Low-level analysis of microarray experiments

Wolfgang Huber[1]*, Anja von Heydebreck[2] and Martin Vingron[3]

[1] European Bioinformatics Institute
European Molecular Biology Laboratory, Cambridge CB1 10SD, UK
Tel. +44 1223 494642, Fax Tel. +44 1223 494486, huber@ebi.ac.uk

[2] Department of Bio- and Chemoinformatics, Merck KGaA
Darmstadt, Germany
Tel. +49 6151 723235, Fax +49 6151 723329, Anja.von.Heydebreck@merck.de

[3] Max-Planck-Institute for Molecular Genetics, Department of
Computational Molecular Biology, 14195 Berlin, Germany
Tel. +49 30 8413 1168, Fax +49 30 8413 1152, vingron@molgen.mpg.de

Contents

*to whom correspondence should be addressed

1 Introduction

This article gives an overview over the methods used in the low–level analysis of gene expression data generated using DNA microarrays. This type of experiment allows to determine relative levels of nucleic acid abundance in a set of tissues or cell populations for thousands of transcripts or loci simultaneously. Careful statistical design and analysis are essential to improve the efficiency and reliability of microarray experiments throughout the data acquisition and analysis process. This includes the design of probes, the experimental design, the image analysis of microarray scanned images, the normalization of fluorescence intensities, the assessment of the quality of microarray data and incorporation of quality information in subsequent analyses, the combination of information across arrays and across sets of experiments, the discovery and recognition of patterns in expression at the single gene and multiple gene levels, and the assessment of significance of these findings, considering the fact that there is a lot of noise and thus random features in the data. For all of these components, access to a flexible and efficient statistical computing environment is an essential aspect.

1.1 Microarray technology

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 nucleic acid levels in particular cells or tissues for many genes or loci at once. Nucleic acids of interest can be polyadenylated RNA, total RNA, or DNA. We will in the following use the term *gene* loosely to denote any unit of nucleic acid of interest. Single strands of complementary DNA for the genes to be considered are immobilized on spots arranged in a grid (”array”) on a support which will typically be a glass slide or a quartz wafer. The number of spots can range from dozens to millions. From a sample of interest, e.g. a tumor biopsy, the nucleic acid is extracted, labeled and hybridized to the array. Measuring the amount of label on each spot then yields an intensity measurement that should be correlated to the abundance of the corresponding gene in the sample. Chapter 25 goes into more detail regarding the experimental technology, therefore we only give a short summary here.

Two schemes of fluorescent labeling are in common use today. One variant labels a single sample. For example, the company Affymetrix synthesizes sets of short oligomers on a glass wafer and uses a single fluorescent label ([26], see also www.affymetrix.com). Alternatively, two samples are labeled with a green and a red fluorescent dye, respectively. The mixture of the two nucleic acid preparations is then hybridized simultaneously to a common array on a glass slide. In the case where the probes are PCR products from cDNA clones that are spotted on the array, this technology is usually refered to as the Stanford technology [12]. On the other hand, companies like Agilent have immobilized long oligomers of 60 to 70 basepairs length and used two-color labeling. The hybridization is quantified by a laser scanner that determines the intensities of each of the two labels over the entire array.

The parallelism in microarray experiments lies in the hybridization of nucleic acids extracted from a single sample to many genes simultaneously. The measured abundances, though, are usually 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.

Traditionally, one or a few probes were selected for each gene, based on known information on its sequence and structure. More recently, it has become possible to produce probes for the complete sequence content of a whole genome or for significant parts of it [3, 6, 32].

1.2 Prerequisites

A number of steps are involved in the generation of the raw data. The *experimental design* includes the choice and collection of samples (tissue biopsies or cell lines exposed to different treatments), the choice of probes and array platform, the choice of controls, RNA extraction, amplification, labeling, and hybridization procedures, the allocation of replicates, and the scheduling of the experiments. Careful planning is needed, as the quality of the experimental design determines to a large extent the utility of the data [7, 23, 42]. A fundamental guideline is the avoidance of *confounding* between different biological factors of interest, or between a biological factor of interest and a technical factor that is anticipated to affect the measurements.

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. In a cohort study, different conditions like healthy/diseased or different disease types may be studied. In designed factorial experiments, one or several factors, for example treatment with a drug, genetic background, and/or tissue type, are varied in a controlled manner. 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 challenging program.

1.3 Preprocessing

Preprocessing is the link between the raw experiment data and the higher-level statistical analysis. The five tasks of preprocessing can be summarized as follows: data import, background adjustment, normalization, summarization of multiple probes per transcript, and quality control.

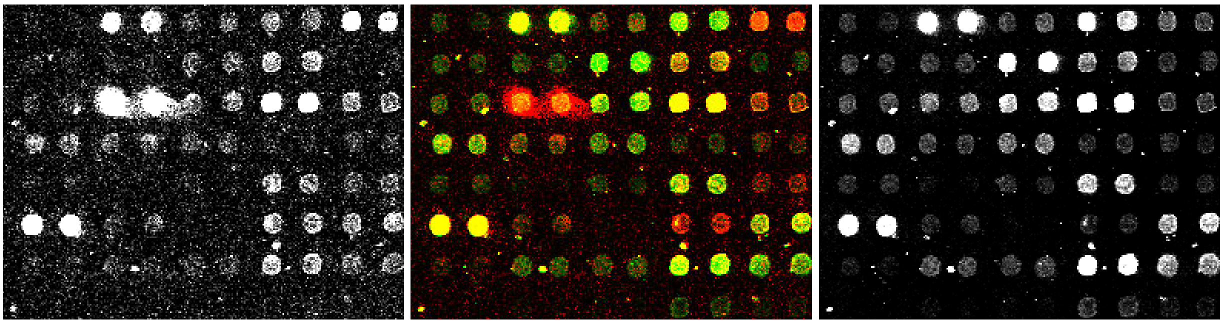


Figure 1: The detected intensity distributions from a cDNA microarray for a region comprising 40 probes spotted in duplicate. The total number of probes on an array may range from a few dozens to tens of thousands. Left panel: grey-scale representation of the detected label fluorescence at 635 nm (red), corresponding to mRNA sample A. Right panel: label fluorescence at 532 nm (green), corresponding to mRNA sample B. 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.

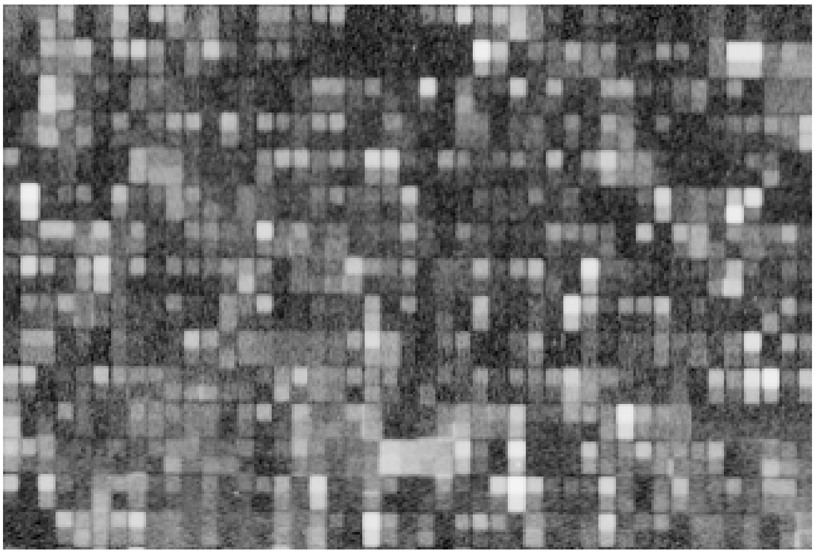


Figure 2: Gray-scale representation of the intensity distribution from a small sector of an Affymetrix HG-U133A genechip.

They are driven by the properties of microarray technology. The data come in different formats, and are often scattered across a number of files (or possibly, database tables), from which they need to be extracted and unified. Part of the hybridization is non-specific and the measured intensities are affected by noise in the optical detection. Therefore, the observed intensities need to be adjusted to give accurate measurements of specific hybridization. We refer to this aspect of pre-processing as *background adjustment*. Different efficiencies of reverse transcription, labeling or hybridization reactions among different arrays cause systematic technical biases and need to be corrected. We call the task of manipulating data to make measurements from different arrays comparable *normalization*. On some platforms, genes are represented with more than one probe. *Summarizing* the data is necessary when we want to reduce the measurements from various probes into one quantity that estimates the amount of RNA transcript. The reproducibility of measurements is limited by random fluctuations or measurement error. Basically, we can distinguish between two types of fluctuations: those that affect individual measurements, and follow a localized distribution; and those that affect whole groups of measurements and are often drastic, large, and irregular. The former type can be described with *error models*, while the latter is best dealt with by *quality control* procedures that try to detect and eliminate the affected measurements.

2 Visualization and exploration of the raw data

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. 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 quartz wafer. The sample molecules are labeled through 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 [12] involves the simultaneous hybridization of two samples labeled with different fluorescent dyes, and detection at the two corresponding wavelengths. Figure 1 shows an example.

2.1 Image analysis

In the *image analysis* step we extract probe intensities out of the scanned images, such as shown in Figures 1 and 2. The images are scanned by the detector at a high spatial resolution, such that each probe 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"

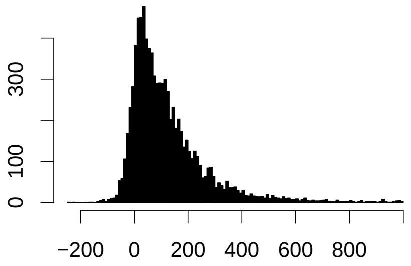


Figure 3: Histogram of probe intensities at the green wavelength for a cDNA microarray similar to the one depicted in Figure 1. The intensities were determined, in arbitrary units, by an image quantification method, and "local background" intensities were subtracted. Due to measurement noise, this lead 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.

intensity, or spot quality measures.

Various software packages offer a variety of segmentation and quantification methods. 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 and how the probes were attached to it (e. g. quill-pen type printing of PCR-product, in situ oligonucleotide synthesis by ink jetting, in situ synthesis by photolithography). 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 is described by Yang et al. [40].

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 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 Figure 3. 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

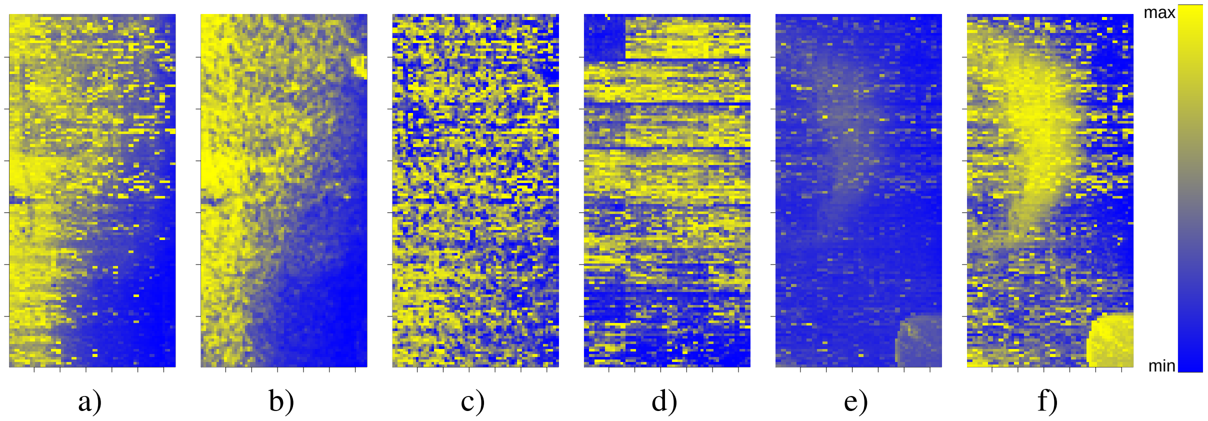


Figure 4: False color representations of the spatial intensity distributions from three different 64×136 spot cDNA microarrays from one experiment series. The color scale is shown in the panel on the right. a) probe intensities in the red color channel, b) local background intensities, c) background-subtracted 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).

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 Figure 4. A more sophisticated and more sensitive method to detect subtle artifacts is to look at the residuals of a probe–level model fitted for a set of arrays instead of the probe intensities themselves [5].

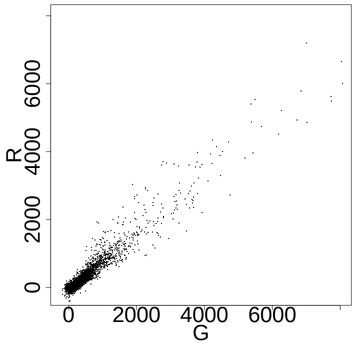


Figure 5: Scatterplot of probe intensities in the red and the green color channel from a cDNA array containing 8000 probes.

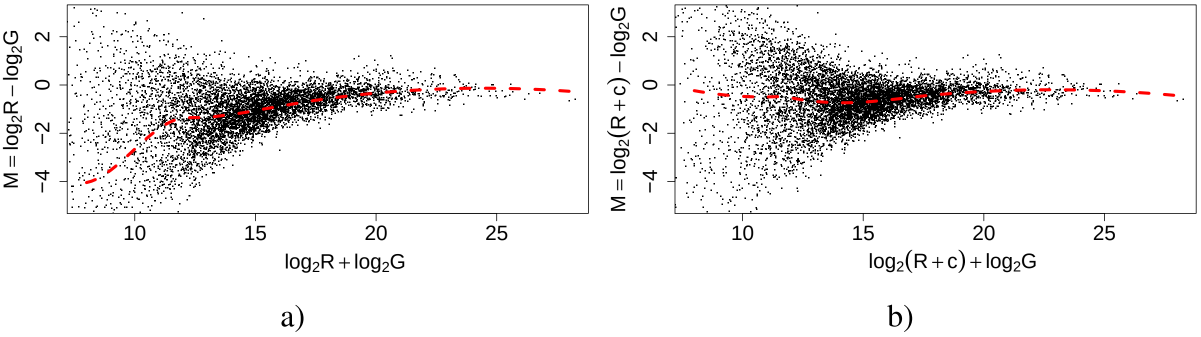
2.3 Scatterplot

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 the same tissue type, possibly subject to different disease conditions. Visually, this can be examined from the scatterplot of the probe intensities for a pair of samples. An example is shown in Figure 5.

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 [33]. 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 Figure 5 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 Figure 6. It is customary to transform to new variables $A = (\log R + \log G)/2$, $M = \log R - \log G$ [11]. Up to a scale factor of $\sqrt{2}$, this corresponds to a 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 in the vertical direction may be considered a measure of the random variation. Figure 6a 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 Figure 6b, the regression line may be modeled sufficiently well by a constant $M_0(A) = M_0$ if an appropriate offset is added to the R values before taking the logarithm. Note that a

[1] We used `loess` [8] with default parameters span=0.75, degree=2.



a) b)

Figure 6: a) the same data as in Figure 5, 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.

horizontal line at $M = M_0$ in Figure 6b corresponds to a straight line of slope 2^{M_0} and with intercept c in Figure 5.

Figure 6 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 Figure 5, 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 5.2.

Two extensions of the scatterplot are shown in Figures 7 and 8. Rather than plotting a symbol for every data point, they use a density representation, which may be useful for larger arrays. For example, Figure 7 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 Figure 8. 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

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

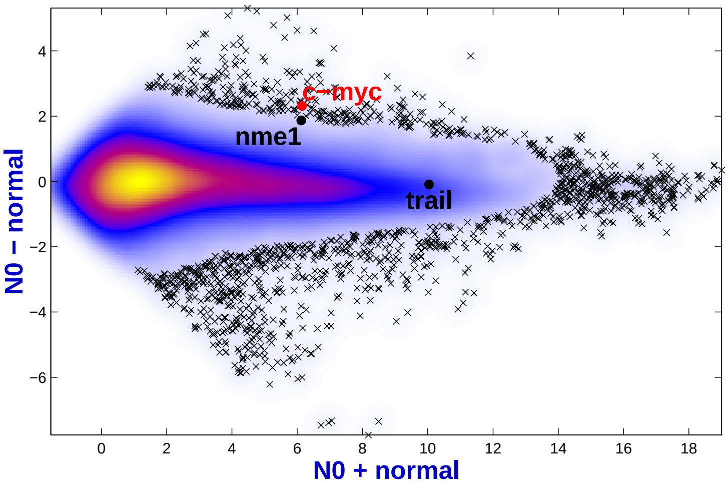


Figure 7: Scatterplot of a pairwise comparison of non-cancerous colon tissue and a colorectal tumor. Individual probes are represented by 'x' symbols. The x-coordinate is the average of the appropriately calibrated and transformed intensities (see Section 5.2). The y-coordinate is their difference, and 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.

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 [33]:

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. Figure 4d or [11]). 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 correlated variations in concentration and quality. An example is shown in Figure 9.

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

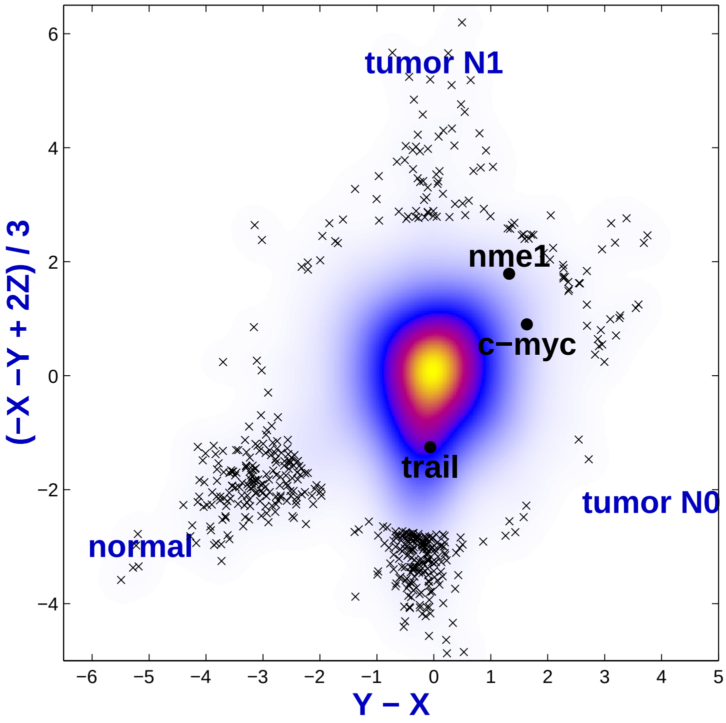


Figure 8: Scatterplot of a triple comparison between non-cancerous colon tissue, a lymph-node negative colorectal tumor (N0), and a lymph-node positive tumor (N1). The measurements from each probe correspond to a point in three-dimensional space, and are projected orthogonally on a plane perpendicular 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.

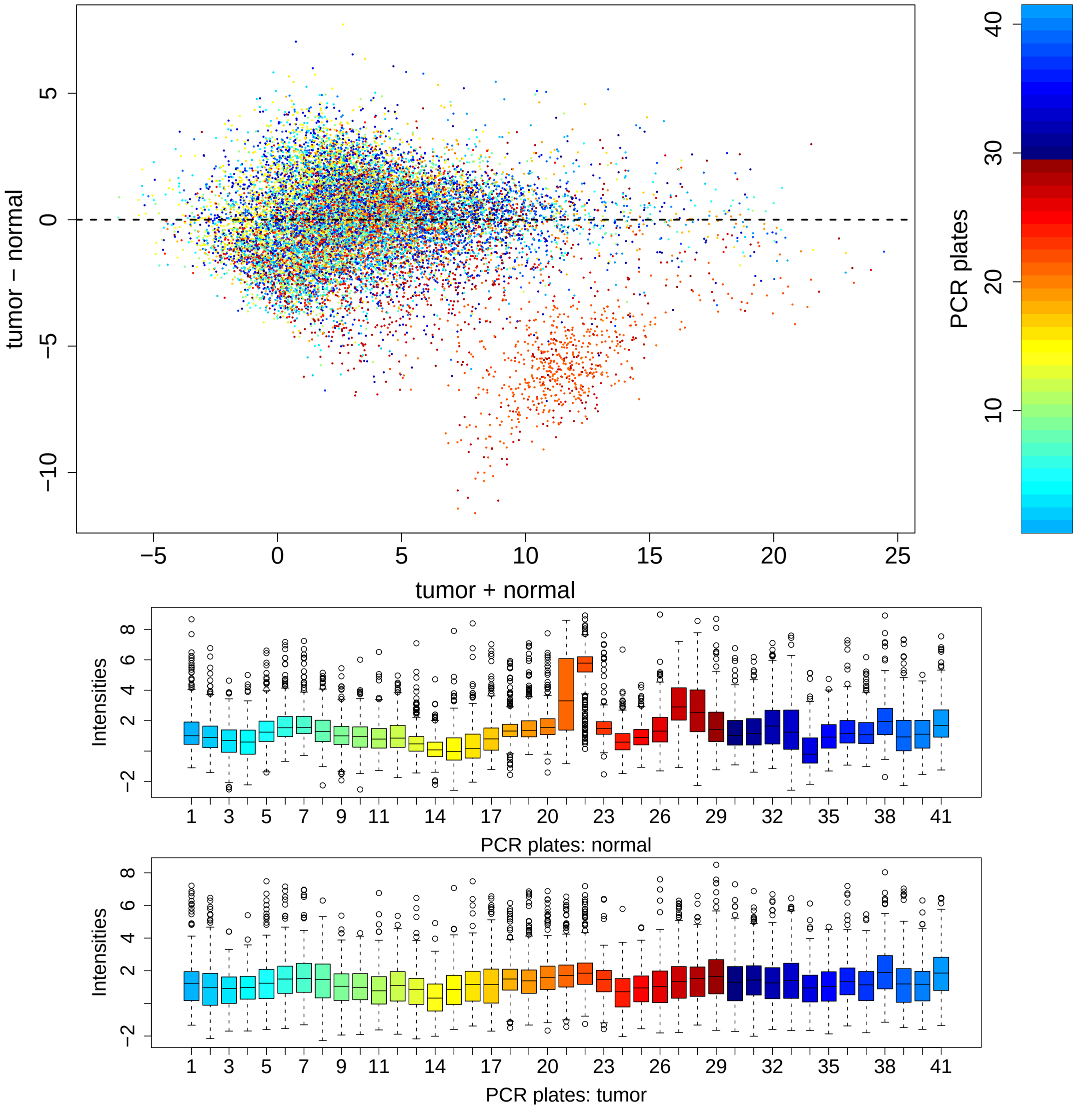


Figure 9:

Top panel: scatterplot of logarithmized intensities from a pair of single-color cDNA arrays, comparing renal cell carcinoma to matched non-cancerous kidney tissue. Similar to Figure 7, 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 facilitated by coloring the dots in the scatterplot by attributes such as PCR plate of origin or spotting pin. 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.

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:* different scanners can produce slightly different intensity images even from identical slides, and the performance of the same scanner can drift over time. Different image analysis programs can use different algorithms to calculate probe summaries, and the same program, in particular when it requires human interaction, can produce different results from the same image.

These considerations have important consequences for the experimental design: first, any variation that can be avoided by any means 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 [12]. 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 do so, which may, for example, be attributed to the fact that observed intensities are the sum of probe-specific signal and unspecific background [43]. 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 show up later in the data as one of the most pronounced sources of variation. A simple and instructive visual tool for exploring 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 $\mathrm{corr}(\vec{Y}_i, \vec{Y}_j)$, sort its rows and columns according to different experimental factors, and visualize the resulting false color images.

2.5 Along chromosome plots

Visualization of microarray data along genomic coordinates can be useful for many purposes, for example, to detect genomic aberrations (deletions, insertions) or regulatory mechanisms that act at the level of genomic regions [35]. Here we show an example from the application of a genome tiling microarray to transcription analysis.

While conventional microarrays contain only a preselected set of one or a few probes for each of a set of known or putative transcripts, more recent microarray designs provide probes for the complete genomic sequence content of an organism. Rather than relying on a manufacturer's assignment of probes to genes, or more exactly, target transcripts, it can become part of the

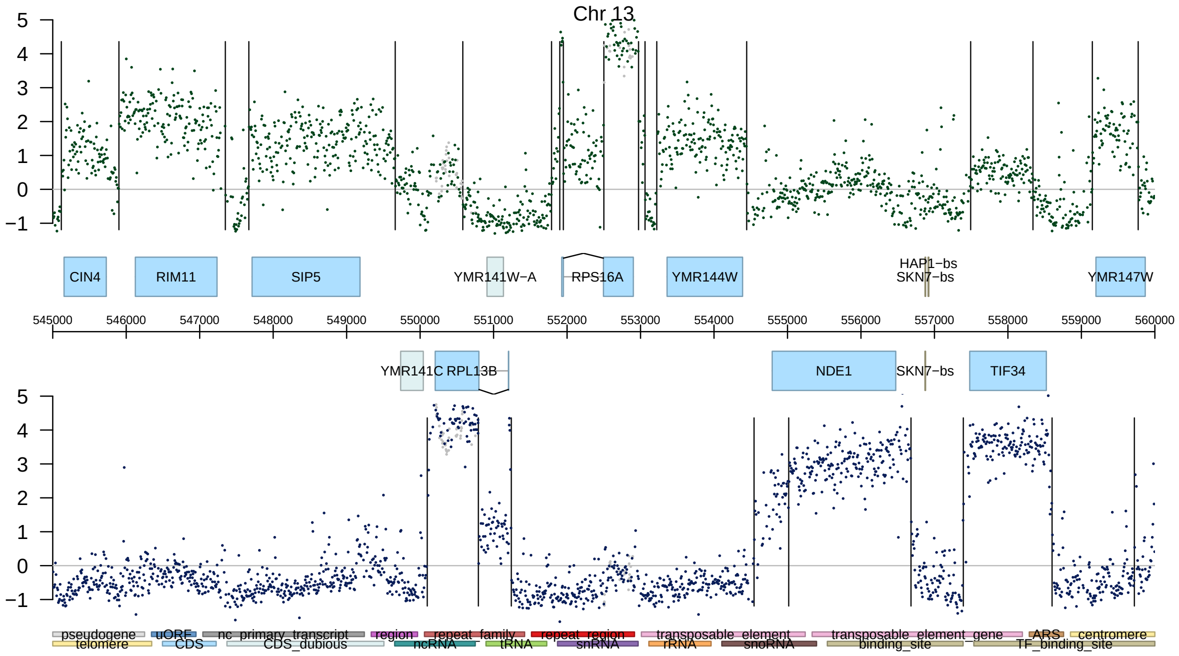


Figure 10: Along chromosome plot of the data from an Affymetrix genechip that contains 25-mer oligonucleotide probes covering the whole genome of Saccharomyces cerevisiae in steps of 8 bases, on both strands. The displayed values are the base 2 logarithms of the ratios between intensities from hybridization with a poly-A RNA sample and with a genomic DNA sample. Also shown are genomic coordinates and annotated genomic features. The vertical bars show the segmentation of the intensity signal into an approximately piecewise constant function [28]. The data allows for the mapping of 5' and 3' untranslated regions, the deconvolution of populations of overlapping transcripts of different lengths, and the detection of novel transcripts.

analysis to make the assignment on the basis of the data themselves. An example is shown in Figure 10.

2.6 Sensitivity and specificity of probes

The probes on a microarray are intended to measure the abundance of the particular transcript or locus that they are assigned to. However, probes may differ in terms of their sensitivity and specificity. Here, sensitivity means that a probe's fluorescence signal indeed responds to changes in the abundance of its target; specificity, that it does not respond to other targets or other types of perturbations.

Probes may lack sensitivity. Some probes initially identified with a gene do not actually hybridize to any of its products. Some probes will have been developed from information that has been superseded. In some cases, the probe may correspond to a different gene or it may in fact not represent any gene. In other cases, a probe may match only certain transcript variants of a given gene, which makes it more complicated to derive statements on the gene's expression (see the examples of NDE1 and CIN4 in Figure 10). There is also the possibility of human error [15, 24].

A potential problem especially with short oligonucleotide technology is that the probes may not be specific, that is, in addition to matching the intended transcript, they may also match others. In this case, we expect the observed intensity to be a composite from all matching transcripts. Note that, particularly in the case of higher eukaryotes, we are limited by the current state of knowledge of the transcriptomes. As our knowledge improves, the information about specificity of probes should also improve.

3 Error models

3.1 Motivation

With a microarray experiment, we aim to make statements about the abundances of specific molecules in a set of biological samples. However, the quantities that we measure are the fluoresence intensities of the different elements of the array. The measurement process consists of a cascade of biochemical reactions and an optical detection system with a laser scanner or a CCD camera. Biochemical reactions and detection are performed in parallel, allowing millions of measurements on one array. Subtle variations between arrays, the reagents used, and the environmental conditions lead to slightly different measurements even for the same sample.

The effects of these variations may be grouped in two classes: *systematic effects*, which affect a large number of measurements (for example, the measurements for all probes on one array; or the measurements from one probe across several arrays) simultaneously. Such effects can be estimated and, to good approximation, be removed. Other kinds of effects are random, with no well-understood pattern. These effects are commonly called *stochastic effects* or *noise*. This classification is not a property of the variations *per se*, but rather, reflects our understanding of them and our modeling effort. The same kind of variation can be considered stochastic in one analysis, and systematic in another.

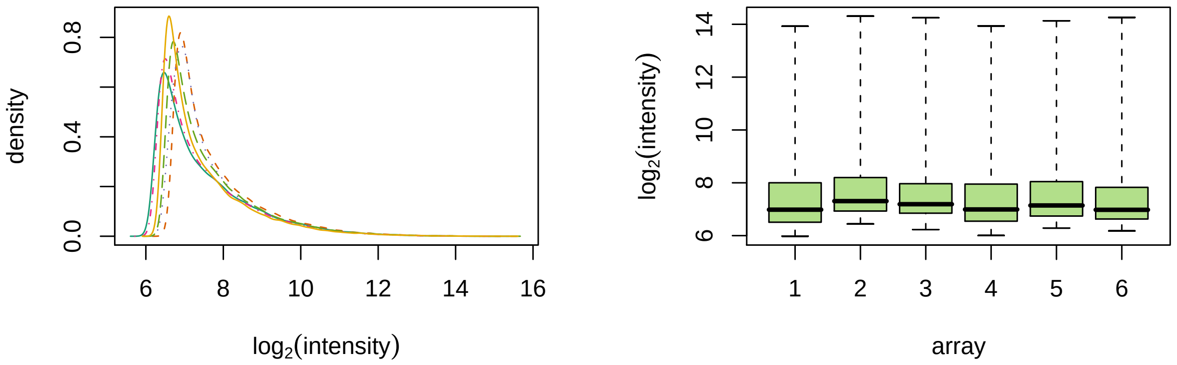
So what is the purpose of constructing error models for microarrays? There are three aspects:

3.1.1 Obtaining optimal estimates

Stochastic models are useful for preprocessing because they permit us to find *optimal* estimates of the systematic effects. We are interested in estimates that are precise and accurate. However, given the noise structure of the data we sometimes have to sacrifice accuracy for better precision and vice-versa. An appropriate stochastic model will aid in understanding the accuracy-precision, or bias-variance, trade off.

3.1.2 Biological inference

Stochastic models are also useful for statistical inference from experimental data. Consider an experiment in which we want to compare gene expression in the colons of mice that were treated



a) b)

Figure 11: a) Density estimates of probe intensity data from six replicate Affymetrix arrays. The x-axis is on a logarithmic scale (base 2). b) Box-plots of the same data.

with a substance and mice that were not. If we have many measurements, we can simply compare their empirical distributions. For example, if the values from ten replicate measurements for the DMBT1 gene in the treated condition are all larger than ten measurements from the untreated condition, the Wilcoxon test tells us that with a p-value of 10^{-5} the level of the transcript is really elevated in the treated mice. But often it is not possible, too expensive, or unethical, to obtain so many replicate measurements for all genes and for all conditions of interest. Often, it is also not necessary. If we have some confidence in a model, we are able to draw significant conclusions from fewer replicates.

3.1.3 Quality control

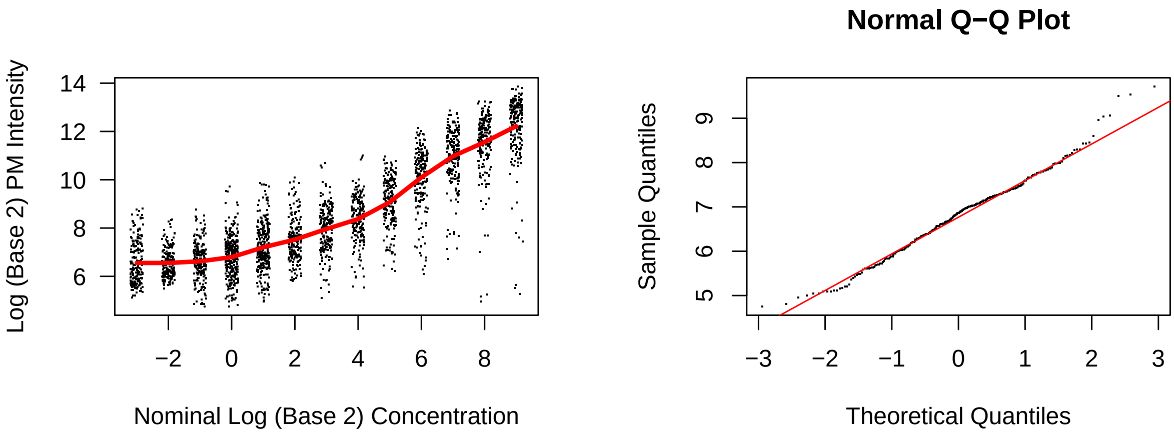
Quality control is yet another example of the usefulness of stochastic models: if the distribution of a new set of data greatly deviates from the model, this may direct our attention to quality issues with these data.

3.2 The additive-multiplicative error model

3.2.1 Induction from data

Different hybridizations will result in more or less different signal intensities even if the biological sample is the same. To see this, let us look in Figure 11 at the empirical distribution of the intensities from six replicate Affymetrix genechips. The data are part of the *Latin Square Data for Expression Algorithm Assessment* provided by Affymetrix[3].

[3] http://www.affymetrix.com/support/technical/sample_data/datasets.affx



a) b)

Figure 12: a) Plot of observed against nominal concentrations. Both axes are on the logarithmic scale (base 2). The curve represents the average value of all probes at each nominal concentration. Nominal concentrations are measured in picomoles. b) Normal quantile-quantile plot of the logarithmic (base 2) intensities for all probes with the same nominal concentration of 1 picomol.

One task of error modeling is to deal with background noise. Notice in Figure 11 that the smallest values attained are around 64, with slight differences between the arrays. We know that many of the probes are not supposed to be hybridizing to anything (as not all genes are expressed), so many measurements should indeed be zero. A bottom line effect of not removing background noise is that estimates of differential expression are biased. Specifically, the ratios are attenuated toward 1. This can be seen using the Affymetrix spike-in experiment, where genes were spiked in at known concentrations. Figure 12a shows the observed concentrations versus nominal concentrations of the spiked-in genes. Measurements with smaller nominal concentrations appear to be affected by attenuation bias. To see why, notice that the curve has a slope of about 1 for high nominal concentrations but gets flat as the nominal concentration gets closer to 0. This is consistent with the additive background noise model which we will discuss in the next section. Mathematically, it is easy to see that if s_1/s_2 is the true ratio and b_1 and b_2 are approximately equal positive numbers, then $(s_1 + b_1)/(s_2 + b_2)$ is closer to 1 than the true ratio, and the more so the smaller the s_i are compared to the b_i.

Figure 12b shows a normal quantile-quantile plot of logarithmic intensities of probes for genes with the same nominal concentration. Note that these appear to roughly follow a normal distribution. Figure 12 supports the multiplicative error assumption of the model that we formulate in the next section.

3.2.2 A theoretical deduction

Consider the generic observation equation $z = f(x, y)$, where z is the outcome of the measurement, x is the true underlying quantity that we want to measure, the function f represents the measurement apparatus, and $y = (y_1, \ldots, y_n)$ is a vector that contains all other parameters on which the functioning of the apparatus may depend. The functional dependence of f on some of the y_i may be known, on others it may not. Some of the y_i are explicitly controlled by the experimenter, some are not. For a well-constructed measurement apparatus, f is a well-behaved, smooth function, and we can rewrite the observation equation as

$$z = f(0, y) + f'(0, y)\, x + O(x^2), \tag{1}$$

where $f(0, y)$ is the baseline value that is measured if x is zero, f' is the derivative of f with respect to x, $f'(0, y)$ is a gain factor, and $O(x^2)$ represents non-linear efffects. By proper design of the experiment, the non-linear terms can be made negligibly small within the relevant range of x. Examples for the parameters y in the case of microarrays are the efficiencies of mRNA extraction, reverse transcription, labeling and hybridization reactions, amount and quality of probe DNA on the array, unspecific hybridization, dye quantum yield, scanner gain, and background fluorescence of the array.

Ideally, the parameters y could be fixed once and forever exactly at some value $\bar{y} = (\bar{y}_1, \ldots, \bar{y}_n)$. In practice, they will fluctuate around $\bar{y}$ between repeated experiments. If the fluctuations are not too large, we can expand

$$f(0, y) \approx f(0, \bar{y}) + \sum_{i=1}^{n} \frac{\partial f(0, \bar{y})}{\partial y_i} (y_i - \bar{y}_i) \tag{2}$$

$$f'(0, y) \approx f'(0, \bar{y}) + \sum_{i=1}^{n} \frac{\partial f'(0, \bar{y})}{\partial y_i} (y_i - \bar{y}_i). \tag{3}$$

The sums on the right hand sides of Eqns. (2) and (3) are linear combinations of a large number n of random variables with mean zero. Thus, it is a reasonable approximation to model $f(0, y)$ and $f'(0, y)$ as normally distributed random variables with means $a = f(0, \bar{y})$ and $b = f'(0, \bar{y})$ and variances σ_a^2 and σ_b^2, respectively. Thus, omitting the non-linear term, Equation (1) leads to

$$z = a + \varepsilon + b\, x(1 + \eta), \tag{4}$$

with $\varepsilon \sim N(0, \sigma_a^2)$ and $\eta \sim N(0, \sigma_b^2/b^2)$. This is the *additive-multiplicative error model* for microarray data, which was proposed by Ideker et al. [19]. Rocke and Durbin [29] proposed it in the form

$$z = a + \varepsilon + b\, x \exp(\eta), \tag{5}$$

which is equivalent to Equation (4) up to first order terms in η. Models (4) and (5) differ significantly only if the coefficient of variation σ_b/b is large. For microarray data, it is typically smaller than 0.2, thus the difference is of little practical relevance.

One of the main predictions of the error model (4) is the form of the dependence of the variance of z on its mean $\mathrm{E}(z)$:

$$\mathrm{Var}(z) = v_0^2 + \frac{\sigma_b^2}{b^2}\,(\mathrm{E}(z) - z_0)^2\,, \tag{6}$$

that is, a strictly positive quadratic function. In the following we will assume that the correlation between ε and η is negligible. Then the parameters of Equation (6) are related to those of Equation (4) via $v_0^2 = \sigma_a^2$ and $z_0 = a$. If the correlation is not negligible, the relationship is slightly more complicated, but the form of Equation (6) remains the same.

4 Normalization

A parametrization of Equation (5) that captures the main factors that play a role in current experiments is

$$z_{ip} = a_{i,s(p)} + \varepsilon_{ip} + b_{i,s(p)} B_p\, x_{j(i),k(p)}\, \exp(\eta_{ip}). \tag{7}$$

Let us dissect this equation: the index p labels the different probes on the array, and $k = k(p)$ is the transcript or locus that probe p maps to. Each probe is intended to map to exactly one k, but one transcript or locus may be represented by several probes. B_p is the probe-specific gain factor of the p-th probe. i counts over the arrays and, if applicable, over the different dyes. $j = j(i)$ labels the biological conditions (e. g. normal/diseased). $a_{i,s(p)}$ and $b_{i,s(p)}$ are normalization offsets and scale factors that may be different for each i and possibly for different groups ("strata") of probes $s = s(p)$. Probes can be stratified according to their physico-chemical properties [39] or array manufacturing parameters such as print-tip [41] or spatial location. In the simplest case, $a_{i,s(p)} = a_i$ and $b_{i,s(p)} = b_i$ are the same for all probes on an array. The noise terms ε and η are as above.

On an abstract level, much of the literature on normalization can be viewed as an application of Equation (7) to data, employing various choices for probe stratification, making simplifying assumptions on some of its parameters, rearranging the equation, and using different, more or less robust algorithms to estimate its parameters [2, 16, 17, 18, 21, 22, 25, 33, 38].

There is an alternative approach to normalization, which focuses on non-parametric methods and the algorithmic aspects. In this approach, one identifies those statistics (properties) of the data that one would like to be the same, say, between different arrays, but observes empirically in the raw data that they are not. One then designs an algorithm that transforms the data so that the desired statistics are made the same in the normalized data. The intention is that the interesting, biological signal is kept intact in the process [4, 31, 41].

For example, the *loess* normalization [41] calculates log–ratios M between the red and the green intensities on one array, plots them versus A, the logarithm of the geometric mean (see Figures 6 and 7), and postulates that a non-parametric regression line, calculated by a so-called `loess` scatterplot smoother [8] ought to look straight. In order to achieve this, the loess-fitted regression values for M are subtracted from the observed values, and the residuals are kept as the normalized data.

In *quantile* normalization, one plots the histogram of log-transformed intensities for each array and postulates that they all should look the same. Bolstad et al. [4] have introduced an algorithm that achieves this by rank-transforming the data and then mapping the ranks back to a consensus distribution. The result is a monotonous non-linear transformation for each array which assures that the distribution function of the transformed data is the same for all arrays.

These non-parametric methods are popular because they always "work", by construction, and usually in a fully automatic manner. In contrast, in model-based approaches it may turn out that a given set of data does not fit. Also, the assessment of goodness of fit is not easily automated, and often requires some human interaction. If the fit is bad, the data cannot be normalized, thus not be further analysed, and an expensive and time-consuming experiment would be left hanging.

However, there is a caveat: experiments may contain failed hybridizations, degraded samples, and non-functioning probes. The goodness of fit criteria from a model-based normalization method can serve as relevant criteria to detect these. With a method that "always works", there is the risk of overlooking these aspects of the data, to normalize them away, and move on to further analysis pretending that everything was fine. Conversely, one needs a sophisticated and largely non-automatic quality control step. So, if we consider normalization and quality control together as one task, the balance between model-based and non-parametric methods is more even.

Furthermore, parametric methods have, if they are appropriate, better power than non-parametric ones. They provide better sensitivity and specificity in the application of detecting differentially expressed genes. Given the typically small sample size and the expense of microarray experiments, this is a consequential point. It has been verified in comparison studies [9, 16].

5 Detection of differentially expressed genes

5.1 Step-wise versus integrated approaches

Most commonly used is the *stepwise* approach to microarray data analysis. It takes a collection of raw data as input and produces an *expression matrix* as output. In this matrix, rows correspond to gene transcripts and columns to conditions. Each matrix element represents the abundance, in certain units, of a transcript under a condition. Subsequent biological analyses work off the expression matrix and generally do not consider the raw data. The preprocessing itself is largely independent of the subsequent biological analysis. In some cases, the preprocessing is further subdivided into a set of sequential instructions, for example: subtract the background, then normalize the intensities, then summarize replicate probes, then summarize replicate arrays. By its *modularity*, the stepwise approach allows to structure the analysis workflow. Software, data structures, and methodology can be more easily re-used. For example, the same machine learning algorithm can be applied to an expression matrix irrespective of whether the raw data were obtained on Affymetrix chips or on spotted cDNA arrays. A potential disadvantage of the stepwise approach is that each step is optimized for itself, and that the results of subsequent steps have no influence on the previous ones. For example, the normalization procedure has to deal with whatever the preceding background correction procedure produced, and has no chance to ask it to reconsider. This can and does lead to inefficiencies.

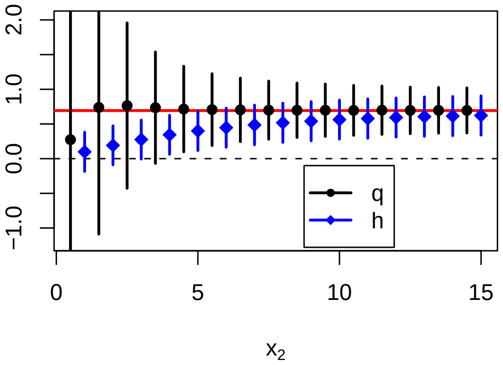


Figure 13: The shrinkage property of the generalized log–ratio h. Blue diamonds and error bars correspond to mean and standard deviation of $h(z_1, z_2)$, cf. Equation (8), black dots and error bars to $q(z_1, z_2)$, cf. Equation (9). Data were generated according to Equation (5) with $x_2 = 0.5, \ldots, 15$, $x_1 = 2x_2$, $a = 0$, $\sigma_a = 1$, $b = 1$, $\sigma_b = 0.1$. The horizontal line corresponds to the true log–ratio $\log(2) \approx 0.693$. For intensities x_2 that are larger than about ten times the additive noise level σ_a, h and q are approximately equal. For smaller intensities, we can see a *variance-bias trade-off*: q has no bias but a huge variance, thus an estimate of the fold change based on a limited set of data can be arbitrarily off. In contrast, h keeps a constant variance – for the price of systematically underestimating the true fold change.

In contrast, *integrated* approaches try to gain sensitivity by doing as much as possible at once, and therefore using the available data more efficiently. For example, rather than calculating an expression matrix, one might fit an ANOVA-type linear model that includes both technical covariates, such as dye and sample effects, and biological covariates, such as treatment [22], to the raw data. In Ben Bolstad's `affyPLM` method [5], the weighting and summarization of the multiple probes per transcript on Affymetrix chips is integrated with the detection of differential expression. Another example is the vsn method [16], which integrates background subtraction and normalization.

Stepwise approaches are often presented as modular data processing pipelines; integrated approaches as statistical models whose parameters are to be fitted to the data. In practice, data analysts will often choose to use a combination of both approaches, maybe starting with the stepwise approach, do a first round of high-level analyses, and then turn back to the raw data to answer specific questions that arise. Good software tools allow to use and explore both stepwise and integrated methods and to freely adapt and combine them.

5.2 Measures of differential expression: The variance-bias trade off

What is a good statistic to compare two (or several) measurements from the same probe on a microarray, taken from hybridizations with different biological targets?

Plausible choices include the difference, the ratio, and the logarithm of the ratio. To understand the problem more systematically, we return to the notation of Section 3.2.2. Let z_1 and z_2 denote two measurements from the same probe, and assume that they are distributed according to Equation (5) with the same parameters a, b, σ_a, and σ_b, but possibly with different values of x_1, x_2, corresponding to different levels of the target in the biological samples of interest. We want to find a function $h(z_1, z_2)$ that fulfills the following two conditions: *antisymmetry*, $h(z_1, z_2) = -h(z_2, z_1)$ for all x_1, x_2, and *homoskedasticity*, constant variance of $h(z_1, z_2)$ independent of x_1, x_2. An approximate solution is given by [17]

$$h(z_1, z_2) = \operatorname{arsinh}\left(\frac{z_1 - a}{\beta}\right) - \operatorname{arsinh}\left(\frac{z_2 - a}{\beta}\right) \tag{8}$$

with $\beta = \sigma_a b/\sigma_b$. If both z_1 and z_2 are large, this expression approaches the log–ratio

$$q(z_1, z_2) = \log(z_1 - a) - \log(z_2 - a)\,. \tag{9}$$

However, for $z_i \to a$, the log–ratio $q(z_1, z_2)$ has a large, diverging variance, a singularity at $z_i = a$, and is not defined in the range of real numbers for $z_i < a$. These unpleasant properties are important for applications: many genes are not expressed or only weakly expressed in some, but not all conditions of interest. That means, we need to compare conditions in which, for example, x_1 is large and x_2 is small. The log–ratio (9) is not a useful quantity for this purpose, since the second term will wildly fluctuate and be sensitive to small errors in the estimation of the parameter a. In contrast, the statistic (8), which is called the *generalized log–ratio* [30], is well-defined everywhere and robust against small errors in a. It is always smaller in magnitude than the log–ratio (see also Figure 13),

$$\begin{array}{rcll} |h(z_1, z_2)| & < & |q(z_1, z_2)| & \forall z_1, z_2, \\ h(z_1, z_2) & \approx & q(z_1, z_2) & \text{for } z_1, z_2 \gg a + \beta. \end{array} \tag{10}$$

The exponentiated value

$$\widehat{FC} = \exp(h(z_1, z_2)) \tag{11}$$

can be interpreted as a shrinkage estimator for the *fold-change* x_1/x_2. It is more specific, i. e. leads to fewer false positives in the detection of differentially expressed genes, than the naive estimator $(z_1 - a)/(z_2 - a)$ [13, 16].

5.3 Identifying differentially expressed genes from replicated measurements

One of the main motivations for doing microarray studies is the need to identify genes whose patterns of expression differ according to phenotype or experimental condition. Gene expression is a well coordinated system, and hence measurements on different genes are in general not independent. Given more complete knowledge of the specific interactions and transcriptional controls it is conceivable that meaningful comparisons between samples can be made by considering the joint distribution of specific sets of genes. However, the high dimension of gene expression space prohibits a comprehensive exploration, while the fact that our understanding of biological systems is only in its infancy means that in many cases we do not know which relationships are important and should be studied. In current practice, differential expression analysis will therefore at least start with a gene-by-gene approach, ignoring the dependencies between genes.

A simple approach in the comparison of different conditions is to rank genes by the difference of means of appropriately transformed intensities in the sense of Section 5.2. This may be the only possibility in cases where no, or very few replicates, are available. An analysis solely based on a difference of means statistic however does not allow the assessment of significance of expression differences in the presence of biological and experimental variation, which may differ from gene to gene. This is the main reason for using statistical tests to assess differential expression. Generally, one might look at various properties of the distributions of a gene's expression levels under different conditions, though most often location parameters of these distributions, such as the mean or the median, are considered. One may distinguish between parametric tests, such as the t–test, and non-parametric tests, such as the Mann–Whitney test or permutation tests. Parametric tests usually have a higher power if the underlying model assumptions, such as normality in the case of the t–test, are at least approximately fulfilled. Non–parametric tests do have the advantage of making less stringent assumptions on the data–generating distribution. In many microarray studies however, a small sample size leads to insufficient power for non–parametric tests and, as discussed in Section 3.1, increasing the sample size might be uneconomical or unethical if parametric alternatives are feasible. A pragmatic approach in these situations is to employ parametric tests, but to use the resulting p–values cautiously to rank genes by their evidence for differential expression, rather than taking them for the truth.

A generalized log–transformation of intensity data as described in Section 5.2 can be beneficial not only when using a difference of means statistic, but also for parametric statistical tests. Typically it will make the distribution of replicated measurements per gene roughly symmetric and more or less close to Normal. The variance stabilization achieved by the transformation can be advantageous for gene–wise statistical tests that rely on variance homogeneity, because it diminishes differences in variance between experimental conditions that are due to differences in the intensity level — however of course differences in variance between conditions may also have gene–specific biological reasons, and these will remain untouched by the transformation.

One or two group t-test comparisons, multiple group ANOVAs, and more general trend tests are all instances of linear models that are frequently used for assessing differential gene expression. As a parametric method, linear modeling is subject to the caveats discussed above, but the convenient interpretability of the model parameters often makes it the method of choice for microarray analysis. Due to the aforementioned lack of information regarding coregulation of genes, linear models are generally computed for each gene separately. When the genes of interest are identified, investigators can hopefully begin to study their coordinated regulation for more sophisticated modeling of their joint behavior.

The approach of conducting a statistical test for each gene is popular, largely because it is relatively straightforward and a standard repertoire of methods can be applied. However, the approach has a number of drawbacks: most important is the fact that a large number of hypothesis tests is carried out, potentially leading to a large number of falsely significant results. Multiple testing procedures allow to assess the overall significance of the results of a family of hypothesis tests. They focus on specificity by controlling type I (false positive) error rates such as the *family–wise error rate* or the *false discovery rate* [10]. Still, multiple hypothesis testing remains a problem, because an increase in specificity, as provided by p–value adjustment methods, is coupled with a loss of sensitivity, that is, a reduced chance of detecting true positives. Furthermore, the genes with the most drastic changes in expression are not necessarily the "key players" in the relevant biological processes [37]. This problem can only be addressed by incorporating prior biological knowledge into the analysis of microarray data, which may lead to focusing the analysis on a specific set of genes. Also if such a biologically motivated preselection is not feasible, the number of hypotheses to be tested can often be reasonably reduced by non–specific filtering procedures, discarding e.g. genes with consistently low intensity values or low variance across the samples. This is especially relevant in the case of genome–wide arrays, as often only a minority of all genes will be expressed at all in the cell type under consideration.

Many microarray experiments involve only few replicates per condition, which makes it difficult to estimate the gene-specific variances that are used e.g. in the t–test. Different methods have been developed to exploit the variance information provided by the data of all genes [1, 20, 27, 36]. In [34], an Empirical Bayes approach is implemented that employs a global variance estimator s_0^2 computed on the basis of all genes' variances. The resulting test statistic is a moderated t–statistic, where instead of the single–gene estimated variances s_g^2, a weighted average of s_g^2 and s_0^2 is used. Under certain distributional assumptions, this test statistic can be shown to follow a t-distribution under the null hypothesis with the degrees of freedom depending on the data.

6 Software

Many of the algorithms and visualizations discussed in this chapter are available through the Bioconductor project [14]. This project is an initiative for the collaborative creation of extensible software for computational biology and bioinformatics. Its goals include fostering development and widespread use of innovative software, reducing barriers to entry into interdisciplinary scientific research, and promoting the achievement of remote reproducibility of research results.

The software produced by the Bioconductor project is organized into packages, each of which is written and maintained relatively autonomously by its authors, who come from many different institutions around the world, and which are held together through a common language platform, R, a set of common data structures, a uniform structure of package organization and documentation, and a lively user community.

Results of this project are available on the website *http://www.bioconductor.org*. Among the packages that are most relevant for the subject of this chapter are `affy` (preprocessing of Affymetrix genechip data), `vsn` (affine-linear parametric normalization and variance stabilizing normalization), `marray` (two-color preprocessing), and `limma` (differential expression with linear models). Some further aspects are represented by `arrayMagic` (high-throughput quality control and preprocessing) and `tilingArray` (along chromosome plots and segmentation).

7 Acknowledgements

We thank Günther Sawitzki, Robert Gentleman, Dirk Buschmann, and Andreas Buneß for fruitful discussions. We are grateful to all contributors to the R and Bioconductor projects for sharing their software. Holger Sültmann, Annemarie Poustka, Anke Schroth, Lars Steinmetz and Lior David are gratefully acknowledged for sharing experimental data.

References

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

[2] **Beißbarth, T., K. Fellenberg, B. Brors, R. Arribas-Prat, J. M. Boer, N. C. Hauser, M. Scheideler, J. D. Hoheisel, G. Schütz, A. Poustka and M. Vingron.** 2000. Processing and quality control of DNA array hybridization data. Bioinformatics **16**:1014–1022.

[3] **Bertone, P., V. Stolc, T. E. Royce, J. S. Rozowsky, A. E. Urban, X. Zhu, J. L. Rinn, W. Tongprasit, M. Samanta, S. Weissman, M. Gerstein and M. Snyder.** 2004. Global identification of human transcribed sequences with genome tiling arrays. Science **306**:2242–2246.

[4] **Bolstad, B. M., R. A. Irizarry, M. Astrand and others.** 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics **19**:185–193.

[5] **Bolstad, B. M.** 2005. affyPLM: Fitting Probe Level Models. Bioconductor Vignettes **Release 1.7**:affyPLM.

[6] **Cheng, J., P. Kapranov, J. Drenkow, S. Dike, S. Brubaker, S. Patel, J. Long, D. Stern, H. Tammana, G. Helt, V. Sementchenko, A. Piccolboni, S. Bekiranov, D. K. Bailey, M. Ganesh, S. Ghosh, I. Bell, D. Gerhard and T. Gingeras.** 2005. Transcriptional Maps of 10 Human Chromosomes at 5-Nucleotide Resolution. Science **308**:1149–1154.

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

[8] **Cleveland, W., E. Grosse and W. Shyu.** Local regression models. Chapter 8. J.M. Chambers and T.J. Hastie Statistical Models in S. Wadsworth & Brooks 1992.

[9] **Cope, L. M., R. A. Irizarry, H. A. Jaffee, Z. Wu and T. P. Speed.** 2004. A benchmark for Affymetrix GeneChip expression measures. Bioinformatics **20**:323–331.

[10] **Dudoit, S., J. P. Shaffer and J. C. Boldrick.** 2003. Multiple hypothesis testing in microarray experiments. Statistical Science **18**:71-103.

[11] **Dudoit, S., Y. H. Yang, T. P. Speed and M. J. Callow.** 2002. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica **12**:111–139.

[12] **Duggan, D. J., M. Bittner, Y. Chen, P. Meltzer and J. M. Trent.** 1999. Expression profiling using cDNA microarrays. Nature Genetics **21 (Suppl 1)**:10–14.

[13] **Durbin, B. P., J. S. Hardin, D. M. Hawkins and D. M. Rocke.** 2002. A Variance-Stabilizing Transformation for Gene-expression Microarray Data. Bioinformatics **18 Suppl. 1**:S105–S110.

[14] **Gentleman, R. C., V. J. Carey, D. J. Bates, B. M. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. K. Smyth, L. Tierney, Y. H. Yang and J. Zhang.** 2004. Bioconductor: Open software development for computational biology and bioinformatics. Genome Biology **5**:R80.

[15] **Halgren, R. G., M. R. Fielden, C. J. Fong and T. R. Zacharewski.** 2001. Assessment of clone identity and sequence fidelity for 1189 IMAGE cDNA clones. Nucleic Acids Res **29**:582–588.

[16] **Huber, W., A. von Heydebreck, H. Sültmann, A. Poustka and M. Vingron.** 2002. Variance Stabilization Applied to Microarray Data Calibration and to the Quantification of Differential Expression. Bioinformatics **18 Suppl. 1**:S96-S104.

[17] **Huber, W., A. von Heydebreck, H. Sültmann, A. Poustka and M. Vingron.** 2003. Parameter estimation for the calibration and variance stabilization of microarray data. Statistical Applications in Genetics and Molecular Biology **2**:Article 3.

[18] **Huber, W.** 2005. Robust calibration and variance stabilization with vsn. Bioconductor Vignettes **Release 1.7**:vsn.

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

[20] **Kendziorski, C., M. Newton, H. Lan and M. Gould.** 2003. On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. Statistics in Medicine **22**:3899-3914.

[21] **Kepler, T. B., L. Crosby and K. T. Morgan.** 2002. Normalization and analysis of DNA microarray data by self-consistency and local regression. Genome Biology **3**:research0037.1–0037.12.

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

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

[24] **Knight, J.** 2001. When the chips are down. Nature **410**:860–861.

[25] **Li, C. and W. H. Wong.** 2001. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. PNAS **98**:31–36.

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

[27] **Lönnstedt, I. and T. P. Speed.** 2002. Replicated microarray data. Statistica Sinica **12**:31-46.

[28] **Picard, F., S. Robin, M. Lavielle, C. Vaisse and J.-J. Daudin.** 2005. A statistical approach for array CGH data analysis. BMC Bioinformatics **6**:27.

[29] **Rocke, D. M. and B. Durbin.** 2001. A model for measurement error for gene expression arrays. Journal of Computational Biology **8**:557–569.

[30] **Rocke, D. M. and B. Durbin.** 2003. Approximate variance-stabilizing transformations for gene-expression microarray data. Bioinformatics **19**:966–972.

[31] **Schadt, E. E., C. Li, B. Ellis and W. H. Wong.** 2001. Feature Extraction and Normalization Algorithms for High-Density Oligonucleotide Gene Expression Array Data. Journal of Cellular Biochemistry **Supplement 37**:120–125.

[32] **Schadt, E. E., S. W. Edwards, D. GuhaThakurta, D. Holder, L. Ying, V. Svetnik, A. Leonardson, K. W. Hart, A. Russell, G. Li, G. Cavet, J. Castle, P. McDonagh, Z. Kan, R. Chen, A. Kasarskis, M. Margarint, R. M. Caceres, J. M. Johnson, C. D. Armour, P. W. Garrett-Engele, N. F. Tsinoremas and D. D. Shoemaker.** 2004. A comprehensive transcript index of the human genome generated using microarrays and computational approaches. Genome Biol **5**:R73.

[33] **Schuchhardt, J., D. Beule, A. Malik, E. Wolski, H. Eickhoff, H. Lehrach and H. Herzel.** 2000. Normalization strategies for cDNA microarrays. Nucleic Acids Research **28**:e47.

[34] **Smyth, G.** 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology **3**:Article 3.

[35] **Su, A. I., T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker and J. B. Hogenesch.** 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc. Natl. Acad. Sci. of the U.S.A. **101**:6062–6067.

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

[37] **von Heydebreck, A., W. Huber and R. Gentleman.** Differential Expression with the Bioconductor Project. 0. Shankar Subramaniam Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics. Wiley 2004.

[38] **Wu, Z., R. Irizarry, R. Gentleman, F. Martinez Murillo and F. Spencer.** 2004. A Model Based Background Adjustement for Oligonucleotide Expression Arrays. Journal of the American Statistical Association **99**:909-917.

[39] **Wu, Z. and R. A. Irizarry.** 2005. Stochastic models inspired by hybridization theory for short oligonucleotide arrays. Journal of Computational Biology **12**:882–893.

[40] **Yang, Y. H., M. J. Buckley, S. Dudoit and T. P. Speed.** 2002. Comparison of methods for image analysis on cDNA microarray data. Journal of Computational and Graphical Statistics **11**:108-136.

[41] **Yang, Y. H., S. Dudoit, P. Luu and T. P. Speed.** 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res **30**:e15.

[42] **Yang, Y. H. and T. P. Speed.** 2002. Design issues for cDNA microarray experiments. Nat. Rev. Gen. **3**:579–588.

[43] **Yue, H., P. S. Eastman, B. B. Wang, J. Minor, M. H. Doctolero, R. L. Nuttall, R. Stack, J. W. Becker, J. R. Montgomery, M. Vainer and R. Johnston.** 2001. An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression. Nucleic Acids Research **29**:e41, 1–9.

[44] **Zien, A., J. Fluck, R. Zimmer and T. Lengauer.** 2003. Microarrays: how many do you need?. Journal of Computational Biology **10**:653–667.