Comp. by: SKarthiProof0000868544 Date:29/9/08 Time:22:28:12 Stage:First Proof File Path://ppdys1108/
Womat3/Production/PRODENV/0000000038/0000010883/0000000005/0000868544.3D Proof by: QC by:

B978-0-12-374419-7.00007-X, 00007

Chapter

7

Au2 c0007

The Human Transcriptome: Implications for the Understanding of Human Disease

Matthias E. Futschik . Wolfgang Kemmner . Reinhold Schäfer . Christine Sers

INTRODUCTION

s0005

p0005 The most fascinating aspect of transcriptomics is that the entire set of messenger RNA (mRNA) molecules or transcripts produced in a population of cells or in tissues can be analyzed simultaneously. The present microarray technology produces devices equivalent to the size of a stamp for gene expression profiling. Analyzing the transcriptome is a challenging task, since the mRNA content of a biological entity is heterogeneous and can vary substantially. The abundance of individual transcripts varies from a few copies to hundreds or thousands of copies per cell [1]. The kinds and copy numbers of individual transcripts expressed at a given time depend on the developmental stage, on external conditions, and environmental stimuli. Quantitative and qualitative alterations of mRNAs can be directly linked to the molecular mechanism of disease or reflect the downstream consequences of these disease processes.

p0010 This chapter outlines the methodological prerequisites for transcriptome analysis and describes typical applications in molecular cell biology and pathology. During the last three decades, technology development and experimental approaches aiming at mRNA analysis were significantly fueled by molecular cancer research. One of the main reasons for progress in this area was the availability of relevant cell lines that could be propagated indefinitely and served as reproducible sources of RNA and of sufficient quantities of normal and diseased tissues. A strong motivation lay in the demand for distinguishing as many transcripts as possible in normal and tumorigenic cells to understand cancer-specific alterations in gene expression. While early work along these lines was mostly related to pathogenesis, more recent applications deal with diagnostic issues such as tumor outcome, prognosis, and therapy response prediction.

GENE EXPRESSION PROFILING: THE SEARCH FOR CANDIDATE GENES INVOLVED IN PATHOGENESIS

s0010

p0015 To date, microarray-based expression profiling is accepted as the gold standard in transcriptome analysis. Microarray technology has gradually improved over the last decade both in academic and industrial/commercial settings to meet high technical and bioinformatic quality standards. Before microarrays were available for most researchers in sufficient quantity and quality (as well as affordable at reasonable costs), alternative techniques were instrumental in answering questions related to the quantity of transcripts expressed ubiquitously, to identifying tissue-specific expression patterns and candidate genes related to disease.

Early Gene Expression Profiling Studies

s0015

p0020 Intriguingly, the question of how many transcription units distinguish normal from tumor cells, expectedly to be a domain of transcriptomics using microarrays, was addressed nearly at the same time when techniques in molecular biology permitted the identification and thorough analysis of individual mRNAs. In 1977 and 1980, researchers described the northern blot technique for transferring electrophoretically separated RNA from

an agarose gel to paper strips, the coupling of the RNA to the paper surface and the detection of specific RNA bands by hybridization with ^{32}P-labeled DNA probes followed by autoradiography for the first time [2,3]. A report published in 1980 provided evidence for the complexity of cellular transformation at the RNA level, when scientists had studied the RNA pool of chicken embryonic breast muscle cells infected with Rous sarcoma virus (RSV). The authors of the paper compared the hybridization kinetics of nuclear RNA preparations from normal and RSV-transformed cells, respectively, incubated in solution with tracer amounts of labeled single-copy chicken DNA. Based on the assumption that an average transcription unit is about 10 times larger than its corresponding mRNA, the authors concluded that the observed increase in the number of stable transcription products in transformed cells relative to normal cells was equivalent to approximately 1,000 transcription units [4]. Several years later, other scientists used a more sophisticated approach for contrasting mRNA patterns of cellular material obtained from colon tumor biopsies [5]. The researchers took advantage of the molecular cloning techniques that allowed the establishment of a set of complementary DNAs (cDNAs) obtained by reverse transcription of mRNA. The reference cDNA library of some 4,000 clones represented abundant and middle abundant RNA sequences. Replicas of the library were then hybridized to ^{32}P-labeled cDNA probes synthesized from polyadenylated RNA from small biopsies obtained from normal and neoplastic intestinal mucosa. The comparison of normal colonic mucosa with carcinomas showed expression alterations of ~7% of the cloned sequences and was extrapolated to the entire, yet unknown, set of transcripts. The number of alterations was smaller between normal mucosa and benign adenomas indicating that transcriptional changes accumulate during cancer progression.

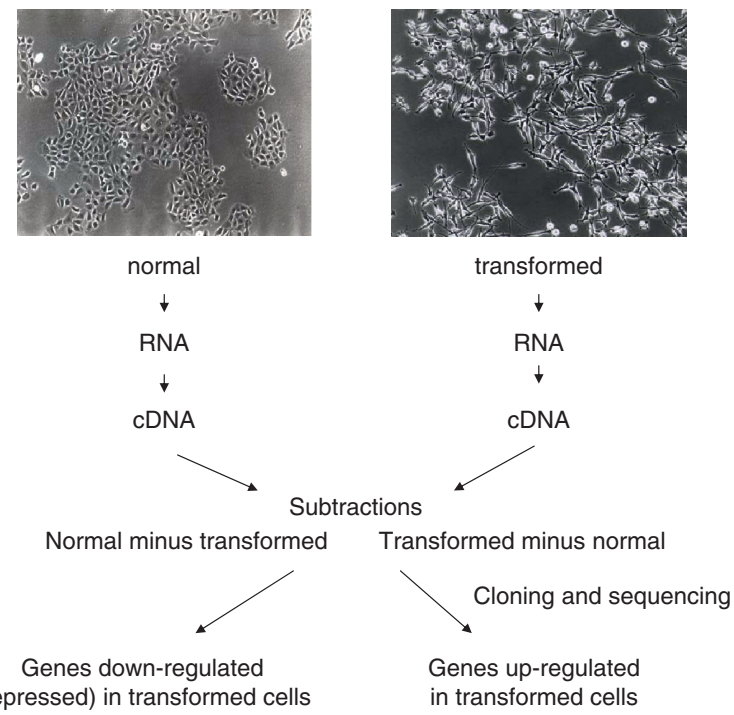
cDNA Libraries and Data Mining

Further advances in deciphering cancer-related transcripts were driven by increased efforts in cDNA cloning, and sequence analysis. Collections of cDNAs were obtained from various normal and diseased tissues, as well as from reference cell lines. The functional characterization of transcribed sequences progressed at the same time. However, due to the complexity of gene function in biological systems, functional information lagged significantly behind sequence information. The large cDNA collections deposited in expression databases often provided only partial sequence information. The corresponding cDNAs known to be expressed in various tissues or cell types analyzed were designated expressed sequence tags (or ESTs). With increasing entries into these EST catalogues, it became feasible to merge overlapping partial sequences and eventually to define full-length open-reading frames (ORFs). As a practical consequence of the global gene expression information provided by cDNA/EST databases, an approach termed the *electronic northern* became feasible. The electronic northern analysis facilitated prediction of expression changes between normal and diseased tissues. Extensive mining of EST databases using stringent statistical tests permitted identification of candidate genes whose altered (stimulated or reduced) expression correlated with the disease state [6].

cDNA Subtraction

The data mining approach was limited by the existing sequence information and the available gene annotations. To circumvent this bias, researchers established several elegant methods that permitted enrichment of mRNA sequences (or cDNAs) associated with special experimental conditions such as cellular stress or oncogenic transformation, or with particular cellular features such as tumorigenicity or metastatic potential. The methods established were cDNA subtraction, differential display PCR (DD), representational difference analysis, and serial analysis of gene expression (SAGE).

In general, cDNA subtraction is a method for separating cDNA molecules that distinguish related cDNA samples, for instance, prepared by reverse transcription of mRNA from normal precursor cells and derived neoplastically transformed cells. The basis of subtraction is that cDNAs prepared from two different cell types to be compared are rendered single-stranded, subsequently mixed, and incubated to allow annealing of sequences common to both cell species. These sequences will hybridize, while sequences unique to one of the cells will stay single-stranded. In the classical subtraction approach, single-stranded and double-stranded cDNAs were separated by hydroxylapatite chromatography [7]. Subsequently, the unique cDNA fragments are cloned and sequenced. The major drawback of this method is that the enrichment of differentially expressed sequences usually does not exceed a factor of 100, that abundant mRNAs (cDNAs) are over-represented due to the lack of normalization, and that rare transcripts are not detected at all. These inherent disadvantages were overcome by development of a method called suppression subtractive hybridization (SSH), a PCR-based subtraction method that combines normalization and subtraction into a single procedure. Differential amplification of unique cDNA fragments is achieved by ligating different primers to each restricted cDNA originating from the cell types to be compared prior to the annealing step and the PCR. The normalization step equalizes the abundance of cDNA fragments within the target population, and the subtraction step excludes sequences that are common to the cell populations being contrasted. Using this method, the probability of recovering differentially expressed cDNAs of low abundance is largely increased (by a factor of 1,000 or more) [8]. For example, SSH was used to report on transformation target genes related to oncogenic RAS signaling on a genome-wide scale [9] (Figure 7.1). Representational difference analysis (RDA) is a technique that combines subtractive hybridization with PCR-mediated kinetic enrichment for the detection of differences between two complex genomes [10]. Later, the protocol was modified to look for differences in transcript expression as well [11].



f0005 **Figure 7.1** A cDNA subtraction approach to identify genes differentially expressed upon conversion from the normal to the transformed state. In this example, immortalized normal epithelial cells (phase contrast microscopy, magnification 100-fold) were transformed by the *KRAS* oncogene.

s0030 Differential Display PCR

p0040 Differential display PCR is a method to separate and clone individual mRNAs that are differentially expressed by means of the polymerase chain reaction. A set of oligonucleotide primers is used, one being anchored to the polyadenylated tail of a subset of mRNAs, the other being short and arbitrary in sequence to allow annealing at different sites relative to the first primer. The mRNA subpopulations defined by the primer pairs are amplified after reverse transcription and the products resolved (displayed) on DNA sequencing gels. Differential display visualizes mRNA compositions of cells by displaying subsets of mRNAs as short cDNAs. The beauty of this approach is that many samples can be run in parallel to reveal differences in mRNA composition [12]. The differentially expressed cDNA fragments can be recovered by cloning techniques. An early application of DD was the identification of genes differentially expressed in breast cancer versus mammary epithelial cells [13].

s0035 Serial Analysis of Gene Expression

p0045 While the previously discussed approaches directly aim at identifying important differences between closely related cell types, the key element of the SAGE method is to represent all transcripts in a given cell type in a quantitative manner. The basic principle of SAGE is that short nucleotide sequence tags of 10 to 14 base pairs contain sufficient information to uniquely identify transcripts. Moreover, concatenation of these short sequence tags permits an efficient analysis of transcripts serially by sequencing of multiple tags within a single cloned element [14]. More recent variants of the method are based on longer sequence tags and integrate microarray technology [15]. Two years after the initial publication of the method, Johns Hopkins University researchers for the first time reported on gene expression profiles in normal and cancer cells based on SAGE [14]. The authors confirmed previous findings on RNA abundance in cells that had been obtained with the help of Rot curves that display RNA-DNA reassociation kinetics [16]. The total number of transcripts varied from approx. 14,000 to 20,000 between cell populations. Most transcripts (86%) were expressed at fewer than 5 copies per cell; however, the bulk of the mRNA mass consisted of more abundant transcripts (more than 5 copies per cell). The relative expression levels of transcripts were determined by dividing the number of tags observed in tumor and normal tissue. Most transcripts were expressed at similar levels. However, 548 of 14,000 to 20,000 transcripts were overrepresented or underrepresented in tumor versus normal cells. The average difference in expression for these transcripts was 15-fold. About 20% of them were less than 3-fold different. The authors also addressed the issue of whether cultured cell lines, frequently used in molecular cancer research, display gene expression patterns that mimic those found in the organ microenvironment. Interestingly, 72% of transcripts expressed at reduced levels in cancer specimens were also expressed at lower levels in cell lines. Likewise, 43% of transcripts exhibiting elevated expression in cancers were also upregulated in cell lines. Useful links and SAGE databases can be found at http://www.sagenet.org/.

p0050 A procedure very similar to SAGE is used to study cellular microRNAs (miRNAs), which are short ~22-nucleotide segments of RNA that have been found to play an important role in gene regulation. Small RNAs are isolated, linkers are added to each of them, and the RNA is converted to cDNA. Afterward the linkers containing internal restriction sites are digested with the appropriate restriction enzyme and the sticky ends are concatamerized. The concatamers are ligated into plasmid vectors and cloned, followed by sequencing. In this way, the expression levels of miRNA can be quantitatively assessed by counting the number of times they are present [17].

s0040 TRANSCRIPTOME ANALYSIS BASED ON MICROARRAYS: TECHNICAL PREREQUISITES

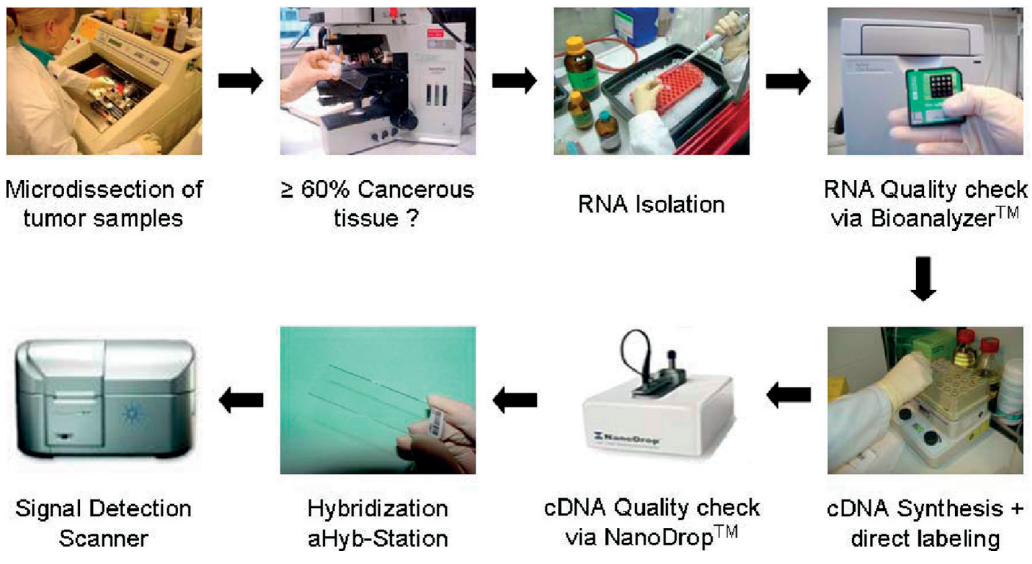
p0055 To date, microarrays are utilized by most researchers in studies related to (i) pathogenic processes at a genome-wide level, (ii) examination of drug effects, and (iii) elucidation of clinical features of disease that cannot be recognized by currently available molecular techniques or conventional histopathology and immunopathology. Microarray technology was pioneered by Pat Brown and colleagues at Stanford University. These researchers not only published the first applications of microarray to study biological questions, but also described the necessary technical devices in detail [18]. In this way, they contributed to the rapid dissemination of the technology. In parallel, microarray technology was developed at Affymetrix (Santa Clara, CA) [19]. The central element that is in common to the various forms of these techniques is that DNA-molecules, cDNA fragments or oligonucleotides are arrayed and immobilized at defined positions on a solid support or matrix. This method extends the existing technique of membrane-based arrays that were interrogated using radioactively labeled cDNA. The probes assembled on solid supports are hybridized with complementary and fluorescent dye-labeled RNA or DNA molecules (targets) derived from biological specimens such as cells, tissues, or blood. Fluorescent dye staining intensity after hybridization obtained within the position of the probe is a measure of the abundance of the corresponding nucleotide sequence in the complex mixture of RNA/cDNA targets. The different kinds of microarrays available today are distinguished by the number, density, design, and size of oligonucleotides or cDNA probes; the manner of chip manufacturing; and the experimental protocols for target hybridization.

Typical Workflow of a Microarray Experiment Starting from Tissue Samples s0045

p0060 Frozen tissue samples are dissected, fixed on glass slides, and stained. Histological characterization reveals the composition of the tissue, including the cell types of interest (and their frequency), the extent of necrotic areas (which contain degraded RNA), and presence of fatty tissue (from which RNA extraction is difficult). Optionally, laser capture microdissection can be used to precisely dissect the cells and tissue areas of interest. Subsequently, RNA isolation is performed, RNA yield is determined, and RNA quality is checked by electrophoresis. Isolated RNA is used for synthesis of labeled sample nucleic acid, mostly cDNA or antisense RNA (aRNA), which is quality-controlled by absorbance measurements. Most commonly, the last step is hybridization of the fluorescent dye-labeled sample nucleic acid to the probe DNA on the microarray (Figure 7.2).

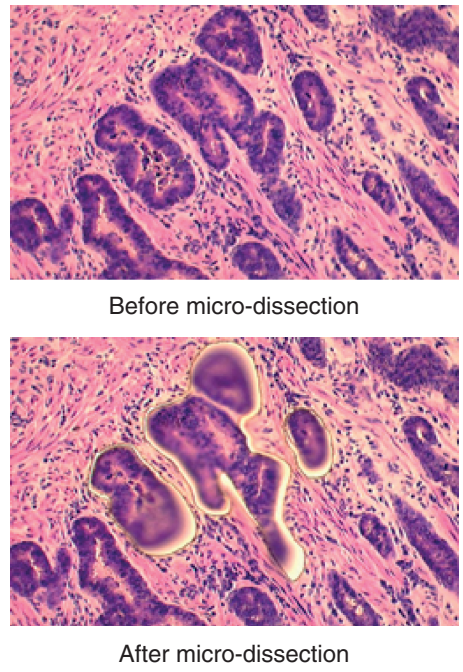
Production of Microarrays s0050

p0065 Microarrays represent a solid support (typically a glass slide or silicon surface) onto which probes are covalently linked using a chemical matrix (via epoxy-silane or amino-silane). The probes are dispensed either by



f0010 **Figure 7.2** Laboratory workflow of a typical microarray experiment.

contact spotting or applied as micro-droplets by techniques resembling ink-jet procedures used in printing. One of the industrial suppliers (Affymetrix Inc.) produces microarrays using photolithographic methods as in silicone chip production. The procedure allows production of high-density arrays containing millions of probes covered in a partially transparent hybridization chamber. In the original fabric, the probes are short sequences of 25 nucleotides. Each potential target sequence is represented with up to 11 different complementary oligonucleotides and 11 paired mismatch probes. A mismatch probe contains a single mismatch located directly in the middle of the 25-base probe sequence. While the perfect match probe shows fluorescence only when a sample nucleic acid binds to it, the paired mismatch probe is used to detect and eliminate any false or contaminating fluorescence within that measurement. Other industrial suppliers (such as Agilent, Illumina, Milteny, and others) and academic facilities use oligonucleotides of $\geq$60 nucleotides or cDNA fragments to improve hybridization specificity. These microarrays are manufactured as open slides and are handled openly or in special hybridization chambers during the entire chip processing.



f0020 **Figure 7.4** Laser microdissection of tissue samples.

s0055 Preparation of Target RNA

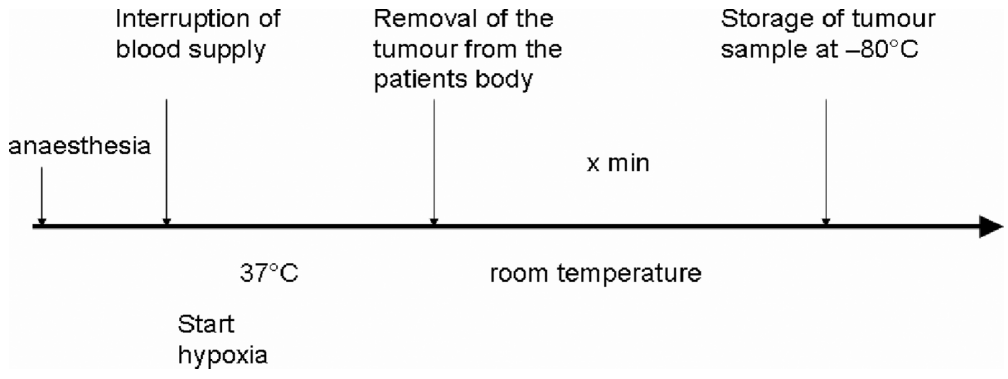
p0070 During surgical removal of malignant tumors, one of the first steps is the interruption of the arterial blood supply. From this moment on, the tumor tissue is exposed to hypoxia at body temperature (Figure 7.3). The duration between artery ligation and the final removal of the tumor can vary considerably and is not subject to standardization under clinical conditions. Following tumor resection, logistical constraints may lead to further considerable delay before the tumor material is finally shock-frozen at –80°C. Thus, this lengthy process might lead to a considerable extent of target RNA degradation.

s0060 *Laser Microdissection of Tumor Tissue*

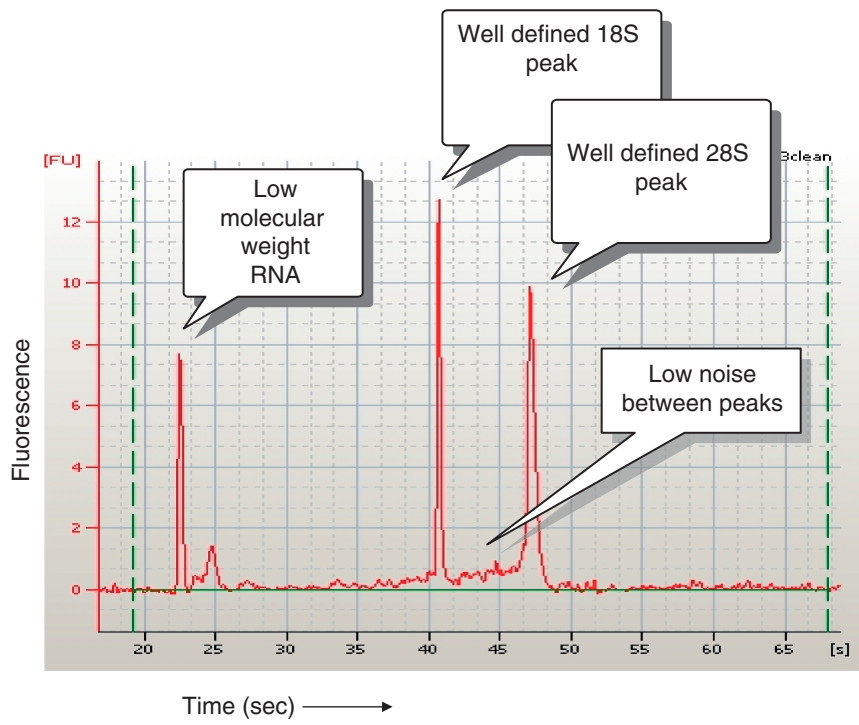
p0075 If the sample material is contaminated with nontumor tissue (such as stromal cells and lymphocytes) or if necrotic areas (with cellular debris) occur, data analysis will be severely hampered. Therefore, researchers often try to obtain homogeneous sample material. A convenient (although potentially laborious) approach is laser-assisted microdissection (Figure 7.4). Starting from a complex tissue architecture, areas with carcinoma cells only, stromal material, or any other area of interest, such as material located at the invasion front of the tumor, are obtained. In the example shown, colorectal carcinoma cells have been microdissected with the use of a laser beam. For each sample, 5 µm tissue sections were microdissected and RNA was extracted by a column-based procedure including DNAase digestion. Microdissection of 5×10^6 µm^2 per specimen yields about 10–20 ng RNA in about 2 hours of working time.

s0065 *RNA Quality Control, Labeling, and Target Amplification*

p0080 Figure 7.5 shows an electropherogram reflecting RNA quality according to the following criteria: (i) clear



f0015 **Figure 7.3** Preservation of tumor tissue during surgery.

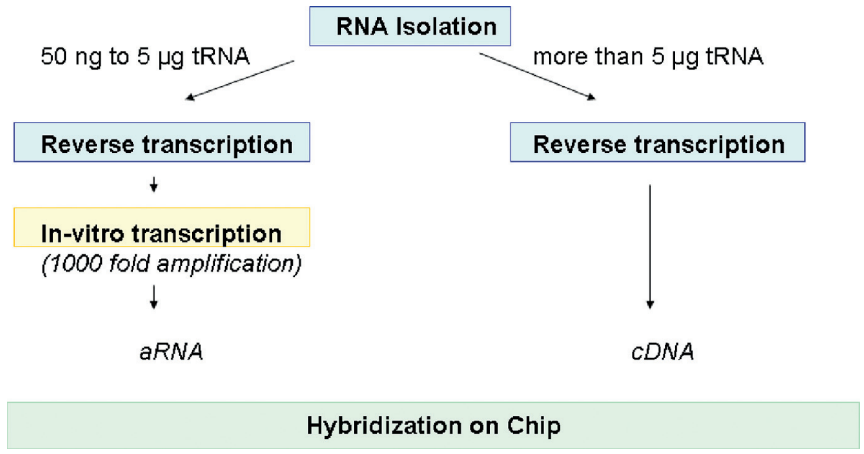


f0025 **Figure 7.5** RNA quality assessment.

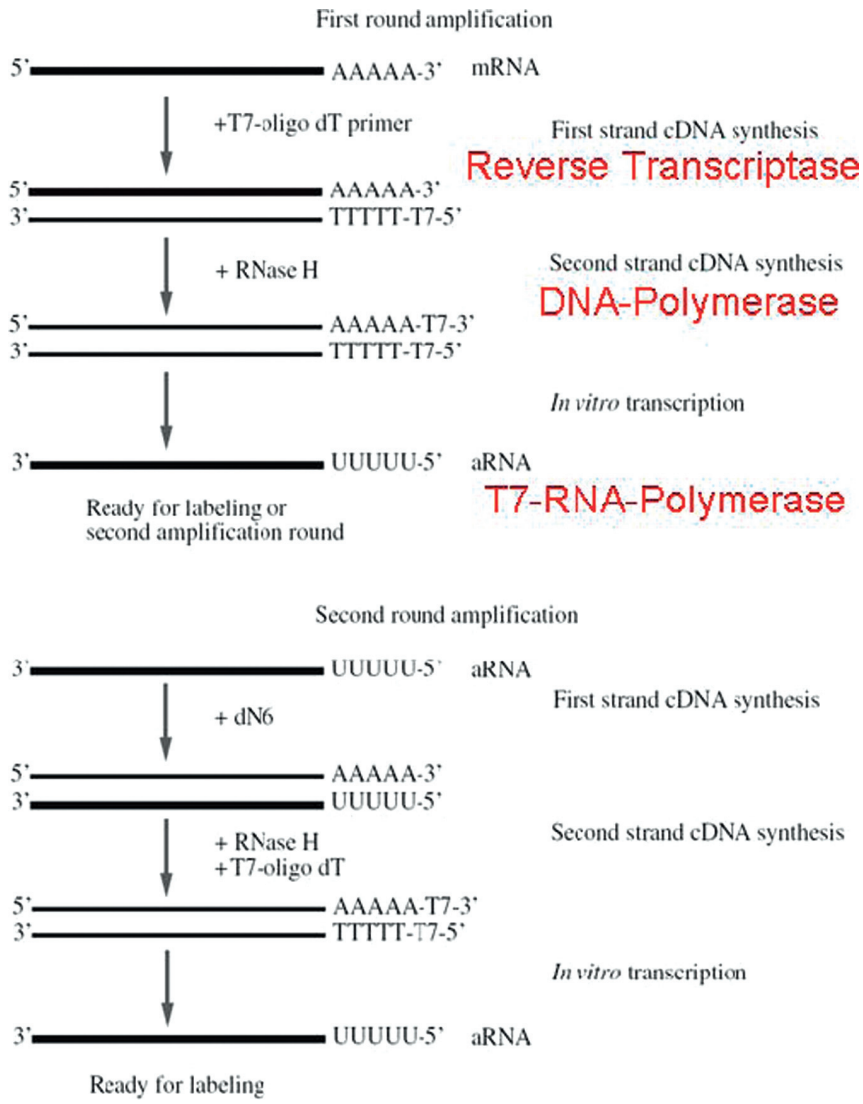
and well-defined 18 S and 28 S peaks of ribosomal RNA, (ii) low noise between the peaks, and (iii) no or only minimal evidence for low molecular weight material. For hybridization with the probe nucleic acids on the microarray, a labeled target sample nucleic acid is needed. RNA extracted from clinical specimens is used for synthesis of labeled cDNA without amplification of the sample RNA, or for production of aRNA (Figure 7.6). The process of aRNA synthesis allows high amplification of the sample material, which is of relevance if only small amounts of sample material are available, such as after laser microdissection. Whether amplification changes the outcome of the experiment by asymmetric amplification of high-abundance and low-abundance genes is still a matter of discussion. Common amplification procedures utilize Bacteriophage T7, T3, or SP6 RNA polymerases to transcribe RNA from a DNA template (Figure 7.7). The DNA template must have an appropriate polymerase binding site (called T7 in the figure) in its sequence, upstream of the region to be transcribed. A complex of this binding sequence ~20 base pairs in length linked to an oligo-dT sequence is incorporated into the cDNA by reverse transcription of the sample RNA (first strand synthesis). The RNA is then degraded by RNase-treatment, and the second strand is fabricated by DNA-polymerase. The resulting double-stranded cDNA serves as the template for the T7-RNA-polymerase producing RNA in antisense direction (aRNA), compared to the orientation of the template RNA. The entire procedure can be repeated resulting in a 1,000-fold or higher amplification of the RNA. By including labeled nucleotides (NTP) in the *in vitro* transcription reaction, one can incorporate labels into the synthesized RNA (using biotin-labeled UTP to generate biotin-labeled RNA or any kind of UTP-bound fluorescent dye to generate fluorescently labeled RNA). Because signal intensity of red and green fluorophores might not be identical, it is mandatory to invert or swap the fluorescent dyes used for labeling. For instance, in the case of Cy3/Cy5 fluorophores, the green signal intensity is often stronger than the red one. To compensate for this, the labeling reactions are exchanged between the two targets and microarray hybridization is repeated.

Microarray Hybridization: Two-Color Experiment s0070

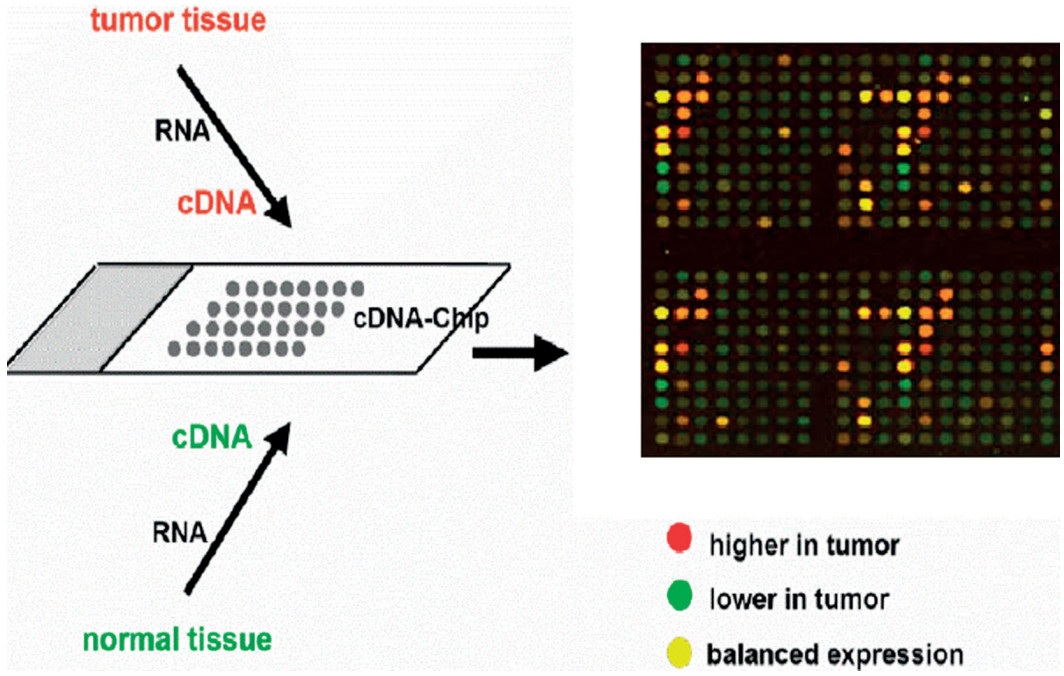
Hybridization of the target nucleic acid molecule to the probe DNA on the chip is most commonly detected and quantified by fluorescence. This requires a target molecule labeled with a fluorophore such as Cy3 or Cy5. The aim is to determine the relative abundance of the target molecule within the sample solution. Spotting of the probe molecules to the surface is a critical step. In order to account for spot-to-spot variations due to differing amounts of probe molecules in the spots, two-color experiments are used and the ratio of the two fluorophores in any single spot is determined (Figure 7.8). In a typical two-color experiment, RNA is extracted from tumor tissue and neighboring normal tissue. The RNA samples are labeled with different fluorophores, for instance, tumor RNA with a red fluorophore, normal RNA with a green one. Both samples are hybridized together on the microarray. If the spot then appears in red in the p0085



f0030 **Figure 7.6** Overview of RNA processing.



f0035 **Figure 7.7** RNA amplification (T7 *in vitro* labeling). Antisense RNA is generated using T7 polymerase. The microarray is populated with sense oligonucleotides corresponding to the genes of interest.



f0040 **Figure 7.8** Two-color microarray experiment.

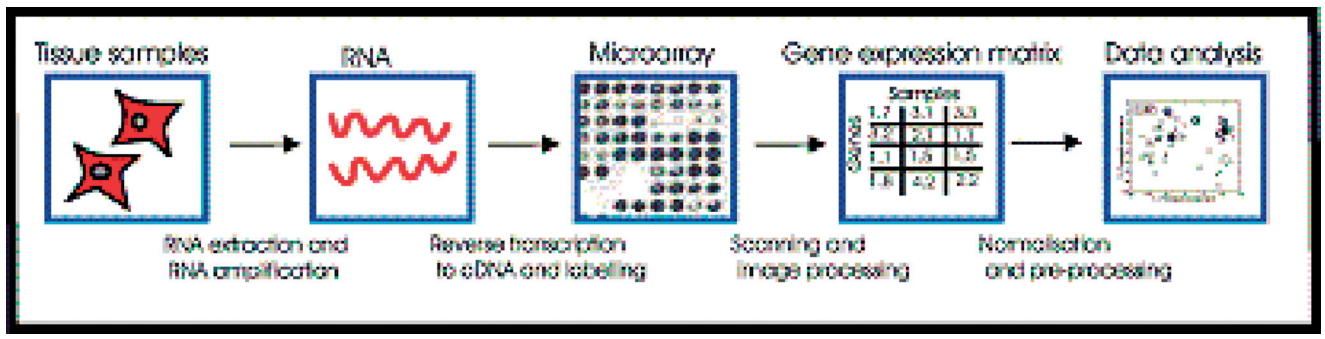
microarray scanner, this means higher expression (upregulation) of the corresponding gene in tumor tissue compared to normal tissue. If the spot looks green, higher expression of that gene in normal tissue compared to tumor tissue has been detected. If the gene is equally expressed in tumor versus normal tissue, the combination of the two fluorophores will produce a yellow color.

s0075 Image Analysis and Data Processing

p0090 Microarrays assess gene activities indirectly by measuring the fluorescence intensities of labeled target cDNA hybridized to cDNA or oligonucleotide probes on the array. There are different detection methods based on light emission (fluorescence), such as confocal laser scanning used by the scanners produced by Affymetrix, Agilent, Axon, HP, or CCD Imaging (Axon, Applied Precision). Other methods for detection of gene expression on microarrays include electrochemically based detection (Motorola) or radiolabeling (Molecular Devices, Hitachi). Most commonly, microarray scanners use laser light for excitation and matching filters and photomultiplier tubes for detection. During scanning, excitation light from the laser source hits the spots on the microarray. Fluorescent probes on the array emit Stokes-shifted light in response to the excitation light, and the emission light is collected by the photomultiplier tube. Scanning plays a pivotal role in the DNA microarray processing workflow and can profoundly affect the quality and reliability of microarray data. Typical sources of error from a microarray scanner include (i) noise in the background light, (ii) nonuniformity of the scan field, (iii) variations in laser brightness and detector gain, and (iv) spectral cross-talk between dye channels. Scanning of the hybridized microarray leads to an intensity picture displaying bright and dark spots (left side). While high resolution scanning (5 μm–10 μm) is the standard, some scanners are capable of scanning with 2 or 3 μm resolution. Using image analysis programs, the raw fluorescence intensity signals are transformed into numerical values for gene expression. This can involve several procedures to ensure the reliability of data. For example, spots which show defects due to printing errors, scratches, and the influence of dust particles should be excluded. Additionally, spot intensities might have to be corrected for any background fluorescence due to nonspecific hybridization. Finally, the obtained measures for gene expression can be analyzed with bioinformatic and statistical methods.

MICROARRAYS: BIOINFORMATIC ANALYSIS s0080

p0095 Finding meaningful structures and information in an ocean of numerical values obtained in microarray experiments is a formidable task and demands various approaches of data processing and analysis. In fact, microarray data analysis poses major challenges due to the sheer enormous lots of data produced. Although the type of data analysis naturally depends on the research questions posed, common steps in the analysis include (i) data preprocessing and normalization, (ii) detection of genes with significant fold changes, (iii) clustering and classification of expression profiles, and (iv) functional profiling (Figure 7.9) [20]. These steps are only partially separated. For example, the choice of preprocessing and normalization procedures can have



f0045 **Figure 7.9** Principal steps of a microarray experiment [20].

considerable impact on the results of clustering and classification. Also, the analysis methods to use might depend on the choice of microarray technology. Here we focus the data analysis for two-color spotted microarray. Other microarray platforms might require different bioinformatic approaches, especially for data-preprocessing and normalization.

s0085 Preprocessing, Visualization, and Normalization

s0090 *Preprocessing*

p0100 A first preprocessing step for two-channel microarray data is commonly the logarithmic transformation of signal ratios, which offers several advantages. First, fold changes of the same order of magnitude become symmetrical around zero for upregulaton and downregulation. For example, using $\log_2$ transformation, a positive or negative fold change of two is displayed as 1 or –1, respectively. Second, the spot intensities are usually more equally distributed along the scale, which enables an easier detection of intensity bias or saturation effects (Figure 7.10). Third, the variance of intensities is more homogenous with respect to a log intensity scale compared to a linear one. A homogenous variance is often required for statistical tests.

s0095 *Data Visualization*

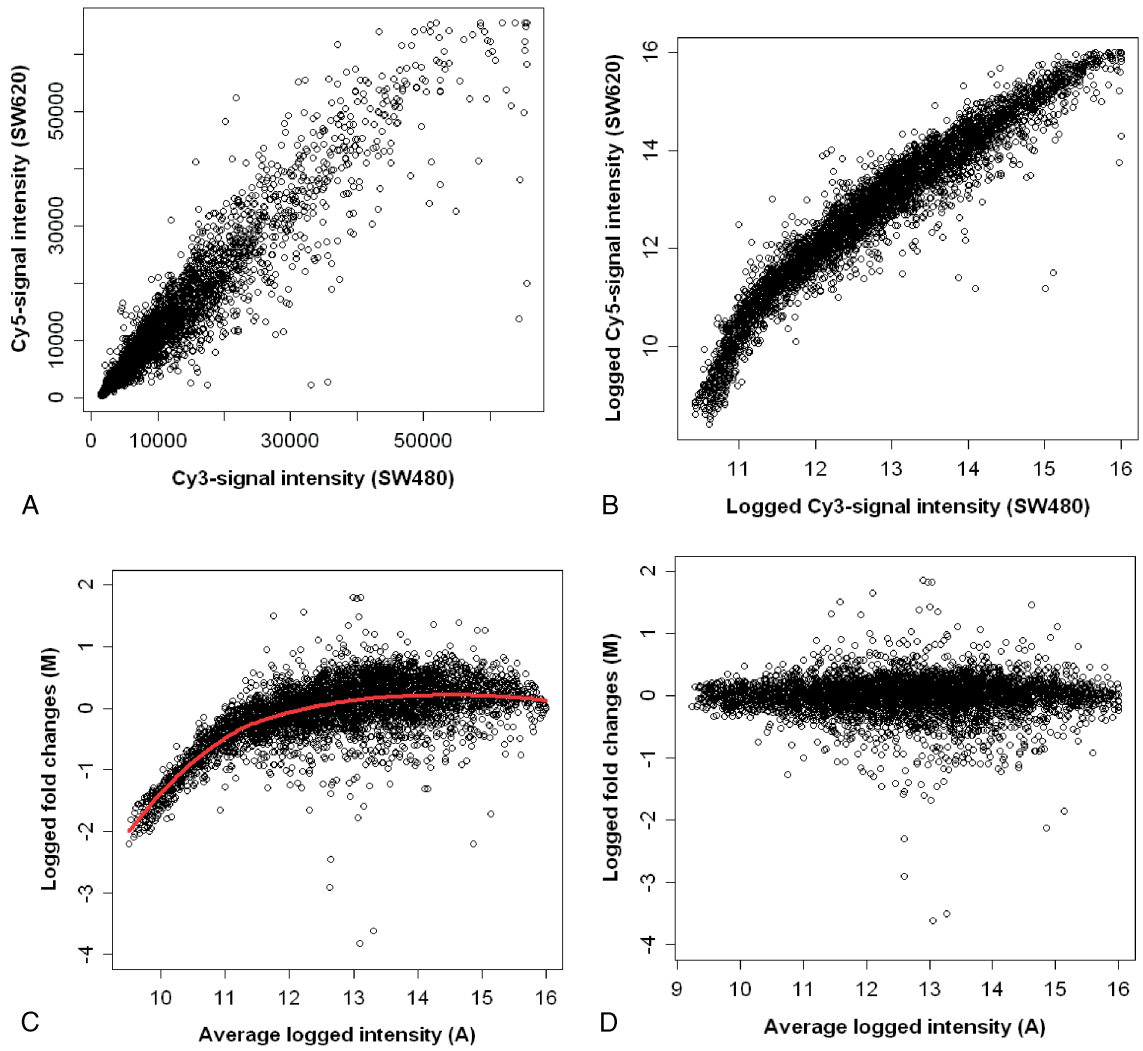
p0105 Plot representations are simple but very helpful tools to detect artifacts or other trends in microarray data. The most basic plots present the two channel intensities versus each other on a linear or log scales (Figure 7.10A and Figure 7.10B). More recently, *MA*-plots have become a popular tool for displaying the logged intensity ratio (M) versus the mean logged intensities (A). Although *MA*-plots basically are only a 45° rotation with a subsequent scaling, they reveal intensity-dependent patterns more clearly than the original plot (Figure 7.10C) [21].

s0100 *Normalization*

p0110 Raw microarray data are often compromised by systematic errors, such as differences in detection efficiencies, dye labeling, and fluorescence yields. Such signals are corrected by normalization procedures [22]. Although normalization is only an intermediate step in the analysis, it considerably influences the final results. Depending on the experimental design and microarray techniques applied, two main normalization schemes are used: (i) between-slide normalization (to compare signal intensities between different microarrays), and (ii) within-slide normalization (for adjustment of signals of a single microarray). While between-slide normalization is commonly used for one-color chip technology, within-slide normalization is applied mainly to two-color arrays for balancing both channels. An approach referred to as simple global normalization (a within-slide procedure) assumes that the majority of assayed genes are not differentially expressed and that the total amount of transcripts remains constant. Therefore, the ratios can be linearly scaled to the same constant median value in both channels. Alternatively, a set of so-called housekeeping genes can be selected, which are thought to be equally expressed in both samples. The median of these genes can then be taken to adjust the intensity in both channels by a linear transformation, so that the intensity medians of the housekeeping genes are the same. If a dye bias is suspected, the use of an intensity-dependent normalization procedure might be justified. A widespread method is to locally regress the logged signal ratios M with respect to the logged intensities A and to subtract the regressed ratios from the raw ratios. The derived residuals of the regression provide the normalized fold changes (Figure 7.10C). Additional normalization procedures are required, if measured spot intensity ratios show a spatial bias across the array.

Detection of Differential Gene Expression s0105

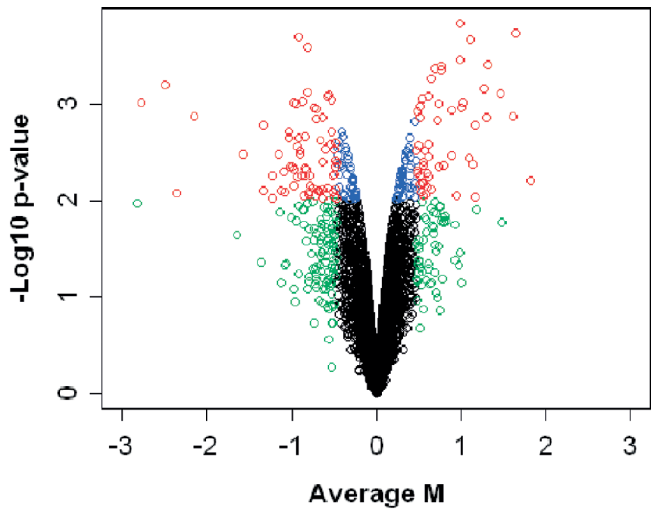
p0115 The standard task in microarray data analysis is the detection of gene expression changes. In early microarray studies, a fixed threshold for fold changes (such as two-fold) was arbitrarily defined to identify differentially expressed genes. However, the setting of fixed thresholds may yield a large number of false positives. Since the measured intensity signals usually are noisy, genes may show differential expression purely due to random signal fluctuations. Particularly, signals related



f0050 **Figure 7.10** Plot representations for signal intensities of a two-color array comparing colorectal cancer cell lines derived from primary carcinoma (labeled by Cy3) and from a metastasis (labeled by Cy5) [21]. The spot intensities in both fluorescence channels are shown using linear (A) and log2-scale (B). The use of log2-scale reveals nonlinear behavior, reflecting a dye bias toward Cy3 for low-intensity spots. The MA-plot presents this dye bias even more clearly and also a saturation effect in the Cy5 channel for large intensities. (C) To correct the dye bias, one can perform a local regression (red line) of M (D). The obtained residuals of the local regression, i.e., normalized logged fold changes, are well balanced around zero in MA-plot.

to weakly expressed genes are affected by high background noise and therefore require selection based on a larger threshold than strongly expressed genes. To distinguish more stringently noise from meaningful changes in gene expression, statistical tests are nowadays commonly used. Such tests assess the statistical significance of changes based on a set of assumptions about the distribution of the random errors. These errors are not correlated with any experimental variable and unlike systematic errors cannot be corrected by normalization. Random errors also set a limit of detectable changes of gene expression in microarray experiments. To estimate the random errors, experimental replicates are essential. After the random error is estimated, statistical significance can be assigned to changes in gene expression in the framework of a statistical test. Replication of microarray analysis provides also a valuable index of the overall quality of the experiment. Ideally, the goal is a high degree of consistency between different replicates. For subsequent visualization of the results of statistical tests, so-called volcano plots have become a popular mean. They offer the advantage of displaying both significance and fold changes observed (Figure 7.11).

p0120 Statistical testing is based on the assessment of the validity of explicitly formulated hypotheses. In general, a null hypothesis H_0 (for instance, that a gene is not differentially expressed) and a contradictory alternative hypothesis H_a (for instance, that a gene is differentially expressed) is set up. The alternative hypothesis is supported if there is evidence against the null hypothesis. The steps in hypothesis testing are as follows: (i) setting up H_0 and H_a, (ii) use of a test statistic to compare the observed values with the values predicted by



f0055 **Figure 7.11** The volcano plot is a graph that shows both fold changes and statistical significance of recovered genes. The graph displays negative log10–transformed *p*-values against the log2-fold changes (M). Volcano plots can be used for the selection of significant genes with a minimal required fold change. Data taken from the experiment described in Figure 7.10 [21]. Genes (displayed in blue and red) having statistically significant differential expression ($p < 0.01$) lie above a horizontal line. Genes (displayed in green and red) with larger fold changes than 1.6 lie outside a pair of vertical lines. Genes which fulfill both criteria are highlighted in red.

H_0, and (iii) definition of a region for the test statistic for which H_0 is rejected in favor of H_a. The level of significance of a test is the probability that the test statistic falls in the rejection region, if H_0 is true. The incorrect rejection of H_0 is called a type I error (in contrast to type II errors where H_0 is not rejected although it is false). The probability p that H_0 is true given the observed test statistic is called the *p*-value of the test.

p0125 A variety of statistical tests have been proposed for the identification of changes in gene expression. A classical test for comparing the mean gene expression values in two biological samples is the Student's t-test. Note that this test assumes the independence and normality of the expression values. The null hypothesis is that the mean value of both samples is equal. Depending on the alternative hypothesis, two types of t-test exist. For one-tailed t-tests, the alternative hypothesis includes the sign of the differences, whereas for the two-tailed test, positive and negative differences are treated equally. Alternatively, permutation tests are used that do not assume any particular data distribution. Permutation tests rely solely on the observed data examples and can be applied with a variety of test statistics. The basic idea of a permutation test is simple. Given labeled data, all permutations of the labels should be equally likely. Evaluating a chosen test statistic for all permutations, an empirical distribution of the test statistic can be derived. The percentage of random permutations that score higher than the actual observed case gives the significance level. However, a major restriction is that permutation tests can be computationally very intensive.

p0130 It is the nature of a microarray experiment that generally thousands of genes, if not the transcriptome as a whole, are tested for differential expression. If multiple tests are performed in parallel, the level of significance for the whole set of tests does not equal the level of significance for the single tests. For example, the probability *P* of rejecting a true null hypothesis in at least one of 1,000 simultaneous tests with a significance level of 0.001 is 63%. Therefore, an adjustment of the overall significance level and the *p*-values is necessary. A popular approach to circumvent the problematic interpretation of *p*-values in multiple testing is the calculation of the false discovery rate (*FDR*), which is defined as the proportion of false positives among significantly regulated genes. For instance, a *FDR* of 0.2 indicates that 20% of significant genes are likely to be false positives.

Classification

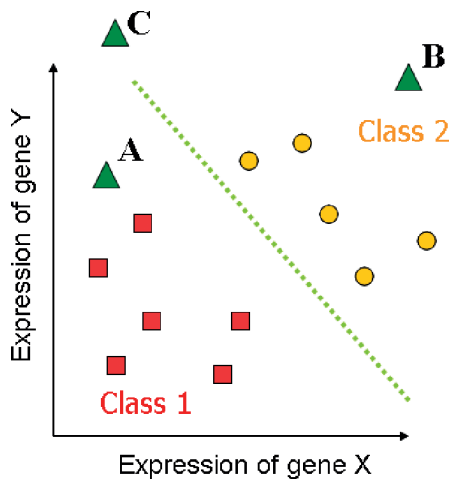
s0110

p0135 In its widest definition, classification is the assignment of a set of objects to a set of classes. In microarray data analysis, classification is commonly used to assign RNA specimens to different classes, for instance, those that distinguish types of tumors. Classification can be performed in an unsupervised and supervised manner. If class labels are not known in advance, the process is called unsupervised. If class labels exist, the process of classification is called supervised. However, note that in the context of microarray data analysis, the term *classification* usually refers to supervised classification, whereas unsupervised classification is generally referred to as clustering. The aim of supervised classification methods is to correctly assign new examples based on a set of examples of known classes. Thus, a classifier should generalize from known class examples to new unclassified examples. In microarray data analysis, the objects are provided as gene expression profiles and the classifiers have to identify decision boundaries between classes based on these given profiles (Figure 7.12). To achieve this goal, classifiers are optimized in a so-called (supervised) learning or training phase. After the optimization, the accuracy of the classifier can be tested using new examples of known class origin.

Challenges

s0115

p0140 Classification of tissue samples based on microarray experiments faces several major challenges. First, microarray data can contain a high level of noise. Experimental procedures such as tissue handling, RNA extraction, labeling, amplification, and hybridization can introduce additional variability in the measured expression levels. Furthermore, oligonucleotide and cDNA probes may not be specific to one gene, and several different genes may hybridize to the same probes (cross-hybridization). Second, in the typical microarray experiment thousands



f0060 **Figure 7.12** Extrapolation in classification: A classifier is trained on the sample from classes 1 and 2 based on the expression values of the two genes X and Y. The dashed line represents the border line derived by the classifier between the classes. Thus, new examples (represented by the dashed line) will be classified according to their gene expression values for X and Y. Thus, example A will be assigned to class 1, whereas example B will be assigned to class 2. The classification of C remains problematic, since it is located close to the border line and different to previously seen examples. Further tests would be advisable in this case.

of genes are monitored, while the number of RNA samples examined is usually restricted to hundreds or less. It is well known that classifiers generally perform poorly when the number of examples is small compared to the number of genes used for classification. Third, tissue samples are frequently heterogeneous in their composition. Thus, different cell types are represented in a single tissue sample used for RNA extraction. This heterogeneity can cloud the separation of the classes of interest, such as the distinction of cancer and normal tissue.

s0120 ## *Gene Selection*

p0145 Generally, large numbers of genes without changes in mRNA abundance introduce noise and may yield a poor classification performance. Gene selection aims at improving classification by excluding noninformative genes and thereby reducing the number of genes for the classifier. Genes are excluded if they only weakly contribute to the classification or not at all. Gene selection can be incorporated in a classification system in two different ways. First, gene selection and classification can be treated separately from the classification model. Genes are selected with respect to predefined criteria such as Pearson correlation or the significance in the Student t-test. This approach often has the advantage of being computationally inexpensive and easy to process. However, the selected genes are frequently highly correlated to each other and are likely to be redundant. Alternatively, the selection of genes is determined by the classification methods themselves in an iterative manner. This constitutes an integrated approach since an optimal set of features depends on the choice of the classifier.

Classification Methods s0125

p0150 Numerous methods for classification have been applied to microarray data. One of the most basic methods is the *k*-nearest neighbor method (with *k* as a positive integer). The classification rule is simple: A new example is assigned to the class most common among its *k* nearest neighbors. The distances of the examples are calculated based on their similarity in the expression profiles. For instance, if *k* is 1, then the example is simply assigned to the class of its nearest neighbor. Other currently popular classifiers are support vector machines based on statistical learning theory and belonging to the class of kernel-based methods. The basic concept of support vector machines is the transformation of input vectors into a highly dimensional feature space, where a linear separation may be possible between the positive and negative class members. In this feature space the support vector learning algorithm maximizes the margin between positive and negative class members of the training set in order to achieve a good generalization.

Cross-Validation s0130

p0155 Biological samples included in microarray analysis generally constitute only a small fraction of a larger sample cohort of interest. However, if a classifier is optimized based on a small number of examples, it will frequently show decreased performance on new data, a phenomenon usually called overfitting. An approach to prevent overfitting is *k*-fold cross-validation. It splits the data into *k* segments of which *k* – *1* segments are used for the training and one segment for the testing of the classifier. This is repeated *k* times, so that every segment is used for testing. The classification error in the validation procedure is then the sum over the error in the *k* tests. This approach has the advantage that a large part of the data can be retained for the training of the classifier, while the validation error is evaluated using all data examples equally. In the extreme case that *k* equals the number of data objects, the cross-validation is also referred to as the leave-one-out or jackknife method. If different models are compared by cross-validation, the model yielding the lowest validation error is generally selected.

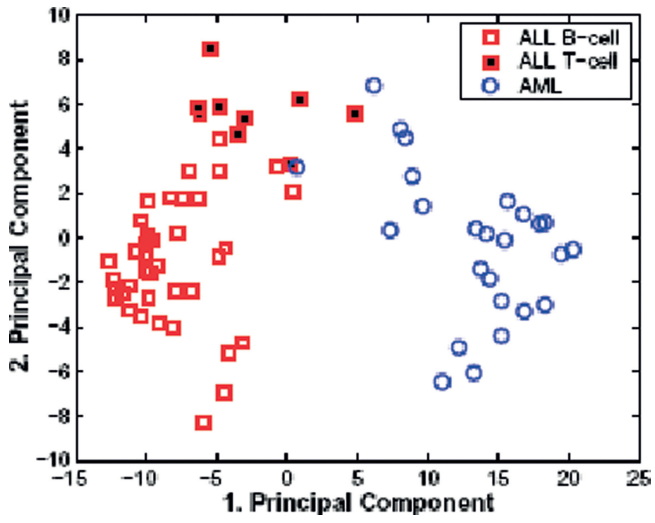
Visualization s0135

p0160 Data visualization is also an important component in the assessment of class distributions. It provides a global picture of the separation of samples and helps to identify potential outliers. However, a major challenge is the accurate representation of high-dimensional microarray data, where samples are defined by the expression values of thousands of genes. In contrast, data plots

are restricted to two or three dimensions. A standard method for representing high-dimensional microarray data is based on principal component analysis (PCA). The goal of this method is to find an optimal linear projection to a lower dimensional space. Practically, PCA leads a projection from the original gene expression values to an orthogonal basis of principal components. The principal components give the directions of the maximal variance in the data (Figure 7.13).

Cluster Analysis

p0165 Clustering, or unsupervised classification, has been studied for many decades in pattern recognition and related fields. Clustering methods generally aim at identifying subsets (clusters) in data sets based on the similarity between single objects. Similar objects are assigned to the same cluster, while dissimilar objects are assigned to different clusters. Cluster analysis, which can be understood as exploratory data analysis, is applied to search for patterns that may reveal relationships between individual examples. Frequently, the data structures detected by cluster analysis can give first insights into the underlying data-producing mechanisms. It is especially useful if prior knowledge is little or nonexistent since it requires minimal prior assumptions. This feature has made clustering a widely applied tool in microarray data analysis. One of the main purposes of clustering is to infer the function of novel genes by grouping them with genes of well-known functionality. This method is based on the observation that genes with similar expression patterns (co-expressed genes) are often functionally related and are controlled by the same regulatory mechanisms (co-regulated genes). Therefore, expression clusters are frequently enriched by genes of certain related functions. If a novel gene of unknown function falls into such a cluster associated with a certain biological function, it seems likely that this gene also plays a role in the same process. This guilt-by-association principle enables assigning possible functions to a large number of genes by clustering of co-expressed genes [23].



f0065 **Figure 7.13** Principal component analysis of leukemia samples based on 100 genes that have the largest squared Pearson correlation with the two classes of leukemia, ALL and AML [36]. The first two principal components include 63.3% of the total variance of the data. Most ALL and AML samples can be separated based on the first two principal components. However, note that the AML outlier makes a perfect separation difficult [21].

p0170 Clustering methods can be divided into hierarchical and partitional clustering. Hierarchical clustering creates a set of nested partitions, so that partitions on a higher hierarchical level comprise partitions on lower levels. The sequential partitioning is conventionally presented as a dendrogram. A dendrogram displays the clusters in a tree structure. The length of the branches represents the similarity between clusters. The shorter the branches, the more similar the clusters are. Usually, hierarchical clustering is performed in a stepwise agglomerative manner, starting with the single objects as singular clusters and gradually merging the clusters until all objects belong to a single cluster. To decide which clusters to merge, one calculates their similarity at every step of the clustering procedure. Clusters which show the largest similarity are subsequently joined. In microarray data analysis, both genes analyzed and biological samples can be hierarchically clustered. If these tasks are performed simultaneously, the procedure is also referred to as two-way clustering. Examples of such two-way clustering are shown in Figure 7.18 and Figure 7.20. As an alternative approach, partitional clustering splits the data in several separate clusters without the definition of a cluster hierarchy. It is commonly used to detect temporal gene expression patterns in time series experiments. The most popular methods for partitional clustering is k-means clustering. It starts with k randomly initiated cluster centers and splits the data in k partitions with a given integer k abased on the distance to the nearest cluster center. By repeated recalculation of the cluster centers and partitioning of the objects, this method aims to iteratively minimize the within-cluster variation.

p0175 Before a cluster analysis is performed, it is important to standardize the expression values, as co-expressed genes frequently show similar changes in expression but may differ in the overall expression rate. Therefore, the expression values of genes are usually adjusted to have a mean value of zero and a standard deviation of one. This ensures that genes with similar changes in expression have similar standardized expression values and thus will tend to cluster together.

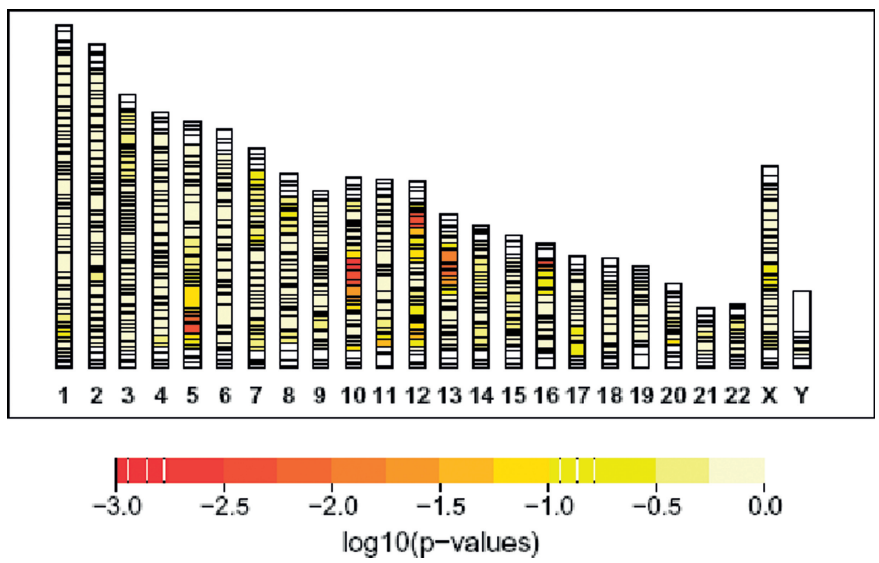
p0180 A crucial question is how many clusters can be retrieved from microarray data. This is generally difficult to answer for gene expression data as the detected clusters frequently are inhomogeneous and may show substructures, which can be interpreted as clusters themselves. While hierarchical clustering is able to indicate the different levels of clustering in the resulting dendrogram, partitional clustering algorithms lack the ability to indicate substructures in clusters. It is also important to note that common clustering methods always produce clusters due to the underlying

algorithms. For one to critically assess reliability of the clusters that result from this analysis, several measures for cluster validity have been introduced. Many of them assess the quality of clusters based on criteria such as compactness and isolation. Alternatively, clusters that emerge from a given analysis can be examined based on their robustness relative to the noise in the data set. For this approach, one would artificially add noise to the data before clustering and compare the newly identified clusters with the original ones. Clusters that remain the same despite added noise are likely to be more reliable than clusters which vanish in the presence of noise.

s0145 Functional Profiling and Other Enrichment Analyses

p0185 Frequently, large numbers of differentially expressed genes are detected in microarray experiments, making the overall interpretation of the results difficult. If further research is not focused on a few candidate genes, a helpful tool for understanding the complexity of the data set is functional profiling. This approach aims at identifying biologically informative classes of genes that are likely to be affected in the experiment. The underlying framework is given by Gene Ontology (GO), a popular database providing gene annotations in a systematic manner for various species [24]. In GO, genes are assigned to a defined set of categories describing molecular functions, biological processes, and cellular compartments. The categories themselves are placed in a tree-like structure with parent-child relationships. Categories at low levels are fairly general (for instance, those related to cell death) in contrast to more specific categories at higher levels (such as those that function in the regulation of caspases). Since GO is computer-accessible, the assignment of annotations to a list of genes has become much easier and rapid. After automatic gene annotation, functional profiling is performed by determining which GO category is represented more frequently than expected in the list of differentially expressed genes. Collecting involved GO categories provides a more holistic picture than the inspection of individual genes. Nowadays, numerous software tools are available for functional profiling of microarray experiments. Besides the list of differentially expressed genes, the list of genes represented on the microarray is a necessary prerequisite for the analysis. Comparing the functional composition of both lists, one can calculate the statistical significance for enrichment of differentially expressed genes in a biological process. The user typically obtains a list of significantly enriched GO categories associated with a particular experimental condition or disease state. However, there are important caveats with respect to using GO. Results can vary considerably when using different software tools. In addition, while there is a considerable number of manually curated gene annotations in GO, the majority of human genes have been annotated solely by computational means [25].

p0190 The concept of functional profiling to examine enrichment of genes belonging to defined functional categories can be applied in a general way. Another example of enrichment analysis is the examination of the chromosomal location of differentially expressed genes. This strategy yields a first indication for potential underlying changes in the chromosomal structure, such as copy number alterations or deletions, integrating transcriptomics and genomics (Figure 7.14).



f0070 **Figure 7.14** Chromosomal localization of genes exhibiting differential expression. The statistical significance for local enrichment of upregulated genes in a metastatic colorectal cancer cell line compared to a primary carcinoma line (SW480) is shown. To detect possible changes in the chromosomal structure of the two related cell lines, researchers differentially expressed genes to their corresponding chromosomal locus. Subsequent enrichment analysis using a sliding window technique indicated several potential chromosomal alterations.

s0150 ## Microarray Databases

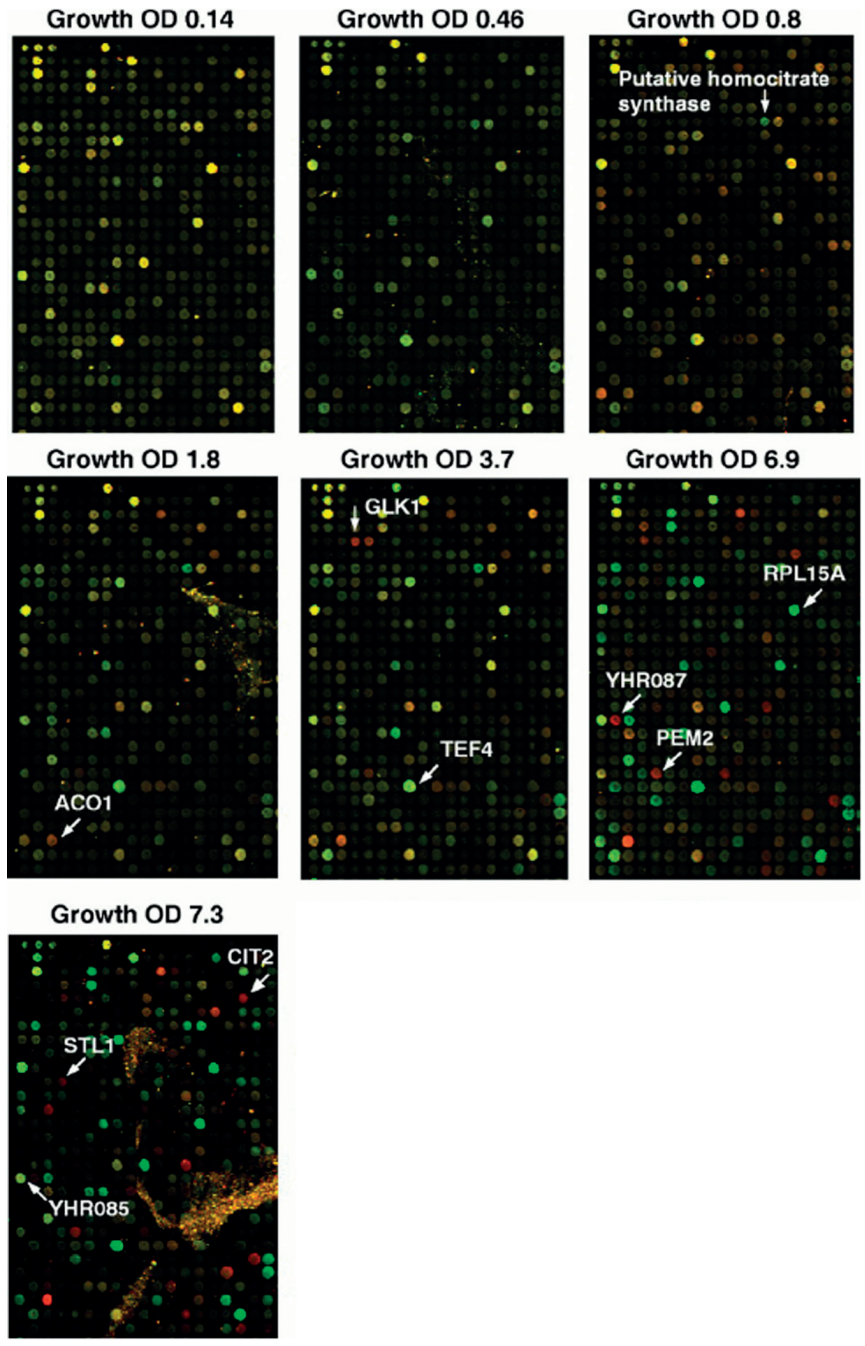
p0195 Microarray experiments produce massive quantities of gene expression data. Therefore, it has become good practice to deposit generated microarray data in publicly accessible databases. This practice is typically requested by journal editors prior to publication of the data. This allows independent researchers not only to scrutinize data obtained by others for their own interests, but also to validate the original analyses. In fact, the practice of sharing microarray data has allowed the community of bioinformaticians and statisticians to develop new methods and compare them with existing ones based on publicly accessible data sets. Such comparisons have been extremely valuable, since results from microarray experiments rely not only on the raw data, but to a substantial part on the applied computational methods. However, the interpretation of microarray experiments requires a common forum providing various types of information on the examined samples and experimental conditions, arrayed genes, microarray platforms, and applied computational approaches. Therefore, standards for publishing microarray data have been established. The most important one is the Minimum Information About a Microarray Experiment (MIAME) standard [26]. This standard requires deposition of raw microarray data, normalized data, sample annotation, experimental design, description of the microarray, and experimental conditions. Additionally, the development of large central microarray databases has facilitated data sharing. One of the first repositories was the Stanford Microarray Database (http://genome-www5.stanford.edu/), including a large collection of two-color array experiments. Currently, the two major public microarray databases are Gene Expression Omnibus provided by the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/geo/) and Array Express (http://www.ebi.ac.uk/microarray-as/ae/) provided by the European Bioinformatics Institute. Both databases follow the MIAME standard and provide several options to users for depositing their own microarray data and for accessing information from others.

s0155 # MICROARRAYS: APPLICATIONS IN BASIC RESEARCH AND TRANSLATIONAL MEDICINE

s0160 ## An Early Example for Microarray-Based Gene Expression Profiling Aimed at Understanding Metabolism

p0200 Ten years ago, microarray analysis was still in its infancy. Most bioinformatic tools available today had not been developed. Here we present one of the early applications of microarray technology published in 1997 by Pat Brown's group at Stanford University [27] as a paradigmatic example. The Brown group used a microarray representing 6,600 yeast genes to study a process known as diauxic shift. In glucose-rich medium, yeast cells generate energy by fermentation and convert the substrate glucose to acetaldehyde, which is then reduced to ethanol by alcohol dehydrogenase. When glucose is consumed, cells switch from fermentation to respiration and utilize the produced ethanol as a carbon source to generate glycogen. To study gene activity during this process, this group labeled cDNA obtained from cells before reaching exponential growth phase with the red fluorescent dye Cy3 as a reference. RNA was prepared at several time points during growth phase and substrate shift, reverse transcribed into cDNA, and labeled with the green fluorescent dye Cy5. Then the Cy5-labeled cDNA (RNA) targets and the Cy3-labeled reference were hybridized to the arrays, and the relative intensities of Cy3 versus Cy5 were measured for each time point. With increasing yeast cell growth indicated by enhancement of the optical density of the cultures, the number of differentially expressed genes increased, as did the level of differential expression indicated by the intensity of red and green staining (Figure 7.15). While in sparse culture, only 0.3% of the genes were altered and the maximal difference in expression was 2.7-fold, 30% of the genes were altered at the final time point of the experiment. More than 300 genes exhibited a differential expression of more than 4-fold. This experiment confirmed that alterations of expression can be efficiently determined in a time-resolved manner by microarray analysis. It also suggested that besides the genes, whose biochemical function was well known already, a number of genes that had not been characterized, approximately 400 at the time of the analysis, could potentially play a role in the diauxic shift, growth control, and energy generation. In summary, these candidate genes were placed into a potential functional framework. This became one of the major goals of microarray experiments in subsequent microarray studies, not only in yeast but also in mammalian systems including human cells and tissues.

p0205 In the yeast microarray experiment, the Stanford researchers went one step further and asked the question, if co-expressed genes are regulated in a similar fashion. Several distinct gene clusters comprising elements that exhibit the same expression pattern of upregulation or downregulation over time were identified. When the gene promoters of the co-expressed genes were analyzed, common regulatory sequences were recovered. For example, all but one gene (*IDp2*) contained a regulatory element named CSRE—carbon source responsive element (Figure 7.16). The CSRE is required to activate transcription of the genes involved in gluconeogenesis and the glyoxylate cycle in yeast. And indeed, all of the genes found in this cluster play a role in the glyoxylate cycle (*MLS1*, *IDP2*, *ICL1*), in the conversion of acetate to acetyl-CoA (*ACR1*), and in the production of fructose-6-phosphate (*FBP1*). In summary, the basic conclusions from the yeast experiment were (i) similar function is associated with co-regulation, (ii) co-regulation provides a way to define novel functional modules, (iii) co-regulation provides a way to define potential functions for unknown genes, (iv) co-regulation is based on similar transcriptional regulatory factors, and (v) co-regulation is a basis for the identification of regulatory mechanisms.



f0075 **Figure 7.15** Gene expression changes associated with increased culture density over time [27]. In each of the arrays used to analyze gene expression during the diauxic shift, red spots represent genes that were induced relative to the initial time point, and green spots represent genes that were repressed. Note that distinct sets of genes are induced and repressed in the different experiments. Cell density as measured by optical density (OD) at 600 nm was used to monitor the growth of the culture.

s0165 ## Elucidating the Transcriptional Basis of the Serum Response in Human Cells

p0210 Diploid human fibroblasts, like most other cell cultures, require the presence of serum growth factors in their culture medium. Routinely, these factors are supplied by adding fetal calf serum to the culture medium. Cultured cells can be made quiescent by serum deprivation. When fetal calf serum is added to such cells, they quickly resume cell cycle progression and proliferation. This cellular reaction is called the serum response, which was chosen as another early example to demonstrate the power of microarray analysis [28]. This time the Stanford researchers obtained RNA from serum-starved cultures and prepared target cDNA labeled with Cy3. RNA from all other time

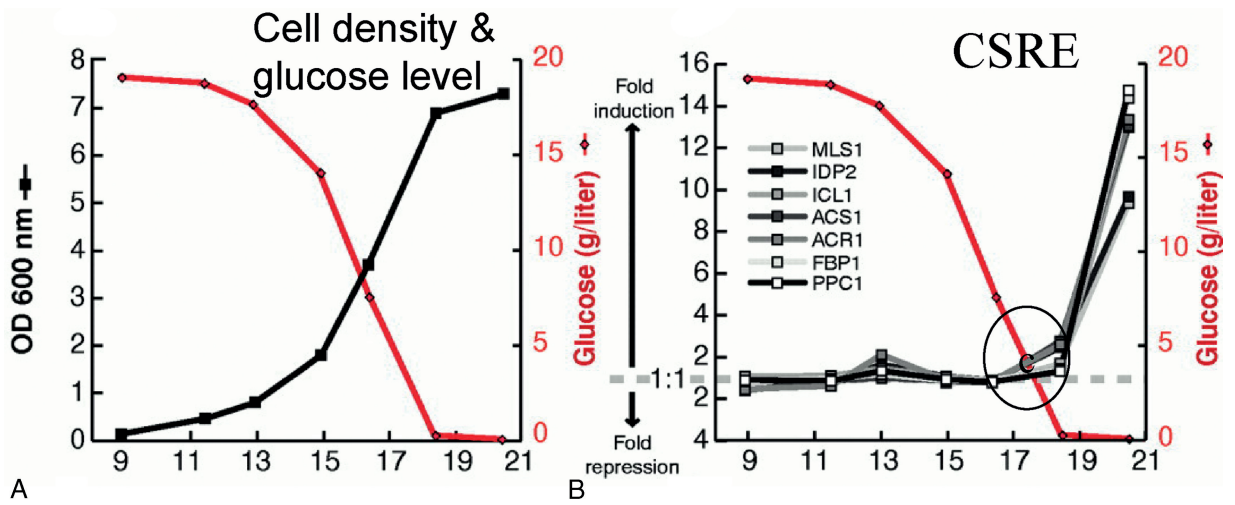
points following serum stimulation was used to prepare Cy5-labeled targets. In two-color microarray experiments, the targets were hybridized to a human cDNA array representing 8,600 human sequences. About 4,000 of them were known human genes, 2,000 sequences were related to these annotated genes, while the remaining genes were ESTs without known function. Figure 7.17 shows a subset of genes ($n = 517$) whose expression changed up to 8-fold during serum stimulation.

p0215 The transcriptional response toward serum stimulation is very rapid; the earliest changes can be observed as early as after 15 minutes. The genes can be divided into several clusters, which exhibit a common regulatory scheme. Some clusters show a characteristic pattern of upregulation followed by downregulation (cluster C) or the reverse pattern (clusters E and B). Based on current knowledge in molecular cell biology and data mining for gene functions, the Stanford researchers performed a functional gene clustering. This analysis was done at a time before the Gene Ontology became available. One functional cluster of co-regulated genes included transcription factors including the ones known to be involved in the immediate early gene response, permitting rapid responses without the need for protein synthesis. Another cluster included phosphatases. Their functional relevance was not known at the time of the analysis. Today it is well established that the phosphatases limit signaling kinase activity, which is rapidly stimulated upon growth stimulation, by negative feedback. Not surprisingly, the researchers recovered genes encoding cell cycle regulatory proteins. Inhibitory genes were quickly downregulated, paving the way for re-entry of the serum-starved cells into the cell cycle. With a short delay, cell cycle stimulatory genes were upregulated, among them cyclin D1 and DNA topoisomerase, which is required for chromosome segregation at mitosis. A more surprising feature of the analysis was the appearance of genes with known functions in wound healing. This referred not only to genes whose products function intracellularly, but also to genes whose products play a role in remodeling clot structure and the extracellular matrix, as well as in intercellular signaling. While previous studies had aimed at elucidating intracellular events in wound healing, gene expression profiling of the serum response indicated the relevance of extracellular events during the first 24 hours in this process.

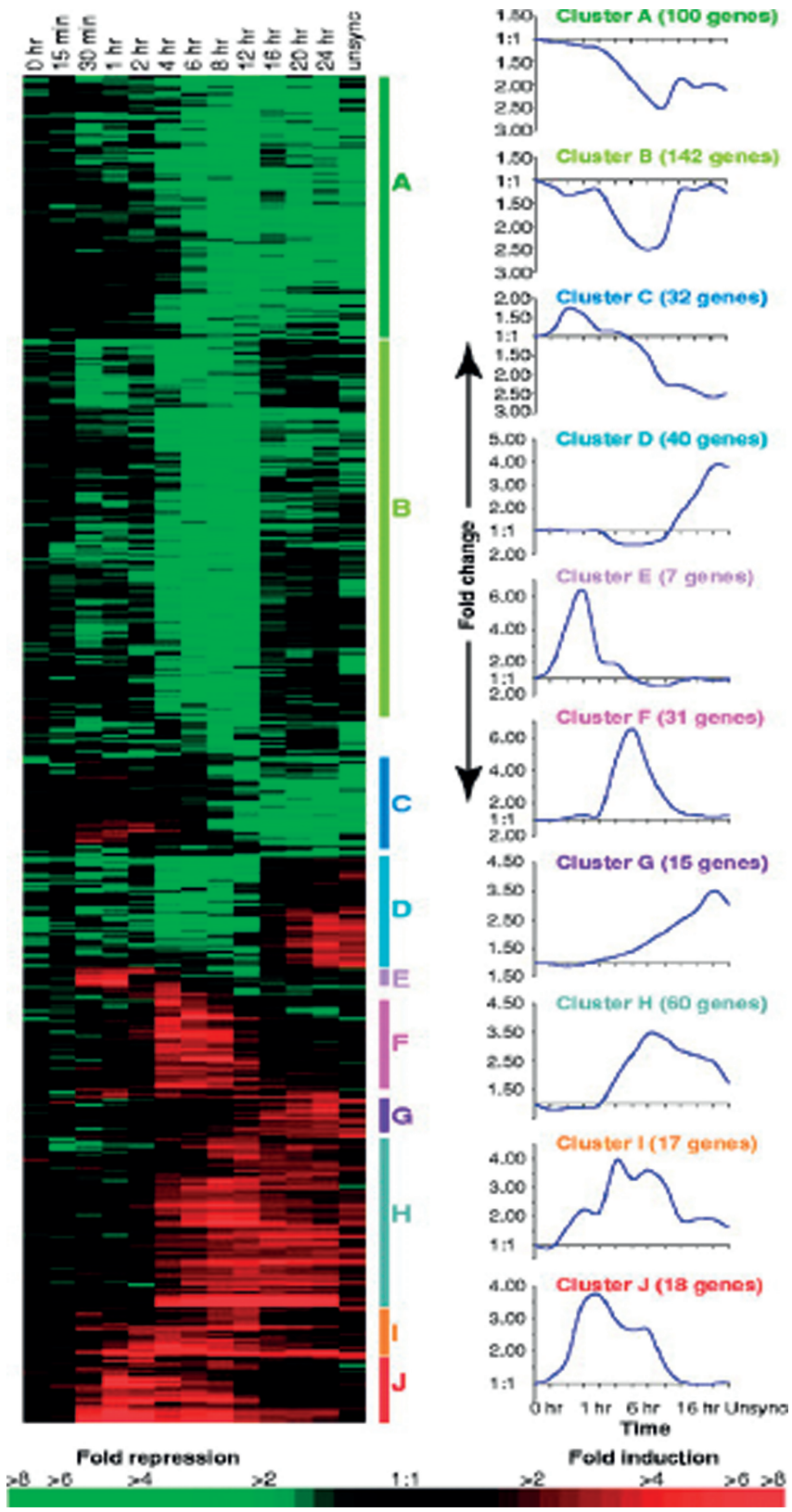
Microarray Applications in Cancer Pathogenesis and Diagnosis s0170

A recent PubMed search revealed that the majority of p0220
microarray and gene expression profiling studies in medicine are devoted to some aspect of cancer. Cancer studies far outnumber similar studies in cardiovascular diseases, neurodegenerative diseases, infection, inflammation, and other diseases (Table 7.1). Therefore, we have chosen some prominent applications of microarrays in the field of cancer as paradigms to demonstrate the power of transcriptome analysis.

The current themes of transcriptomics in cancer p0225
analysis are related to the mechanisms of pathogenesis, cancer classification, and outcome prediction. To elucidate the mechanisms of tumorigenesis and metastasis, particularly to study the complexity of the underlying processes, researchers frequently use microarrays. Cancer classification based on microarray studies aims at identifying characteristics beyond anatomical site and histopathology. Outcome prediction tries to overcome the limitations of current diagnostic procedures by establishing gene-based criteria to indicate and predict tumor prognosis and therapy response, even for individual cancer patients. Basically, there are three types of microarray-based approaches: (i) class comparison,



f0080 **Figure 7.16** Analysis of regulatory modules within the promoters of co-regulated genes associated with the diauxic shift [27]. (A) A group of genes exhibited an CSRE (carbon source element) within their promoter regions. (B) The growth curve, shown as increasing optical density (black line) and decreasing glucose level (red line), allows determination of a glucose threshold for the onset of gene expression due to the CSRE.



f0085 **Figure 7.17** Hierarchical clustering of genes induced or repressed during serum response in human fibroblasts [28]. Ten gene clusters (A–J) harboring 517 genes, which show significant alterations in gene expression over time, are depicted. For each gene, the ratio of mRNA levels in fibroblasts at the indicated time intervals after serum stimulation compared to their level in the serum-deprived (time zero) fibroblasts is represented by a color code, according to the scale for fold-induction and fold-repression shown at the bottom. The diagram at the right of each cluster depicts the overall tendency of the gene expression pattern within this cluster. The term *unsync* denotes exponentially growing cells.

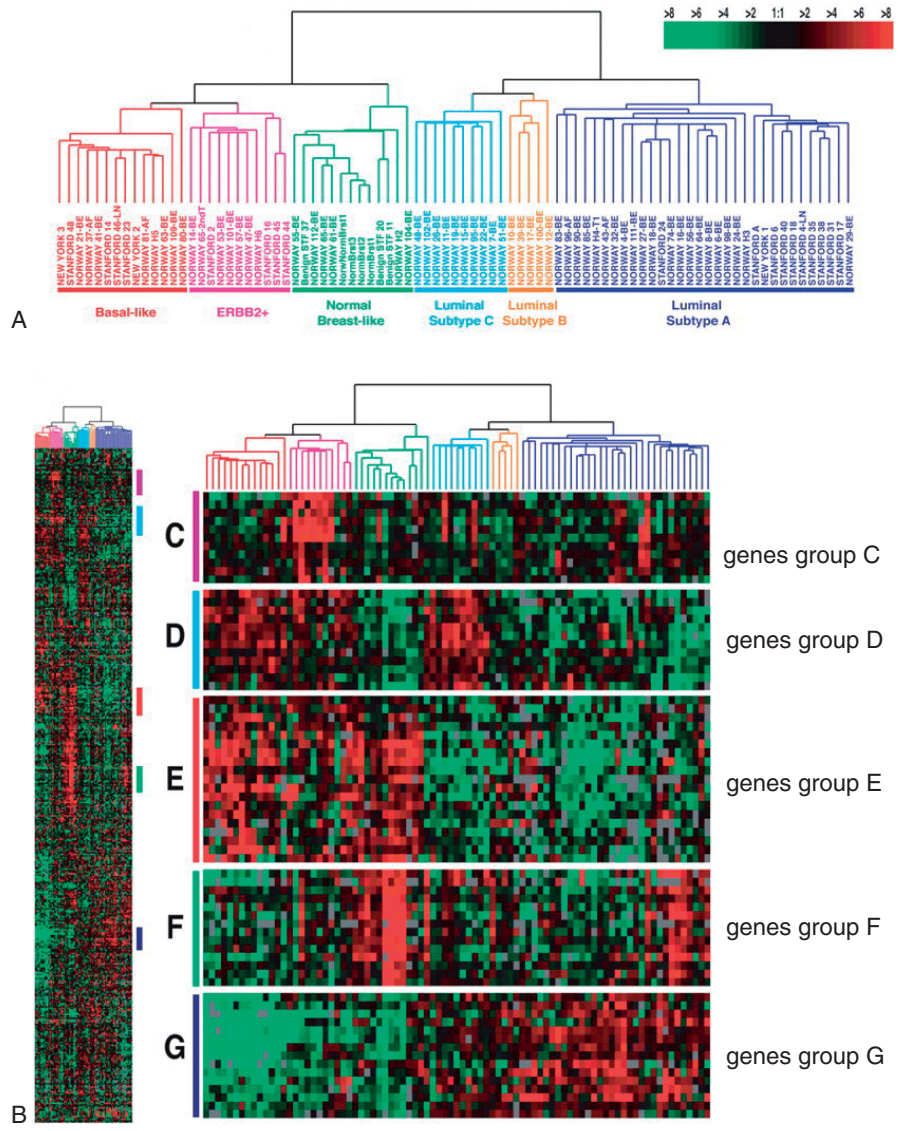
t0005 **Table 7.1** Number of Published Microarray and Gene Expression Profiling Applications in Research. Results of a PubMed Search Dated July 13, 2008, Using Single Keywords and Combinations of Two Keywords Without Limits to Publication Years.

	2nd Keyword		
1st keyword	**None**	**Gene Expression Profiling**	**Microarray**
None	-	35,712	25,841
Pharmacology	4,244,442	8,358	5,668
Diseases	3,520,603	8,169	5,904
Cancer	*2,172,278*	*11,513*	*8,657*
Pathology	1,813,867	7,790	5,708
Cardiovascular diseases	*1,460,061*	*1,149*	*616*
Development	1,266,779	8,529	5,491
Immunology	1,094,163	3,021	2,078
Infection	*892,647*	*1,665*	*1,254*
Nutrition	404,775	528	371
Drug development	281,029	1,798	1,206
Inflammation	*280,552*	*1,318*	*1,001*
Neurodegenerative diseases	*152,735*	*457*	*277*
Toxicology	82,952	649	460

(ii) class discovery, and (iii) class prediction. Using class comparison, one tries to compare the expression profiles of two (or more) predefined classes. For example, two tissue samples, normal versus malignant cells or tissues, different developmental stages, or cells treated with drugs under different conditions. Using class discovery, one tries to identify novel subtypes within an apparently homogenous population. In this case, microarray analysis is used to identify features that cannot be distinguished by other available tools. The starting point usually is a homogenous group of specimens, in which a concealed proportion behaves aberrantly or exhibits invisible or unknown features. The problem of cancer treatment falls into this category, since patients who are stratified into treatment groups according to standard histopathological criteria often respond differently to therapy. We will see that microarray studies can help to successfully address this urgent clinical problem. Class prediction means to find a set of features that are predictive for a certain, predefined class. This is perhaps the most sophisticated type of microarray application. It is usually based on class discovery, but now the characterization of a novel class is intended. Rather, the idea is to establish a classifier. A classifier is a set of features, like genes, proteins, micro-RNAs that are surrogate markers for a certain class. This is the common approach to identify predictive gene sets or gene signatures that can predict clinical outcome or therapy response.

s0175 Identification of Hidden Subtypes Within Apparently Homogenous Cancers

p0230 The group of T. Sorlie identified 456 genes out of 8,000 genes on a microarray that discriminated between tumor subclasses in a cohort of 65 tumors from 42 breast cancer patients [29]. Gene expression patterns of breast carcinomas helped to distinguish tumor subclasses with clinical implications. Using hierarchical clustering, the Sorlie group distinguished five distinct tumor groups characterized by their gene expression pattern: the basal epithelial cancer type, the luminal epithelial cancer types A–C, a group displaying expression of the breast cancer oncogene *ERBB2* (*HER2*), and a group without any known feature. There was yet another group showing features of normal breast epithelial cells (Figure 7.18). In the next step of the analysis, the researchers addressed the question as to whether these different groups are characterized by distinct clinical parameters. Therefore, they compared the groups by certain statistical methods, among others by univariate statistical analysis, for either overall survival or relapse-free survival monitored for up to 4 years (Figure 7.19). The patient groups that were *ERBB2*-positive or were characterized as basal epithelial breast tumors had the shortest survival times. While this information was not new for the *ERBB2*-positive tumors, the basal epithelial breast cancers belong to a novel group with an obviously bad prognosis. One characteristic of this tumor type is the high frequency of *TP53* mutations. The tumor suppressor gene *TP53*, well known as the guardian of the genome, is lost or mutated in more than 50% of all advanced human cancers, and might be responsible for the bad prognosis. There was also a difference in clinical outcome between the luminal-type breast cancers. Most strikingly, luminal A tumors exhibited a very good performance at least within 4 years, while luminal B or luminal C tumors were intermediate. In conclusion, this study opened the door to further screen many tumors for gene signatures indicative of the clinical performance of breast cancer patients. With respect to cancer treatment, the most important issue is to find gene sets predictive for the susceptibility or resistance to therapy, particularly to chemotherapy, and to clinical outcome in the absence of other

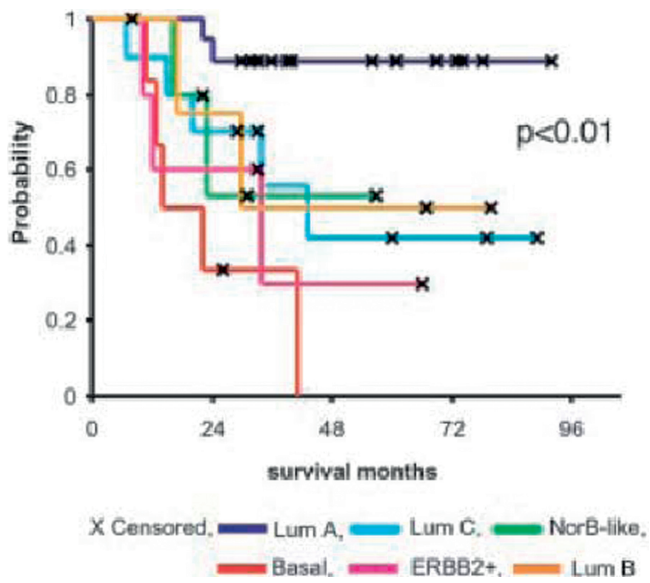


f0090 **Figure 7.18** Differential breast cancer gene expression [29]. Gene expression patterns of 85 experimental samples (78 carcinomas, 3 benign tumors, 4 normal tissues) analyzed by hierarchical clustering using a set of 476 cDNA clones. (A) Tumor specimens were divided into 6 subtypes based on their differences in gene expression: luminal subtype A, dark blue; luminal subtype B, yellow; luminal subtype C, light blue; normal breast-like, green; basal-like, red; and ERBB2+, pink. (B) The full cluster diagram obtained after two-dimensional clustering of tumors and genes. The colored bars on the right represent the characteristic gene groups named C to G and are shown enlarged in the right part of the graph: (C) ERBB2 amplification cluster, (D) novel unknown cluster, (E) basal epithelial cell-enriched cluster, (F) normal breast epithelial-like cluster, (G) luminal epithelial gene cluster containing ER.

conventional indicators. So far, questions related to chemotherapy resistance and drug sensitivity have also been addressed by microarray studies, but have not been advanced to the clinical level.

s0180 Gene Expression Profiling Can Predict Clinical Outcome of Breast Cancer

p0235 Breast cancer patients with the same stage of disease exhibit markedly different treatment responses and overall outcome. However, histopathological assessment of these cancers does not have sufficient power to discriminate which patients will perform well versus those that will not. The strongest predictors for metastases (such as lymph node status and histological grade) fail to classify accurately breast tumors according to their clinical behavior. None of the signatures of breast cancer gene expression reported to date allow for patient-tailored therapy strategies. The study published by van't Veer et al. [30] in the Netherlands has pioneered gene array-based breast cancer diagnostics. The study was based on a well-characterized cohort of breast cancer patients ($n = 117$). This included 78 sporadic primary



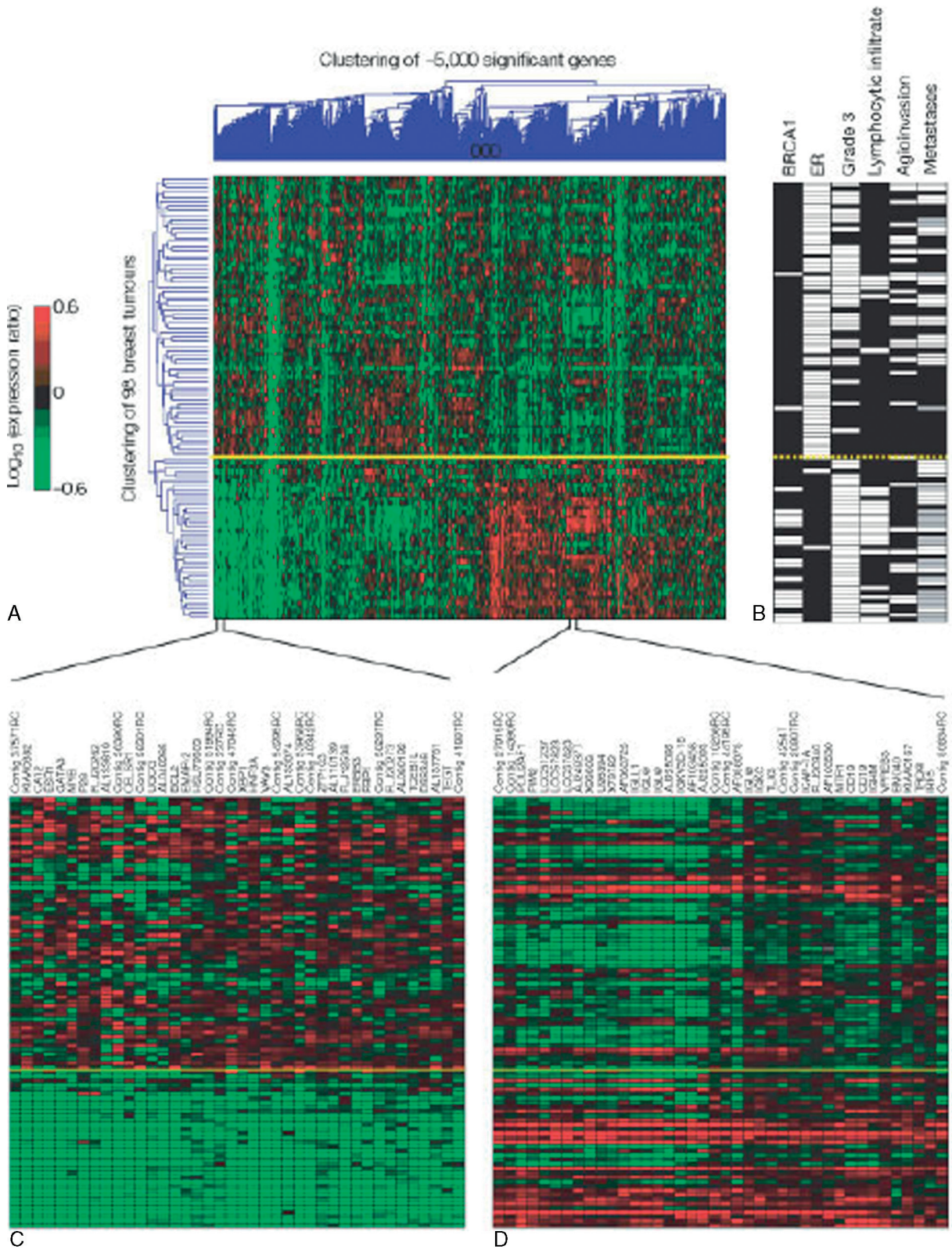
f0095 **Figure 7.19** Survival analysis (Kaplan-Meier plot) of patient groups distinguished according to gene expression profiling [29]. The Y-axis shows the survival probability for each individual group; the X-axis represents the time scale according to patient follow-up data. All groups identified by gene expression profiling are shown. Luminal type A, dark blue; luminal type B, yellow; luminal type C, light blue; normal type, green; ErbB2-like type, pink; and basal type, red. Patients with ErbB2-like or basal type tumors had the shortest survival times; luminal-type A patients had the prognosis. All others showed an intermediate probability and were not clearly distinguishable.

invasive ductal and lobular breast carcinomas of less than 5 cm in size. The tumor stages were T1 or T2, nodal status N0 (without axilliary metastases), patient age <55 years at diagnosis without a history of previous malignancies. The patients received surgical treatment followed by radiotherapy, but no adjuvant chemotherapy (except for 5 patients). The follow-up period of the patient cohort was 5 years. Tissue samples contained more than 50% tumor cells by pathological inspection; estrogen receptor (ER) and progesterone receptor (PR) status were known. The cohort was supplemented by 20 hereditary tumors carrying *BRCA1/BRCA2* mutations that were of similar histology to the sporadic cancers. Target RNA/cDNA was labeled and hybridized to an oligonucleotide array representing more than 24,000 human sequences and more than 1,000 control sequences. The reference target used in this system was a pooled cRNA derived from an RNA mixture of all patients. This means that gene expression of each sample was determined relative to the pool of all samples. The hybridizations were performed in duplicate and ~5,000 genes appeared significantly regulated more than 2-fold with a *p*-value of less than 0.01.

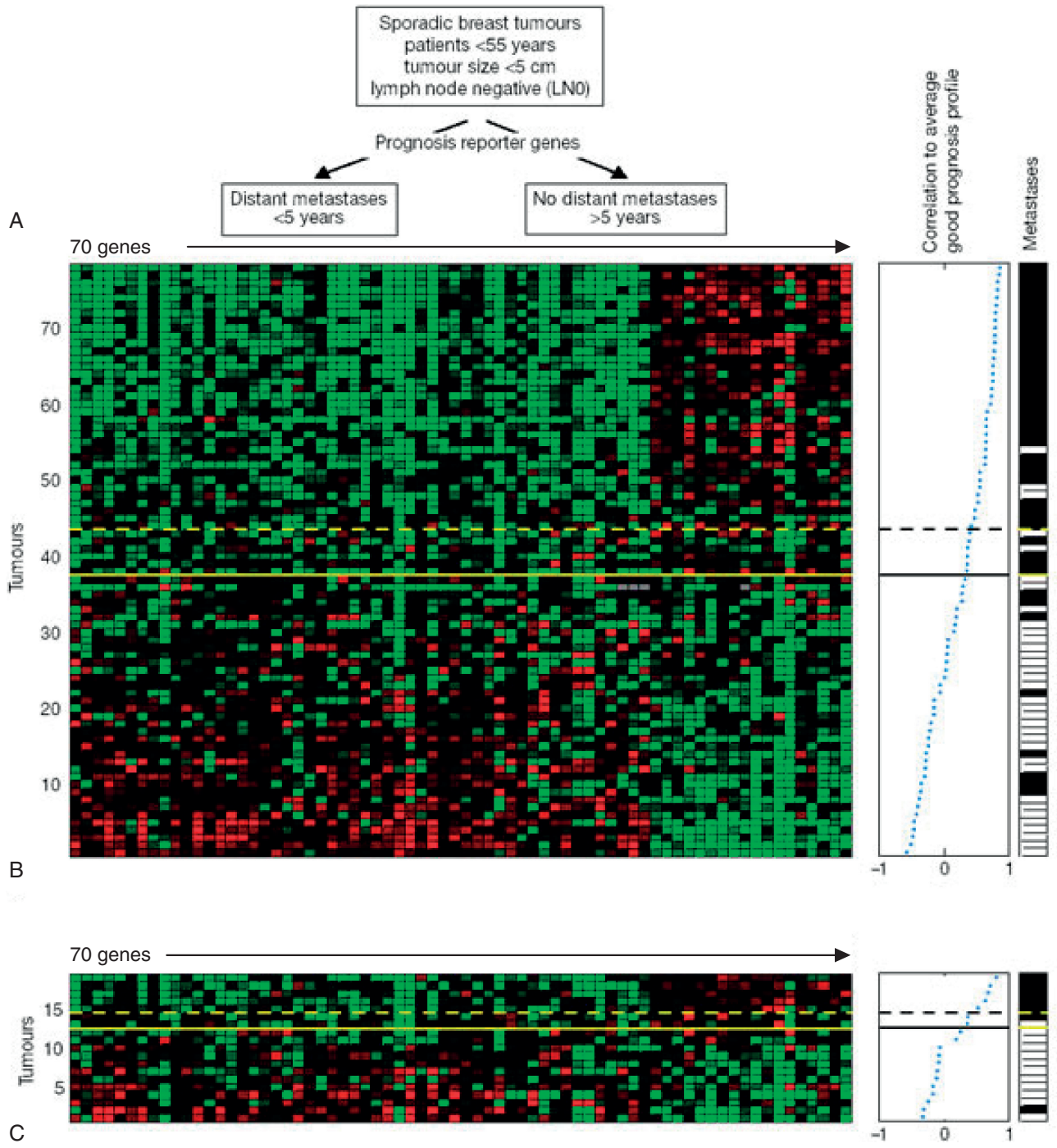
p0240 In the first step of bioinformatic analysis, expression profiles of 98 cancer samples analyzed were clustered hierarchically according to similarities among the 5,000 genes (Figure 7.20A and Figure 7.20B). This revealed two distinct groups of tumors. In the upper group, 34% had developed distant metastasis within 5 years, while in the lower group 70% exhibited metastatic spread. There was also a clear association with ER expression, which when lacking indicates a bad prognosis. Therefore, they filtered out ER-negative tumors that did not express ER and also some of the known ER targets (Figure 7.20C). In addition, the second group of tumors expressed a B-cell and T-cell gene signature. The tumors were thus characterized by a lymphocyte infiltration and clearly separated from the ER-negative group (Figure 7.20D).

p0245 In a supervised classification procedure, the researchers from the Netherlands used the gene expression profiles obtained from the sporadic tumors only. In the first step of the classification procedure, the ~5,000 genes that were significantly regulated in more than 3 of 78 tumors were selected from the 25,000 genes represented on the array. The correlation of each gene expression profile with the clinical outcome of patients was calculated, and 231 genes were found to be significantly associated with disease progression. In the second step, the 231 informative genes were rank-ordered according to their correlation coefficient. In the third step, the number of genes in this preliminary prognosis classifier was optimized by cross-validation, particularly by the leave-one-out procedure. The final result was a signature of 70 genes, which predict the clinical outcome—distant metastasis within 5 years—with an accuracy of 83% (Figure 7.21). This means that of 78 patients, 65 were assigned to the right category, poor prognosis (Figure 7.21, cluster below the yellow line) or good prognosis (above). Five patients with poor prognosis and 8 patients with good prognosis were misclassified. van't Veer et al. used an independent set of 19 lymph-node negative breast tumors (Figure 7.21C) to validate their classifier. This time, 2 of 19 patients were assigned to the wrong group. Thus, the classifier predictive of a short interval to distant metastases (poor prognosis signature) in patients without tumor cells in local lymph nodes at diagnosis (lymph node negative patients) showed a similar performance on this test set of tumors as compared to the training set.

p0250 Today, three gene expression-based prognostic breast cancer tests have been licensed for use. These are MammaPrint (Agendia BV, Amsterdam, the Netherlands; based on the work described above), Oncotype DX (Genomic Health, Redwood City, California), and H/I (AvariaDX, Carlsbad, California). However, a recent comparative study showed that for all tests offered, the relationship of predicted to observed risk in different patient populations and their incremental contribution over conventional predictors, optimal implementation, and relevance to patients receiving current therapies need further study [31]. A particular caveat on the currently available predictors was also provided in a paper published in 2005 [32]. The authors re-evaluated data from 8 different microarray-based studies with more than 800 tumor samples. The results suggested that the list of genes identified as predictors was highly unstable



f0100 **Figure 7.20** Microarray-based prediction of breast cancer prognosis [30]. Two-dimensional clustering of 98 tumor samples based on approx. 5,000 significantly regulated genes. (A) Clustering, (B) molecular characteristics of tumors, BRCA1 mutation and estrogen receptor status (ER), grade, lymphocyte infiltration, blood vessel count, and distant metastases occurring within 5 years following diagnosis. The group above the yellow line is defined as the good prognosis group (34% of patients developed distant metastasis), the group below as the bad prognosis group (70%). (C) Expression pattern of subgroup associated with estrogen receptor expression, (D) subgroup exhibiting lymphocytic infiltration.



f0105 **Figure 7.21** Identification of the prognostic breast cancer gene set using a supervised approach [30]. The 231 genes identified as being most significantly correlated to disease outcome were used to recluster, as described in the text. Each row represents a tumor and each column a gene. The genes are ordered according to their correlation coefficient with the two prognostic groups. The tumors are ordered according to their correlation to the average profile of the good prognosis group. The solid line marks the prognostic classifier showing optimal accuracy; the dashed line, the classifier showing optimized sensitivity. Patients above the dashed line have a good prognosis signature, while patients below the dashed line have a poor prognosis signature. The metastasis status for each patient is shown at the right. White bars indicate patients who developed distant metastases within 5 years after the primary diagnosis; black indicates disease-free patients.

and that the molecular signatures strongly depended on the selection of patients in the training set. Notably, 5 of 7 studies re-evaluated did not classify patients better than chance.

s0185 ## From Gene Expression Signatures to Simple Gene Predictors

p0255 In 1999, a group of scientists from highly ranked medical schools in the United States assembled a specialized microarray representing genes preferentially expressed in lymphoid cells. The so-called lympho-chip harbored more than 17,000 cDNA probes derived from libraries specific for germinal center B-cells, diffuse large B-cell lymphoma (DLBCL), follicular lymphoma, mantle cell lymphoma, chronic lymphatic leukemia (CLL), genes induced or repressed in T-cell or B-cell activation, supplemented by lymphocyte-specific genes and cancer genes [33]. The consortium interrogated these chips using targets prepared from normal cells and tumors to define signatures for the different immune cell types, under different conditions and developmental stages. Particularly, the researchers analyzed the most prevalent adult lymphomas using the lympho-chip. They identified signatures for distinct types of diffuse large B-cell lymphoma (DLBCL) exhibiting a bad prognosis, follicular lymphoma (FL) exhibiting a low proliferation rate, and for chronic lymphatic leukemia (CLL) with slow progression (> 20 years). In addition, profiles were obtained from normal lymphocytes (tonsil, lymph node) as well as from several lymphoma and leukemia cell lines. Clustering analysis placed the CLL and FL profiles close to those of resting B-cells, while genes of the so-called proliferation signature were weakly expressed in these tumors. DLBCL, the highly proliferative, more aggressive disease, had higher expression levels of proliferation-associated genes. An additional signature characterized germinal center B-cells, which was clearly different from the resting blood B-cells and from the *in vitro* activated B-cells. This indicated that that germinal center B-cells represent a distinct stage of B-cells and do not simply resemble activated B-cells located in the lymph node.

p0260 When the scientists reclustered all DLBCL cases, particularly considering the genes that define the germinal center B-cells, they could clearly separate two different subclasses of DLBCL. One of them strictly showed the signature of the germinal center B-cells, while the other one was clearly distinct. These data suggested that a certain class of DLBCL was derived from germinal center B-cells and retained its differentiation signature even after malignant transformation. By investigating the genes exclusively expressed in either of the DLBCL types and reclustering, the authors defined two signatures representative of either the germinal center-type (GC-like) and what they called the activated-type DLBCL (Figure 7.22). Analysis of the clinical follow-up showed that the GC-like tumors have a much better prognosis than the activated type of DLBCL (Figure 7.23). Did the result of the microarray study provide novel information up to this stage of investigation? When the authors compared the microarray-based classification to the standard classifiers that define high and low clinical risk, there was obviously no significant classification progress (Figure 7.23B). However, when the low-risk patients initially classified conventionally are further stratified by subgrouping them into the GC and activated-type DLBCL types, the molecular classifier was superior. Subsequent functional classification of genes associated with activated-type DLBCL revealed an NFkB pathway signature that comprises several anti-apoptotic genes. The functional studies culminated in the finding that inhibition of that pathway affected growth of activated-type DLBCL, while CG-DLBCL cells were insensitive [34].

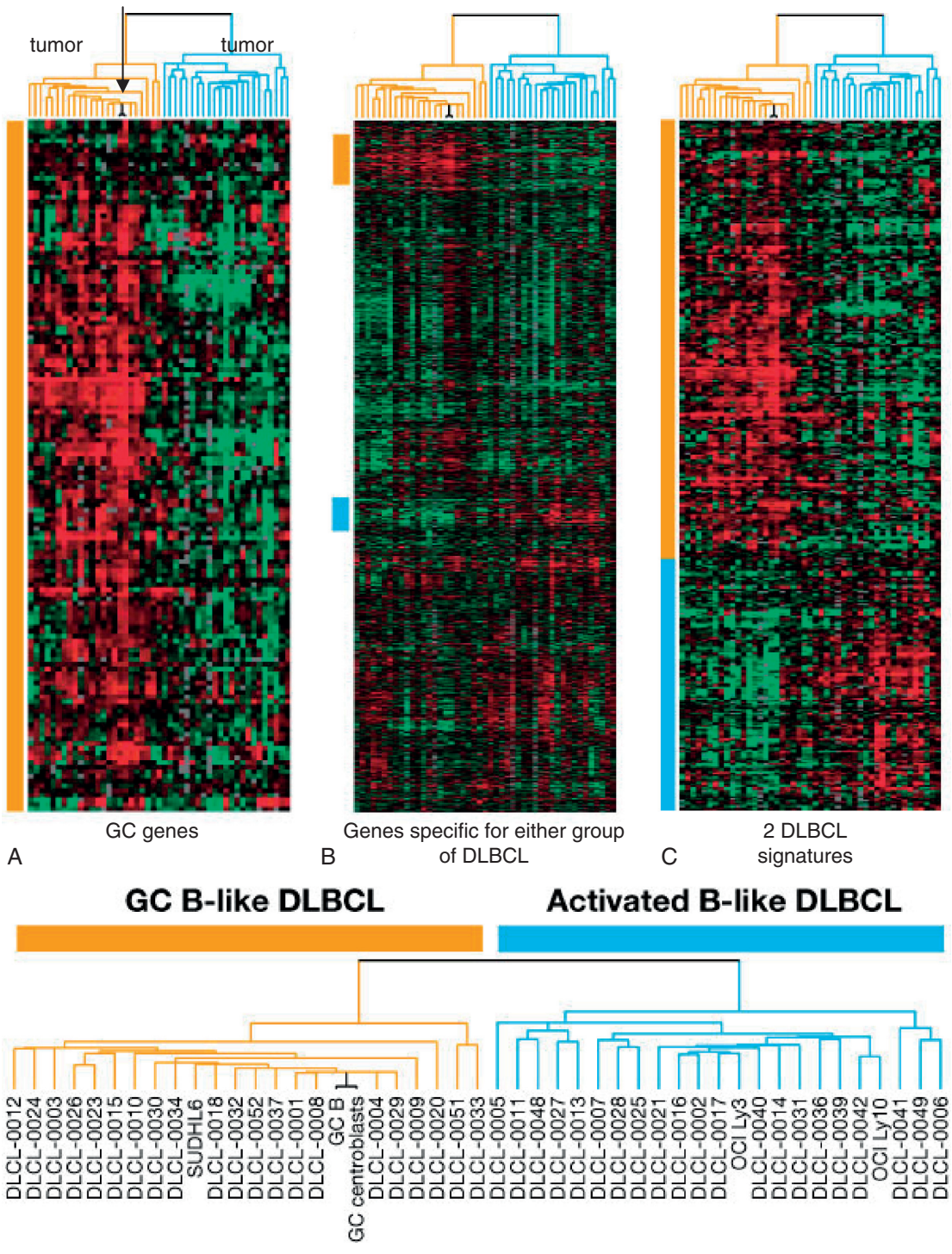
p0265 Several further microarray studies confirmed that gene signatures were associated with clinical outcome of diffuse B-cell lymphoma. However, among these studies there were disparities with regard to the number and the nature of informative genes. A recent study tried to circumvent the technical and bioinformatic issues of microarray analysis by using quantitative real-time polymerase-chain-reaction. Scientists from Miami and Stanford studied the expression of 36 genes that had previously been reported to be of predictive value among 66 lymphoma patients. The prediction of survival could be based on only 6 genes [35]. This result opens the interesting perspective that selecting informative genes that have been filtered through genome-wide microarray studies may permit the application of conventional methods in the future and may obviate microarray applications in routine clinical testing.

PERSPECTIVES

s0190

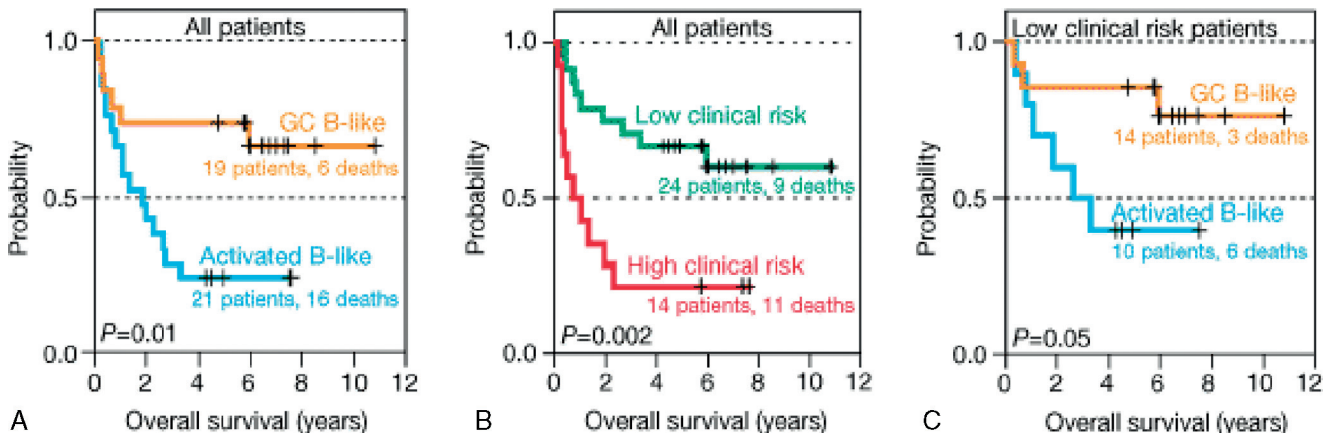
p0270 Microarrays have developed into an indispensable tool for transcriptome analysis in basic research, translational studies, and clinical investigations. In experimental pathology, gene expression activities under various conditions can be assessed at an unprecedented quantity, speed, and precision. Commercial microarray platforms exhibit a high degree of standardization allowing service laboratories and academic core facilities to offer the technology to users from industry and academia, respectively, who do not have the means to develop their own specific expertise in this field. Together with other –omics technologies, transcriptomics will be an essential component in worldwide efforts to understand normalcy and disease at the systems level. Already now, transcriptomic approaches are a standard strategy for data collection in systems biology and systems medicine.

p0275 In the clinical situation, the current instabilities of predictive gene signatures will probably be scrutinized by enforcing standard operating procedures and efforts aiming at the general standardization of diagnostic approaches, as was the case in the optimization period of microarray technology. The strong need for predictive markers in the clinic, the issue of personalized medicine, and the requirement to study the effects of old and novel drugs at the genome level are expected to



f0110 **Figure 7.22** Gene signatures representing GC-like DLBCL and activated B-like DLBCL [33]. (A) Genes characteristic for normal germinal center B-cells were used to cluster the tumor samples. This process defines two distinct classes of B-cell lymphomas: GC-like DLBCL and activated B-like DLBCL. (B) Genes that were selectively expressed either in GC-like DLBCL (yellow bar) or activated B-like DLBCL (blue bar) were identified in the tumor samples. (C) Result of hierarchical clustering that generated GC-like and activated B-cell-like DLBCL gene signatures.

increase the use of microarray technologies even further. Alternative high-throughput approaches such as proteomic profiling combined with mass spectroscopy or deep sequencing will probably not be regarded as competitive approaches. Rather, these techniques will further increase our knowledge on complex biological phenomena and pathogenic mechanisms. New types of microarrays have become available that allow analysis of alternative splicing at the level of the transcriptome, as well as to analyze the expression of microRNAs, a



f0115 **Figure 7.23** Survival analysis of DLBCL patients distinguishable according to gene expression profiling, conventional clinical criteria, and a combination of both sets of criteria [33]. (A) DLBCL patients grouped on the basis of gene expression profiling. The GC-like and the activated B-cell-like show clearly different survival probabilities. (B) DLBCL patients grouped according to the International Prognostic Index (IPI) form two groups with clearly different survival, independent of gene expression profiling. Low clinical risk patients (IPI score 0–2) and high clinical risk patients (IPI score 3–5) are plotted separately. (C) Low clinical risk DLBCL patients (IPI score 0–2) shown in B were grouped on the basis of their gene expression profiles and exhibited two distinct groups with different survival probabilities.

novel class of gene expression regulators in development, normal physiology, and disease. Rapid progress will also be made in understanding the molecular basis for the transcriptional alterations that can be assessed by microarrays, by combining chromatin immuno-precipitation (ChIP) and microarray analysis (ChIP-on-chip). Last but not least, efforts are being made to develop chip technologies that permit a truly quantitative estimation of mRNA expression. It is tempting to speculate that these novel chip technologies will gradually replace the currently available microarrays, facilitate transcriptome analysis with even higher precision, and obviate extensive validation (based on real-time PCR, immunohistochemistry, or other methods) and quantification procedures. Finally, microRNAs that have been overlooked for many years in transcriptomics have started to demonstrate their impact on gene regulation and possibly also predictive power in tumor analysis. Since the microRNAome is much smaller than the transcriptome, a current challenge is to understand the regulatory relationships between small RNAs and their mRNA targets. After all, the race is open for deciphering the protein and RNA master regulators of the transcriptome.

REFERENCES

1. Velculescu VE, Madden SL, Zhang L, et al. Analysis of human transcriptomes. *Nature Genetics.* 1999;23:387–388.
2. Alwine JC, Kemp DJ, Stark GR. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc Natl Acad Sci USA.* 1977;74:5350–5354.
3. Thomas PS. Hybridization of denatured RNA and small DNA fragments transferred to nitrocellulose. *Proc Natl Acad Sci USA.* 1980;77:5201–5205.
4. Groudine M, Weintraub H. Activation of cellular genes by avian RNA tumor viruses. *Proc Natl Acad Sci USA.* 1980;77:5351–5354.
5. Augenlicht LH, Wahrman MZ, Halsey H, Anderson L, Taylor J, Lipkin M. Expression of cloned sequences in biopsies of human colonic tissue and in colonic carcinoma cells induced to differentiate in vitro. *Cancer Res.* 1987;47:6017–6021.
6. Schmitt AO, Specht T, Beckmann G, et al. Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues. *Nucleic Acids Res.* 1999;27:4251–4260.
7. Fargnoli J, Holbrook NJ, Fornace AJ Jr. Low-ratio hybridization subtraction. *Anal Biochem.* 1990;187:364–373.
8. Diatchenko L, Lau YF, Campbell AP, et al. Suppression subtractive hybridization: A method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proc Natl Acad Sci USA.* 1996;93:6025–6030.
9. Zuber J, Tchernitsa OI, Hinzmann B, et al. A genome-wide survey of RAS transformation targets. *Nature Genetics.* 2000;24:144–152.
10. Lisitsyn N, Wigler M. Cloning the differences between 2 complex genomes. *Science.* 1993;259:946–951.
11. Hubank M, Schatz DG. Identifying differences in mRNA expression by representational difference analysis of cDNA. *Nucleic Acids Res.* 1994;22:5640–5648.
12. Liang P, Pardee AB. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science.* 1992;257:967–971.
13. Liang P, Averboukh L, Keyomarsi K, Sager R, Pardee AB. Differential display and cloning of messenger RNAs from human breast cancer versus mammary epithelial cells. *Cancer Res.* 1992;52:6966–6968.
14. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. *Science.* 1995;370:484–487.
15. Matsumura H, Reuter M, Kruger DH, Winter P, Kahl G, Terauchi R. SuperSAGE. *Methods Mol Biol.* 2008;387:55–70
16. Bishop JO, Morton JG, Rosbash M, Richardson M. Three abundance classes in HeLa cell messenger RNA. *Nature.* 1974;250:199–204.
17. Lau NC, Lim LP, Weinstein EG, Bartel DP. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans.* *Science.* 2001;294:858–862.
18. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science.* 1995;370:467–470.
19. Lipshutz RJ, Morris D, Chee M, et al. Using oligonucleotide probe arrays to access genetic diversity. *Biotechniques.* 1995;19:442–447.
20. Futschik ME, Reeve A, Kasabov N. Evolving connectionist systems for knowledge discovery from gene expression data of cancer tissue. *Artif Intell Med.* 2003;28:165–189.

21. Futschik ME. Gene expression profiling of metastatic and non-metastatic colorectal cancer cell lines. *Genome Letters*. 2002; 1:26–34.
22. Quackenbush J. Microarray data normalization and transformation. *Nature Genetics*. 2002;32:496–501.
23. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA*. 1998;95:14863–14868.
24. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*. 2000;25:25–29.
25. Rhee SY, Wood V, Dolinski K, Draghici S. Use and misuse of the gene ontology annotations. *Nature Rev Genetics*. 2008;9:509–515.
26. Brazma A, Hingamp P, Quackenbush J, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics*. 2001;29:365–371.
27. DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*. 1997;278:680–686.
28. Iyer VR, Eisen MB, Ross DT, et al. The transcriptional program in the response of human fibroblasts to serum. *Science*. 1999;283:83–87.
29. Sorlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA*. 2001;98:10869–10874.
30. van't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 2002;415:530–536.
31. Marchionni L, Wilson RF, Wolff AC, et al. Systematic review: Gene expression profiling assays in early-stage breast cancer. *Ann Intern Med*. 2008;148:358–369.
32. Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: A multiple random validation strategy. *Lancet*. 2005;365:488–492.
33. Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*. 2000;403:503–511.
34. Davis RE, Brown KD, Siebenlist U, Staudt LM. Constitutive nuclear factor kappaB activity is required for survival of activated B cell-like diffuse large B cell lymphoma cells. *J Exp Med*. 2001;194:1861–1874.
35. Lossos IS, Czerwinski DK, Alizadeh AA, et al. Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes. *N Engl J Med*. 2004;350:1828–1837.
36. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*. 1999;286:531–537.

Comp. by: SKarthiProof0000868544 Date:29/9/08 Time:22:29:18 Stage:First Proof File Path://ppdys1108/
Womat3/Production/PRODENV/0000000038/0000010883/0000000005/0000868544.3D Proof by: QC by:

Author Query Form

Book: Molecular Pathology
Chapter No:10007

Query Refs.	Details Required	Author's response
AU1	OK as edited?	
AU2	Please confirm figures permissions have been received and the captions are update.	