Analysis of microarray gene expression data

Wolfgang Huber
w.huber@dkfz.de
German Cancer Research Center
Division of Molecular Genome Analysis
69120 Heidelberg

Anja von Heydebreck
anja.heydebreck@molgen.mpg.de
Max-Planck-Institute for Molecular Genetics
14195 Berlin

Martin Vingron
martin.vingron@molgen.mpg.de
Max-Planck-Institute for Molecular Genetics
14195 Berlin

April 2, 2003

Abstract

This article reviews the methods utilized in processing and analysis of gene expression data generated using DNA microarrays. This type of experiment allows to determine relative levels of mRNA abundance in a set of tissues or cell populations for thousands of genes simultaneously. Naturally, such an experiment requires computational and statistical analysis techniques. At the outset of the processing pipeline, the computational procedures are largely determined by the technology and experimental setup that are used. Subsequently, as more reliable intensity values for genes emerge, pattern discovery methods come into play. The most striking peculiarity of this kind of data is that one usually obtains measurements for thousands of genes for only a much smaller number of conditions. This is at the root of several of the statistical questions discussed here.

1 Introduction

In the context of the human genome project, new technologies emerged that facilitate the parallel execution of experiments on a large number of genes simultaneously. The so-called DNA

microarrays, or DNA chips, constitute a prominent example. This technology aims at the measurement of mRNA levels in particular cells or tissues for many genes at once. To this end, single strands of complementary DNA for the genes of interest - which can be many thousands - are immobilized on spots arranged in a grid ("array") on a support which will typically be a glass slide, a quartz wafer, or a nylon membrane. From a sample of interest, e.g. a tumor biopsy, the mRNA is extracted, labeled and hybridized to the array. Measuring the quantity of label on each spot then yields an intensity value that should be correlated to the abundance of the corresponding RNA transcript in the sample.

Two schemes of labeling are in common use today. One variant labels a single sample, either radioactively or fluorescently. Radioactive labeling is used, e.g., in conjunction with hybridization on nylon membranes [1]. The company Affymetrix synthesizes sets of short oligomers on a glass wafer and uses a single fluorescent label ([2], see also www.affymetrix.com). Alternatively, two samples are labeled with a green and a red fluorescent dye, respectively. The mixture of the two mRNA preparations is then hybridized simultaneously to a common array on a glass slide. This technology is usually refered to as the Stanford technology [3]. Quantification utilizes a laser scanner that determines the intensities of each of the two labels over the entire array. Recently, companies like Agilent have immobilized long oligomers of 60 to 70 basepairs length and used two-color labeling.

The parallelism in this kind of experiment lies in the hybridization of mRNA extracted from a single sample to many genes simultaneously. The measured abundances, though, are not obtained on an absolute scale. This is because they depend on many hard to control factors such as the efficiencies of the various chemical reactions involved in the sample preparation, as well as on the amount of immobilized DNA available for hybridization.

The class of transcripts that is probed by a spot may differ in different applications. Most commonly, each spot is meant to probe a particular gene. The representative sequence of DNA on the spot may be either a carefully selected fragment of cDNA, a more arbitrary PCR product amplified from a clone matching the gene, or one of a set of oligonucleotides specific for the gene. Another level of sophistication is reached when a spot represents, e.g., a particular transcript of a gene. In this case, or for the distinction of mRNA abundances of genes from closely related gene families, careful design and/or selection of the immobilized DNA is required. Likewise, the selection of samples to study and to compare to each other using DNA microarrays requires careful planning as will become clear upon consideration of the statistical questions arising from this technology [4, 5, 6].

There are many different ways for the outline of a microarray experiment. In many cases, a development in time is studied leading to a series of hybridizations following each other. Alternatively, different conditions like healthy/diseased or different disease types may be studied. We generally refer to a time point or a state as a condition and typically for each condition several replicate hybridizations are performed. The replicates should provide the information necessary to judge the significance of the conclusions one wishes to draw from the comparison of the different conditions. When going deeper into the subject it soon becomes clear that this simple outline constitutes a rather challenging program.

This article is organized along the various steps of analysis of a microarray experiment. Statistical problems arise firstly as a consequence of the many technical peculiarities and their solution is a

prerequisite to any meaningful subsequent interpretation of the experiment. Section 2 describes some of the issues related to quality control. Visualization methods are introduced because they may greatly help both in detecting and removing obviously failed measurements, as well as in finding more subtle systematic biases associated with variations in experimental conditions.
Microarray measurements are subject to multiple sources of experimental variation, the mathematical treatment of which are discussed in Section 3. Some of the variations are *systematic*, and may be explicitly corrected for, others are *random*, and may be accounted for through an error model. The correction for systematic effects is refered to as calibration or normalization. We will discuss two error models: One involving a constant coefficient of variation, i. e. a purely multiplicative noise term, and one allowing for a more general variance-to-mean dependence, with a noise term that has both multiplicative and additive components. From these models we derive *measures of relative abundance* of mRNA.
The goal of many microarray experiments is to identify genes that are differentially transcribed with respect to different biological conditions of cell cultures or tissue samples. Section 4 focuses on these issues, paying particular attention to the notoriously low numbers of repeated hybridizations per condition in relation to the high numbers of genes about which one wants to make conclusions. Section 5 proceeds to highlight some of the issues in pattern discovery in microarray data. Here, again, classical methods of data analysis need to be carefully evaluated with respect to their applicability to the particular type of data at hand. A short summary will be given of the methods that have so far been successfully applied. Emphasis is given to exploratory approaches that allow the subsequent formulation of hypothesis that can be tested either through further analysis or further experiments.

2 Data visualization and quality control

A microarray experiment consists of the following components: a set of *probes*, an *array* on which these probes are immobilised at specified locations, a *sample* containing a complex mixture of labeled biomolecules that can bind to the probes, and a *detector* that is able to measure the spatially resolved distribution of label after it has bound to the array [7]. The probes are chosen such that they bind to specific sample molecules; for DNA arrays, this is ensured by the high sequence-specificity of the hybridization reaction between complementary DNA strands. The array is typically a glass slide or a nylon membrane. The sample molecules may be labeled through the incorporation of radioactive markers, such as ^{33}P, or of fluorescent dyes, such as phycoerythrin, Cy3, or Cy5. After exposure of the array to the sample, the abundance of individual species of sample molecules can be quantified through the signal intensity at the matching probe sites. To facilitate direct comparison, the spotted array technology developed in Stanford [3] involves the simultaneous hybridization of two samples labeled with different fluorescent dyes, and detection at the two corresponding wavelengths. Fig. 1 shows an example.

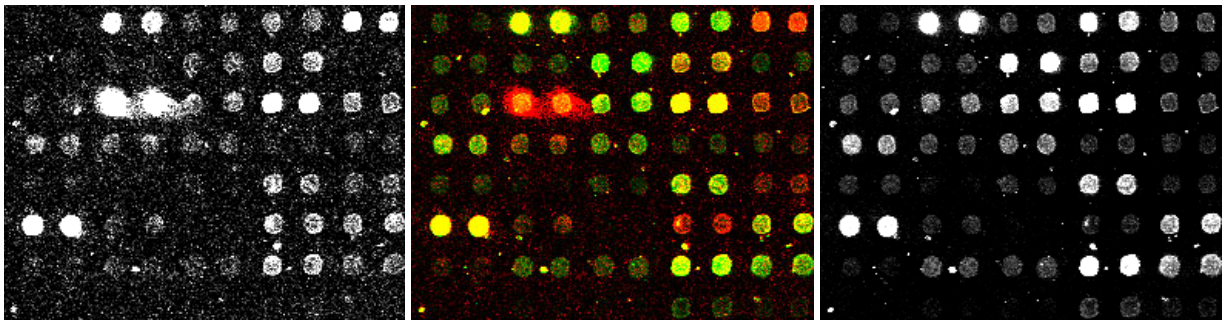


Figure 1: The detected intensity distributions from a cDNA microarray for a region comprising around 80 probes. The total number of probes on an array may range from a few dozens to tens of thousands. Left panel: grey-scale representation of the detected label fluorescence at 635 nm (red), corresponding to mRNA sample A. Right panel: label fluorescence at 532 nm (green), corresponding to mRNA sample B. Spots that light up in only one of the two images correspond to genes that are only transcribed in one of the two samples. Middle panel: false-color overlay image from the two intensity distributions. The spots are red, green, or yellow, depending on whether the gene is transcribed only in sample A, sample B, or both.

2.1 Image quantification.

The intensity images are scanned by the detector at a high spatial resolution, such that each probe spot is represented by many pixels. In order to obtain a single overall intensity value for each probe, the corresponding pixels need to be identified (segmentation), and the intensities need to be summarized (quantification). In addition to the overall probe intensity, further auxiliary quantities may be calculated, such as an estimate of apparent unspecific "local background" intensity, or a spot quality measure. A variety of segmentation and quantification methods is implemented in available software packages. They differ in their robustness against irregularities and in the amount of human interaction that they require. Different types of irregularities may occur in different types of microarray technology, and a segmentation or quantification algorithm that is good for one platform is not necessarily suitable for another. For instance, the variation of spot shapes and positions that the segmentation has to deal with depends on the properties of the support (e. g. glass or nylon), on the probe delivery mechanism (e. g. quill-pen type, pin and ring systems, ink jetting), and on the detection method (optical or radioactive). Furthermore, larger variations in the spot positioning from array to array can be expected in home-made arrays than in mass produced ones. An evaluation of image analysis methods for spotted cDNA arrays was reported by Yang et al. [8].

For a microarray project, the image quantification marks the transition in the work flow from "wet lab" procedures to computational ones. Hence, this is a good point to spend some effort looking at the quality and plausibility of the data. This has several aspects: confirm that positive and negative controls behave as expected; verify that replicates yield measurements close to each other; and check for the occurrence of artifacts, biases, or errors. In the following we present a

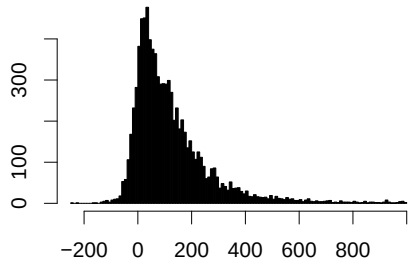


Figure 2: Histogram of probe intensities at the green wavelength for a cDNA microarray similar to the one depicted in Fig. 1. The intensities were determined, in arbitrary units, by an image quantification method, and "local background" intensities were subtracted. Due to measurement noise, this lead to non-positive probe intensities for part of the genes with low or zero abundance. The x-axis has been cut off at the 99% quantile of the distribution. The maximum value is about 4000.

number of data exploration and visualization methods that may be useful for these tasks.

2.2 Dynamic range and spatial effects

A simple and fundamental property of the data is the dynamic range and the distribution of intensities. Since many experimental problems occur at the level of a whole array or the sample preparation, it is instructive to look at the histogram of intensities from each sample. An example is shown in Fig. 2. Typically, for arrays that contain quasi-random gene selections, one observes a unimodal distribution with most of its mass at small intensities, corresponding to genes that are not or only weakly transcribed in the sample, and a long tail to the right, corresponding to genes that are transcribed at various levels. In most cases, the occurence of multiple peaks in the histogram indicates an experimental artifact. To get an overview over multiple arrays, it is instructive to look at the box plots of the intensities from each sample. Problematic arrays should be excluded from further analysis.
Crude artifacts, such as scratches or spatial inhomogeneities, will usually be noticed already from the scanner image at the stage of the image quantification. Nevertheless, to get a quick and potentially more sensitive view of spatial effects, a false color representation of the probe intensities as a function of their spatial coordinates can be useful. There are different options for the intensity scaling, among them the linear, logarithmic, and rank scales. Each one will highlight different features of the spatial distribution. Examples are shown in Fig. 3.

2.3 Scatterplot

Usually, the samples hybridized to a series of arrays are biologically related, such that the transcription levels of a large fraction of genes are approximately the same across the samples. This can be expected e. g. for cell cultures exposed to different conditions or for cells from biopsies of

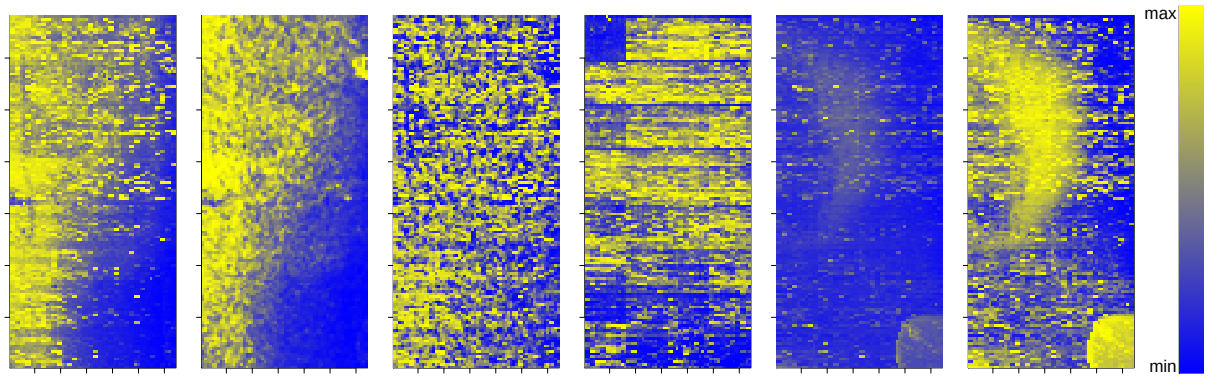


Figure 3: False color representations of the spatial intensity distributions from three different 64×136 spot cDNA microarrays from one experiment series. a) probe intensities in the red color channel, b) local background intensities, c) background-subtracted probe intensities. In a) and b), there is an artifactual intensity gradient, which is mostly removed in c). For visualization, the color scale was chosen in each image to be proportional to the ranks of the intensities. d) For a second array, probe intensities in the green color channel. There is a rectangular region of low intensity in the top left corner, corresponding to one print-tip. Apparently, there was a sporadic failure of the tip for this particular array. Panels e) and f) show the probe intensities in the green color channel from a third array. The color scale was chosen proportional to the logarithms of intensities in e) and proportional to the ranks in f). Here, the latter provides better contrast. Interestingly, the bright blob in the lower right corner appears only in the green color channel, while the half moon shaped region appears both in green and red (not shown).

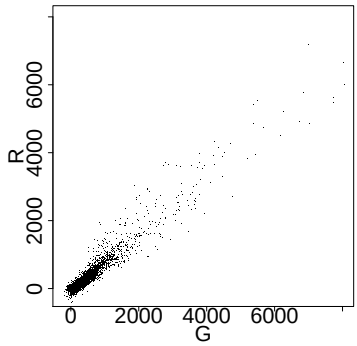
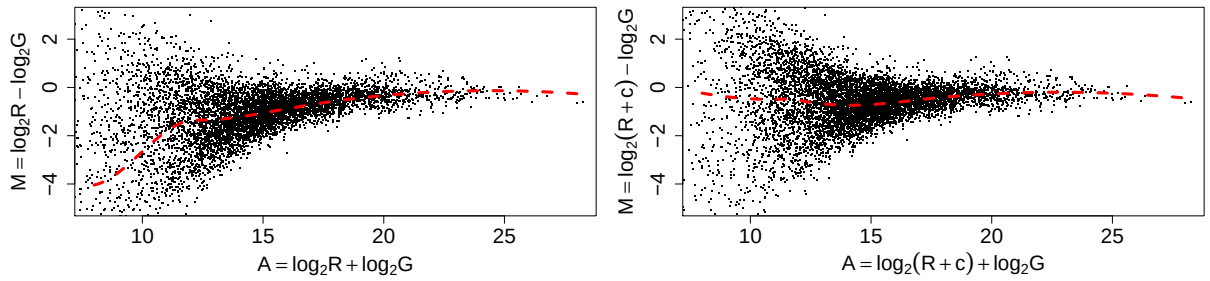


Figure 4: Scatterplot of probe intensities in the red and the green color channel from a cDNA array containing 8000 probes.



a) b)

Figure 5: a) the same data as in Fig. 4, after logarithmic transformation and clockwise rotation by $45°$. The dashed line shows a local regression estimate of the systematic effect $M_0(A)$, see text. b) similar to a), however a constant value $c = 42$ has been added to the red intensities before log transformation. After this, the estimated curve for the systematic effect $M_0(A)$ is approximately constant.

the same tissue type, possibly subject to different disease conditions. We call this the *majority of genes unchanged* property. Visually, it can be verified from the scatterplot of the probe intensities for a pair of samples. An example is shown in Fig. 4.

The scatterplot allows to assess both measurement noise and systematic biases. Ideally, the data from the majority of the genes that are unchanged should lie on the bisector of the scatterplot. In reality, there are both systematic and random deviations from this [9]. For instance, if the label incorporation rate and photoefficiency of the red dye were systematically lower than that of the green dye by a factor of 0.75, the data would be expected not to lie on the bisector, but rather on the line $y = 0.75x$.

Most of the data in Fig. 4 is squeezed into a tiny corner in the bottom left of the plot. More informative displays may be obtained from other axis scalings. A frequently used choice is the double-logarithmic scale. An example is shown in Fig. 5. It is customary to transform to new

variables $A = \log R + \log G$, $M = \log R - \log G$ [10]. Up to a scale factor of $\sqrt{2}$, this corresponds to a clockwise coordinate system rotation by $45°$. The horizontal coordinate A is a measure of average transcription level, while the *log-ratio* M is a measure for differential transcription. If the majority of genes are not differentially transcribed, the scatter of the data points in the vertical direction may be considered a measure of the random variation. Fig. 5a also shows a systematic deviation of the observed values of M from the line $M = 0$, estimated through a local regression line[1]. There is an apparent dependence $M_0(A)$ of this deviation on the mean intensity A. However, this is most likely an artifact of applying the logarithmic transformation: as shown in Fig. 5b, the deviation may be explained sufficiently well through a constant $M_0(A) = M_0$ if an appropriate offset is added to the R values before taking the logarithm. Note that a horizontal line at $M = M_0$ in Fig. 5b corresponds to a straight line of slope 2^{M_0} and with intercept c in Fig. 4.

Fig. 5 shows the *heteroskedasticity* of log-ratios: while the variance of M is relatively small and approximately constant for large average intensities A, it becomes larger as A decreases. Conversely, examination of the differences $R - G$, for example through plots like in Fig. 4, shows that their variance is smallest for small values of the average intensity $R+G$ and increases with $R + G$. Sometimes, one wishes to visualize the data in a manner such that the variance is constant along the whole dynamic range. A data transformation that achieves this goal is called a variance-stabilizing transformation. In fact, *homoskedastic* representations of the data are not only useful for visualization, but also for further statistical analyses. This will be discussed in more detail in Section 3.

Two extensions of the scatterplot are shown in Figs. 6 and 7. Rather than plotting a symbol for every data point, they use a density representation, which may be useful for larger arrays. For example, Fig. 6 shows the scatterplot from the comparison of two tissue samples based on 152,000 probes[2]. The point density in the central region of the plot is estimated by a kernel density estimator. Three-way comparisons may be performed through a projection such as in Fig. 7. This uses the fact that the $(1, 1, 1)$-component of a three-way microarray measurement corresponds to average intensity, and hence is not directly informative with respect to differential transcription. Note that if the plotted data was pre-processed through a variance-stabilizing transformation, its variance does not depend on the $(1, 1, 1)$-component.

2.4 Batch effects

Present day microarray technology measures abundances only in terms of relative probe intensities, and generally provides no calibration to absolute physical units. Hence, the comparison of measurements between different studies is difficult. Moreover, even within a single study, the measurements are highly susceptible to *batch effects*. By this term, we refer to experimental factors that (i) add systematic biases to the measurements, and (ii) may vary between different subsets or stages of an experiment. Some examples are [9]:

[1] We used `loess` [11] with default parameters span=0.75, degree=2.

[2] The arrays used were RZPD Unigene-II arrays (www.rzpd.de).

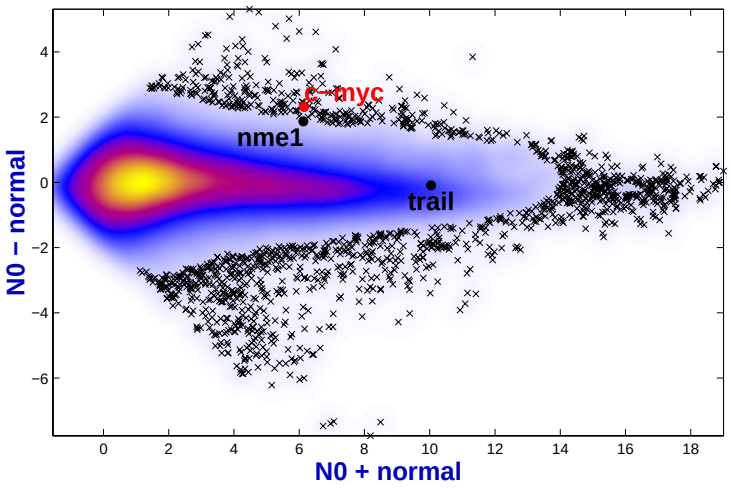


Figure 6: Scatterplot of a pairwise comparison of non-cancerous colon tissue and a colorectal tumor. Individual transcripts are represented by 'x' symbols. The x-coordinate is the average of the appropriately calibrated and transformed intensities (cf. Section 3). The y-coordinate is their difference, and is a measure of differential transcription. The array used in this experiment contained 152,000 probes representing around 70,000 different clones. Since plotting all of these would lead to an uninformative solid black blob in the centre of the plot, the point density is visualized by a color scale, and only 1500 data points in sparser regions are individually plotted.

1. *spotting:* to manufacture spotted microarrays, the probe DNA is deposited on the surface through spotting pins. Usually, the robot works with multiple pins in parallel, and the efficiency of their probe delivery may be quite different (e. g. Fig. 3d or [10]). Furthermore, the efficiency of a pin may change over time through mechanical wear, and the quality of the spotting process as a whole may be different at different times, due to varying temperature and humidity conditions.

2. *PCR amplification:* for cDNA arrays, the probes are synthesized through PCR, whose yield varies from instance to instance. Typically, the reactions are carried out in parallel in 384-well plates, and probes that have been synthesized in the same plate tend to have highly correlated variations in concentration and quality. An example is shown in Fig. 8.

3. *sample preparation protocols:* The reverse transcription and the labeling are complex biochemical reactions, whose efficiencies are variable and may depend sensitively on a number of hard-to-control circumstances. Furthermore, RNA can quickly degrade, hence the outcome of the experiment can depend sensitively on when and how conditions that prevent RNA degradation are applied to the tissue samples.

4. *array coating:* both the efficiency of the probe fixation on the array, as well as the amount of unspecific background fluorescence strongly depend on the array coating.

5. *scanner and image analysis*

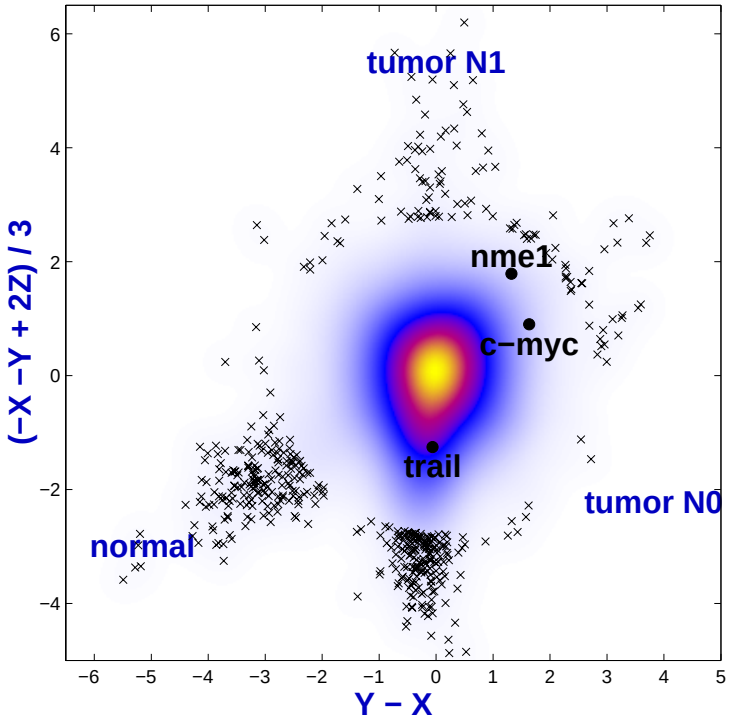


Figure 7: Scatterplot of a triple comparison between non-cancerous colon tissue, a lymph-node negative colorectal tumor (N0), and a lymph-node positive tumor (N1). The measurements from each probe correspond to a point in three-dimensional space, and are projected orthogonally on a plane perpendicular to the (1,1,1)-axis. The three coordinate axes of the data space correspond to the vectors from the origin of the plot to the three labels “normal”, “tumor N0”, and “tumor N1”. The (1,1,1)-axis corresponds to average intensity, while differences between the three tissues are represented by the position of the measurements in the two-dimensional plot plane. For instance, both c-myc and nme1 are higher transcribed in the N0 and in the N1 tumor, compared to the non-cancerous tissue. However, while the increase is approximately balanced for c-myc in the two tumors, nme1 (nucleoside diphosphate kinase A) is more upregulated in the N1 tumor than in the N0 tumor, a behavior that is consistent with a gene involved in tumor progression. On the other side, the apoptosis inducing receptor trail-r2 is down-regulated specifically in the N1 tumors, while it has about the same intermediate-high transcription level in the non-cancerous tissue and the N0 tumor. Similar behavior of these genes was observed over repeated experiments.

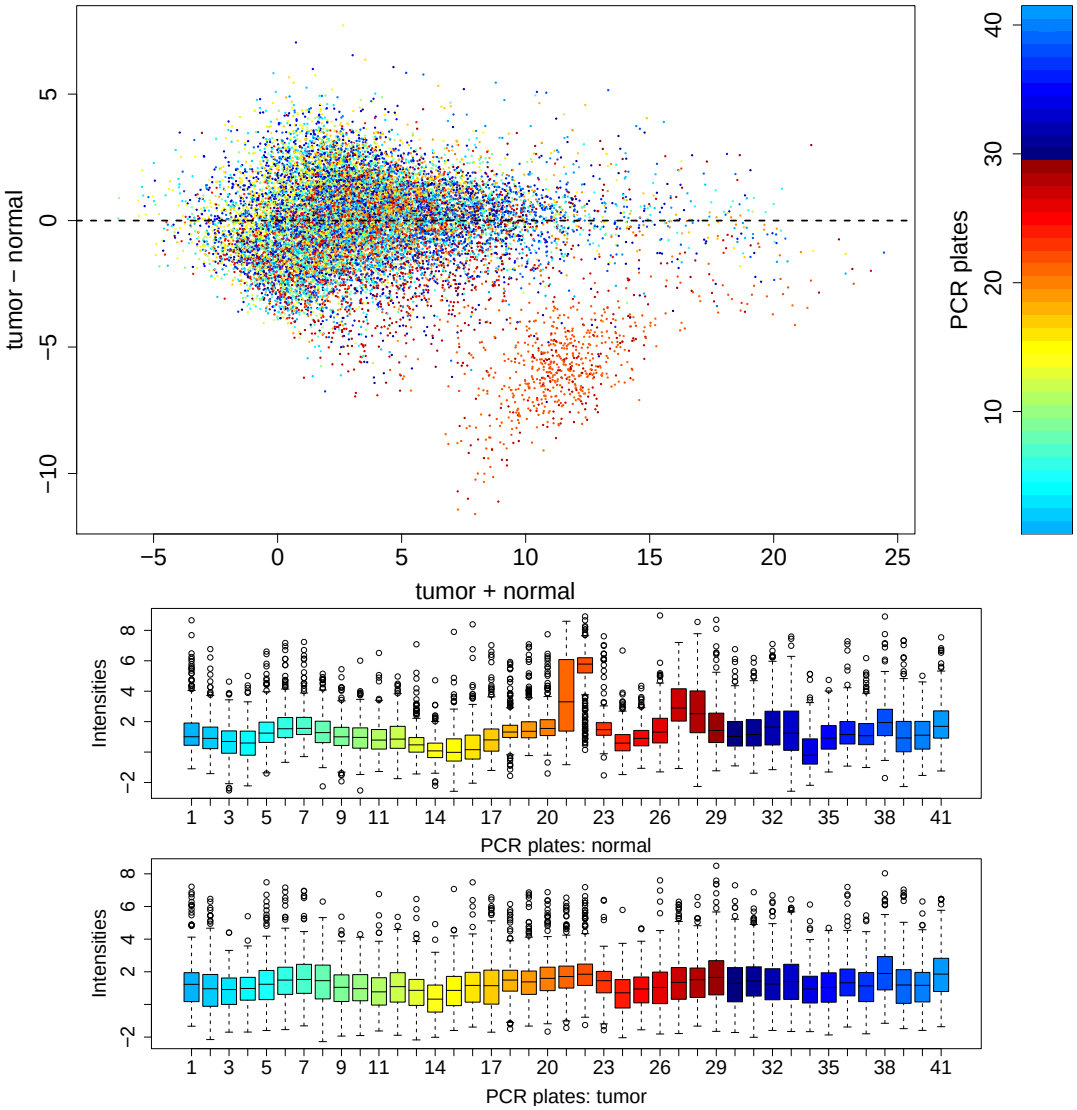


Figure 8:

Top panel: scatterplot of intensities from a pair of cDNA arrays, comparing renal cell carcinoma to matched non-cancerous kidney tissue. Similar to Fig. 6, the x-coordinate represents average, and the y-coordinate differential signal. In the bottom of the plot, there is a cloud of probes that appear to represent a cluster of strongly down-regulated genes. However, closer scrutiny reveals that this is an experimental artifact: the bottom panels show the boxplots of the intensities for the two arrays, separately for each of the 41 PCR plates (see text). Probes from plates no. 21, 22, 27, and 28 have extraordinarily high intensities on one of the arrays, but not on the other. Since the clone selection was quasi-random, this points to a defect in the probe synthesis that affected one array, but not the other. The discovery of such artifacts may be helped by coloring the dots in the scatterplot by attributes such as PCR plate of origin or spotting pin (for technical reasons, the print version of this figure is shown in grey-scale). While the example presented here is an extreme one, caution towards batch artifacts is warranted whenever arrays from different manufacturing lots are used in a single study.

These considerations have important consequences for the experimental design: first, any variation that can at any means be avoided within an experiment should be avoided. Second, any variation that cannot be avoided should be organized in such a manner that it does not confound the biological question of interest. Clearly, when looking for differences between two tumor types, it would not be wise to have samples of one tumor type processed by one laboratory, and samples of the other type by another laboratory.
Points 1 and 2 are specific for spotted cDNA arrays. To be less sensitive against these variations, the two-color labeling protocol is used, which employs the simultaneous hybridization of two samples to the same array [3]. Ideally, if only ratios of intensities between the two color channels are considered, variations in probe abundance should cancel out. Empirically, they do not quite, which may, for example, be attributed to the fact that observed intensities are the sum of probe-specific signal and unspecific background [12]. Furthermore, in the extreme case of total failure of the PCR amplification or the DNA deposition for probes on some, but not all arrays in an experimental series, artifactual results are hardly avoidable.
If any of the factors 3–5 is changed within an experiment, there is a good chance that this will later show up in the data as one of the most pronounced sources of variation. A simple and instructive visual tool to explore such variations is the correlation plot: Given a set of d arrays, each represented through a high-dimensional vector $\vec{Y}_i$ of suitably transformed and filtered probe intensities, calculate the $d \times d$ correlation matrix corr$(\vec{Y}_i, \vec{Y}_j)$, sort its rows and columns according to different experimental factors, and visualize the resulting false color images.

3 Error models, calibration and measures of differential expression

The relation between a measured intensity y_{ki} of probe k and the true abundance x_{ki} of molecule type k in sample i may be described as

$$y_{ki} = a_{ki} + b_{ki}\, x_{ki}. \tag{1}$$

The gain factor b_{ki} represents the net result of the various experimental effects that come between the count of molecules per cell in the sample and the final readout of the probe intensity, such as: number of cells in the sample, the mean number of label molecules attaching to a sample molecule, hybridization efficiency, label efficiency, and detector gain. The additive term a_{ki} accounts for that part of the measured intensity that does not result from x_{ki}, but from effects such as unspecific hybridization, background fluorescence, stray signal from neighboring probes, and detector offset.
The parameters a_{ki} and b_{ki} are different for each probe k and for each hybridization i. It is not practical to determine them exactly, but neither is it necessary. Rather, one is content with obtaining *statistical* statements about *relative* abundances. To this end, one may build stochastic models for the effects a_{ki} and b_{ki}. Different variations on this theme have been proposed, as will be presented below.
First, however, we would like to discuss the functional form of Eqn. (1), whose major statement is that when the true abundance x_{ki} increases, the measured signal y_{ki} increases proportionally.

Could it be necessary to consider more complex non-linear relationships? Clearly, this cannot be ruled out for all possible experiments, or for future technologies. However, a linear operating range over several orders of magnitude has been reported by a number of authors for current microarray technologies (e. g. [13, 14, 15]). At the lower end, this range is limited only by the requirement that x_{ki} be non-negative. At the upper end, the linear range is limited by saturation effects such as quenching, limited probe abundance, and detector saturation. However, for realistic concentrations of sample molecules, the upper limit is not reached in well-conducted experiments.

3.1 Multiplicative calibration and noise

In a seminal paper in 1997, Chen et al. [16] introduced a decomposition of the multiplicative effect (cf. Eqn. (1)),

$$b_{ki} = b_i \beta_k (1 + \varepsilon_{ki}). \tag{2}$$

Here, β_k is a probe-specific coefficient, the same for all samples. For each sample i, the normalization factor b_i is applied across all probes. The remaining variation in b_{ki} that cannot be accounted for by β_k and b_i is absorbed by ε_{ki}. Furthermore, since the measured intensities y_{ki} are already "background-corrected" by the image analysis software's local background estimation, Chen et al. assumed the additive effects a_{ki} to be negligibly small. They further simplified the problem in two steps:
First, they noted that one is mainly interested in relative comparisons between the levels of the same gene under different conditions, i. e. , in the ratios x_{ki}/x_{kj}. Hence the probe-specific effects β_k can be absorbed, $\mu_{ki} = \beta_k x_{ki}$, simply rescaling the units in which molecule abundances are measured, and need not be determined.
Second, they turned to a stochastic description, and modeled ε_{ki} as a normally distributed noise term with mean zero and standard deviation c, independent of i and k. Thus, in the model of Chen et al. the measured intensity Y_{ki} is a random variable and depends on the true level μ_{ki} as follows:

$$Y_{ki} = b_i \, \mu_{ki} \, (1 + \varepsilon_{ki}), \qquad \varepsilon_{ki} \sim \mathrm{N}(0, c^2). \tag{3}$$

Note that Y_{ki} has constant coefficient of variation c.
Chen et al. specifically considered two-color cDNA microarrays, where $i = 1, 2$ represents the red and the green color channel, respectively. For a given true ratio μ_{k1}/μ_{k2}, Chen et al. derived the distribution of the observed, normalized ratio $M_k = Y_{k2}/Y_{k1} \times b_1/b_2$. It only depends on the values of c and b_1/b_2, and Chen et al. gave an algorithm for the estimation of these parameters from the data. Based on this, they were able to formulate a statistical test for differential expression, i. e. a test against the hypothesis $\mu_{k1} = \mu_{k2}$. Hence, the ratios M_k were regarded as a sufficient summary of the results from a single microarray slide, and they, or their logarithms, would then be used as the input for further higher level analyses of data from multiple slides.
To allow for a more systematic analysis of multiple slide experiments, Kerr et al. proposed an approach based on the ANOVA technique [17]. They modeled the measured intensity Y_{kjlm} of probe k on slide j, in the color channel of dye l, from a sample that received treatment m, as

$$\log Y_{kjlm} = g_k + s_j + d_l + v_m + [gs]_{kj} + [gv]_{km} + \varepsilon_{kjlm}. \tag{4}$$

Here, g_k, s_j, d_l, v_m are main effects for probe, array, dye, and treatment, respectively. The probe-array interaction $[gs]_{kj}$ accounts for variations of probe quality in the array manufacture, and the probe-treatment interaction $[gv]_{km}$ for differential levels of transcription of gene g between different treatment groups m. The noise terms ε_{kjlm} account for all other variations and are assumed to be independent and identically distributed. The ANOVA model (4) is related to Eqn. (1) by setting

$$a_{ki} = 0 \tag{5}$$

$$\log b_{ki} + \log x_{ki} = (s_j + d_l + [gs]_{kj}) + (g_k + v_m + [gv]_{km}) + \varepsilon_{kjlm} \tag{6}$$

where $j \equiv j(i)$, $l \equiv l(i)$ and $m \equiv m(i)$ are slide, dye, and treatment associated with sample i, respectively. The terms in the first pair of parentheses on the right hand side of Eqn. (6) may be attributed to the measurement gain b_{ki}, and the terms in the second pair to the actual abundance x_{ki}, but generally such a decomposition is not unique.
Both the models of Chen et al. and of Kerr et al. were formulated with reference to the two-color cDNA array technology. However, Eqn. (4) can be adapted (in fact, simplified) in a straightforward manner to data from one-color array technologies, such as Affymetrix genechips or cDNA membranes with radioactive labeling. Furthermore, to represent more complex experimental designs than simple two-way comparisons, more detailed terms than the single factor v_m can be introduced into (4), and the efficiencies of different designs can be compared using standard techniques for linear models [4].

3.2 Limitations

The concepts of Section 3.1 have been widely used for microarray data analysis. However, it has also become clear that, for many data sets that are encountered in practice, they are not sufficient. The following points are worth noting:

1. **Robustness.** In order to make model (3) identifiable, Chen et al. assumed that the transcription levels of all genes were unchanged, and set $\mu_{k1} = \mu_{k2}$ for all k. Thus, their model is misspecified for the part of the data arising from truely differentially transcribed genes, which act as outliers. However, their parameter estimation is based on least-squares-criteria, and may be sensitive to the presence of such outliers. Besides that, outliers may be caused by technical artifacts.

2. **Heteroskedasticity.** The significance of log-ratios depends on the absolute values of the intensities in the numerator and denominator [18, 19, 20, 21, 22]. Typically, the variance of log-transformed intensities increases as their mean decreases.

3. **Apparent non-linearities.** According to the above models, the data from a pair of samples should lie along a straight line in the scatterplot of the log-transformed intensities. However, in real data, several authors have observed data that follows a curved line, e. g. Fig. 5, [18, 20, 10].

4. **Negative values.** While the image quantification's estimates for probe "foreground" and "background" intensities are generally positive, this is usually not true for their difference. If a gene is weakly or not expressed, it can happen by chance that the background estimate is larger than the foreground estimate (cf. Fig. 2). However, non-positive values make sense neither for ratios nor for the log-transformation.

To address these problems, various fixes have been proposed. We give a brief and incomplete review.

1. Robustness. Robust estimation techniques in the context of microarray data have been described by many authors (e. g. [18, 20, 10, 23, 24]). A general overview is given in [25].

2. Heteroskedasticity. It is often observed that the variance of the log-ratio is a monotonously decreasing function of the mean intensity. One common practice has been to discard the log-ratios calculated from intensities below a certain threshold and to treat the rest as if they were homoskedastic.

Newton et al. [19] proposed a *shrinkage estimator*

$$\frac{y'_{k1} + \nu}{y'_{k2} + \nu} \tag{7}$$

to replace the naive ratio y'_{k1}/y'_{k2}. Here, $y'_{ki} = y_{ki}/b_i$ are the normalized intensities. Similar to Chen et al., they neglected the additive terms a_{ki} and used a model of the measurement error with a constant coefficient of variation. To arrive at (7), they enclosed this in a hierarchical Bayesian model, using a prior distribution for the mRNA abundances, and, in particular, their positivity. The form of this distribution is reflected by the shrinkage parameter ν, which is estimated from the data. To infer differential transcription, they derived "posterior odds of change", which, however, are no simple function of the log-ratio or of (7).

Several authors have addressed the problem of heteroskedasticity by estimating the variance of the log-ratios or of log-transformed intensities separately for each gene (e. g. [10, 23]). However, in many applications the number of samples available is too small for reliable estimates of gene-specific variance, hence it has been proposed to estimate the variance as a non-parametric smooth function of the mean intensity, through a local regression. The log-ratios may then be *studentized* by dividing them by their locally estimated standard deviation [20]. Baggerly et al. [22] provided some theoretical foundation for this from models of the measurement error for different levels of replication. According to these, the variance of the log-ratio is largest for small intensities and exponentially decreases towards an asymptotic positive value as the intensity increases.

3. Apparent non-linearities. To correct for the curved appearence of the scatterplot of log-transformed data, Dudoit et al. [10] proposed to replace the normalization factor b_1/b_2 in Eqn. (3) by a smooth function $M_0(A)$ (cf. Fig. 5). It is estimated through robust local regression [11] and, by construction, this correction makes the scatterplot look straight.

A similar correction was proposed by Kepler et al. [20], in the framework of a model similar to Eqn. (4). In their approach, the terms $s_j + d_l$ (slide and dye effects) are replaced by smooth functions of g_k (mean logarithmic abundance of gene k), which are again estimated through robust local regression.

4. Negative values. In order to be able to calculate ratios and logarithms from real microarray data, different fixes have been proposed to deal with non-positive values: mark them as invalid or missing; replace them by a fixed, small positive value; use an imputation algorithm to replace them by a more acceptable value; add pseudocounts, such that the whole set of intensities becomes positive; ignore the local background estimate (cDNA arrays) or the mismatch probes (Affymetrix genechips) and use only the strictly positive foreground or perfect match intensities. All of these approaches seem to reflect the common wisdom that molecule abundances are not negative. However, probe intensities are only *measurements* of abundance, and in the presence of an additive component of the measurement noise negative measurements may well be consistent with zero or positive abundance. In any experiment, a certain proportion of genes will have zero or low abundance in some samples but not in others, hence the treatment of the non-positive intensity measurements may affect a large and potentially informative fraction of the data.

3.3 Multiplicative and additive calibration and noise

Interestingly, points 2.-4. of the previous section can be related to a rather basic assumption of the models (3) and (4), and it appears that in many cases the associated problems can be resolved by using a more general model. Chen et al. as well as Kerr et al. assumed that the additive terms a_{ki} in Eqn. (1) were negligible, or at least sufficiently accounted for by the image quantification's local background estimation algorithm. One way to arrive at a more realistic model is to set

$$a_{ki} = a_i + b_i \eta_{ki}, \tag{8}$$
$$b_{ki} = b_i \beta_k (1 + \varepsilon_{ki}), \tag{9}$$

where the decomposition of the multiplicative effect (9) is the same as in Eqn. (2), a_i is a sample-specific additive parameter, and η_{ki} are independent and normally distributed random variables with zero mean and common variance. Hence, model (3) is replaced by

$$\frac{Y_{ki} - a_i}{b_i} = \mu_{ki}\, e^{\varepsilon_{ki}} + \eta_{ki}, \qquad \varepsilon_{ki} \sim \mathrm{N}(0, c^2),\ \eta_{ki} \sim \mathrm{N}(0, s^2). \tag{10}$$

Model (10) was proposed by Rocke and Durbin [26] and, using different distributional assumptions, by Ideker et al. [13]. The latter authors used a more detailed parameterization of the noise terms, allowing for different values of the standard deviations c and s for the red and green color channels $i = 1, 2$ and for correlation between ε_{k1} and ε_{k2}, as well as between η_{k1} and η_{k2}. In both cases, the authors did not try to estimate the calibration parameters a_i, b_i, but rather assumed that a calibration had already been performed through some other means.

Consequences. First, the intensities Y_{ki} are no longer supposed to have a constant coefficient of variation. Rather, they obey a variance-to-mean dependence

$$v(u) = c^2 (u - a_i)^2 + b_i^2 s^2 \tag{11}$$

where, in a slight abuse of notation, $v \equiv \mathbf{Var}(Y_{ki})$ and $u \equiv \mathbf{E}(Y_{ki})$, and the equation holds for all probes k for sample i. Recall that a constant coefficient of variation corresponds to a dependence

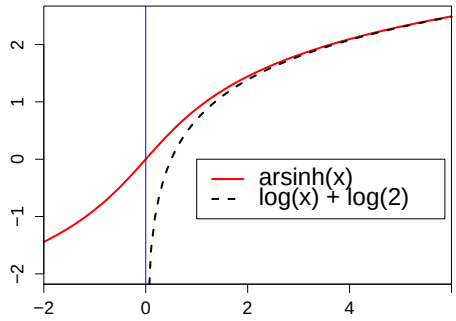


Figure 9: Solid line: graph of the function $\mathrm{arsinh}(x)$. Dashed line: graph of $\log(2x)$. The vertical line marks the singularity of the logarithm function at $x = 0$. The arsinh function is symmetric, $\mathrm{arsinh}(x) = -\mathrm{arsinh}(-x)$, however, most relevant for Eqn. (12) is its behaviour in an x-range as depicted here.

$v(u) = c^2u^2$, which is a special case of (11) for $a_i = s = 0$. In this case, the logarithm is a variance-stabilizing transformation, i. e. the log-transformed data have approximately constant variance. For the more general variance-to-mean dependence (11), such a transformation can also be found, as will be explained below.
Second, the ratio of intensities y_{k1}/y_{k2} is no longer the best estimator for the true fold change μ_{k1}/μ_{k2}. This was addressed by Dror et al. [27], who estimated $\log(\mu_{k1}/\mu_{k2})$ by the posterior mean of a hierarchical model that consists of Eqn. (10) together with an empirical prior for the distribution of μ_{ki}. Their estimator coincides with the log-ratio if both y_{k1} and y_{k2} are large, and remains well-behaved for small or non-positive values of y_{k1} and y_{k2}.
The appropriate variance stabilizing transformation was described by Huber et al. [24] and by Rocke and Durbin [28]. It has the form

$$h_i(y_{ki}) = \mathrm{arsinh}\left(\frac{c}{s}\cdot\frac{y_{ki}-a_i}{b_i}\right). \tag{12}$$

The parameters a_i and b_i may be interpreted as array-specific calibration parameters, while the coefficient of variation c and the background noise level s parametrize the overall error model. The graph of the arsinh function is depicted in Fig. 9. The following two relations hold between the arsinh and the $\log$ function:

$$\mathrm{arsinh}(x) = \log(x+\sqrt{x^2+1}) \tag{13}$$

$$\lim_{x\to\infty}\{\mathrm{arsinh}(x)-\log(2x)\} = 0. \tag{14}$$

In the framework of Section 3.1, log-ratios, the differences between the logarithms of normalized intensities, were the appropriate measure of differential transcription to be used in downstream analyses. In analogy, we define [24]

$$\Delta h_{k;ij} = h_i(y_{ki}) - h_j(y_{kj}). \tag{15}$$

For intensities that are much larger than the additive noise level, (15) becomes equivalent to the log-ratio, as is seen from Eqn. (14). But, in contrast to the log-ratio, $\Delta h_{k;ij}$ is well-defined and has constant variance c^2 across the whole range of intensities. In fact, $\Delta h_{k;ij}/c$ may be thought of as a "studentized log-ratio".
To estimate the model and transformation parameters, one could directly fit model (10) to the data, using the *majority of genes unchanged* assumption $\mu_{ki} = \mu_k$ for most genes k. A computationally simpler approach is to fit the model

$$h_i(Y_{ki}) = \tilde{\mu}_k + \tilde{\varepsilon}_{ki}, \qquad \tilde{\varepsilon}_{ki} \sim \mathrm{N}(0, c^2). \tag{16}$$

Up to first and second moments, models (10) and (16) are equivalent. Parameter estimates can be obtained from a robust variant of the maximum likelihood estimator. A robust estimator with high breakdown point is needed not only because there may be technical outliers, but also because the assumption $\mu_{ki} = \mu_k$ does not hold for a minority of genes that have biologically different transcription levels in different samples [24].
The identification of differentially transcribed genes through statistical tests on Δh_k values was shown to have higher sensitivity and specificity than that through tests on log-ratios [24]. This may be explained by the fact that for non-differentially transcribed genes the Δh_k values have unimodal distributions with mean zero and variances independent of the genes' mean transcription levels. Hence, within the limits of the error model, all available information with respect to differential transcription of gene k is contained in the values of $\Delta h_{k;ij}$. On the other side, the distributions of log-ratios may have, even for some of the non-differentially transcribed genes, mean values different from zero due to sensitive dependence on calibration errors, they may have variances that strongly depend on the mean transcription levels, and they involve missing values, if there are non-positive net probe intensities. These points are illustrated in Fig. 10.

Probe set summaries. A gene transcript may be represented by multiple probes on an array. To obtain an overall measure of abundance per gene, a straightforward approach is to take the average of the corresponding calibrated and transformed probe intensities (12). If additional information on the reliability of the probe measurements is available, a weighted average may be used. This has been investigated most extensively in the context of Affymetrix genechip data [2]. On these chips, each transcript is represented by 16 to 20 pairs of oligonucleotide probes referred to as probe sets. Each probe pair consists of an oligonucleotide of 25 bases that exactly matches the target sequence, and of one that has a mismatch in the middle. The mismatch probes are thought to provide estimates of unspecific contributions to the signal measured from the perfect match probes. A good overview, with many further references, was given by Irizarry et al. [15].

4 Identification of differentially expressed genes

One of the basic goals in the analysis of microarray gene expression data is the identification of differentially expressed genes in the comparison of different types of cell or tissue samples. In order to control the biological and experimental variability of the measurements, statistical inference has to be based on an adequate number of replicate experiments. Here one may distinguish

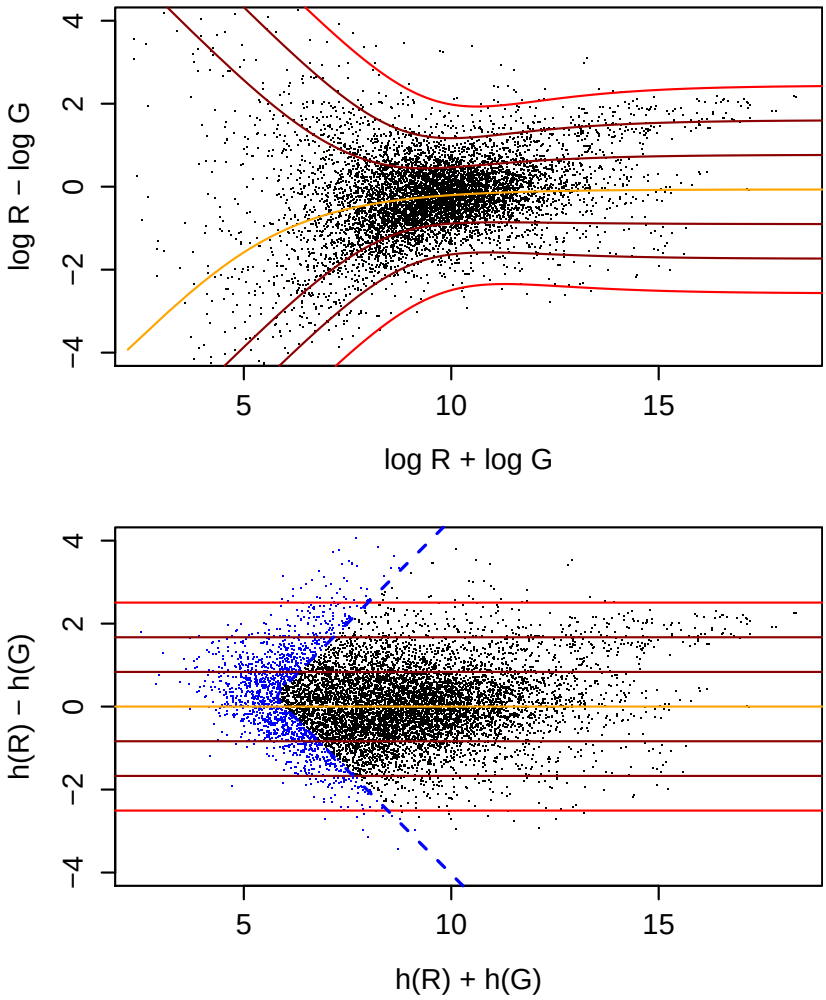


Figure 10: Scatterplot of differential versus total intensities from a two-color-cDNA array, using two different transformations: logarithmic, upper panel, and Eqn. (12), lower panel. The horizontal lines correspond to the z-score $\Delta h/\hat{c} = 0, \pm 1, \pm 2, \pm 3$. The z-score of a pair of red and green probe intensities is their difference divided by its expected standard deviation according to the variance-versus-mean function $v(u)$. The z-score is a statistical measure for how strongly an observed pair of intensities is indicative of true differential abundance. While the contours of the z-score are functions of both log-ratio and total intensity (upper panel), they are independent of total intensity in the coordinate system of the lower panel. Due to a local background subtraction, this data set contained small and negative net intensities. The top panel shows measurements with $R, G > 0$ and $\log(RG) > 2.5$. All data is shown in the lower panel, with the subset of the upper panel to the right of the dashed line.

between cases in which one wishes to make statements on a particular cell population and cases in which one wants to make statements that hold in the presence of biological variability, such as with biopsy studies of diseased tissues. In the first case, independent replications can be obtained on the level of multiple mRNA isolations, in the second, they may be provided by samples from different patients.

For the following, we assume that the data are given either as absolute intensities or as relative values with respect to a common reference sample, and have been calibrated (see Section 3).

To identify differentially expressed genes with respect to a certain biological question, a suitable statistical test may be performed for each gene [29]. The choice of the test statistic depends on the biological question and on the nature of the available experimental data. In the simplest case, one asks for genes that show different transcript abundance between two conditions. In more complex situations, one may look for genes whose abundance is associated with multiple factor levels of one or more sample characteristics. Furthermore, one may consider continuous-valued sample characteristics and test for genes which show non-zero coefficients in a regression model, such as a linear model or a Cox proportional hazards model on patient survival data.

Different statistical tests may make more or less strong assumptions on the distributions of replicate measurements. Important questions are whether the distributions are symmetric, how similar or dissimilar they are to normal distributions, what their behavior at the tails is like, and whether or how their variance (or another appropriate measure of scale) varies between different genes or between different conditions. Such differences in the variance may occur for several reasons: In Section 3, we have discussed the dependence of the variance on the mean, an example for which is given in Eqn. (11). There may be other technological effects that can influence the variance of the measurement distributions in a gene- or condition-dependent manner, such as GC-content, or probe length. Finally, there may be genuine biological differences, such as different tightness of the regulatory control for different genes or for the same gene under different physiological or disease conditions.

Data transformations, such as the logarithmic transformation or a variance stabilizing transformation like Eqn. (12) may be used to make the distributions more symmetric and possibly close to normal, and to remove the systematic dependence of the variance on the mean (see Section 3).

In the comparison of two conditions, one might use Student's t–test or the Wilcoxon rank sum test. Both tests assume that the distributions of the replicate measurements under the two conditions have the same shape and test for differences in the location, with the t–test additionally assuming normal distributions. To account for possibly unequal variances in the two groups, Welch's version of the t–test may be preferred [30, 31]. In order to avoid distributional assumptions, Dudoit et al. proposed to estimate the null distribution of the t–statistic (or, equivalently, of the difference of means) for each gene using permutations of the sample labels. Comparative analyses of different univariate statistical tests in the analysis of gene expression data were presented in [32, 23, 33], however without a conclusive result.

In addition to standard aspects of hypothesis testing, two specific properties of microarray data have motivated the development of novel strategies:

1. Variance estimation: in one extreme, one may estimate the variances of the distributions separately for each gene, and possibly for each condition. This requires a large number of repetitions, which are not always available. In the other extreme, one may use a pooled estimate of the vari-

ance over all conditions and genes. After the application of a variance-stabilizing transformation such as Eqn. (12), the assumption of constant variance may result in tolerable bias and, due to the large number of genes represented on an array, in very low variance of the estimator. This is the case especially if few repetitions are available. In-between the two extremes, a number of methods have been proposed that pool the variance estimation over some genes, but also retain some gene-dependence.
2. Multiple testing: Due to the large number of genes on an array and thus the large number of tests performed, a considerable number of genes may show differential signal intensities simply by chance. Several concepts of assessing the statistical significance of test results obtained from microarray data have been developed.

4.1 Regularized t-statistics

To overcome the instability of the gene-specific variance estimate in the case of few replicate experiments per condition, several authors have proposed methods where a value estimated from a larger set of genes is used to augment the gene-specific standard deviation estimate, thus providing a regularized version of the t-statistic.
Baldi and Long [34] suggested to replace the within-group empirical variance s_k^2 of gene k in the two-sample t-statistic obtained from d observations by an expression of the form

$$\tilde{\sigma}_k^2 = \frac{\nu_0 \sigma_0^2 + (d-1) s_k^2}{\nu_0 + d - 2}.$$

This variance estimate results as the posterior mean from a Bayesian hierarchical model for the measurements of each gene under an experimental condition. The measured values are assumed to be normally distributed, and ν_0 and σ_0 are hyperparameters of the prior for the parameters of the normal distribution. For practical purposes, the authors recommended to choose σ_0 as the empirical standard deviation obtained from averaging over all genes within a certain intensity range. If a variance stabilizing transformation has been applied to the data, σ_0 may be obtained from the pooled variance over all genes on the array. The value ν_0 is chosen as an integer determining the weight of σ_0 compared to the gene-specific standard deviation. Thus the large number of genes interrogated is exploited to obtain potentially biased, but more stable variance estimates for each single gene. The resulting regularized t-statistic, used with a t-distribution with $\nu_0 + d - 2$ degrees of freedom as null distribution, is shown to perform better than the standard t-test on real and simulated data when there are less than about 5 replications per condition. A similar approach was pursued by Lönnstedt and Speed [35]. Tusher et al. [36, 37] also proposed to use a regularized version of the t-statistic, where the empirical standard deviation s_k of gene k is replaced by $\tilde{\sigma}_k = s_k + s_0$, with s_0 determined from the data in a heuristic fashion.

4.2 Multiple testing

Assume that for each gene a statistical test for differential expression has been conducted. If one fixes a gene-wise significance level of e.g. $\alpha = 0.05$, on average one in every 20 genes that are

actually not differentially expressed will show a p-value below α just by chance. Due to the large number of genes represented on a microarray, this may lead to a large number of false positive calls. For this reason, Dudoit et al. [10] suggested to choose a procedure that controls the *family–wise error rate* (FWER). The FWER is defined as the probability that the selected set of genes contains at least one false positive. A multiple testing procedure is said to provide *strong control* of the FWER if it controls the FWER for any combination of true and false null hypotheses. If p-values for the test statistics $T_1, \ldots, T_n$ of n genes are available, a simple adjustment that gives strong control of the FWER is the Bonferroni correction, which amounts to multiplying the unadjusted p-values by n. Dudoit et al. [10] described the use of a step-wise p-value adjustment that is due to Westfall and Young [38]. This procedure is less conservative than the Bonferroni correction, and in contrast to the latter, it takes possible dependences between the test statistics into account. The adjusted p-values are estimated by a permutation algorithm.
For many applications however, control of the FWER is too conservative, with the danger of many interesting genes being missed. As microarrays are often used to screen for candidate genes that may then be validated through further experiments, the researcher may be willing to accept a certain fraction of false positives. This demand is addressed by the concept of the *False Discovery Rate* (FDR, [39]). For a family of hypothesis tests, let R denote the number of rejected null hypotheses, and V the number of falsely rejected null hypotheses. The FDR is defined as

$$FDR = \mathbf{E}[\left.\frac{V}{R}\right| R > 0] \cdot \mathbf{Pr}(R > 0).$$

Benjamini and Hochberg described a procedure to control the FDR under the assumption that the test statistics arising from the true null hypotheses are independent. More precisely, given the set of p-values from all individual hypothesis tests and a desired upper bound q for the FDR, they give a bound p^* such that rejecting all null hypotheses with p-value smaller than p^* guarantees an FDR of at most q for any possible combination of true and false null hypotheses.
Another approach based on the FDR was presented by Storey and Tibshirani [40], see also [36]. For a given rejection region of the statistical tests, the authors estimated the FDR and the *positive False Discovery Rate* (pFDR), which is defined as[3]

$$pFDR = \mathbf{E}[\left.\frac{V}{R}\right| R > 0].$$

Rather than computing a rejection region that guarantees an upper bound for the FDR, Storey and Tibshirani assumed that a rejection region was fixed and estimated the FDR on the basis of the distribution of the test statistics. The estimation procedure has been designed for any kind of dependence between the test statistics and does not require p-values for the single hypothesis tests. The algorithm of Storey and Tibshirani works as follows. They assumed that all null hypotheses were identical and that the same rejection region Γ was used for all test statistics $T_1, \ldots, T_n$, leading to a number $R(\Gamma)$ of rejections. Furthermore, they assumed that the joint

[3] In contrast to the method of Benjamini and Hochberg, where arbitrary, but fixed combinations of true and false null hypotheses are allowed, here the null hypotheses are considered as i.i.d. Bernoulli random variables that are true with probability π_0.

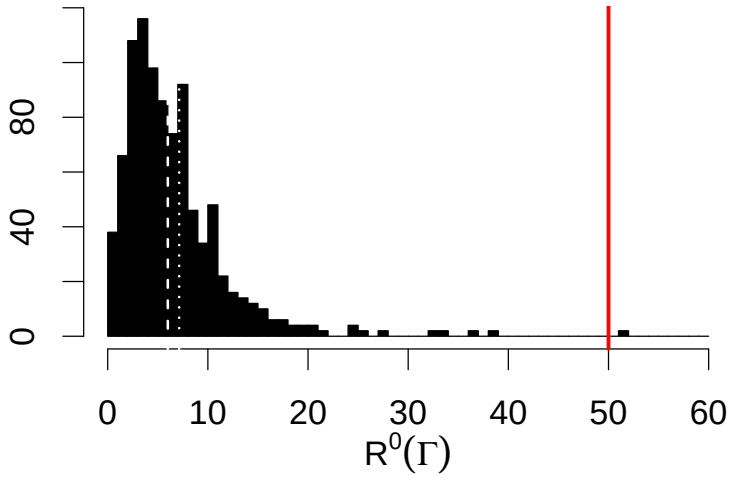


Figure 11: Estimation of the false discovery rate: Using 24 arrays with 32,000 cDNA probes each, 12 pairs of matched breast cancer tissues dissected before and after neoadjuvant chemotherapy were compared. Differentially transcribed genes were selected according to the absolute value of the one-sample t-statistic. The rejection region Γ was fixed such that 50 genes were selected (solid line). The histogram shows the distribution of $R^0(\Gamma)$, estimated from all 924 balanced sign flips. The dashed and dotted lines show median and mean, respectively. The mean may be used as an estimate for $\mathbf{E}[R^0(\Gamma)]$ in Eqns. (17) and (18). Note the skewness of the distribution of $R^0(\Gamma)$.

null distribution of the test statistics could be simulated by permutations of the sample labels. From this, the authors obtained estimates for the expected number of rejections given that all null hypotheses are true,

$$\widehat{\mathbf{E}}[R^0(\Gamma)],$$

as well as for the probability of at least one rejection,

$$\widehat{\mathbf{Pr}}[R^0(\Gamma) > 0].$$

The pFDR is then estimated by

$$\widehat{\text{pFDR}}(\Gamma) = \frac{\hat{\pi}_0 \cdot \widehat{\mathbf{E}}[R^0(\Gamma)]}{\widehat{\mathbf{Pr}}[R^0(\Gamma) > 0] \cdot \max(R(\Gamma), 1))}, \tag{17}$$

and similarly for the FDR:

$$\widehat{\text{FDR}}(\Gamma) = \frac{\hat{\pi}_0 \cdot \widehat{\mathbf{E}}[R^0(\Gamma)]}{\max(R(\Gamma), 1)}. \tag{18}$$

The expected proportion $\hat{\pi}_0$ of true null hypotheses is estimated as follows. Let Γ' be a rejection region whose complement is likely to be achieved mostly for true null hypotheses. The estimate for π_0 is obtained as

$$\hat{\pi}_0 = \frac{n - R(\Gamma')}{\hat{\mathbf{E}}[n - R^0(\Gamma')]}.$$

In order to determine how many falsely significant genes may appear with a certain probability, or how likely it is that *all* genes with test statistics in the rejection region are false positives, it is interesting to estimate not only the pFDR, but also quantiles of the distribution of V/R. This is illustrated in Fig. 11.

Under certain conditions on the dependence structure between the test statistics, it was shown in [40] that for all π_0, the estimates are greater or equal than the true values of the FDR and pFDR in expectation. In [41] (see also [42]), it is shown that in the case of independent test statistics (and asymptotically also for some forms of dependence) the pFDR can be interpreted in a Bayesian framework as the posterior probability that a gene is not differentially expressed, given its test statistic lies in the rejection region:

$$\text{pFDR}(\Gamma) = \mathbf{Pr}(H = 0 | T \in \Gamma).$$

A special property of the approach of Storey and Tibshirani and Tusher et al. is how it makes use of the assumption that the null distributions of the test statistics are identical for all genes: The fact that the estimation procedure is based on the test statistics of *all* genes under permutation of the sample labels gives accurate estimates already for relatively few replicate experiments, while at the same time it preserves the dependence structure between genes. On the other hand, this type of procedure is not able to take possibly unequal variances in the two classes into account.

5 Pattern discovery

Unsupervised as well as supervised learning methods play a central role in the analysis of microarray gene expression data. Supervised methods aim at inferring information from the data with respect to a pre-defined response variable. For instance, in the context of tumor diagnostics one tries to classify mRNA samples obtained from tumor cells with respect to given tumor types. The application of classification methods to microarray data was discussed e. g. in [43, 44, 45, 46]. In the following, we focus on unsupervised methods, which aim at detecting structures in the data without making use of gene or sample annotations. A primary purpose of such methods is to provide a visualization of the data in which conspicuous structures can easily be recognized. These may be relations among genes, among samples, or between genes and samples. The perception of such structures can lead the researcher to develop new hypotheses: e.g. the result of a clustering of genes may indicate the putative involvement of uncharacterized genes in a biological process of interest, whereas a separation of the expression profiles of a set of patient tissue samples into clusters may point to a possible refinement of disease taxonomy. On the other hand, unsupervised methods are often used to confirm known differences between genes or samples on the level of gene expression: If a clustering algorithm groups samples from e.g. two different tumor types into distinct clusters without using prior knowledge, this provides evidence that the tumor types do indeed show clearly detectable differences in their global gene expression profiles.
For all of the following methods, we assume that we have a gene expression data matrix of suitably calibrated and transformed expression levels with, say, the rows corresponding to genes and the columns corresponding to cell or tissue samples.

5.1 Projection methods

An important class of unsupervised methods works through dimension reduction. The row or column vectors of a gene expression data matrix are projected onto a low-dimensional space such that some measure of similarity between the vectors is optimally preserved. The projected data may be visualised through one or more scatterplots, in the hope that these convey important information contained in the data.
In *principal component analysis*, mutually orthogonal linear combinations of the row or column vectors (the principal components) are computed, such that the ith principal component has maximal variance among all vectors orthogonal to the first $i-1$ principal components. In applications, one may hope that the first few principal components carry most of the information contained in the data, which can then be displayed in scatterplots. Alter et al. [47] demonstrated the use of principal component analysis for a gene expression study of the cell cycle in yeast. The first principal component was found to reflect experimental artifacts and was consequently filtered out. After that, the authors found that the first two principal components ("eigengenes") are well described by a sine and cosine function of time, respectively. The interpretation is that these "eigengenes" reflect oscillating gene expression patterns, while the corresponding "eigenarrays" define a two–dimensional coordinate system for the cell cycle phases.

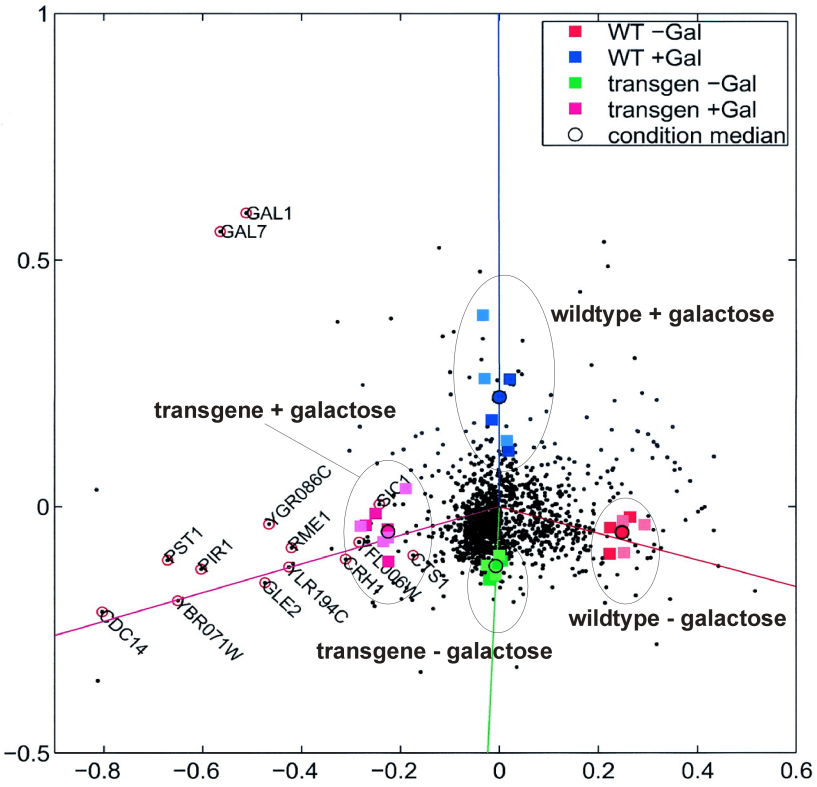


Figure 12: Correspondence analysis applied to an experiment that searched for genes expressed as a consequence of induction of the yeast cell cycle gene CDC14 [48]. A yeast transgene was constructed with the CDC14 gene under a galactose dependent promotor that allows induction of the CDC14 gene through the addition of galactose. As a consequence, one observes upregulation of genes both due to CDC14 induction and due to the natural reaction to galactose. Four conditions were studied: wildtype yeast with and without galactose, and the yeast transgene with and without galactose. For each condition several replicates were made. The correspondence analysis biplot shows an embedding of rows and columns of the entire data matrix with genes depicted as black dots and hybridizations depicted as small squares. Replicates for each condition cluster together, and each of the 4 clusters defines a direction in which the genes that are typical for the condition can be found. The bisection between the two galactose conditions points to two GAL genes, known to be involved in the galactose pathway. Genes in the transgene+galactose condition that are turned on in response to the addition of galactose are attracted also by the wildtype+galactose condition. Thus, the lower left direction highlights genes that are exclusively due to the CDC14 induction. Genes are encircled which show up in a related experiment [49], too, where they are also seen to be linked to CDC14.

In *correspondence analysis*, the rows and the columns of a non-negative data matrix are simultaneously projected onto a low-dimensional space [50]. The method decomposes the deviations from homogeneity between rows and columns, as we will explain now. For the data matrix $\boldsymbol{Y}$, let y_{k+} and y_{+i} be the sum of the kth row and the ith column, respectively, y_{++} the grand total, and $r_k = y_{k+}/y_{++}$ and $c_i = y_{+i}/y_{++}$ the mass of the kth row and the ith column. The matrix $\boldsymbol{S}$ with elements

$$s_{ki} = (y_{ki}/y_{++} - r_k c_i)/\sqrt{r_k c_i}$$

is submitted to singular value decomposition, $\boldsymbol{S} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T$. Note that the sum of the squared elements of $\boldsymbol{S}$ is just the χ^2–statistic of $\boldsymbol{Y}$. Suitably normalized columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ provide the principal coordinates for the rows and the columns of $\boldsymbol{Y}$, respectively:

$$f_{kj} = \lambda_j u_{kj}/\sqrt{r_k}, \qquad g_{ij} = \lambda_j v_{ij}/\sqrt{c_i}.$$

Usually, only the first two or three principal coordinates are used for a simultaneous display of the rows and columns of the data matrix. The corresponding entries of $\boldsymbol{\Lambda}$ reflect which proportion of the χ^2-statistic of $\boldsymbol{Y}$ is represented in the low-dimensional projection. The distances between rows and between columns approximate their χ^2-distances, and moreover, the association between rows and columns is reflected in a biplot: A row and a column that are positively (negatively) associated will approximately lie on the same (opposite) half-ray through the origin, with the distance from the origin reflecting the strength of the association. An example is shown in Fig. 12.

5.2 Cluster algorithms

Cluster algorithms generally aim at grouping objects according to some notion of similarity. An overview of questions and methods in cluster analysis is given in [51]. For microarray data, clustering may be applied to the genes whose expression levels are measured, with the expectation that functionally related or co–regulated genes will show similar expression patterns. On the other hand, one may use clustering to analyse the expression profiles of a set of cell or tissue samples with the hope that samples with similar biological characteristics will be grouped together. Cluster algorithms are explicitly or implicitly based on a quantitative measure of dissimilarity between the objects of interest. In the case of row and column vectors of a gene expression data matrix, typical examples are the Euclidean distance or 1 minus the correlation coefficient. For the clustering of genes based on their expression patterns, the latter is often used because it is invariant under affine–linear transformations of the input vectors and focuses on the pattern of relative changes.

In *hierarchical* clustering algorithms, a tree structure (dendrogram) is computed that contains the objects as leaves. *Agglomerative* methods build the tree starting with the leaves ([52]), whereas in *divisive* methods ([53]), the set of objects is iteratively partitioned into subsets. To obtain a partition of the set of objects into clusters, the resulting dendrogram may be cut at a certain height. In the analysis of microarray data however, hierarchical clustering is often simply used to obtain a linear ordering of both the rows and the columns of a gene expression data matrix such that similar rows or columns are located close to each other. The reordered data matrix is

then displayed using a color map, which may be a powerful visualization tool. However, many implementations of hierarchical clustering do not try to find an optimal linear order of the n leaves of the obtained dendrogram out of the 2^{n-1} possible orders that are compatible with the tree structure. Bar–Joseph et al. [54] described an efficient algorithm to compute an optimal linear order of the set of leaves of a cluster tree compatible with the tree structure in the sense that the sum of distances of pairs of neighbour leaves is minimized.

Non–hierarchical clustering algorithms directly yield a partition of the set of objects into clusters, the number of which has to be fixed in advance in most methods. Examples are k–means clustering, partitioning around medoids [55] and self-organizing maps [56]. A graph–theoretical clustering method that was developed especially for gene expression data is described in [57].

An important, but difficult question in cluster analysis is that of the validity of the results. Cluster analysis assumes that the data are organized in distinguishable clusters. However, a cluster algorithm will usually also produce a set of clusters if this assumption is not fulfilled. Furthermore, the results may be affected by random fluctuations of the data. Thus, it is of interest to estimate the number of clusters present in a data set (if any), as well as to assess the variability of various features of the result.

Dudoit and Fridlyand [55] proposed an approach for estimating the number of clusters in a data set. In the first step, they applied a clustering algorithm to a subset of all observations. Then they analysed how much an assignment of the remaining observations to the clusters by a class prediction method coincided with the partition obtained from clustering these observations. Through comparing the resulting measure of predictability to that obtained under a null model without cluster structure, they arrived at a quality index that can be computed for different numbers of clusters. If none of the predictability values is significant, there is no evidence for clusters in the data, whereas otherwise the number of clusters yielding the highest quality index is chosen. The performance of this approach was demonstrated using simulated data and real gene expression data, where clustering is applied to the samples. However, one might imagine that real data sometimes lie in–between the extreme cases of a common distribution for all objects on one side and distinct clusters on the other side. Furthermore, if the objects are organized in hierarchically nested clusters, such that there are partitions at different levels of granularity, the question of how many clusters are present is not meaningful without further specifications.

Kerr and Churchill [58] used the bootstrap in order to assess the reliability of the results of a cluster analysis. Resampling was performed on the residuals of an analysis of variance model, yielding stability values for the assignment of genes to pre-specified clusters. In a more general context, Pollard and van der Laan [59] analysed the performance of the bootstrap in assessing the variability of the results of a wide class of clustering methods. Assuming that the expression profiles of the biological samples are generated from a mixture of n-variate probability distributions, where n denotes the number of genes, they gave a general definition of clustering methods (for clustering genes or samples, or the simultaneous clustering of genes and samples) as algorithms to estimate certain parameters of the data-generating distribution. Although an underlying probabilistic model is assumed, the cluster algorithms under consideration do not have to be model-based. This framework allows to apply concepts of statistical inference to clustering algorithms. In a simulation study, they analysed both the non-parametric (resampling of sample expression profiles) and the parametric bootstrap (based on normal distributions). The results

indicated that both bootstrapping methods are able to assess the variability of various quantities describing the output of a clustering algorithm. In addition, the authors proposed to test the statistical significance of a clustering of samples via the comparison with data generated from an appropriate null model. A similar approach for the evaluation of temporal gene expression patterns was presented in [60]. Assuming the expression pattern of each gene to be generated from one of several multivariate normal distributions, the authors analysed the error rate of various clustering algorithms in determining the correct cluster membership of genes.
Probabilistic clustering methods assume that each observation belongs to a cluster k with probability π_k, and the observations within each cluster k are generated according to a probability distribution $\mathcal{L}_k$. After the number of clusters and the family of admissible probability distributions have been specified, the model parameters and the most likely cluster assignment of each observation can be estimated by maximum likelihood. This is usually done via the Expectation–Maximization algorithm [61], starting with some initial clustering. In such a probabilistic framework, it is possible to assess the adequacy of different models — concerning the number of clusters, as well as the allowed parameter space for the component distributions — through the Bayesian Information Criterion [62]. The application of model–based clustering to microarray data is described in [63, 64, 65]. Concerning the clustering of genes, the application of normal mixture models, possibly with constraints on the covariance matrices in order to reduce the number of free parameters, is more or less straightforward. On the other hand, the application of model–based methods to the clustering of samples poses problems, because in typical microarray data sets, the number of genes, and thus the number of parameters to be estimated, exceeds by far the number of samples. Ghosh and Chinnaiyan suggested to cluster the samples via a model–based approach using the first few components obtained from a principal component analysis. McLachlan et al. proposed to cluster the samples into a mixture of factor analysis models.

5.3 Local pattern discovery methods

One limitation of clustering methods as described in Section 5.2 lies in the fact that they are based on a global measure of similarity between the rows or the columns of the data matrix. However, there may be biologically relevant situations where tissue samples share similar expression levels of one particular set of genes, e. g. those belonging to a molecular pathway that is active in this group, whereas they differ with respect to the expression of other genes. Also the similarity of the expression levels of a group of genes may be present only under certain biological conditions. We give a brief overview on methods that were developed with the aim of detecting such structures in an unsupervised fashion.
Getz et al. [66] described an algorithm where hierarchical clustering is alternately applied to the rows and columns of submatrices of the data matrix, the rows and columns of which were obtained as stable clusters in previous iterations.
A number of authors have suggested methods to identify interesting submatrices of a gene expression data matrix [67, 68, 69, 70]. The underlying idea is that a set of genes, perhaps belonging to a common molecular pathway, are co–regulated only under certain experimental conditions. This notion is quantified in terms of a score function on submatrices of a gene expression matrix. As the number of submatrices is exponential in the number of genes and the number of

samples, efficient heuristics are applied in order to find high–scoring submatrices. The identified submatrices can be evaluated in terms of their statistical significance.
For the identification of conspicuous class distinctions among a set of tissue samples based on microarray data, special approaches have been proposed [71, 72]. They are based on a score function that quantifies the strength of differential gene expression for any possible bipartition of the set of samples. An optimization algorithm is used in order to find high-scoring bipartitions. As the scoring is not based on global properties of the gene expression profiles, but rather on the presence of subsets of genes that are differentially expressed, several independent bipartitions can be obtained, each being based on a specific subset of differentially expressed genes. In [72], it was shown that this approach is able to detect biologically meaningful class distinctions that are not identified with cluster algorithms based on a global dissimilarity measure.

6 Conclusion

We have described different aspects of microarray gene expression data analysis, from the quality control of the raw probe intensity data, via calibration, error modeling, and the identification of differentially transcribed genes to explorative methods such as clustering or pattern discovery. Yet, statistical analysis is only one part of a microarray experiment. Frequently, data analysis is expected to correct for technological problems or shortcomings in the design of an experiment. Awareness has grown, though, in recent years and interaction between experimentalists and data analysts is improving and, very importantly, starting already early in the planning of an experiment. This raises the hope that statistical analysis will in the future be ever less diverted by trouble shooting, and can increase its focus on generating and validating biological hypotheses from the data.
Unlike with, e.g., sequence data, it is still extremely difficult to relate different experiments to each other and to quantitatively compare their results. Much stricter standardization of the measurement process, which will have to include significant improvements of present technologies or development of new ones, will be necessary to obtain measurements that would be comparable across laboratories. As a result, currently each experiment has to be large enough to be analyzable by itself because it is still not feasible to view one's own experiment as an incremental addition to an existing knowledge base on gene expression. While this is both a technical problem and a data integration problem, suggestions regarding the data standardization aspect are given in [73].
Microarray based experiments are frequently seen as the stronghold of hypothesis-free genome research. While debatable in itself, this assertion simply shifts the responsibility to the computational scientist analyzing the data. In the absence of a clear hypothesis much of the analysis will be of an exploratory nature. Once this leads to a hypothesis further independent verification is needed. This embeds microarray experiments and statistical analysis into a feedback cycle producing new experiments.

Acknowledgments

We thank Holger Sültmann, Anke Schroth, Jörg Schneider from the Department of Molecular Genome Analysis at the DKFZ Heidelberg for sharing the experimental data and for continuing stimulating collaboration, and Annemarie Poustka for providing support. We are grateful to Günther Sawitzki, Dirk Buschmann, and Andreas Buneß for highly fruitful discussions. Kurt Fellenberg kindly provided Fig. 12.

References

[1] G.G. Lennon and H. Lehrach. Hybridization analyses of arrayed cDNA libraries. *Trends in Genetics*, 10:314–317, 1991.

[2] R.J. Lipshutz, S.P. Fodor, T.R. Gingeras, and D.J. Lockhart. High density synthetic oligonucleotide arrays. *Nature Genetics*, 21 (Suppl 1):20–24, 1999.

[3] David J. Duggan, Michael Bittner, Yidong Chen, Paul Meltzer, and Jeffrey M. Trent. Expression profiling using cDNA microarrays. *Nature Genetics*, 21 (Suppl 1):10–14, 1999.

[4] M. Kathleen Kerr and Gary A. Churchill. Statistical design and the analysis of gene expression microarray data. *Genet. Res.*, 77:123–128, 2001.

[5] G.A. Churchill. Fundamentals of experimental design for cDNA microarrays. *Nature Genetics*, 32 Suppl. 2:490–495, 2002.

[6] Yee Hwa Yang and Terence P. Speed. Design issues for cDNA microarray experiments. *Nat. Rev. Gen.*, 3:579–588, 2002.

[7] *The chipping forecast. Special supplement to Nature Genetics*, volume 21, 1999.

[8] Yee Hwa Yang, M. J. Buckley, Sandrine Dudoit, and Terence P. Speed. Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics*, 11:108–136, 2002.

[9] Johannes Schuchhardt, Dieter Beule, Arif Malik, Eryc Wolski, Holger Eickhoff, Hans Lehrach, and Hanspeter Herzel. Normalization strategies for cDNA microarrays. *Nucleic Acids Research*, 28:e47, 2000.

[10] Sandrine Dudoit, Yee Hwa Yang, Terence P. Speed, and Matthew J. Callow. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12:111–139, 2002.

[11] W.S. Cleveland, E. Grosse, and W.M. Shyu. *Statistical Models in S*, chapter Chapter 8: Local regression models. Wadsworth & Brooks, Cole, 1992.

[12] Huibin Yue, P. Scott Eastman, Bruce B. Wang, James Minor, Michael H. Doctolero, Rachel L. Nuttall, Robert Stack, John W. Becker, Julie R. Montgomery, Marina Vainer, and Rick Johnston. An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression. *Nucleic Acids Research*, 29(8):e41, 1–9, 2001.

[13] T. Ideker, V. Thorsson, A.F. Siegel, and L.E. Hood. Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *Journal of Computational Biology*, 7:805–818, 2000.

[14] Latha Ramdas, Kevin R. Coombes, Keith Baggerly, Lynne Abruzzo, W. Edward Highsmith, Tammy Krogmann, Stanley R. Hamilton, and Wei Zhang. Sources of nonlinearity in cDNA microarray expression measurements. *Genome Biology*, 2:research0047.1–0047.7, 2001.

[15] Rafael A. Irizarry, Bridget Hobbs, Francois Collin, Yasmin D. Beazer-Barclay, Kristen J. Antonellis, Uwe Scherf, and Terence P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 2003. Accepted for publication. http://biosun01.biostat.jhsph.edu/∼ririzarr/papers/index.html.

[16] Yidong Chen, Edward R. Dougherty, and Michael L. Bittner. Ratio-based decisions and the quantitave analysis of cDNA microarray images. *Journal of Biomedical Optics*, 2:364–374, 1997.

[17] M. Kathleen Kerr, Mitchell Martin, and Gary A. Churchill. Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, 7:819–837, 2000.

[18] Tim Beißbarth, Kurt Fellenberg, Benedikt Brors, Rose Arribas-Prat, Judith Maria Boer, Nicole C. Hauser, Marcel Scheideler, Jörg D. Hoheisel, Günther Schütz, Annemarie Poustka, and Martin Vingron. Processing and quality control of DNA array hybridization data. *Bioinformatics*, 16:1014–1022, 2000.

[19] M.A. Newton, C.M. Kendziorski, C.S. Richmond, F.R. Blattner, and K.W. Tsui. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, 8(1):37–52, 2001.

[20] Thomas B. Kepler, Lynn Crosby, and Kevin T. Morgan. Normalization and analysis of DNA microarray data by self-consistency and local regression. *Genome Biology*, 3(7):research0037.1–0037.12, 2002.

[21] Joachim Theilhaber, Steven Bushnell, Amanda Jackson, and Rainer Fuchs. Bayesian estimation of fold-changes in the analysis of gene expression: The PFOLD algorithm. *Journal of Computational Biology*, 8:585–614, 2001.

[22] Keith A. Baggerly, Kevin R. Coombes, Kenneth R. Hess, David N. Stivers, Lynne V. Abruzzo, and Wei Zhang. Identifying differentially expressed genes in cDNA microarray experiments. *Journal of Computational Biology*, 8:639–659, 2001.

[23] Jeffrey G. Thomas, James M.Olson, Stephen J. Tapscott, and Lue Ping Zhao. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Research*, 11:1227–1236, 2001.

[24] Wolfgang Huber, Anja von Heydebreck, Holger Sültmann, Annemarie Poustka, and Martin Vingron. Variance stablization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl. 1:S96–S104, 2002. ISMB 2002.

[25] Peter J. Rousseuw and Annick M. Leroy. *Robust Regression and Outlier Detection*. John Wiley and Sons, 1987.

[26] David M. Rocke and Blythe Durbin. A model for measurement error for gene expression analysis. *Journal of Computational Biology*, 8:557–569, 2001.

[27] R.O. Dror, J.G. Murnick, N.J. Rinaldi, V.D. Marinescu, R.M. Rifkin, and R.A. Young. A Bayesian approach to transcript estimation from gene array data: the BEAM technique. In *Proceedings of the Sixth Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, April 2002.

[28] Blythe Durbin, Johanna Hardin, Douglas Hawkins, and David Rocke. A variance-stabilizing transformation from gene-expression microarray data. *Bioinformatics*, ISMB, 2002.

[29] J.M. Claverie. Computational methods for the identification of differential and coordinated gene expression. *Human Molecular Genetics*, 8:1821–1832, 1999.

[30] B.L. Welch. The generalization of Student's problem when several different population variances are involved. *Biometrika*, 34:28–35, 1947.

[31] D.I. Best and C.W. Rayner. Welch's approximate solution for the Behrens–Fisher problem. *Technometrics*, 29:205–220, 1987.

[32] R. Herwig, P. Aanstad, M. Clark, and H. Lehrach. Statistical evaluation of differential expression on cDNA nylon arrays with replicated experiments. *Nucleic Acids Research*, 29:E117, 2001.

[33] W. Pan. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 18:546–554, 2002.

[34] Pierre Baldi and Anthony D. Long. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17:509–519, June 2001.

[35] Ingrid Lönnstedt and Terence P. Speed. Replicated microarray data. Technical report, Department of Statistics, UC Berkeley, http://www.stat.Berkeley.EDU/users/terry/zarray/Html/papersindex.html, 2001. Accepted to Statistica Sinica.

[36] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*, 98:5116–5121, 2001.

[37] Bradley Efron, Robert Tibshirani, Virginia Goss, and Gilbert Chu. Microarrays and their use in a comparative experiment. Technical report, Stanford University, http://www-stat.stanford.edu/∼tibs/research.html, October 2000.

[38] P.H. Westfall and S.S. Young. *Resampling-based multiple testing: examples and methods for p-value adjustment*. John Wiley and Sons, 1993.

[39] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57:289–300, 1995.

[40] J.D. Storey and R. Tibshirani. Estimating false discovery rates under dependence, with applications to DNA microarrays. Technical report, Department of Statistics, Stanford University, http://www.stat.berkeley.edu/∼storey/, 2001.

[41] J.D. Storey. The positive false discovery rate: A Bayesian interpretation and the q-value. Technical report, Department of Statistics, Stanford University, http://www.stat.berkeley.edu/∼storey/, 2001.

[42] B. Efron, R. Tibshirani, J.D. Storey, and V. Tusher. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96:1151–1160, 2001.

[43] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[44] Sandrine Dudoit, Jane Fridlyand, and Terence P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97:77–87, 2002.

[45] R. Spang, C. Blanchette, H. Zuzan, J. R. Marks, J. Nevins, and M. West. Prediction and uncertainty in the analysis of gene expression profiles. In E. Wingender, R. Hofestädt, and I. Liebich, editors, *Proceedings of the German Conference on Bioinformatics GCB 2001*, Braunschweig, 2001.

[46] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. Tissue classification with gene expression profiles. *Journal of Computational Biology*, 7:559–583, 2000.

[47] O. Alter, P. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA*, 97:10101–10106, 2000.

[48] Kurt Fellenberg, Nicole C Hauser, Benedikt Brors, A Neutzner, Jörg Hoheisel, and Martin Vingron. Correspondence analysis applied to microarray data. *Proc. Natl. Acad. Sci. USA*, 98(19):10781–10786, 2001.

[49] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, 9:3273–97, 1998.

[50] M.J. Greenacre. *Theory and applications of correspondence analysis*. Academic Press, London, 1984.

[51] A.K. Jain and R.C. Dubes. *Algorithms for clustering data*. Prentice-Hall, 1988.

[52] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–14868, 1998.

[53] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, 96:6745–6750, 1999.

[54] Z. Bar-Joseph, D.K. Gifford, and T.S. Jaakkola. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, Suppl. 1:S22–29, 2001. ISMB 2001.

[55] S. Dudoit and J. Fridlyand. A prediction–based resampling method for estimating the number of clusters in a dataset. *Genome Biology*, 3(7):research0036.1–21, 2002.

[56] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, and T.R. Golub. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA*, 96:2907–2912, 1999.

[57] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 6:281–297, 1999.

[58] M. Kathleen Kerr and Gary A. Churchill. Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *Proc. Natl. Acad. Sci. USA*, 98:8961–8965, 2001.

[59] K.S. Pollard and M.J. van der Laan. Statistical inference for simultaneous clustering of gene expression data. *Mathematical Biosciences*, 176:99–121, 2002.

[60] E.R. Dougherty, J. Barrera, M. Brun, S. Kim, R.M. Cesar, Y. Chen, M. Bittner, and J.M. Trent. Inference from clustering with application to gene-expression microarrays. *Journal of Computational Biology*, 9:105–126, 2002.

[61] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B*, 39:1–38, 1977.

[62] C. Fraley and A.E. Raftery. How many clusters? Which clustering method? Answers via model–based cluster analysis. *The Computer Journal*, 41:578–588, 1998.

[63] D. Ghosh and A.M. Chinnaiyan. Mixture modelling of gene expression data from microarray experiments. *Bioinformatics*, 18:275–286, 2002.

[64] G.J. McLachlan, R.W. Bean, and D. Peel. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, 18:413–422, 2002.

[65] K.Y. Yeung, C. Fraley, A. Murua, A.E. Raftery, and W.L. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17:977–987, 2001.

[66] G. Getz, E. Levine, and E. Domany. Coupled two-way clustering analysis of gene microarray data. *Proc. Natl. Acad. Sci. USA*, 97:12079–12084, 2000.

[67] Y. Chen and G.M. Church. Biclustering of expression data. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB)*, pages 93–103. AAAI Press, 2000.

[68] A. Califano, G. Stolovitzky, and Y. Tu. Analysis of gene expression microarrays for phenotype classification. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB)*, pages 75–85. AAAI Press, 2000.

[69] A. Ben-Dor, Benny Chor, Richard Karp, and Z. Yakhini. Discovering local structure in gene expression data: The order–preserving submatrix problem. In *Proceedings of the Fifth Annual International Conference on Computational Molecular Biology (RECOMB)*, pages 49–57. ACM Press, 2002.

[70] L. Lazzeroni and A. Owen. Plaid models for gene expression data. *Statistica Sinica*, 12:61–86, 2002.

[71] A. Ben-Dor, N. Friedman, and Z. Yakhini. Class discovery in gene expression data. In *Proceedings of the Fifth Annual International Conference on Computational Molecular Biology (RECOMB)*, pages 31–38. ACM Press, 2001.

[72] Anja von Heydebreck, Wolfgang Huber, Annemarie Poustka, and Martin Vingron. Identifying splits with clear separation: A new class discovery method for gene expression data. *Bioinformatics*, 17 Suppl. 1:S107–114, 2001. ISMB 2001.

[73] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron. Minimum information about a microarray experiment (MIAME) - toward standards for microarray data. *Nat. Genet.*, 29(4):365–371, 2001.