence intensities measured for each probe cell. Microarrays are designed with internal controls, that is, specific probes that are used to evaluate the reliability of the hybridization procedure and to ensure that the laser scanner performed properly. At this stage, the microarray data (i.e., the collection of fluorescence intensities of each probe cell) is normalized to correct for variations (systematic errors) caused by technical factors that contribute to the fluorescence intensities of a probe cell and enable comparison among the microarrays of an experiment. To minimize errors, multiple probe cells for each gene are included on a single microarray, and replicate samples are independently prepared under the same conditions

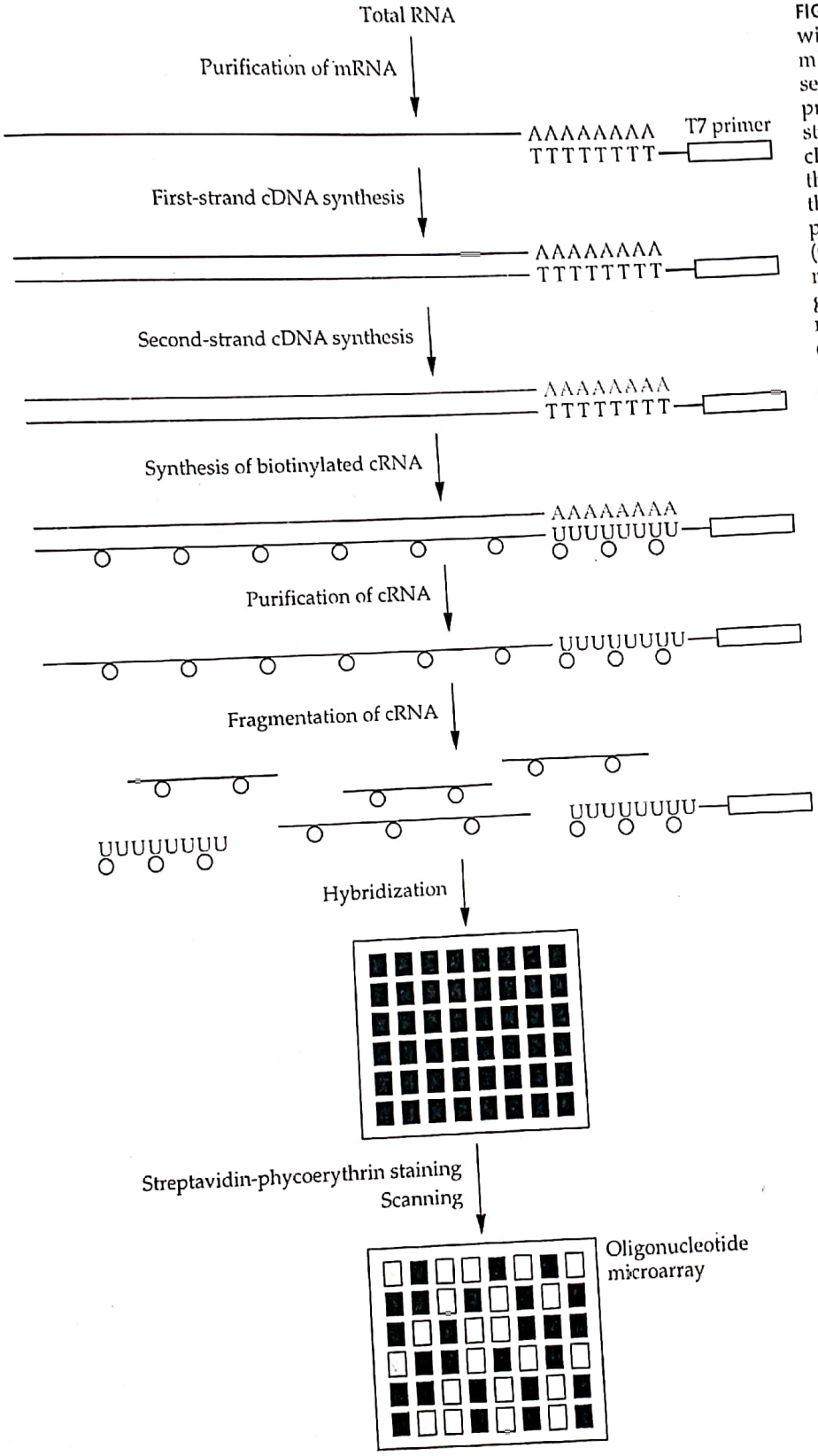



**FIGURE 5.6** Gene expression profiling with an oligonucleotide microarray. mRNA is purified with a poly(dT) sequence that has a T7 RNA polymerase primer sequence extension. After two-stranded cDNA synthesis, the second cDNA strand acts as a template for synthesis of cRNA by T7 RNA polymerase in the presence of ATP, cytidine triphosphate (CTP), guanosine triphosphate (GTP), uridine triphosphate (UTP), biotinylated CTP, and biotinylated UTP. The gray circles represent incorporated biotinylated nucleotides. The biotinylated cRNA is purified, fragmented into pieces from 50 to 100 nucleotides in length, and hybridized to an oligonucleotide microarray. The microarray is treated with streptavidin–phycoerythrin, and the probe cells (black squares) are scanned for emission (yellow) from the biotin-bound streptavidin–phycoerythrin.

and hybridized to different microarrays. Misleading interpretations are caused, in part, by variations in sample preparation, such as growth conditions, RNA extraction, cDNA synthesis and labeling efficiencies, differences in efficiencies of hybridization of the target sequences among replicate microarrays or across a single microarray, variations in the concentrations of probes on different microarrays, or unequal amounts of target sequences applied to different microarrays or unequal distribution of targets on a single microarray. Several methods for normalization are used to calibrate the data among replicate microarrays, such as using the fluorescence intensity of a gene that is not differentially expressed under different conditions as a reference point, including spiked control sequences that are sufficiently different from the target sequences and therefore bind only to a corresponding control probe cell, and adjusting the total fluorescence intensity for each microarray to a similar value under the assumption that a relatively small number of genes are expected to change under different conditions.

One of the major purposes of a microarray experiment is to identify genes whose expression changes in response to a particular biological condition. The response to a biological condition is determined by comparing the fluorescence intensity for each gene (each probe cell), averaged among replicates, under two different conditions and calculating the ratio, commonly expressed as an *n*-fold change. For effective comparisons, the raw data of the dye emissions of each probe cell of a microarray are often converted to $\log_2$ ratios (Table 5.4). The sign indicates the dye with the higher intensity. Generally, positive log ratios represent more Cy5 than Cy3 and, therefore, greater expression of the gene in the test sample than in the reference sample. Negative values (more Cy3 than Cy5) indicate a lower level of expression in the test sample than in the reference sample. The log ratios for all probe cells are compiled into a table called an expression matrix.

Microarray analysis is also used to identify genes that are coexpressed under different conditions or over a period of time, with the goal of determining which gene products function in a given pathway. A number of computational strategies are available that organize the data into related groups (clusters). For a clear presentation of the clustered data, ranges of log ratio values are assigned arbitrary colors. Usually, black is designated for a log ratio of zero, dark to bright red for increasing positive log ratios, and dark to bright green for decreasing negative log ratios. In other words, red is frequently used to denote gene overexpression and green to denote underexpression. A visualized representation of a clustered microarray is called a gene expression profile, where the rows represent the reordered genes and the columns represent either conditions or samples (Fig. 5.7).

The gene expression profile in Fig. 5.7, determined by microarray analysis, clearly shows that different genes are transcribed in patients with cirrhosis of the liver than in healthy individuals and in patients with ethanol-induced cirrhosis than in those with cirrhosis induced by the hepatitis

**TABLE 5.4** Converting Cy3 and Cy5 intensities to $\log_2$ ratios

| Feature (gene) | Cy3 intensity | Cy5 intensity | Cy5/Cy3 | $\log_2$(Cy5/Cy3) |
|---|---|---|---|---|
| 1 | 180 | 10,000 | 56 | +5.81 |
| 2 | 5,400 | 5,400 | 1 | 0 |
| 3 | 8,400 | 400 | 0.05 | −4.39 |

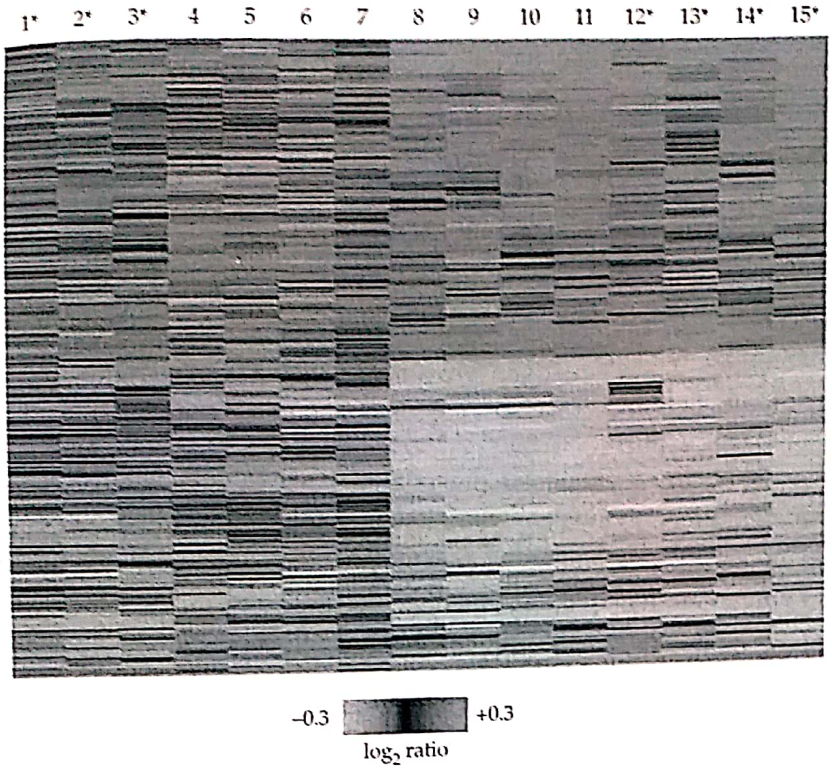
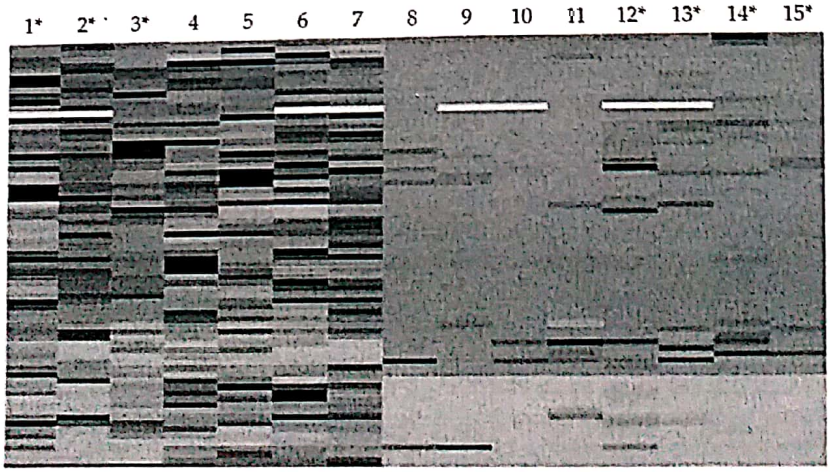



FIGURE 5.7 Gene expression profile of cirrhotic liver tissue. Columns 1 to 7 and 8 to 15 are liver samples from patients with ethanol- and hepatitis C virus-induced cirrhosis of the liver, respectively. Each patient's sample was compared to normal liver tissue. The cluster consists of 2,965 genes. The asterisks denote patients with severe cirrhosis of the liver. Adapted from Lederer et al., *Virol. J.* 3:98, 2006.

FIGURE 5.8 Gene expression profile of lymphocyte-specific genes from cirrhotic liver tissue. Columns 1 to 7 and 8 to 15 are liver samples from the patients shown in Fig. 5.7 with ethanol- and hepatitis C virus-induced cirrhosis of the liver, respectively. Each patient's sample was compared to normal liver tissue. The cluster consists of about 70 genes. The asterisks denote patients with severe cirrhosis of the liver. Adapted from Lederer et al., *Virol. J.* 3:98, 2006.



C virus. Moreover, there is a difference between the genes that are turned on during advanced ethanol-induced liver damage and those in patients with less severe ethanol-induced cirrhosis (Fig. 5.7). No such distinction is evident among individuals with different severities of virus-induced cirrhosis (Fig. 5.7). In addition, information about the transcription of genes that contribute to a particular pathway or cellular activity can be extracted from a gene expression profile. For example, genes that are transcribed during lymphocyte proliferation and activation are highly expressed in virus-induced liver cirrhosis and to a much lesser extent in ethanol-associated cirrhotic samples (Fig. 5.8).

The importance and pervasiveness of DNA microarrays cannot be overstated. In 2007, for example, there were more than 13,000 published journal articles that either used this technology or described methods for enhancing data analysis. Clinical applications for DNA microarrays are being developed. The U.S. Food and Drug Administration granted permission in 2007 for the first commercial diagnostic assay based on a DNA microarray. In this case, a 70-gene expression profile (MammaPrint) distinguishes between patients with breast cancer that is likely to migrate to other sites (metastasize) and those whose cancer has a low risk of metastasis. The reliability and reproducibility of microarrays with different formats and from various laboratories have been major concerns. However, standards for both running microarray experiments and analyzing the data have been proposed by a number of international groups, which should alleviate these problems. Finally, it should be noted that in addition to gene expression studies, DNA microarrays are used to determine the binding sites for DNA-binding proteins (e.g., ChIP-on-chip assays, which use chromatin immunoprecipitation [ChIP] to identify proteins bound to a DNA microarray), the sites where the transcription of genes starts and stops, and many other aspects of genome architecture. This research has shown that a much larger proportion of the eukaryotic genome is transcribed than was previously thought; a number of genes have multiple start and termination sites, some of which are hundreds of kilobases from known sites for many genes; both strands of many genomic regions are transcribed; splicing occurs between RNA molecules; and some transcription factors bind to dozens of sites scattered throughout the genome. In short, whole-genome microarray analysis has revealed greater complexity in the processes that control transcription in a eukaryotic organism than could be predicted through smaller-scale transcriptional analyses.