Molecular Tools, Expression Profiling

17

Angela M. Kaindl and Konrad Oexle

CONTENTS

KEY POINTS

- High-throughput technologies in genomics, epigenomics, transcriptomics, proteomics, and metabolomics may detect specific variation patterns and help to optimize individual medical decisions.
- Transcriptomics and proteomics quantify a large number of RNA and protein species, respectively, by using quantitative hybridization onto chips (microarrays) or signature sequencing of RNA and two-dimensional electrophoresis, mass spectrometry, or antibody array binding of proteins.
- High-throughput analyses are subject to problems of noise and multiple testing and involve the necessity to select reliable, informative, and biologically reasonable subsets.
- In the field of breast cancer, RNA expression profiles have been derived that achieve similar sensitivity but are more specific than are conventional algorithms in predicting distant metastasis, that is, less error-prone in recommending adjuvant systemic therapy.
- Meta-analysis of different prognostic RNA signatures revealed that genes associated with cell proliferation provide the driving force in all of them.
- While proteomics potentially oversees a larger space of expression variation than transcriptomics, proteomic profiling beyond the testing of individual markers has not yet been transferred successfully to the field of breast cancer.

A. M. Kaindl, MD
Klinik für Pädiatrie m. S. Neurologie, Charité – Universitätsmedizin Berlin, Campus Virchow-Klinikum, Augustenburger Platz 1, 13353 Berlin, Germany
and
Laboratoire de Neurologie du Développement, UMR 676 Inserm-Paris 7 & Service de Neuropédiatrie, Hôpital Robert Debré, 48 Blvd. Serurier, 75019 Paris, France

K. Oexle, MD
Institut für Humangenetik, Klinikum Rechts der Isar, Technische Universität München, Trogerstraße 32, 81675 München, Germany

Abstract

High-throughput technologies of modern biology provide "molecular portraits" of tissues and have entered the field of oncology. In the present chapter, we describe tools of high-throughput expression analysis in transcriptomics and proteomics, with an emphasis on microarrays, two-dimensional electrophoresis, and mass spectrometry. Options and limitations in data production, extraction, and interpretation are outlined. Problems of sensitivity, specificity, multiple testing, and noise are discussed. As a concrete example, we review the application of these tools to the field of breast cancer, where expression analyses already contribute to individual treatment decisions.

17.1 Introduction

Therapeutic algorithms in oncology depend on various clinical and pathologic parameters that provide information about the risk of disease recurrence and likelihood of response to specific treatment options. However, the predictive power of these parameters is still limited. Expression profiling offers a possibility to further classify tumor subtypes; to improve the prediction of survival, disease recurrence, and efficiency of therapeutic regimen; and to recognize more precisely the necessity of systemic and aggressive treatment in individual patients (Fig 17.1). Beyond refined disease subtyping and individualized treatment decision, it may provide molecular understanding of the disease and novel targets for therapy. Expression profiling may thus become relevant in all sections of oncology. This includes radiation oncology, e. g., the study of the radiation response of tumors and of normal tissues and the development of biomarkers that predict local disease control and toxicity after radiotherapy (Nuyten et al. 2006).

From a methodological point of view, expression profiling belongs to a group of new high-throughput technologies that provide molecular portraits of cells and tissues (Perou et al. 2000; Sotiriou and Piccart 2007). Such portraits can include data on genomic individuality, i.e., on DNA polymorphisms, mutations, copy number variations (*genomics*), and epigenetic modifications (*epigenomics*), as well as genome-wide quantitative data on gene expression at a specific point in time and under specific environmental circumstances (Fig. 17.2). Expression may be assessed in terms of *transcriptomics*, *proteomics*, or *metabolomics* by quantifying a large, if not exhaustive set of transcripts, proteins, or metabolites, respectively. RNA microarray expression studies look at responses on the transcript level. Thus, this methodical approach does not portrait alternative splicing or co- and posttranslational modifications. Proteomics, which encompasses an analysis of protein populations encoded by single genes, may offer further information at that level.

In this chapter, we review molecular tools of expression profiling that can be assigned to the field of *transcriptomics* and *proteomics*. We describe current techniques applied in these two research fields and, as an example, discuss advances that have been made or can be expected by their application to the field of breast cancer, the most frequently diagnosed cancer in women in Western countries (Miller et al. 2006; Sotiriou and Piccart 2007).

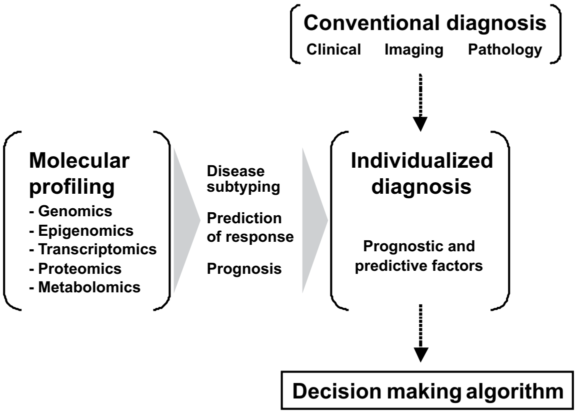

Fig 17.1. Prospect of an individualized therapy in oncology by the use of molecular profiles. Therapy of malignant tumors depends on clinical and pathological data such as tumor size, lymph node invasion, and distant metastasis. However, patients with similar clinicopathological features may have markedly different outcomes. Both the response of the tumor and that of normal tissue may vary substantially. Molecular profiling offers a possibility to further classify tumor subtypes, to improve the prediction of individual patient outcome, and to select the optimal therapeutic regimen

Genomics addresses the set of all genes (the "genome"). The field includes the elucidation of the entire DNA sequence of various organisms and the mapping of phenotypes (that may reveal pleiotropic effects and epistatic interactions of gene loci). In medicine, genomics serves to attribute disease features to common or rare variants with weak or strong effects, respectively. Variants may involve single genes only, submicroscopic copy number changes of chromosomal domains, or visible rearrangements. Microscopic analysis of chromosomes was the first form of genomics that achieved considerable relevance in all parts of genetics including tumor cytogenetics.

Epigenomics is the study of heritable modifications (marks) other than those in the DNA sequence that regulate gene expression, silence the activity of transposable elements, and stabilize adjustments of gene dosage as seen in X-chromosome inactivation and genomic imprinting. Epigenomics encompasses two major modifications of DNA and chromatin: DNA methylation and posttranslational histone modification.

Transcriptomics is the global study of gene expression at the RNA level. Generally, the transcriptome implies the set of all messenger RNA (mRNA) molecules, or "transcripts", produced in one or a population of cells. However, RNA-"omics" may also address the set of all microRNAs, transcripts that have regulative functions but are not translated into proteins.

Proteomics is the study of the entire spectrum of proteins (including co- and posttranslational modifications) of a cell, a tissue, or an organism.

Metabolomics is the study of the entire metabolic content of a cell, a tissue, or an organism addressing quantitatively the set of usually small molecules that are educts, intermediates, or products of metabolic pathways.

Interactomics is the study of interactions among proteins and other molecules within a cell by applying methods of biology, informatics, and engineering.

Fig. 17.2. Different forms of "-omics"

17.2 Transcriptomics

17.2.1 Data Production

Several methods are available to monitor transcription levels of tens of thousands of genes rapidly and simultaneously. The quantification of RNA species by sequence-specific annealing (hybridization) to complementary DNA probes arrayed on a substrate (microarray) was developed by Schena et al. (1995). Sample RNA was submitted to reverse transcription and fluorescent labeling. Thereby, a quantitative parameter was produced that could be measured as a localized signal after hybridization to the arrayed sensors. Comparison by competitive hybridization of two RNA samples labeled with two different dyes (Cy3 and Cy5) resulted in expression ratios of the two sources (e.g., of tumor and nontumor tissue). Whereas sensors were originally taken from libraries of DNA clones (cDNA), present-day microarray technology prefers synthetic oligonucleotides, i.e., oligomers of single-stranded DNA. Oligonucleotides are designed *in silico* and can be synthesized *in situ* by a combinatorial sequence of photolithographic steps applied to the nascent microarray (Fig. 17.3; Pease et al. 1994; Hardiman 2004).

Array technologies rely on representational labeling of the source RNA with reverse transcription and production of labeled or tagged molecules. This process may be coupled with a PCR amplification step. Arrays of oligonucleotides involve either the two-label scheme that results in ratios between two samples or a single-label method that attributes intensities to the RNA targets of a single source. In the two-label scheme, labels may be exchanged (dye swap) in order to neutralize artifacts associated with one of the dyes.

In most types of microarrays, the identity of a sensor is specified by its location and referenced *ex ante* by its Cartesian coordinates (Fig. 17.4). Alternatively, a sensor's identity may be referenced by an optical bar code or an address sequence. The array positions of the sensors may then be chosen randomly and identified

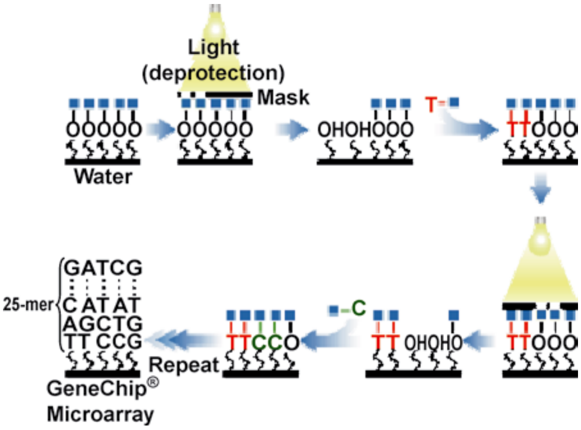

Fig. 17.3. Photolithographic in situ synthesis of an oligonucleotide microarray (courtesy of Affymetrix, Santa Clara, California)

From Computer Desktop Encyclopedia
Reproduced with permission.
© 2007 Affymetrix

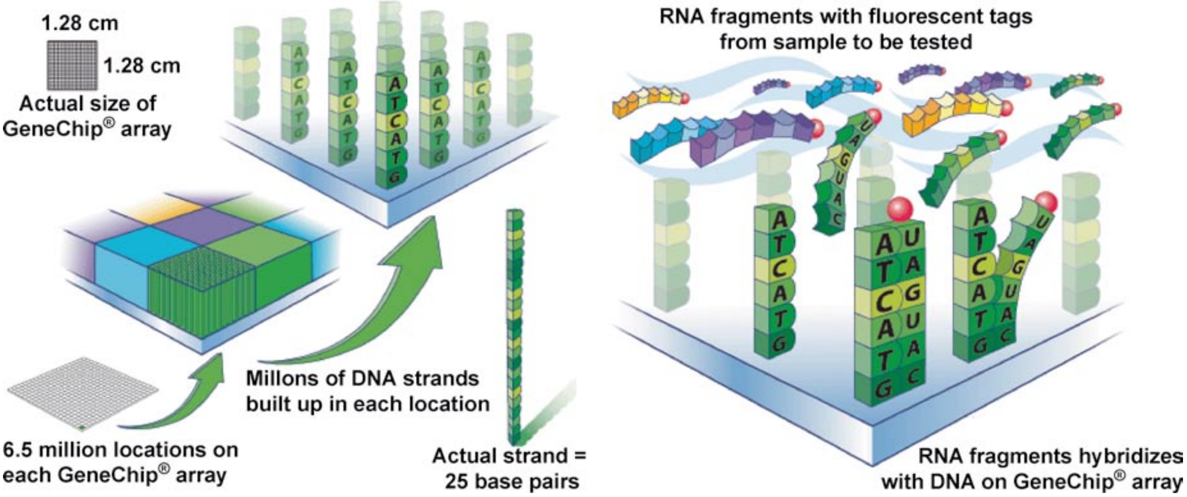

Fig. 17.4. Design and function of a microarray expression chip. DNA oligonucleotides act as sensors. Sensors of the same sequence reside at one location. The hybridization intensity at this location depends on the concentration of complementary RNA in the sample. Quantification is achieved by laser-induced fluorescence of the label (courtesy of Affymetrix, Santa Clara, California)

ex post (Steemers et al. 2000). This method is used in bead technology, which allows for further miniaturization of the arrays.

Other methods of RNA monitoring exist that, in contrast to array techniques, are not based on quantitative hybridization. In serial analysis of gene expression ([SAGE] Velculescu et al. 1995), RNA fragments derived from a sample to be analyzed are ligated and cloned in a vector, which is then sequenced. The number of stretches in the vector sequence that belong to the same RNA species indicates the concentration of this RNA in the original sample. In massively parallel signature sequencing (MPSS), the relative amount of each RNA species in a sample is determined by mass sequencing of reversely transcribed DNA and subsequent counting of identical sequencing data.

17.2.2 Data Extraction

All methods of gene expression monitoring are subject to biological variability and experimental noise. Biological variability is due to endogenous, environmental, periodic, and stochastic causes. While factors such as daytime and feeding status may be controllable, others cannot be predicted. In the plant model *Arabidopsis*, for instance, touching has been shown to induce significant changes in gene expression (Chotikacharoensuk et al. 2006). In general, biological variability is reduced by randomization and replication. Replication of sampling may be more important than repeating the examination of a sample (Breitling 2006). However, experimental noise has to be controlled as well. This can be achieved by a variance stabilization procedure such as log-transformation (Fig. 17.5).

RNA arrays need to be normalized since the distribution of the expression signal varies from array to array. Most simply, each signal y of an array a is replaced by a z-score with $z = (y - \mu_a)/\sigma_a$. Thereafter, all arrays have mean $\mu = 0$ and variance $\sigma^2 = 1$. Expression data of individual genes then can be compared across arrays. In many studies (e.g., Stranger et al. 2005), section-wise normalization (quantile normalization) is performed, as the signal distribution of an array may be affected by skewness or other distortions. The method is motivated by the idea that two data vectors have the same distribution if the quantile–quantile plot is a straight diagonal line. The extension from two to n dimensions (i.e., arrays) is straightforward. Bolstad et al. (2003) provided a stepwise description of quantile normalization with standard spreadsheet software:

1. Given n arrays of p sensors (gene probes), form a spreadsheet X of dimension $p \times n$, where each array is a column.
2. Sort each column of X to give X_{sort}.
3. Take the means across rows of X_{sort} and assign this mean to each element in the row to get X^{*}_{sort}.
4. Get $X_{\mathrm{normalized}}$ by rearranging each column of X^{*}_{sort} to have the same ordering as original X.

After normalization, individual replicates are averaged for each probe resulting in an expression data for each gene in each individual. However, the results have to be regarded cautiously. Thus, for instance, the normalized expression of the testis determining factor (SRY) in lymphoblastoid cells of the CEU and YRI parental HapMap samples (March 2007 release, www.sanger.ac.uk/humgen/genevar/) seems to be higher in women than in men (6.02 ± 0.05 versus 6.00 ± 0.07, $p = 0.03$, two-sided t-test). Of course, this is an artifact revealing that normalized expression levels of 6.02 do not indicate significant expression in this setting. For significance analysis of expression data, t-tests or rank products may be used (Breitling et al. 2004). The seeming significance of the above example highlights the problem of multiple

Analysis of oligonucleotide-based microarray data revealed Poisson-like noise of gene expression data for a large range of expression levels (Tu et al. 2002). This noise was mainly related to the hybridization process. Poisson noise occurs in signals that come about by a sequence of independent probabilistic events (Poisson process). The variance of such signals, i.e., the average squared distance from the mean, equals the mean signal intensity, $\sigma^2 = \mu$, and increases proportionally. Log-transformation of the data, $y(x) = \log(x)$, results in variance stabilization at higher expression levels. This is related to the property of the logarithm to compress distance with increasing number. Approximating y in the region surrounding μ_X by a Taylor expansion, $y(x) \approx \log(\mu_x) + \log'(\mu_x)(x - \mu_x) + \ldots$, with $\log'(\mu_x) = 1/\mu_x$, yields a rough estimate $\mu_y \approx \log(\mu_x)$ of the expectation (i.e., mean) of y if the zero-order approximation is used. The variance, i.e., the expectation of $(y - \mu_x)^2$ is then derived from the first-order approximation as $\sigma^2_Y \approx \sigma^2_x/\mu^2_x$. In the case of a Poisson process where $\sigma^2_x = \mu_x$ logarithmic transformation thus results in a variance $\sigma^2_y \approx 1/\exp(\mu_y)$ that declines with increasing signal intensity. However, this "noise stabilization" is not effective at low expression levels. In the case of a Poisson process, for instance, the variance of the log-transformed signal becomes larger than the mean signal intensity μ_y if the latter is less than 0.57. Different methods of variance stabilization at low levels have been devised. As a most simple procedure started-log transformation has been recommended, i.e., $y = \log(x + b)$, which, in case of a Poisson process, implies an upper variance limit of about $1/(2b)$. Forgoing subtraction of the background from the raw signal may already have the desired effect of noise stabilization in the low expression range (Breitling 2006).

Fig. 17.5. Variance stabilization

testing that occurs in all epidemiological investigations that run a large number of parameters on the same set of probands. In a microarray study that measures the expression of 30,000 genes, about 1,500 spurious results are to be expected purely due to chance if the "usual" significance level of $\alpha = 1/20 = 0.05$ is not corrected for multiple testing.

The classical *Bonferroni* correction divides the significance level acceptable for a single test ($p < \alpha$) by the number n of the tests in the multiplex assay ($p_i < \alpha/n$). This correction is too conservative, however, and implies an unnecessary decline of power. The *Simes* procedure is less conservative. It controls the false discovery rate (FDR), i.e., the expected fraction of false-positive results among all positive results (Benjamini and Hochberg 1995). After listing the n tests in an ascending order according to their p-values, the position with the largest k is identified which satisfies $p_k/k < \alpha/n$ and all tests up to this position are declared to be positive. Thus, the observed distribution of the p-values is taken into account. Multiple tests may be correlated; in case of negative correlations, the limit of the *Simes* procedure can be relaxed even more. Permutation, e.g., random redistribution of proband labels, is another powerful method to control the probability of false-positive results as given by the actual data distributions and test correlations. For that purpose, each unadjusted p-value of the multiplex assay is replaced by the fraction (i.e., relative frequency) of random permutations that, by chance, produced smaller minimal p-values (Westfall and Young 1993). The necessary number of permutations depends on the smallest p-value to be adjusted and may thus imply considerable computation time.

17.2.3 Data Interpretation

Results that are likely to be true positive still need to be interpreted (Breitling 2006; Sotiriou and Piccart 2007). This is the expression-profiling step and involves dimensionality reduction of the large number of posi-

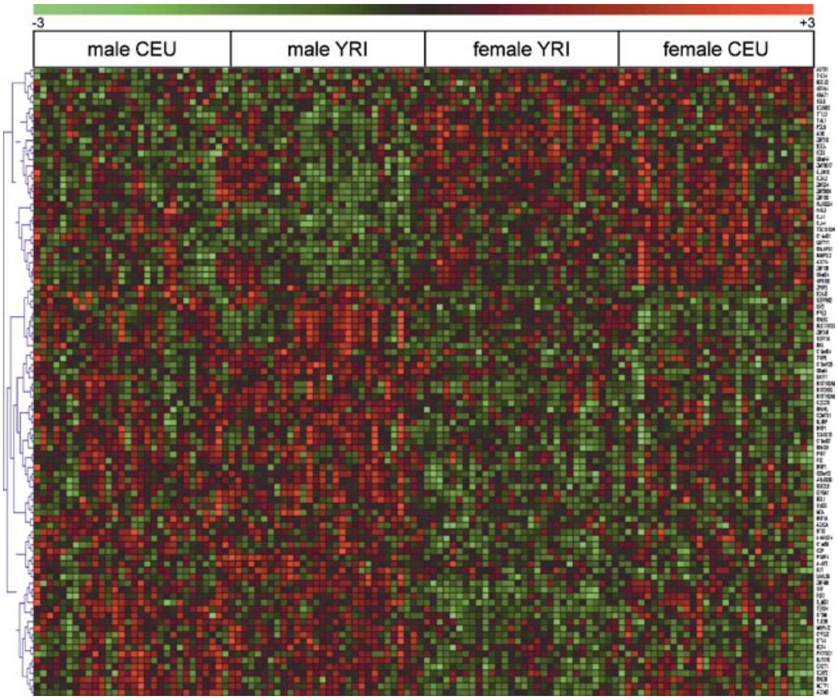

Fig. 17.6. Example of a heat map. Autosomal gene expression in lymphoblastoid cells of 2 × 60 men and women of African (*YRI*) and European (*CEU*) origin as listed in the HapMap gene variation project (log transformed and normalized across all samples; ftp.sanger.ac.uk/pub/genevar). 98 RefSeq-annotated genes (*right panel*) revealed a sex-dependent difference at a significance level of $p < 0.002$. Their average expression ranged from 5.7 to 12.0, with 8 genes not surpassing the average SRY expression in females, indicating substantial artifacts (see text). The heat map shows z-scores, i.e., deviations from the mean in standard deviation (SD) units. Hierarchical clustering of the 98 genes (complete linkage, Genesis®; *left panel*) recapitulates the sex difference

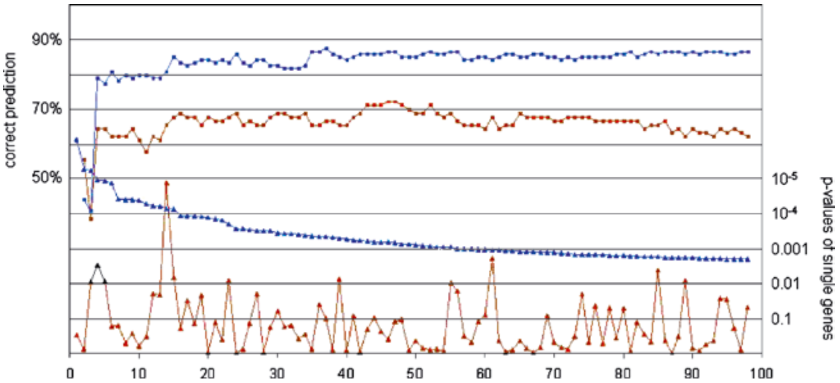

Fig. 17.7. Example of a predictive classifier. The autosomal genes shown in Fig. 17.6 were listed according to their significance level (*blue triangles*). Beginning with expression data (*z*-scores) of the two most significant genes a sex classifier (*blue squares*) was derived from the combined sample of 120 men and women (*CEU-YRI*) by a leave-one-out procedure: The gene expressions in each single individual were compared to the averages in the two groups of the remaining 59 or 60 men and women (Pearson coefficient of correlation between individual and average gene expressions). If the correlation with the group of the same sex was superior, then the prediction was counted as correct. The classifier reached a plateau after the top 40–50 genes had been included. Classification based on male and female averages in the CEU-YRI sample also worked in a sample of 90 Asian men and women (*CHN-JPN*) with a maximal predictive power of 70% (*brown squares*), that is, well above random prediction. However, only for a small number of genes the sex difference was replicated in the Asian sample (*brown triangles*). Technical artifacts, ethnical confounders, and random effects in multiple testing schemes have to be considered

tive results. Profiling may be *supervised*, that is, directed by a known grouping of the samples (e.g., probands versus control). In a “leave-one-out” procedure, an optimal set of genes may thus be selected that classifies correctly the largest number of left-out probands and controls (Figs. 17.6, 17.7). In *unsupervised* profiling, that is, grouping of similar expression patterns in a dataset without using any outside information, cluster analysis is the standard method. Clustering demands a measure of distance such as the correlation coefficient, for instance. Simple hierarchical clustering may proceed in an agglomerative way: Recursively, individuals and/or clusters with the smallest distance, i.e., highest correlation of gene expression, are united into a new cluster until a top cluster is formed that contains all individuals. Vice versa, genes may be clustered according to the correlation of their expression across probands. The following pitfall frequently occurs if clustering is applied in supervised analyses: if an outcome-related selection of genes is spurious, then a claim of correlation between clusters and clinical outcome is also spurious if the clustering is based on the expression of these genes (Dupuy and Simon 2007).

Knowledge-driven dimensionality reduction according to the biological annotation of genes may be done before or after significance analysis, clustering, or classifier extraction. If it is done before, e.g., by focusing on a subset of genes only, it may help to uncover subtle effects that might remain insignificant otherwise. If it is done afterwards, e.g., by comparison with gene ontology databases (Smith et al. 2007), it may help to recognize biological processes and to evaluate clusters or classifiers. Further conceptual integration may involve annealing with genome data and other “-omics” results, cross-species comparison, and network approach in terms of systems biology.

17.3 Proteomics

In the following, we review basic principles of proteomics as well as methods currently applied in this field and discuss the application of proteomic strategies in cancer research. The term *proteome*, a linguistic equivalent to the term *genome*, refers to the entire protein content encoded in the genome of a cell, a tissue, or an organism. In comparison to the genome that is believed to be similar in different cell types, the proteome of an or-

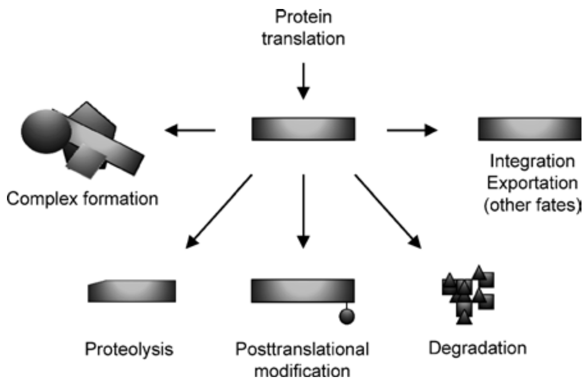

Fig. 17.8. Possible fates of proteins in the cell. The proteome is not stable, as there is constant turnover of proteins with a changing dynamic that depends on environmental and developmental conditions

ganism is a dynamic system that is constantly subject to changes. Protein composition changes from cell type to cell type, within subcellular compartments and between different stages of development and thus represents the functional status of a biological compartment (Fig. 17.8). Proteome research (proteomics) can be defined as the large-scale characterization of proteins expressed by the genome. Unlike the study of a single protein or pathway, proteomic methods enable a systematic overview of expressed protein profiles. An advantage of proteomics over transcriptomics is the ability to study posttranslational modifications. There is limited value, for example, in measuring signal transduction processes at the mRNA levels if they are characterized by protein phosphorylation or acetylation. Moreover, there are several genes with little correlation between RNA and protein expression levels.

Proteomics employs protein electrophoresis, mass spectrometry, and microarrays for the detection, identification, and characterization of proteins. These proteomic tools have their own individual advantages and limitations affecting their ability to assess the protein profile. Currently, the identification and characterization of all proteins in a given sample through high-resolution two-dimensional gel electrophoresis (2-DE) and subsequent analysis with mass spectrometry (MS) are expensive and time-consuming and, thus, not yet amenable to day-to-day use in the clinical setting. Routine approaches for obtaining protein data include enzyme-linked immunosorbent assay (ELISA) and immunohistochemistry. MS techniques have matured rapidly in recent years, due to the invention of two ionization techniques, electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI). Protein arrays are being developed involving up to a few hundred antibodies or based on surface enhanced laser desorption/ionization (SELDI) for a wider coverage of the proteome.

Protein profiles could ultimately improve the diagnosis, prognosis, and management of patients by indicating protein markers of disease similar to the tumor markers already available (Healy et al. 2007), revealing the protein interactions affecting overall tumor progression, and identifying individual cancer profiles which are suitable for tailored chemotherapeutic strategies (Banks and Selby 2003; Alessandro et al. 2005).

In 2001, the Human Proteome Organization (HUPO) was launched. For information on international collaborations and training courses in proteomics, we refer to their webpage: http://www.hupo.org.

17.3.1 2-DE

2-DE, first introduced independently by Klose (1975) and O'Farrell (1975), still represents the most powerful tool for separating complex protein mixtures when combined with staining procedures and mass spectrometry (Fig. 17.8). The principle of 2-DE is to separate proteins according to the two parameters isoelectric point (pI; pH value at which the net charge on a protein is zero) and molecular weight. For this, it combines isoelectric focusing (IEF) in a polyacrylamide gel that has a pH gradient in the first dimension with a separation of proteins on a SDS polyacrylamide gel in the second dimension. After silver staining, protein spots in protein patterns of individual samples are compared among different 2-DE gels. The power of 2-DE lies in its high resolution of up to 10,000 proteins per sample and its ability to detect simultaneously vast amounts of proteins and to visualize co- and posttranslational modifications. Thereby, for instance, disease-associated proteins can be elucidated through subtractive analyses comparing disease with control protein patterns. At the stage of

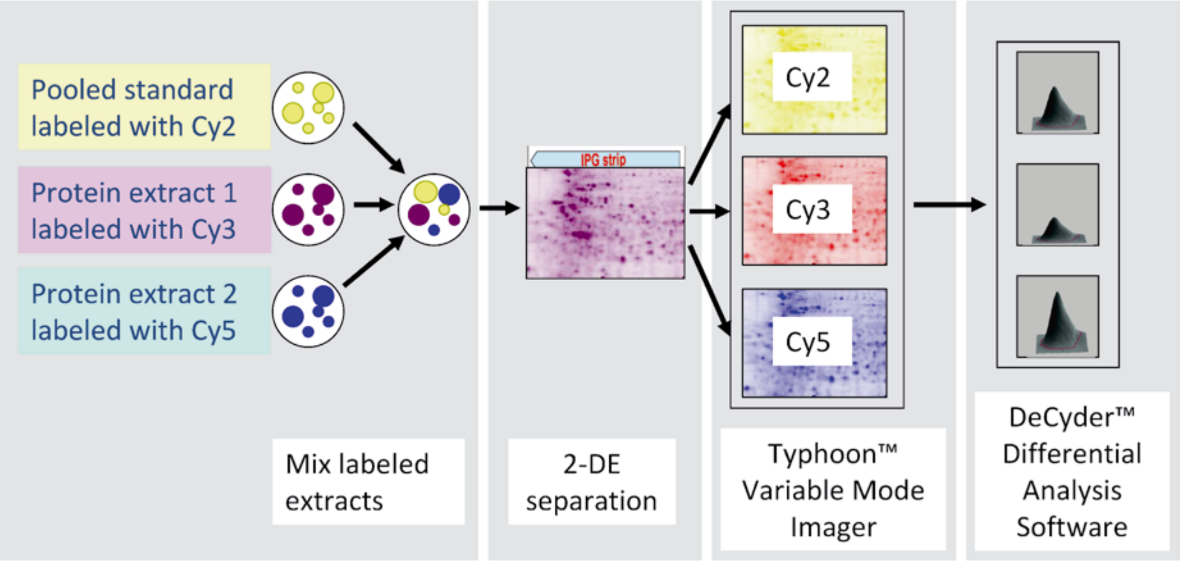
Pooled standard labeled with Cy2
Protein extract 1 labeled with Cy3
Protein extract 2 labeled with Cy5
IPG strip
Cy2
Cy3
Cy5
Mix labeled extracts
2-DE separation
Typhoon™ Variable Mode Imager
DeCyder™ Differential Analysis Software

Fig. 17.9. Workflow for a standard two-dimensional difference gel electrophoresis (DIGE) experiment. After being labeling separately with different dyes, individual samples can be compared on a single gel. Thereby, experimental variation is reduced and spot matching is improved. Using an internal standard, i.e., a pool of all the samples within an experiment, each protein's abundances in different samples can be normalized and compared across different gels. Hence, the number of samples included in an experiment is not limited. Gels are imaged and analyzed quantitatively in order to identify protein differences among different samples (courtesy of GE Healthcare Life Sciences Little Chalfont, UK)

subtractive analysis, the approach has the potential to unravel complex networks of protein interactions. Individual stained protein spots can be digested into peptides, which can be analyzed by mass spectrometry and subsequent protein database searches. However, 2-DE in its current form has a number of serious disadvantages such as its lack of real high-throughput capability, for resolving hydrophobic and very low as well as very high molecular weight proteins.

Two-dimensional difference gel electrophoresis (DIGE) strengthened the 2-DE platform by allowing the detection and quantification of differences between three samples resolved on the same gel, or across multiple gels, when linked by an internal standard (Fig. 17.9; Issaq and Veenstra 2007). Samples (and standard) are labeled separately and then mixed to allow resolution on a single gel. This minimizes experimental variation and improves spot matching. Differentiation and comparison of samples is possible since they are labeled with different dyes (limited lysine labeling with DIGE Fluor Cy2, Cy3, and Cy5). The standard, a pool of all the samples within an experiment, enables normalizing the relative abundance of each protein and comparing abundances across different gels and sets of more than three samples. Protein detection levels span the linear range of 0.125 ng to 10 µg. Image analysis with appropriate software allows for the identification of differences in protein abundance.

In classic 2-DE and DIGE approaches, highly alkaline and highly hydrophobic proteins are underrepresented since (1) in aqueous media, proteins have a minimum of solubility at their isoelectric point, may therefore precipitate there, and subsequently do not migrate into the SDS-PAGE gel; (2) hydrophobic proteins generally do not transfer easily from the first to the second dimension; and (3) non-ionic and zwitterionic detergents commonly used for isoelectric focusing have a lower power of solubilizing membrane proteins than ionic detergents. To bypass these limitations of 2-DE in resolving hydrophobic proteins such as membrane proteins, an alternative technique, the two-dimensional BAC/SDS-PAGE (2-DB), has been developed. Here, the first-dimension separation occurs according to molecular weight in an acidic discontinuous PAGE system (pH 4.0–1.5) using cationic benzyldimethyl-*n*-hexadecylammonium chloride (BAC) as detergent and the second-dimension separation is performed using the anionic detergent SDS (Zahedi et al. 2005, 2007; Braun et al. 2007).

17.3.2 MS-Based Proteomics

MS has become an indispensable analytical tool of proteomics (Sanz-Medel et al. 2008). Mass spectrometers measure the molecular mass of a sample through the following steps:

1. A protein sample is enzymatically digested into its constituent peptides.
2. The peptides of a sample are introduced to the ionization source of the instrument directly or are separated into a series of components, which then enter the mass spectrometer sequentially for individual analysis. Such en route separation can be performed, for example, through high-pressure liquid chromatography (HPLC).
3. Inside the ionization source, the sample molecules are ionized by ESI or MALDI.
4. The charged sample ions are accelerated into the vacuum-maintained mass analyzer region of the mass spectrometer where they are separated according to their mass (m) to charge (z) ratios (m/z). Mass analyzers currently available include quadrupoles and time-of-flight (TOF) analyzers; they differ in the covered m/z range, their mass accuracy, and their resolution.
5. Data on relative abundance and m/z ratios of detected ions are stored in the format of an m/z spectrum.
6. The m/z spectra are analyzed using protein databases and enable protein identification.

Since proteomics began with 2-DE methodology, the application of MS has been driven by the qualitative character of protein identification on a 2-DE gel. Indeed, MS techniques are very convenient for protein identification. However, their application to protein quantification is more complicated since there is no linear dependence between the concentrations of protein or peptides in a sample and the MS signals observed. While there are several promising gel-free MS-based approaches, presently available methods do not fulfill the increasing need for reliable methods of absolute quantification of proteins (Sanz-Medel et al. 2008).

17.3.3 Protein Arrays

Protein microarrays use either multiple capture antibodies dotted separately on a slide (forward microarrays) or multiple tissue/protein samples, again dotted and fixed together on single slides which then are stained with the different antibodies (reverse microarrays; Kopf and Zharhary 2007; Wingren and Borrebaeck 2007). Whereas these methods can detect the presence of numerous proteins or the level of expression in multiple tissue samples in a high-throughput manner, the technique is still limited by the availability of specific and sensitive antibodies. The latter proved to be an issue, for instance, in case of known lung cancer markers such as the cytokeratins (Conrad et al. 2008). Antibody specificity must be validated by immunoblotting, and internal controls may be required if the antibodies do not bind predictably. Detection of low-abundance proteins also remains a problem, as simple methods of multiple protein amplification, analogous to the polymerase chain reaction for DNA amplification, are not available. Moreover, the capacity of protein arrays to detect co- and posttranslational modifications is limited.

17.4 Expression Profiling in Breast Cancer

Expression analysis is applied in various medical fields. Here, we review some developments in the field of transcriptomics and proteomics concerning breast cancer. With a lifetime risk of 13%, breast cancer is the most frequently diagnosed cancer in women of Western countries (Miller et al. 2006). In a minor fraction of cases, the tumor develops due to the constitutional mutation of a breast cancer (*BRCA*) gene, whereas in general the genetic basis of breast cancer is complex and not sufficiently understood. Therapy is based on more or less radical surgery combined with radiation and adjuvant systemic treatment (chemotherapy, receptor-specific drugs). Adjuvant systemic therapy of patients with localized breast cancer reduces the risk of distant metastases by 30%, but 70–80% of these patients would survive without systemic therapy (van't Veer et al. 2002). Conventional clinical and pathological parameters such as age, menopausal status, tumor size, histological grade, lymph node involvement, and status of estrogen receptor (ER) and ERBB2 receptor (Her-2/neu) are used in algorithms such as Adjuvant!Online that prognosticate the course of the disease or provide recommendations for individual treatment decisions such as the St. Gallen criteria. However, the predictive power of these algorithms is limited. Expression profiling produces additional predictive information and, possibly, new treatment options (Rouzier et al. 2005; Sotiriou and Piccart 2007).

17.4.1 RNA Analyses

RNA microarray data of breast tumors have been analyzed in supervised or unsupervised manner. *Supervised* methods use outside information about the experimental condition (e.g., cases with metastases versus cases without) to shape the derivation of a model from the dataset. *Unsupervised* methods use information contained within the RNA data only and usually involve hierarchical clustering (see section on transcriptomics) to detect relationships among tumors, among genes, and connections between specific genes and specific tumors.

In *unsupervised* analyses, breast tumors have been found to cluster in at least four groups with specific composite expression profiles (Perou et al. 2000; Sotiriou and Piccart 2007): Three major types related to a specific receptor status, ER$^-$/ERBB2$^-$, ERBB2$^+$, or ER$^+$, with the last type being subdivided in two groups that showed high or low proliferation resembling the luminal breast cancer subtypes A and B. Beyond this reproduction of the conventional histopathological classification, distinct expression patterns were found. In a subset of ER-negative tumors, for instance, a functional androgen receptor response was detected which eventually might serve as a novel therapeutic target (Doane et al. 2006).

Supervised analyses produce expression classifiers on a set of tumors for which the outcome is known already. The endpoint, i.e., the definition of what is considered as outcome (e.g., metastasis-free survival or response to a specific treatment) may vary from study to study. For evaluation, classifiers are applied to an independent set of tumors and compared to conventional predictive algorithms such as Adjuvant!Online. van't Veer et al. (2002) extracted a classifier from the expression data of 78 lymph-node negative breast cancer patients younger than 55 years of age, of whom 44 remained free of distant metastases for at least 5 years after diagnosis. By evidence of differential regulation and of correlation with disease outcome, 231 of 25,000 genes were selected. A leave-one-out procedure with the 78 samples then yielded an optimal classifier of 70 genes, which had maximal predictive power. Subsequent evaluation showed that the classifier is effective both in lymph node-negative and lymph node-positive patients (van de Vijver et al. 2002). Comparison with conventional predictive algorithms showed that the 70-genes signature (MammaPrint®) has similar sensitivity (>90%) but is more specific, that is, less patients are classified erroneously into the high-risk group where they would receive adjuvant systemic therapy. Especially, patients in the ER-positive subgroup profit from this specification.

Recently, expression of SATB1 was found to have high prognostic value in both node-negative and node-positive breast cancer patients (Han et al. 2008). SATB1 is a nuclear protein that acts as a cell-type-specific genome organizer and gene regulator essential for T-cell differentiation and activation. In breast cancer, SATB1 induces a metastatic gene expression pattern that correlates significantly with the 231 genes selected by van't Veer et al. (2002) (see above) and with expression signatures for lung and for bone metastasis (Han et al. 2008). However, the comparison of breast cancer classifiers among each other (including MammaPrint®, the Rotterdam signature, Onco*type* DX®, and others) revealed little or no overlap although they have similar predictive values and carry similar prognostic information (Wang et al. 2005; Fan et al. 2006; Sotiriou and Piccart 2007). There are several reasons for this disparity. Methodical differences (microarray platforms, hybridization conditions, gene annotations, normalization methods, profiling strategies) have been identified which now are addressed in the US Food and Drug Administration (FDA)-launched microarray quality control (MAQC) project. Moreover, the different classifiers were not derived from the same patient sets. Most importantly, however, sets of selected genes may vary substantially among studies since on the one hand, the expression levels of different genes are correlated and, on the other hand, statistical power of small study groups is limited. Thus, genes from the same pathway that carry similar biological information are likely to rank differently in different expression studies based on relatively small numbers of tumors (Dupuy and Simon 2007).

Supervised analyses may be developed bottom-up, that is, driven by a biological hypothesis and deriving an expression signature from a preselected subset of genes. Several subsets have been applied such as the wound-response signature, which was shown to be expressed in breast cancers of patients with markedly worse clinical outcome (Chang et al. 2004) or the gene-expression grade index (GGI), a signature of 97 genes that consistently differed in expression between low- and high-grade breast cancers and which was used successfully to predict the clinical outcome in intermediate-grade tumors (Sotiriou et al. 2006).

A meta-analysis revealed that genes associated with cell proliferation provide the driving force in all previously reported prognostic signatures (Sotiriou and Piccart 2007). All of these classifiers provide useful information on the intrinsic properties of a tumor. However, tumor size and nodal status retain important prognostic information.

Besides prognostic signatures, predictive classifiers have been developed. Both top-down and bottom-up supervised analyses have been performed in order to find classifiers that—beyond the determination of the ER- and ERBB2-receptor status—predict the response to a specific treatment. Thereby, predictors of anti-estrogen treatment in patients with ER-positive tumors have been derived as well as classifiers that accurately predict the effect of chemotherapeutic agents that target specific pathways (Bild et al. 2006). Prediction of treatment response is complicated, however, by the heterogeneity and evolution of tumors, and by the individual biological properties of the host (Sotiriou and Piccart 2007).

17.4.2 Proteomics

Histological data, especially the receptor expression status, represents the protein level. Therefore, in a general sense, proteomics already has been taken successfully to the clinics of breast cancer. As yet, however, proteomics in the sense of multiprotein pattern analysis by 2-DE and MS has not (Harris et al. 2007). Proteomics is hampered by the heterogeneity of tissue biopsies, variability in time and space, and small volumes in focused sampling procedures such as microdissection or nipple aspiration. While some of these are shared with transcriptomics, proteomics lacks a PCR-like amplification method (Hondermarck et al. 2008). Of course, multiprotein analyses also encounter the multitesting problem, which might result in the generation of spurious results.

Thus, proteomics contributes to the understanding of factors in breast cancer biology such as the chaperone 14-3-3, which is involved in the control of proliferation and differentiation, the ubiquitinating activity of *BRCA1*, and the downstream effects of ERBB2 or tumor growth factor (TGF)β receptor activation, and may eventually reveal new therapeutic targets (Hondermarck et al. 2008). As of 2007, however, the clinical use of proteomic pattern analysis is not reliable enough and has not been recommended (Harris et al. 2007). Classifiers such as a 21-protein-signature of metastasis-free survival derived from unsupervised protein expression profiling (Jacquemier et al. 2005) still require confirmation in larger and well-designed prospective studies.

References

Alessandro R, Fontana S, Kohn E, De Leo G (2005) Proteomic strategies and their application in cancer research. Tumori 91:447–455

Banks R, Selby P (2003) Clinical proteomics—insights into pathologies and benefits for patients. Lancet 362:415–416

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc B 57:289–300

Bild AH, Yao G, Chang JT et al (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 439:353–357

Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19:185–193

Braun RJ, Kinkl N, Beer M, Ueffing M (2007) Two-dimensional electrophoresis of membrane proteins. Anal Bioanal Chem 389:1033–1045

Breitling R (2006) Biological microarray interpretation: the rules of engagement. Biochim Biophys Acta 1759:319–327

Breitling R, Armengaud P, Amtmann A, Herzyk P (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett 573:83–92

Brenner S, Johnson M, Bridgham J et al (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol 18:630–634

Chang HY, Sneddon JB, Alizadeh AA et al (2004) Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. PLoS Biol 2:E7

Chotikacharoensuk T, Arteca RN, Arteca JM (2006) Use of differential display for the identification of touch-induced genes from an ethylene-insensitive *Arabidopsis* mutant and partial characterization of these genes. J Plant Physiol 163:1305–1320

Conrad DH, Goyette J, Thomas PS (2008) Proteomics as a method for early detection of cancer: a review of proteomics, exhaled breath condensate, and lung cancer screening. J Gen Intern Med 23 Suppl 1:78–84

Doane AS, Danso M, Lal P et al (2006) An estrogen receptor-negative breast cancer subset characterized by a hormonally regulated transcriptional program and response to androgen. Oncogene 25:3994–4008

Dupuy A, Simon RM (2007) Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. J Natl Cancer Inst 99:147–157

Fan C, Oh DS, Wessels L et al (2006) Concordance among gene-expression-based predictors for breast cancer. N Engl J Med 355:560–569

Han HJ, Russo J, Kohwi Y, Kohwi-Shigematsu T (2008) SATB1 reprogrammes gene expression to promote breast tumour growth and metastasis. Nature 452:187–193

Hardiman G (2004) Microarray platforms—comparisons and contrasts. Pharmacogenomics 5:487–502

Harris L, Fritsche H, Mennel R et al (2007) American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. J Clin Oncol 25:5287–5312

Healy DA, Hayes CJ, Leonard P et al (2007) Biosensor developments: application to prostate-specific antigen detection. Trends Biotechnol 25:125–131

Hondermarck H, Tastet C, El Yazidi-Belkoura I et al (2008) Proteomics of breast cancer: the quest for markers and therapeutic targets. J Proteome Res 7:1403–1411

Issaq HJ, Veenstra TD (2007) The role of electrophoresis in disease biomarker discovery. Electrophoresis 28:1980–1988

Jacquemier J, Ginestier C, Rougemont J et al (2005) Protein expression profiling identifies subclasses of breast cancer and predicts prognosis. Cancer Res 65:767–779

Kikuchi T, Carbone DP (2007) Proteomics analysis in lung cancer: challenges and opportunities. Respirology 12:22–28

Klose J (1975) Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues: a novel approach to testing for induced point mutations in mammals. Humangenetik 26:231–243

Kopf E, Zharhary D (2007) Antibody arrays—an emerging tool in cancer proteomics. Int J Biochem Cell Biol 39:1305–1317

Miller BA, Scoppa SM, Feuer EJ (2006) Racial/ethnic patterns in lifetime and age-conditional risk estimates for selected cancers. Cancer 106:670–682

Nuyten DS, Kreike B, Hart AA et al (2006) Predicting a local recurrence after breast-conserving therapy by gene expression profiling. Breast Cancer Res 8:R62

O'Farrell PH (1975) High resolution two-dimensional electrophoresis of proteins. J Biol Chem250:4007–4021

Pease AC, Solas D, Sullivan EJ et al (1994) Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proc Natl Acad Sci USA: 5022–5026

Perou CM, Sorlie T, Eisen MB et al (2000) Molecular portraits of human breast tumours. Nature 406:747–752

Rouzier R, Rajan R, Wagner P et al (2005) Microtubule-associated protein tau: a marker of paclitaxel sensitivity in breast cancer. Proc Natl Acad Sci USA 102:8315–8320

Sanz-Medel A, Montes-Bayon M, del Rosario Fernandez de la Campa M et al (2008) Elemental mass spectrometry for quantitative proteomics. Anal Bioanal Chem 390:3–16

Schena M, Shalon D, Davis RW, Brown PO (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467–470

Smith JC, Lambert JP, Elisma F, Figeys D (2007) Proteomics in 2005/2006: developments, applications and challenges. Anal Chem 79:4325–4343

Sotiriou C, Piccart MJ (2007) Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? Nat Rev Cancer 7:545–553

Sotiriou C, Wirapati P, Loi S et al (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst 98:262–272

Steemers FJ, Ferguson JA, Walt DR (2000) Screening unlabeled DNA targets with randomly ordered fiber-optic gene arrays. Nat Biotechnol 18:91–94

Stranger BE, Forrest MS, Clark AG et al (2005) Genome-wide associations of gene expression variation in humans. PLoS Genet 1:e78

Tu Y, Stolovitzky G, Klein U (2002) Quantitative noise analysis for gene expression microarray experiments. Proc Natl Acad Sci U S A 99:14031–14036

van de Vijver MJ, He YD, van't Veer LJ et al (2002) A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347:1999–2009

van't Veer LJ, Dai H, van de Vijver MJ et al (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415:530–536

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Scienc. 270:484–487

Wang Y, Klijn JG, Zhang Y et al (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 365:671–679

Westfall PH, Young SS (1993) Resampling-based multiple testing: examples and methods for *p*-value adjustment. Wiley, New York

Wingren C, Borrebaeck CA (2007) Progress in miniaturization of protein arrays—a step closer to high-density nanoarrays. Drug Discov Today 12:813–819

Zahedi RP, Meisinger C, Sickmann A (2005) Two-dimensional benzyldimethyl-n-hexadecylammonium chloride/SDS-PAGE for membrane proteomics. Proteomics 5:3581–3588

Zahedi RP, Moebius J, Sickmann A (2007) Two-dimensional BAC/SDS-PAGE for membrane proteomics. Subcell Biochem 43:13–20