BIOINFORMATICS

Vol. 18 Suppl. 1 2002
Pages S96–S104

Variance stabilization applied to microarray data calibration and to the quantification of differential expression

Wolfgang Huber[1], Anja von Heydebreck[2], Holger Sültmann[1], Annemarie Poustka[1] and Martin Vingron[2]

[1]Department of Molecular Genome Analysis, German Cancer Research Center, INF 280, Heidelberg, 69120, Germany and [2]Department of Computational Molecular Biology, Max-Planck-Institute for Molecular Genetics, Dahlem, Berlin, 14195, Germany

Received on January 24, 2002; revised and accepted on March 30, 2002

ABSTRACT

We introduce a statistical model for microarray gene expression data that comprises data calibration, the quantification of differential expression, and the quantification of measurement error. In particular, we derive a transformation h for intensity measurements, and a difference statistic Δh whose variance is approximately constant along the whole intensity range. This forms a basis for statistical inference from microarray data, and provides a rational data pre-processing strategy for multivariate analyses. For the transformation h, the parametric form $h(x) = \text{arsinh}(a + bx)$ is derived from a model of the variance-versus-mean dependence for microarray intensity data, using the method of variance stabilizing transformations. For large intensities, h coincides with the logarithmic transformation, and Δh with the log-ratio. The parameters of h together with those of the calibration between experiments are estimated with a robust variant of maximum-likelihood estimation. We demonstrate our approach on data sets from different experimental platforms, including two-colour cDNA arrays and a series of Affymetrix oligonucleotide arrays.

Availability: Software is freely available for academic use as an R package at http://www.dkfz.de/abt0840/whuber

Contact: w.huber@dkfz.de

INTRODUCTION

Microarrays simultaneously measure transcript abundances for thousands of genes in a cell population or tissue sample. The measurement is performed by quantitating the fluorescence intensities from labeled sample cDNA that has hybridized to the probes on the array. Multiple samples of interest are processed either by labeling them with different dyes, and letting them hybridize simultaneously against a single array, or by labeling them with the same dye, and letting them hybridize separately against multiple arrays. In each case, statements about the relative abundance of a gene transcript in these samples can be made by comparing the corresponding fluorescence intensities. Due to variations in sample treatment, labeling, dye efficiency and detection, the fluorescence intensities can in general not be compared directly, but only after appropriate calibration, which is sometimes also called ‘normalization’. One way of quantifying relative transcript abundance is the fold-change, that is the ratio of calibrated intensities. As the intensities are associated with measurement error, the usefulness of the fold-change or of any other measure of relative abundance depends on knowing its error distribution: one needs to know whether, for example, a calculated ratio of 1.5 is noteworthy, or whether it is most likely just a chance fluctuation. To understand the error distribution, it is necessary to first consider that of the original spot intensities.

The analysis of replicate microarray data typically shows that the variance of the measured spot intensities increases with their mean. For high intensities, the coefficient of variation is approximately constant, that is, the standard deviation increases roughly linearly with the mean. In a pioneering paper Chen *et al.* (1997) built a model based on the assumption of a constant coefficient of variation, and derived the distribution of the ratios of intensities. The distribution has one parameter, the coefficient of variation, and according to the model is the same for all probes on the array. To fit their model to the intensity data from a two colour cDNA array, they used a multiplicative calibration, which is estimated along with the coefficient of variation in an iterative algorithm. The model of Chen *et al.* (1997) motivates the use of logarithm-transformed intensities: ratios in the original data correspond to differences in the transformed data, the calibration amounts to a simple shift, and the constant coefficient of variation in the original data corresponds to an approximately constant standard deviation in the transformed data.

These concepts have been widely used in microarray data analysis. However, it has also become clear that for many data sets that are encountered in practice they are insufficient (e.g. Beißbarth *et al.* (2000); Hughes *et al.* (2000); Rocke and Durbin (2001); Newton *et al.* (2001); Baldi and Long (2001); Baggerly *et al.* (2001); Theilhaber *et al.* (2001)). The limitations mostly affect the data from weakly expressed genes. The significance of a ratio of, say, 1.5, is higher when it is observed in the high intensity range, than when it is observed in the low intensity range. Furthermore, many image quantization methods produce a certain fraction of non-positive intensities, for which ratios make no sense and the (real-valued) logarithm is not defined. Often, measurements below a threshold are dismissed, but it is unclear where to set the threshold and what to do with the missing values in the subsequent analysis. At the root of these problems lies the fact that with real microarray data the relationship between variance and mean typically is of a different form than that assumed by the model of Chen *et al.* Another limitation is that Chen *et al.* consider only linear calibration transformations. With this, the data should lie along a straight line in the scatterplot of the log-transformed data. In many data sets, however, one observes deviations from the straight line, resulting in, for example, 'banana-shaped' scatter plots.

In order to overcome these limitations, we generalize the approach of Chen *et al.* (1997). A major component is a model for the distribution of measurement error that has been proposed by Rocke and Durbin (2001), which leads to a quadratic variance-versus-mean dependence. Based on this, we derive a parametric family of transformations of the measured intensities, such that the variance of the transformed intensities becomes approximately independent of the mean. Together with the calibration transformations, these are incorporated into a statistical model which allows for maximum likelihood estimation of its parameters. Moreover, this generalized model is formulated for an arbitrary number d of replicates, extending the setup of Chen *et al.* (1997), who considered the case of $d = 2$, with the two-colour cDNA array technology in mind. The case of $d > 2$ is relevant for the analysis of series of one-colour arrays, such as Affymetrix arrays, or cDNA membranes, and could also be useful for multi-colour slides, should this technology emerge.

The utility of the model is twofold: First, it allows the construction of a 'difference statistic' Δh whose variance does not depend on the mean intensity, and whose value is a measure of differential expression. Δh may be viewed as a generalization of the log-ratio, and the two coincide for the highly expressed genes. Second, our approach provides the calibration as part of the model fitting. This offers a model-based, interpretable solution to the problem of normalization and may be preferable to commonly used ad hoc procedures.

The estimation of the model parameters uses replicate data from either the different colour channels of one array, or from a series of one-colour arrays. The different samples need *not* be exact biological replicates. Rather, the samples should be biologically related closely enough that the expression of most genes does not change. We use a robust estimation technique, which seeks to ignore the differentially expressed genes, and fits the model only to that subset of data points (typically, 50 to 90%) that is closest to the model mean.

We validate our approach on experimental data. First we provide evidence for the claimed form of the variance-versus-mean dependence. After this, we look at the distribution of the proposed difference statistic Δh as a function of the mean spot intensity. We find that it is centered around zero, and has constant width along the whole intensity range. Finally, we evaluate our approach with respect to the identification of differentially expressed genes. This is accomplished by comparing how the power of standard statistical tests depends on the method used for calibration and quantification of differential expression.

THE MODEL

A microarray data set may be pictured as a rectangular table y_{ki} of real numbers. The rows k correspond to the probes on the arrays, representing genes, and the columns i to the samples. The number n of probes may range from a few hundred to tens of thousands. The number of columns is $d = 2$ for the two-colour glass chip technology, and may range up to dozens or a few hundred for series of one-colour arrays. The values y_{ki}, with $k = 1, \ldots, n$ and $i = 1, \ldots, d$ are the intensity data as produced by the image quantization software. Many programs estimate local background intensities, which may be subtracted.

Due to variations in experimental factors such as amount of sample mRNA, or labeling and hybridization efficiencies, the values y_{ki} cannot directly be compared. We assume that the different columns (samples) can be brought on the same scale through affine-linear mappings, parametrized by the $2d - 2$ real-valued parameters $o_2, \ldots, o_d$ and $s_2, \ldots, s_d > 0$:

$$y_{ki} \mapsto \tilde{y}_{ki} = o_i + s_i \, y_{ki} \qquad (1)$$

where $i = 1, \ldots, d$, and $o_1 = 0$, $s_1 = 1$ without loss of generality. After this, one can calculate measures of differential expression, quantifying how much the intensity of a certain probe is different in one sample from another. For example, one may consider the difference between calibrated intensities, or the ratio. We use the general term *difference statistic* for such measures.

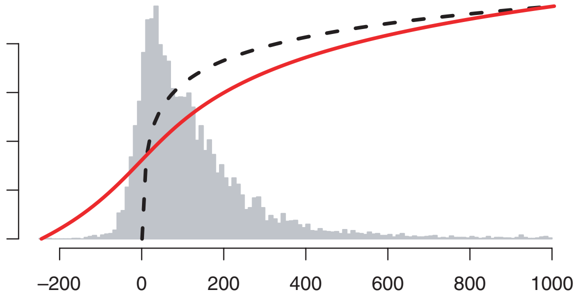

Fig. 1. Graph of the variance stabilizing transformation (4) (solid line), and of the logarithm function (dashed line). The histogram shows the intensity distribution of one colour channel on an 8400-element cDNA microarray. The parameters of the transformation (4) were estimated from the comparison with the intensities from the other colour channel.

For a non-differentially expressed gene, the values $\tilde{y}_{ki}$ for $i = 1, \dots, d$ scatter around the true value according to the distribution of the measurement error of probe k. We may thus regard $\tilde{y}_{ki}$ as realizations of the random variables Y_k with mean $\mathrm{E}(Y_k) = u_k$ and variance $\mathrm{Var}(Y_k) = v_k$. We assume that v_k only depends on k through a quadratic function of the mean u_k of the following form:

$$v_k = v(u_k) = (c_1 u_k + c_2)^2 + c_3, \qquad \text{with } c_3 > 0. \quad (2)$$

We will discuss the motivations for this assumption in the next section. The method of variance stabilization can be used to derive a transformation h such that the variance $\mathrm{Var}(h(Y_k))$ is approximately independent of the mean $\mathrm{E}(h(Y_k))$. An expression for h is given by (Tibshirani, 1988)

$$h(y) = \int^{y} 1/\sqrt{v(u)}\, du, \quad (3)$$

and results from a linear approximation of $h(Y_k)$ around $h(u_k)$ ('Delta method'). Inserting (2) into (3) yields

$$h(y) = \gamma \operatorname{arsinh}(a + by), \quad (4)$$

where the parameters of h are related to those of (2) through $\gamma = c_1^{-1}$, $a = c_2/\sqrt{c_3}$, and $b = c_1/\sqrt{c_3}$.

A graph of the arsinh function is shown by the solid line in Figure 1. Relationships between the arsinh function and the logarithm are given by

$$\operatorname{arsinh}(x) = \log(x + \sqrt{x^2 + 1}),$$
$$\lim_{x\to\infty} (\operatorname{arsinh} x - \log x - \log 2) = 0. \quad (5)$$

Hence, for large intensities the transformation (4) becomes equivalent to the usual logarithmic transformation. However, unlike the logarithm, it does not have a singularity at zero, and continues to be smooth and real-valued in the range of small or negative intensities.

Now we apply the variance stabilizing transformation (4) to the calibrated data $\tilde{y}_{ki}$ from Equation (1) to obtain the transformation $y_{ki} \mapsto h(\tilde{y}_{ki}) = \operatorname{arsinh}(a + b(o_i + s_i\, y_{ki}))$. The parameter γ may be omitted since it is merely an overall scaling factor. Setting $a_i = a + b\, o_i$, $b_i = b\, s_i$, and

$$h_i(y_{ki}) = \operatorname{arsinh}(a_i + b_i\, y_{ki}), \quad (6)$$

we can incorporate the calibration transformation (1), as well as the variance-versus-mean dependence (2) of the y_{ki} both together in the following statistical model:

$$h_i(Y_{ki}) = \mu_k + \varepsilon_{ki}, \qquad k \in K. \quad (7)$$

Here, K denotes the set of probes representing not differentially expressed genes, $\mu_k = \mathrm{E}(h(Y_{ki}))$ is the mean, and the variance of the error term is constant,

$$\mathrm{E}(\varepsilon_{ki}) = 0, \qquad \mathrm{Var}(\varepsilon_{ki}) = \sigma^2. \quad (8)$$

The condition $\mathrm{E}(\varepsilon_{ki}) = 0$ reflects the goal of calibration, whereas the common variance σ^2 of the error term is aimed at by variance stabilization. We fix the higher moments by assuming that the ε_{ki} are i. i. d. normal. In Section *Parameter Estimation* we will provide a robust variant of the maximum likelihood estimator for the $2d$ parameters a_i and b_i. Using the estimated transformations $\hat{h}_i$, the difference statistic that quantifies the change in expression between samples i and j of a gene represented by probe k is

$$\Delta h_{k;ij} = \hat{h}_i(y_{ki}) - \hat{h}_j(y_{kj}), \qquad k = 1, \dots, n. \quad (9)$$

One may express Equation (9) in terms of the arsinh function:

$$\begin{aligned}\Delta h_{k;ij} &= \operatorname{arsinh}(\hat{z}_{ki}) - \operatorname{arsinh}(\hat{z}_{kj})\\ &= \log \frac{\hat{z}_{ki} + \sqrt{\hat{z}_{ki}^2 + 1}}{\hat{z}_{kj} + \sqrt{\hat{z}_{kj}^2 + 1}}.\end{aligned}$$

where $\hat{z}_{ki} = \hat{a}_i + \hat{b}_i\, y_{ki}$ and $\hat{z}_{kj} = \hat{a}_j + \hat{b}_j\, y_{kj}$ are the calibrated intensities. This shows that in the limit of large intensities, $\Delta h_{k;ij}$ coincides with the log-ratio, whereas for near-zero intensities, it approaches the difference $\hat{z}_{ki} - \hat{z}_{kj}$.

THE VARIANCE-VERSUS-MEAN DEPENDENCE

The basic assumption underlying the results of the previous section is that the variance v_k depends on k as in

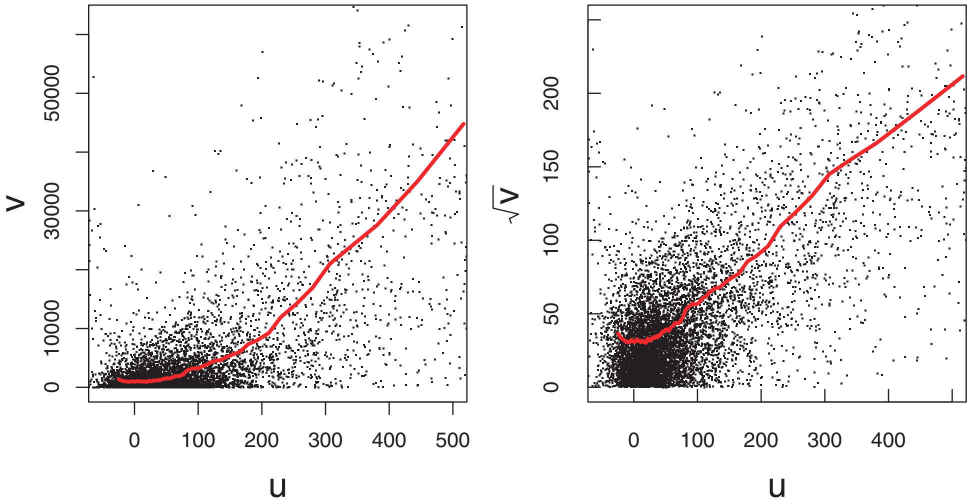

Fig. 2. Variance-versus-mean dependence $v(u)$ in microarray data. Shown is the data from one mRNA sample, labeled both in red and green and hybridized against an 8400-element cDNA slide. The plots show the variance versus the mean (left), and the standard deviation versus the mean (right). The dots correspond to single-spot estimates $\hat{v}_k = (y_{1k} - y_{2k})^2/2$, $\hat{u}_k = (y_{1k} + y_{2k})/2$, the solid lines show a moving average. The axis units are arbitrary.

Equation (2). First, this assumption implies that v_k depends on k mainly through the mean intensity u_k, and that other factors such as sequence-specific effects, or effects associated with array geometry or the production process may be neglected. This would have to be verified from case to case, but appears to be plausible in many experiments. Second, we make a particular parametric ansatz, namely a quadratic function of the form (2). There are several motivations for this. One is provided by the following model for the measurement error of gene expression arrays (Rocke and Durbin, 2001):

$$Y = \alpha + \beta e^{\eta} + \nu, \tag{10}$$

where β is the expression level in arbitrary units, α is an offset, and ν and η are additive and multiplicative error terms, respectively. ν and η are assumed to be independent and normally distributed with mean zero. This leads to

$$\mathrm{E}(Y) = \alpha + m_\eta \beta \tag{11}$$

$$\mathrm{Var}(Y) = s_\eta^2 \beta^2 + s_\nu^2 \tag{12}$$

where m_η and s_η^2 are mean and variance of e^η, and s_ν^2 is the variance of ν. Inserting Equation (11) into Equation (12) yields a quadratic expression of the form of Equation (2), and the relation between the parameters of model (10) and those of the variance stabilizing transformation (4) is given by $a = -\alpha s_\eta/(m_\eta s_\nu)$, $b = s_\eta/(m_\eta s_\nu)$, $\gamma = m_\eta/s_\eta$.

A further motivation for the quadratic ansatz (2) is provided by estimating $v(u)$ directly from microarray data. A typical example is shown in Figure 2. The right plot shows how the assumption of constant coefficient of variation breaks down in the low intensity range: the curve has a non-zero intercept, that is, $v(0) > 0$, and its convexity is in agreement with the assumption that $c_3 > 0$ in Equation (2). Similar curves have been observed for many slides, and also for other levels of replication, e.g. with data from replicate spots on one array, or from replicate arrays. The essential features of these curves may be captured by parametrizing $v(u)$ as a quadratic function of the form (2).

PARAMETER ESTIMATION

The parameters of the model (7) are estimated from data with a robust variant of maximum likelihood estimation. The detailed derivation, as well as results on convergence and identifiability are described in (Huber *et al.*, 2002). Given the data (y_{ki}), $k \in K$, $i = 1, \dots, d$, the profile log-likelihood (Murphy and van der Vaart, 2000) for the parameters $a_1, b_1, \dots, a_d, b_d$ is

$$-\frac{|K|d}{2}\log\left(\sum_{k\in K}\sum_{i=1}^{d}(h_i(y_{ki}) - \hat{\mu}_k)^2\right) + \sum_{k\in K}\sum_{i=1}^{d}\log h_i'(y_{ki}), \tag{13}$$

with h_i as in Equation (6). For a fixed set of probes K, we maximize (13) numerically under the constraints $b_i > 0$. The set of probes K is determined iteratively by a version of *least trimmed sum of squares (LTS) regression* (Rousseuw and Leroy, 1987). Briefly, K consists of those probes for which $r_k = \sum_{i=1}^{d}(\hat{h}_i(y_{ki}) - \hat{\mu}_k)^2$ is smaller than an appropriate quantile of the r_k. The LTS regression addresses the fact that the data distribution of y_{ki} is produced by a mixture from genes that are differentially expressed, and ones that are not.

VALIDATION

In this section we investigate how the variance stabilization and calibration work on real data, and how useful the resulting difference statistic Δh is to quantify differential gene expression.

To visualize the variance-versus-mean dependence, we plot the difference between the gene expression data for a pair of samples against the rank of their mean. Plotting against the rank distributes the data evenly along the x-axis and thus facilitates the visualization of variance heterogeneity. We look at data from a cDNA microarray experiment where samples from closely neighbouring parts of a kidney tumor were labeled with green and red fluorescent dyes, respectively. The expression levels of almost all of the genes in the two samples are expected to be unchanged, so that observed differences should represent the distribution of Δh in the absence of differential expression.

Figure 3 shows such plots for six different types of data transformation. First, Figure 3a corresponds to applying no transformation at all. Clearly, the width of the difference distribution increases with the signal average. After applying the logarithm transformation, the data looks as in Figure 3b. Here, all intensities below 1 have been replaced by 1 before taking the logarithm. This results in the two bands of points on the left side, corresponding to probes where one of two values was below 1, and one was not. If instead we dismiss the non-positive measurements, we obtain Figure 3c. In addition, a non-linear calibration (Yang *et al.*, 2001) has been applied. It has been argued that the problem of small, or non-positive intensity values is an artifact of the image analysis' local background estimation, and hence one might consider using the spot intensities without local background subtraction. The result is depicted in Figure 3d. In fact, over a wide range the difference distribution now happens to have a practically constant width. However, the distribution no longer follows a horizontal line. Instead of the logarithm transformation, another plausible choice is the rank transformation. This is shown in Figure 3e. Finally, Figure 3f shows the difference statistic Δh, as defined in Equation (9). The distribution is centered around the x-axis, and its width is constant along the whole range. Similarly good results have been observed with microarray expression data from many different sources.

We now turn to the question how well the values of $\Delta h_{k;ij}$, for a given probe k and measured over many pairs of samples i and j, reflect potential differential expression of the gene represented by that probe. We compare this to other commonly used difference statistics: the log-ratio together with different normalization methods, and the difference of ranks. As test data, we consider two data sets that contain highly replicated expression data. Both data sets compare two biological conditions, the first one clear cell renal cell cancer with non-cancerous kidney cortex tissue, and the second one acute myeloid with acute lymphoblastic leukemia (Golub *et al.*, 1999). Given that there is a large number of genes differentially expressed between the two conditions, we determined the number of those that were detected by a statistical test on the values of $\Delta h_{k;ij}$ and, in comparison, on the other difference statistics. Since the permutation test we used allows to control the type I error, the number of genes detected indicates how well the various difference statistics do in fact represent differential expression.

The first data set was produced at the Department of Molecular Genome Analysis at the German Cancer Research Center, using cDNA slides with about 4200 clones spotted in duplicate. Paired cancerous and non-cancerous tissue samples from 19 patients were used, and each tissue pair was hybridized against two slides, with the dyes swapped between repetitions, resulting in a total of 38 slides. From this data, we calculated (i) the difference statistic $\Delta h_{k;ij}$, as well as log-ratios. For the latter, in order to deal with the negative intensity values produced by subtracting the image analysis software's background estimates, four different rules were tried: (ii) ignore the background estimates, (iii) replace the negative values by 1 before taking the logarithm, (iv) subtract the 5%-quantile, then replace the remaining negative values by 1, and (v) flag them as missing values, For (ii)–(iv), a multiplicative calibration was estimated by the midpoint of the shorth of the uncalibrated log-ratios. The shorth of a univariate distribution is defined as the shortest interval containing half of the values, and for a unimodal distribution, its midpoint is a robust estimator of its mode. For (v), we used the local regression proposed by (Yang *et al.*, 2001), using the implementation in the R package `sma` (http://www.r-project.org) with default parameters. Finally, we calculated (vi) the rank differences. Each of the difference statistics (i)–(vi) was averaged over the two replicate spots, and over the two replicate arrays, resulting in one value per gene per patient, and hence in a matrix with 4200 rows, for the clones, and 19 columns, for the patients. The mean of each row was compared against its permutation distribution, obtained from performing random column-wise sign flips. We counted the number of genes that were at the extremes of their respective permutation distribution, as a function of the quantile α. The result is shown in Figures 4a and b. The test based

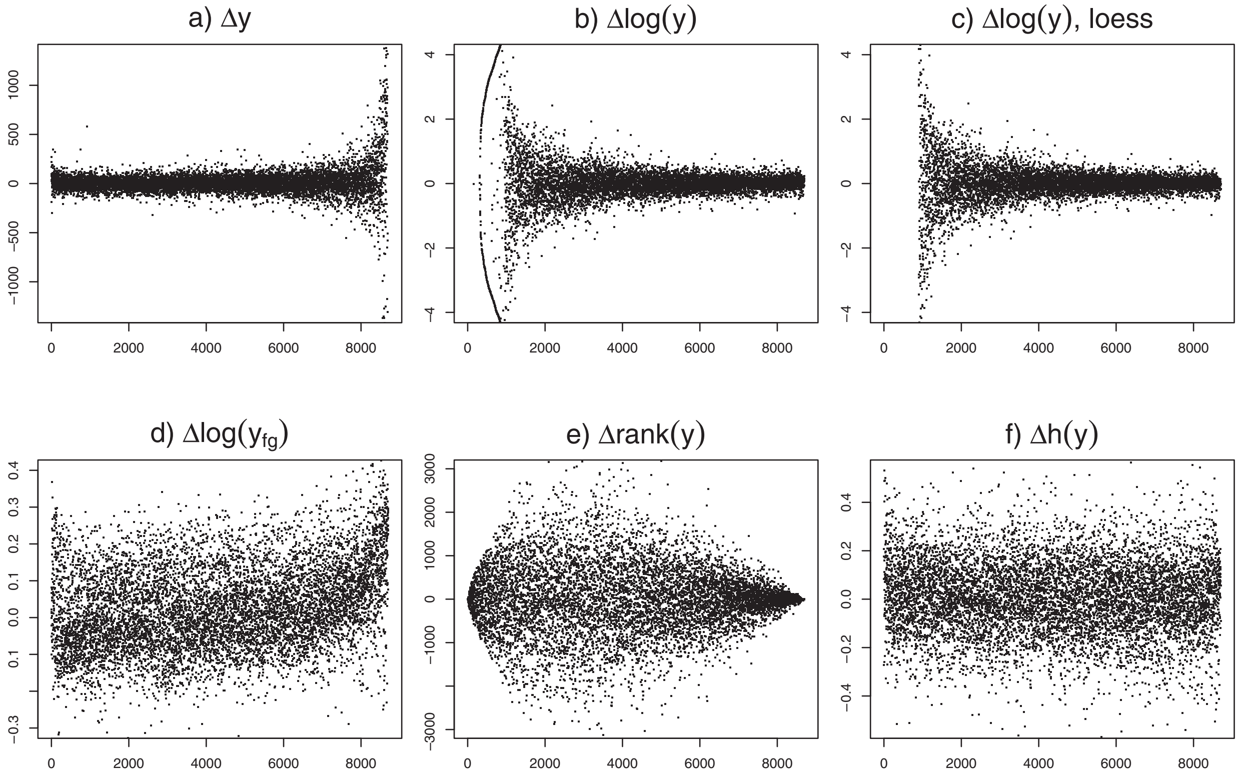

Fig. 3. The difference between the two colour channels of a cDNA microarray versus the rank of their average. Plot a) shows the untransformed intensity data, plots b-f) show the effect of five different transformations (see text). The y-axes of plots b-d) correspond to the usual 'log ratio', the y-axis of plot f) to the difference statistic Δh as proposed in this article.

on the difference statistic Δh uniformly had the highest power.

Figures 4a and b correspond to two one-sided tests, testing the row mean of the expression matrix against the hypothesis that it is less or equal to zero, or greater or equal to zero, respectively. We chose this procedure in order to make the comparison insensitive to potential subtle biases in the estimation of the calibration parameters. Such biases could be caused by a difference in the number of up- and down-regulated genes, and could consequently lead to biases in any of the difference statistics (i)-(vi). However, they would have opposite effects on the number of detected genes in the two tests. The fact that the difference statistic Δh detects more genes in both one-sided tests verifies that its better performance is not related to such potential calibration errors.

To evaluate our method with data from a different technological platform and experimental design, we used an expression data set measured on Affymetrix oligonucleotide arrays. It comprises 47 samples of acute myeloid leukemia and 25 samples of acute lymphoblastic leukemia (Golub *et al.*, 1999). From the data matrix provided at Golub *et al.*'s (1999) website (http://www-genome.wi.mit.edu/mpr) we calculated calibrated and transformed data $h_i(y_{ki})$, with $k = 1, \ldots, 7129$ and $i = 1, \ldots, 72$. We used the data as is, with no further selection or tresholding, and ignored the A/M/P-flags that the Affymetrix software associated with each value. The simultaneous estimation of the $2d = 144$ parameters posed no particular problem. In contrast, Golub *et al.* (1999) used a calibration method based on a linear regression, which in a pairwise fashion referenced arrays $2 \ldots 38$ to array 1, and arrays $40 \ldots 72$ to array 39. We used a two-sample permutation t-test to detect genes differentially expressed between AML and ALL. The result is shown in Figures 4c and d. Again, the test based on Δh has higher power.

Finally, an example for how the difference statistic Δh leads to more easily interpretable data displays is depicted in Figure 5. Since the distribution of Δh is independent of the mean intensity, observed values can directly be compared to the marginal empirical distribution, shown

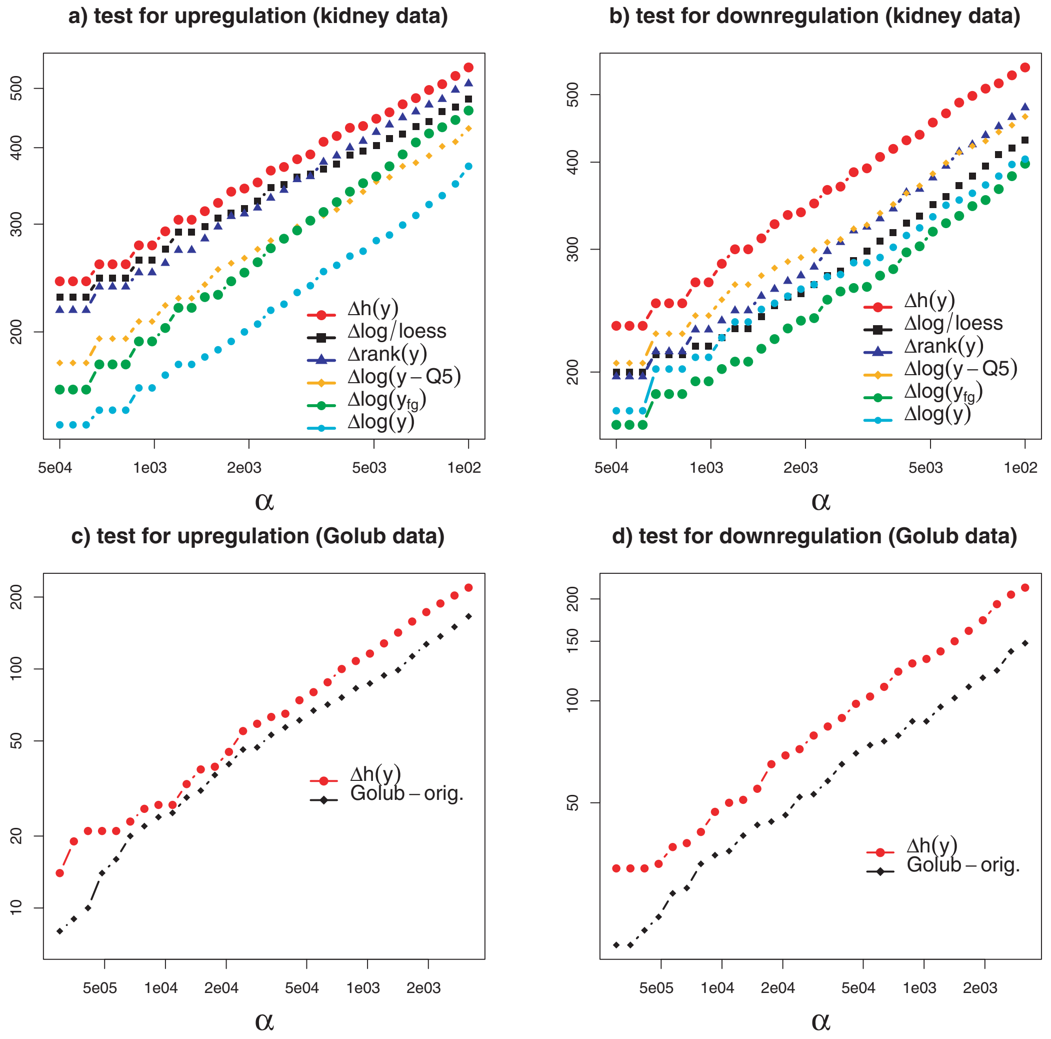

Fig. 4. Sensitivity and specificity of different methods for the quantification of differential expression. Top row: comparison of Δh to 4 methods based on log-ratios, and to the rank difference, on two-colour cDNA glass chip data. Bottom row: comparison of Δh to the procedure used in (Golub *et al.*, 1999) on the AML/ALL data. The plots show the number of genes selected by permutation tests against the significance level α. The test based on the difference statistic Δh uniformly has the best power.

in the histogram to the right. A scale on the Δh axis may be defined through a robust measure of width $\sigma_{\Delta h}$ of the empirical distribution, as indicated in Figure 5. Note, however, that in general the null distribution of Δh is not known, and in the presence of an unknown subset of differentially expressed genes, it is also not easy to estimate it.

DISCUSSION

A long-standing problem in the analysis of microarray gene expression experiments is how to take into account the dependence of the standard deviation of a spot intensity of its mean. In a seminal paper by Chen *et al.* (1997), this relationship has been modelled as a linear function. A main consequence is the use of logarithmic ratios as a measure of differential expression. Here, we have shown that their approach, although alleviating the problem, does not solve it entirely. The main limitation of the log-ratio as a measure of differential expression is the dependence of its variability on the intensity. To address this fact, we propose the general approach of applying a variance stabilizing transformation in order to achieve a constant signal-to-noise ratio. This results in the difference statistic Δh which displays approximately constant variance independent of the spot intensity, and replaces the log-ratio as a measure of differential gene expression.

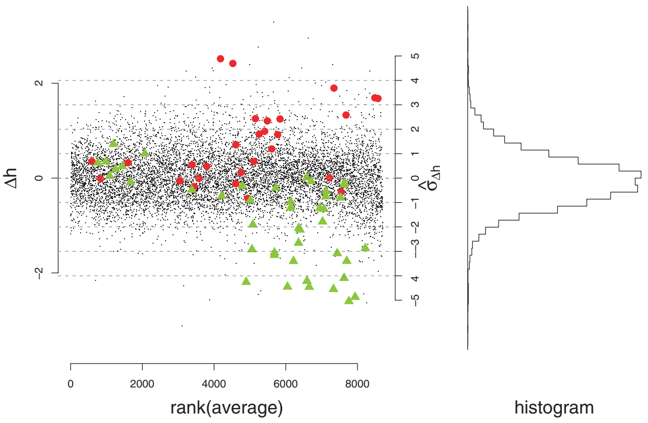

Fig. 5. Display of the data from a two-colour cDNA slide, taken from the kidney data set. Analogous to Figure 3, the difference statistic Δh is plotted along the y-axis, and the rank of the average spot intensity along the x-axis. The variance of the measurement error is constant over the whole intensity range, and horizontal dotted lines are plotted at multiples of its estimated standard deviation. Note that the figure shows the complete intensity data from the slide, without any thresholding or truncation. The circles and triangles represent genes that have been found up- and down-regulated, respectively, in renal cell cancer in a previous study (Boer *et al.*, 2001). Most of these are verified by the present slide, while for some, possibly due to biological or experimental variation, the value of Δh is close to zero.

Alternative approaches to this problem (e.g. Hughes *et al.* (2000); Baggerly *et al.* (2001); Theilhaber *et al.* (2001)) have also put forward quantitative models for this intensity-dependence, and propose to augment log-ratio values by ‘significance values’ calculated from such models.

Additionally, however, our approach takes into account the problem of calibration. Due to differential behaviour of dyes, or variations between samples and arrays, intensity measurements need to be brought on a common scale before they can be compared. In alternative approaches, the estimation of the calibration parameters is complicated by the non-constant variances. Furthermore, the resulting ratios may sensitively depend on the calibration. These problems are overcome in our approach: the estimation of the calibration parameters is simplified through the use of a variance stabilizing transformation, and is an integral part of the overall model fitting. Furthermore, our approach is parsimonious with parameters. No non-parametric curve estimation is required, which helps to provide robustness and avoid overfitting. Finally, since the difference statistic Δh is simply obtained as the difference between the transformed data of the individual samples, our approach not only allows for the comparison of two samples, but without any further effort can also be used for multivariate analyses comparing more than two samples.

In our model, the variance stabilizing transformation h turns out to be an arsinh function. This generalizes earlier results, as follows. Using a quadratic ansatz, the variance-versus-mean dependence at the basis of our approach has three parameters, which may be related to those of the model of Rocke and Durbin (2001). If in their model the additive noise component vanishes, the resulting limiting case turns out to be the logarithmic transformation with pseudocounts, $h_{\mathrm{pc}}(y) = \log(y + y_0)$, which has been used by various authors to overcome limitations of the logarithmic transformation (e.g. Beißbarth *et al.* (2000); Newton *et al.* (2001)). Furthermore, if both the constant and the linear term in the quadratic function vanish, our model turns into that of Chen *et al.* (1997)

Our approach is based on the following main assumptions: First, the variance of the measurements on a probe mainly depends on the mean intensity, and the relationship may be described by a second order polynomial with negative discriminant. This is grounded in the analysis of a large number of experiments and is in agreement with the model of Rocke and Durbin (2001). Second, we assume that the relationship of measurements between samples is captured by an affine-linear transformation. While non-linear behaviour may be observed under certain conditions, it has been demonstrated (e.g. Ramdas *et al.* (2001)) that current day microarray technology has a large, practically accessible working range in which intensities increase linearly with mRNA concentrations. It appears to us that in many cases apparent non-linearities that have been observed in the logarithmic plot (for an example, see Figure 3d) are an artifact of the logarithmic transformation, and disappear when using the appropriate affine-linear calibration. However, the general approach we proposed can easily be modified to incorporate a different class of calibration transformations or a different form of the variance-versus-mean dependence.

A third assumption concerns the statistical distribution of the intensity measurements. The variance-stabilized intensities per spot are assumed to be normally distributed. The parameter estimation draws on this assumption in particular near the center of the distribution, but because of our use of a robust regression procedure, it should not be affected by possible deviations from normality in the tails of the distribution.

A crucial point in modelling and parameter estimation from data is identifiability. The transformations h_i have $2d$ parameters (cf. Equations (6) and (13)), which we need to determine from nd data points. d ranges from $d = 2$ up to a few dozen or hundred, while n is typically in the order of several thousands. Given this generally favourable relation between the amount of data and number of parameters, and according to our experience with simulations and jackknife sampling, the transformations are well identifiable.

One might see a drawback of our method in the fact that it measures expression differences in terms of a function h with two estimated, experiment-specific parameters a and b, while the log-ratio can be calculated directly from the calibrated data, with no further parameters, and is easily interpreted as a fold-change. However, for large intensities, the values of Δh and of the log-ratio coincide (cf. Equation (5)), irrespective of the values of the experiment-specific parameters, and hence Δh may just as well be interpreted as the logarithm of a fold change. For small intensities that are near the detection limit of the experiment, the values of Δh are contracted towards 0 in comparison to those of the log-ratio. The onset and magnitude of this contraction are parametrized by the parameters a and b, which in this way encode the intensity dependent measurement error distribution of the experiment. We note that corresponding intensity-dependent thresholds are also used in the analysis of log-ratios, albeit usually in a less systematic manner.

Finally, and perhaps most importantly, our method also proves successful in the application to real data. It can typically be used off-the-shelf, without any particular tuning, and has been applied to different platforms, such as two-colour slides, Affymetrix chips, and radioactive membranes. Like in the ANOVA approach by Kerr *et al.* (2000), calibration is done not necessarily for pairs of samples but simultaneously for a whole set. The simple error distribution of the transformed intensities h_i makes them particularly suitable as input for clustering or other multivariate analysis methods. Software is provided as an R package, which is freely available for academic use.

ACKNOWLEDGEMENTS

We thank Günther Sawitzki and Dirk Buschmann for fruitful discussions, and Rainer Spang for critical reading of the manuscript. Bernd Korn has supplied the *RZPD Oncoset* of clones for the slides used in the kidney experiment, and Stephanie Süß provided expert technical assistance.

REFERENCES

Baggerly,K.A., Coombes,R.R., Hess,K.R., Stivers,D.N., Abruzzo,L.V. and Zhang,W. (2001) Identifying differentially expressed genes in cDNA microarray experiments. *J. Comput. Biol.*, **8**, 639–659.

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

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

Boer,J.M., Huber,W., Sültmann,H., Wilmer,F., von Heydebreck, A., Haas,S., Korn,B., Gunawan,B., Vente,A., Füzesi,L., Vingron,M. and Poustka,A. (2001) Identification and classification of differentially expressed genes in renal cell carcinoma by expression profiling on a global human 31 500-element cDNA array. *Genome Res.*, **11**, 1861–1870.

Chen,Y., Dougherty,E.R. and Bittner,M.L. (1997) Ratio-based decisions and the quantitave analysis of cDNA microarray images.. *J. Biomed. Opt.*, **2**, 364–374.

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

Huber,W., von Heydebreck,A., Sültmann,H., Poustka,A. and Vingron,M. A model for the calibration of microarray data and the quantification of differential expression. *Preprint available on request*.

Hughes,T.R., Marton,M.J., Jones,A.R., Roberts,C.J., Stoughton,R., Armour,C.D., Bennett,H.A., Coffey,E., Dai,H., He,Y.D., Kidd,M.J., King,A.M., Meyer,M.R., Slade,D., Lum,P.Y.Y., Stepaniants,S.B., Shoemaker,D.D., Gachotte,D., Chakraburtty,K., Simon,J., Bard,M. and Friend,S.H. (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

Kerr,K.M., Martin,M. and Churchill,G.A. (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, **7**, 819–837.

Murphy,S.A. and van der Vaart,A.W. (2000) On profile likelihood. *J. Amer. Stat. Assoc.*, **95**, 449–465.

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

Ramdas,L., Coombes,K.R., Baggerly,K., Abruzzo,L.V., Highsmith,W.E., Krogmann,T., Hamilton,S.R. and Zhang,W. (2001) Sources of nonlinearity in cDNA microarray expression measurements. *Genome Biol.*, **2**, research0047.1–research0047.7.

Rocke,D.M. and Durbin,B. (2001) A model for measurement error for gene expression analysis. *J. Comput. Biol.*, **8**, 557–569.

Rousseuw,P.J. and Leroy,A.M. (1987) *Robust Regression and Outlier Detection*. Wiley.

Theilhaber,J., Bushnell,S., Jackson,A. and Fuchs,R. (2001) Bayesian estimation of fold-changes in the analysis of gene expression: The PFOLD algorithm. *J. Comput. Biol.*, **8**, 585–614.

Tibshirani,R. (1988) Estimating transformations for regression via additivity and variance stabilization. *J. Amer. Stat. Assoc.*, **83**, 394–405.

Yang,Y.H., Dudoit,S., Luu,P., Lin,D.M., Peng,V., Ngai,J. and Speed,T.P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15:1–e15:11.