Statistical Applications in Genetics and Molecular Biology

Volume 2, *Issue* 1 2003 *Article* 3

Parameter estimation for the calibration and variance stabilization of microarray data

Wolfgang Huber* Anja von Heydebreck† Holger Sueltmann‡ Annemarie Poustka** Martin Vingron††

*German Cancer Research Center, Heidelberg, Germany, w.huber@dkfz-heidelberg.de

†Max-Planck-Institute for Molecular Genetics, Berlin, Germany, heydebre@molgen.mpg.de

‡German Cancer Research Center, Heidelberg, Germany, h.sueltmann@dkfz.de

**German Cancer Research Center, Heidelberg, Germany, a.poustka@dkfz.de

††Max-Planck-Institute for Molecular Genetics, Berlin, Germany, Martin.Vingron@molgen.mpg.de

Parameter estimation for the calibration and variance stabilization of microarray data

Wolfgang Huber, Anja von Heydebreck, Holger Sueltmann, Annemarie Poustka, and Martin Vingron

Abstract

We derive and validate an estimator for the parameters of a transformation for the joint calibration (normalization) and variance stabilization of microarray intensity data. With this, the variances of the transformed intensities become approximately independent of their expected values. The transformation is similar to the logarithm in the high intensity range, but has a smaller slope for intensities close to zero. Applications have shown better sensitivity and specificity for the detection of differentially expressed genes. In this paper, we describe the theoretical aspects of the method. We incorporate calibration and variance-mean dependence into a statistical model and use a robust variant of the maximum-likelihood method to estimate the transformation parameters. Using simulations, we investigate the size of the estimation error and its dependence on sample size and the presence of outliers. We find that the error decreases with the square root of the number of probes per array and that the estimation is robust against the presence of differentially expressed genes. Software is publicly available as an R package through the Bioconductor project (http://www.bioconductor.org).

KEYWORDS: microarrays, error model, variance stabilizing transformation, resistant regression, robust estimation, maximum likelihood, simulation

1 Introduction

Two important topics in the analysis of microarray data are the calibration (normalization) of data from different samples and the problem of variance inhomogeneity, in the sense that the variance of the measured intensities depends on their expected value. A family of transformations has been proposed that makes the variance of transformed intensities roughly independent of their expectation value [1, 2, 3]. In the following, we describe and validate an approach to the estimation of the parameters of a transformation for calibration and variance stabilization. In Section 2, we formulate our assumptions about the data in terms of a statistical model. It is a version of the multiplicative-additive error model that has been introduced in the context of microarray data by Ideker et al. [4] and Rocke and Durbin [5]. The framework encompasses two-color data from spotted cDNA arrays, Affymetrix genechip probe intensities, and radioactive intensities from nylon membranes. In Section 3, we use the method of approximate variance stabilization to simplify the model. In Section 4, we derive a robust estimator for the parameters of the calibration and the variance stabilizing transformation. Finally, in Section 5, we use simulations to investigate the validity of the variance stabilizing transformation and to evaluate the estimator for different sample sizes and parameter regimes.
Applications to real data, demonstrating increased sensitivity and specificity for the detection of differentially expressed genes compared to other methods, have previously been described [3]. The comparison method and program code are provided in a Bioconductor vignette [6], with the intention to make it easy for other scientists to reproduce this type of comparison.

2 The model

A microarray consists of a set of probes immobilised on a solid support. The probes are chosen such that they bind to specific sample molecules; for DNA arrays, this is ensured by the sequence-specificity of the hybridization reaction between complementary DNA strands. The interesting fraction from the biological sample is prepared in solution, labeled with fluorescent dye and allowed to bind to the array. The abundance of sample molecules can then be compared through comparing the fluorescence intensities at the matching probe sites.
The measured intensity y_{ki} of probe $k = 1, \ldots, n$ for sample $i = 1, \ldots, d$ may be decomposed into a specific and an unspecific part,

$$y_{ki} = \alpha_{ki} + \beta_{ki} x_{ki}. \tag{1}$$

Here, x_{ki} is the abundance of the transcript represented by probe k in the sample i, β_{ki} is a proportionality factor, and α_{ki} subsumes unspecific signal contri-

butions which may be caused by effects such as non-specific hybridization, cross-hybridization or background fluorescence.
The offsets α_{ki} and gain factors β_{ki} are usually not known, but microarray technologies are designed in such a way that their values for different k and i are related. This makes it possible to infer statements about the concentrations x_{ki} from the measured data y_{ki}. Relations between the offsets and gain factors for different k and i can be expressed in terms of a further decomposition,

$$\beta_{ki} = \beta_i \, \gamma_k \, e^{\eta_{ki}}, \tag{2}$$

$$\alpha_{ki} = a_i + \bar{\nu}_{ki}. \tag{3}$$

Thus, the gain factor is the product of a probe affinity γ_k, which is the same for all measurements involving probes of type k, times a normalization factor β_i, which applies to all measurements from sample i. The remainder $\beta_{ki}/(\beta_i\gamma_k)$ is accounted for by $e^{\eta_{ki}}$. One can choose the units of β_i and γ_k such that $\sum_k \eta_{ki} = \sum_i \eta_{ki} = 0$. The unspecific signal contribution α_{ki} can be decomposed into a per-sample offset a_i and a remainder $\bar{\nu}_{ki}$ with $\sum_k \bar{\nu}_{ki} = 0$.
The probe affinity γ_k may depend, for example, on the probe sequence, secondary structure and the abundance of probe molecules on the array. The normalization factor β_i may depend, for example, on the amount of mRNA in the sample, on the labeling efficiency, and on dye quantum yield. The idea behind the decompositions (1)–(3) is that while the individual values of η_{ki} and $\bar{\nu}_{ki}$ may fluctuate around zero, they do so in an unsystematic, random manner. Thus, for example, we assume that there are no systematic non-linear effects, which would imply trends in the η_{ki} or $\bar{\nu}_{ki}$ dependent on the value of x_{ki}.
Now one can reduce the parameter complexity of Eqn. (1) through the following three modeling steps:

1. Do not try to explicitly determine the probe affinities γ_k. They can be absorbed into $m_{ki} = \gamma_k x_{ki}$, which may be considered a measure of the abundance of transcript k in sample i in probe-specific units.

2. Treat η_{ki} and $\bar{\nu}_{ki}$ as "noise terms" coming from appropriate probability distributions.

3. Estimate the values of the normalization factors β_i and offsets a_i, as well as parameters of the probability distributions from the data.

Thus, Eqn. (1) leads to the following stochastic model:

$$\frac{Y_{ki} - a_i}{\beta_i} = m_{ki} \, e^{\eta_{ki}} + \nu_{ki}, \qquad \eta_{ki} \overset{\text{iid}}{\sim} \mathcal{L}_\eta, \; \nu_{ki} \overset{\text{iid}}{\sim} \mathcal{L}_\nu. \tag{4}$$

Here, $\nu_{ki} = \bar{\nu}_{ki}/\beta_i$ is the additive noise scaled by the normalization factor β_i. The right hand side of Eqn. (4) is a combination of an additive and a multiplicative error term. It was proposed by Rocke and Durbin [5], using normal distributions $\mathcal{L}_\eta = N(0, \sigma_\eta^2)$ and $\mathcal{L}_\nu = N(0, \sigma_\nu^2)$. In the following, we will consider distributions $\mathcal{L}_\eta$ and $\mathcal{L}_\nu$ that are unimodal, roughly symmetric, and have mean zero and variances σ_η^2 and σ_ν^2, respectively, but we do not rely on the assumption of a normal distribution. The left hand side describes the calibration of the microarray intensities Y_{ki} through subtraction of offsets a_i and scaling by normalization factors β_i [7, 3].
According to Eqn. (4), the variance of the random variable Y_{ki} is related to its mean through

$$\mathrm{Var}(Y_{ki}) = c^2 \left(\mathrm{E}(Y_{ki}) - a_i\right)^2 + \beta_i^2 \sigma_\nu^2, \tag{5}$$

where $c^2 = \mathrm{Var}(e^\eta)/\mathrm{E}^2(e^\eta)$ is a parameter of the distribution of $\eta \sim \mathcal{L}_\eta$. In the log-normal case, $c^2 = \exp(\sigma_\eta^2) - 1$. Thus, the relationship of the variance to the mean is a strictly positive, quadratic function. For a highly expressed gene, the variance $\mathrm{Var}(Y_{ki})$ is dominated by the quadratic term and the coefficient of variation of Y_{ki} is approximately c, independent of k and i. For a weakly expressed or unexpressed gene, the variance $\mathrm{Var}(Y_{ki})$ is dominated by the constant term and the standard deviation of Y_{ki} is approximately $\beta_i \sigma_\nu$, which may be interpreted as the background noise level for the i-th sample, and is independent of k.

3 Variance stabilizing transformations

Consider a random variable X with expectation value 0 and a differentiable function h defined on the range of X. Then

$$h(X) = h(0) + h'(0)\, X + r(X)\, X, \tag{6}$$

where r is a continuous function with $r(0) = 0$ and

$$\mathrm{Var}(h(X)) = h'(0)^2\, \mathrm{Var}(X) + \mathrm{Var}(r(X)\, X) + 2h'(0)\, \mathrm{E}(r(X)\, X^2). \tag{7}$$

If h does not deviate from linearity too strongly within the range of typical values of X, then $r(X)$ is small and the terms involving $r(X)$ on the right hand side of Eqn. (7) are negligible. Thus, for a family of random variables Y_u with expectation values $\mathrm{E}(Y_u) = u$ and variances $\mathrm{Var}(Y_u) = v(u)$

$$\mathrm{Var}(h(Y_u)) \approx h'(u)^2\, v(u). \tag{8}$$

An approximately variance-stabilizing transformation can be obtained by finding a function h for which the right hand side is constant, that is, by integrating $h'(u) =$

$v^{-1/2}(u)$,

$$h(y) = \int^{y} 1/\sqrt{v(u)}\, du. \tag{9}$$

Note that if h is approximately variance-stabilizing, then so is $\gamma_1 h + \gamma_2$ with $\gamma_1, \gamma_2 \in \mathbb{R}$.
For a variance-mean dependence as in Eqn. (5), this leads to

$$h_i(y) = \operatorname{arsinh} \frac{y - a_i}{b_i}, \tag{10}$$

with $b_i = \beta_i \sigma_\nu / c$. Thus, on the transformed scale the model (4) takes the simple form

$$\operatorname{arsinh} \frac{Y_{ki} - a_i}{b_i} = \mu_{ki} + \varepsilon_{ki}, \qquad \varepsilon_{ki} \overset{\text{iid}}{\sim} \mathcal{L}_\varepsilon, \tag{11}$$

where μ_{ki} represents the true abundance on the transformed scale of gene k in sample i and $\mathcal{L}_\varepsilon$ has mean zero and variance c^2. The relation between μ_{ki} and m_{ki} is $\mu_{ki} = \mathrm{E}(\operatorname{arsinh}(\frac{c}{\sigma_\nu}(m_{ki} e^{\eta_{ki}} + \nu_{ki}))) \approx \operatorname{arsinh}(\frac{c}{\sigma_\nu} m_{ki})$.

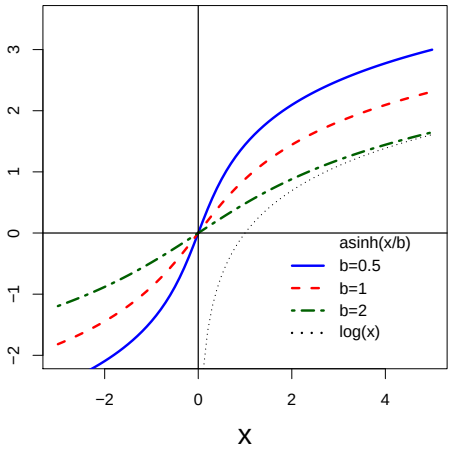

Figure 1: Graph of the function (10) for $a_i = 0$ and three different values of b_i. For comparison, the dotted line shows the graph of the logarithm function.

The graph of the function (10) is shown in Fig. 1. The inverse hyperbolic sine and the logarithm are related to each other via

$$\operatorname{arsinh}(x) = \log(x + \sqrt{x^2 + 1}) \tag{12}$$

$$\log(x) = \operatorname{arsinh} \frac{1}{2}\left(x - \frac{1}{x}\right) \tag{13}$$

$$\lim_{x\to\infty} \left(\operatorname{arsinh} x - \log x - \log 2\right) \quad = \quad 0. \tag{14}$$

Functions that may have, within the range of the data, a graph similar to that of $\operatorname{arsinh}((y-a)/b)$ have been proposed, among them the shifted logarithm and the linlog-transformation [8, 9]

$$\tilde{h}(y) \quad = \quad \log \frac{y-\tilde{a}}{\tilde{b}} \tag{15}$$

$$\bar{h}(y) \quad = \quad \begin{cases} \log(y/\bar{b}), & y \geq \bar{a} \\ y/\bar{a} + \log(\bar{a}/\bar{b}) - 1, & y < a \end{cases} \tag{16}$$

While transformation (10) corresponds, via Eqn. (9), to a variance-mean dependence of the form $v(u) \propto (u-a)^2 + \text{const.}$, the two transformations (15) and (16) correspond to variance-mean dependences of the form

$$\begin{aligned} \tilde{v}(u) \quad &\propto \quad (u-\tilde{a})^2, \\ \bar{v}(u) \quad &\propto \quad \begin{cases} u^2, & u \geq \bar{a} \\ \bar{a}^2, & u < \bar{a} \end{cases} \end{aligned}$$

respectively. All of these may fit the data, but in the following we choose transformation (10) both for computational convenience and for its interpretability in terms of the error model (4).

4 Parameter estimation

Model (11) relates the measured intensities Y_{ki} to the true expression values μ_{ki} in terms of the calibration and variance stabilization parameters a_i and b_i and the noise distribution $\mathcal{L}_\varepsilon$. We would like to estimate the parameters a_i and b_i. To this end, we consider the non-differentially expressed genes, for which $\mu_{ki} = \mu_k$ for all i. If in addition we require that $\mathcal{L}_\varepsilon$ be normal, we obtain

$$\operatorname{arsinh} \frac{Y_{ki}-a_i}{b_i} = \mu_k + \varepsilon_{ki}, \qquad \varepsilon_{ki} \overset{\text{iid}}{\sim} N(0, c^2). \tag{17}$$

In the following, we derive the maximum likelihood (ML) estimator for the parameters a_i and b_i. Then, by using a robust procedure similar to *least trimmed sum of squares regression* [10], we extend its validity to situations in which there is a minority fraction of differentially expressed genes, for which model (17) is misspecified, and to distributions $\mathcal{L}_\varepsilon$ that are approximately normal in the center, but may have different tails.

4.1 Maximum likelihood estimation

According to Eqn. (17), the probability of observing a value y_{ki} within an interval $[y_{ki}^{\alpha}, y_{ki}^{\omega}]$ is

$$\begin{aligned} P(y_{ki} \in [y_{ki}^{\alpha}, y_{ki}^{\omega}]) &= \int_{h_i(y_{ki}^{\alpha})}^{h_i(y_{ki}^{\omega})} \rho\left(\frac{h_i(y_{ki}) - \mu_k}{c}\right) dh_i(y_{ki}) \\ &= \int_{y_{ki}^{\alpha}}^{y_{ki}^{\omega}} \rho\left(\frac{h_i(y_{ki}) - \mu_k}{c}\right) h_i'(y_{ki})\, dy_{ki}. \end{aligned} \tag{18}$$

Here, ρ denotes the density of the standard normal distribution. The ML estimates of the model parameters $\{a_i\}, \{b_i\}, c, \{\mu_k\}$ are those that maximize the likelihood

$$\prod_{k=1}^{n} \prod_{i=1}^{d} \rho\left(\frac{h_i(y_{ki}) - \mu_k}{c}\right) h_i'(y_{ki}). \tag{19}$$

Now the ML estimates of the parameters of h_i are those that maximize the *profile likelihood* $\mathrm{pl}(a_1, b_1, \ldots, a_d, b_d)$, which is obtained by replacing c^2 and μ_k by their maximizing values [11]

$$\begin{aligned} \hat{\mu}_k &= \frac{1}{d} \sum_{i=1}^{d} h_i(y_{ki}) \\ \hat{c}^2 &= \frac{1}{nd} \sum_{k=1}^{n} \sum_{i=1}^{d} (h_i(y_{ki}) - \hat{\mu}_k)^2. \end{aligned}$$

This results in

$$\begin{aligned} &\mathrm{pl}(a_1, b_1, \ldots, a_d, b_d) = \\ &= \prod_{k=1}^{n} \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}\hat{c}} \exp\left(\frac{(h_i(y_{ki}) - \hat{\mu}_k)^2}{\frac{2}{nd} \sum_{k=1}^{n} \sum_{i=1}^{d} (h_i(y_{ki}) - \hat{\mu}_k)^2} \right) h_i'(y_{ki}) \\ &= \frac{e^{nd/2}}{(2\pi)^{nd/2}\, \hat{c}^{nd}} \prod_{k=1}^{n} \prod_{i=1}^{d} h_i'(y_{ki}). \end{aligned} \tag{20}$$

The profile log-likelihood is, up to a constant, given by

$$\begin{aligned}\mathrm{pll}(a_1, b_1, \ldots, a_d, b_d) &= \\ &= -nd \log \hat{c} + \sum_{k=1}^{n}\sum_{i=1}^{d} \log h_i'(y_{ki}) \\ &= -\frac{nd}{2} \log\left(\sum_{k=1}^{n}\sum_{i=1}^{d}(h_i(y_{ki}) - \hat{\mu}_k)^2\right) + \sum_{k=1}^{n}\sum_{i=1}^{d} \log h_i'(y_{ki}), \qquad (21)\end{aligned}$$

and can be maximized numerically under the constraints $b_i > 0$. The first term on the right hand side involves the sum of squared residuals, while the second term contains the Jacobi determinant of the transformation. The optimization landscape is not too different from that of an ordinary quadratic cost function and we have never encountered multiple local maxima. Our implementation uses the R function `optim` with the method `L-BFGS-B`.

4.2 Resistant regression

The maximum likelihood estimator that is obtained by maximizing (21) is sensitive to deviations from normality and to the presence of differentially expressed genes, for which $\mu_{ki} = \mu_k$ does not hold. In the following, we consider a modification which makes it more robust against outliers.

In *least sum of squares (LS) regression*, the fitted parameters $\hat{\boldsymbol{a}}$ are those that minimize the sum of squared residuals,

$$\hat{\boldsymbol{a}}_{\mathrm{ls}} = \operatorname*{argmin}_{\boldsymbol{a}} \sum_{k=1}^{n} r_k(\boldsymbol{a})^2, \qquad (22)$$

where $r_k(\boldsymbol{a})$ is the residual between fit and the k-th data point and n is the number of data points. In *least trimmed sum of squares (LTS) regression* [10], this is replaced by

$$\hat{\boldsymbol{a}}_{\mathrm{lts}} = \operatorname*{argmin}_{\boldsymbol{a}} \min_{K} \sum_{k\in K} r_k(\boldsymbol{a})^2. \qquad (23)$$

The minimization over K runs over all subsets of $\{1, \ldots, n\}$ of size $\lceil n q_{\mathrm{lts}} \rceil$, where $0.5 < q_{\mathrm{lts}} \leq 1$ and $\lceil x \rceil$ is the smallest integer greater or equal to x. While the LS estimate can be arbitrarily off due to even a single outlier, the LTS estimator has a breakdown point of approximately $1 - q_{\mathrm{lts}}$, that means, the estimation error does not become too large as long as the fraction of outliers is less than $1 - q_{\mathrm{lts}}$.

Practically, the exact solution of (23) is possible only for small n. We use the following heuristic iterative procedure. In the expression (21) for the profile log-likelihood, replace the sums over $k = 1, \ldots, n$ by sums over $k \in K$. Start the

iteration with $K = \{1, \ldots, n\}$. This results in an initial parameter estimate $\hat{\boldsymbol{a}} = (\hat{a}_1, \hat{b}_1, \ldots, \hat{a}_d, \hat{b}_d)$ and residuals

$$r_k = \sum_{i=1}^{d} (\hat{h}_i(y_{ki}) - \hat{\mu}_k)^2, \qquad k = 1, \ldots, n. \tag{24}$$

Now partition the set $\{1, \ldots, n\}$ into 10 slices T_s, such that T_1 contains those k for which $\hat{\mu}_k$ is smaller than the 10%-quantile of the $\hat{\mu}_k$, T_2 those for which $\hat{\mu}_k$ is between the 10%- and 20%-quantile, and so on. Let Q_s be the q_{lts}-quantile of $\{r_k | k \in T_s\}$. A new selection K of putative non-outlying probes is found through

$$K = \bigcup_s \{k \in T_s | r_k \leq Q_s\}. \tag{25}$$

With this, compute new parameter estimates, and repeat the iteration. Stable estimates are usually reached after less than 10 iterations.

The iteration procedure is a simple version of the method proposed in [12]. In addition, rather than permitting for K any subset of $\{1, \ldots, n\}$ of size $\lceil n q_{\text{lts}} \rceil$, we require that K contains an equal amount of data from every slice, according to Eqn. (25). This has the following advantage. Robust estimation involves the weighting of data points as more or less reliable and thus implies a trade-off between being diverted by outliers, and ignoring important data. To find a good compromise, it can be useful to consider further context information. Here, we have invoked the notion that the fraction of outliers should be roughly the same across the whole range of average intensities μ_k.

5 Results

5.1 Properties of the variance stabilizing transformation

The derivation of the variance stabilizing transformation (10) involves the approximation (8). Fig. 2 investigates how well this approximation holds for a family Y_m of random variables distributed according to

$$Y_m = m e^{\eta} + \nu, \qquad \eta \sim N(0, \sigma_\eta^2),\ \nu \sim N(0, 1) \tag{26}$$

for parameters $m, \sigma_\eta > 0$. This corresponds to the right hand side of Eqn. (4). In the notation of Eqn. (26), the variance stabilizing transformation (10) takes the form $h(y) = \operatorname{arsinh}(cy)$ with $c^2 = \exp(\sigma_\eta^2) - 1$. If m is large, Y_m is dominated by the multiplicative term, $Y_m \approx m e^{\eta}$, and $h(Y_m) \approx \log(Y_m)$. The asymptotic standard deviation of $h(Y_m)$ for $m \to \infty$ is thus $\mathrm{Sd}(\log(m e^{\eta})) = \sigma_\eta$. In Fig. 2, the

behavior of $\mathrm{Sd}(h(Y_m))$ for finite values of m is compared against the asymptotic value for different choices of the parameter σ_η. For each plot, the function was numerically evaluated at 60 values of m through Monte Carlo integration with 10^6 samples. Even in the case of $\sigma_\eta = 0.4$, which corresponds to a probability of 5% of observing a relative error larger than $e^{2 \cdot 0.4} \approx 2.2$, the standard deviation $\mathrm{Sd}(h(Y_m))$ does not depart from the asymptotic value by more than a factor of 1.035.

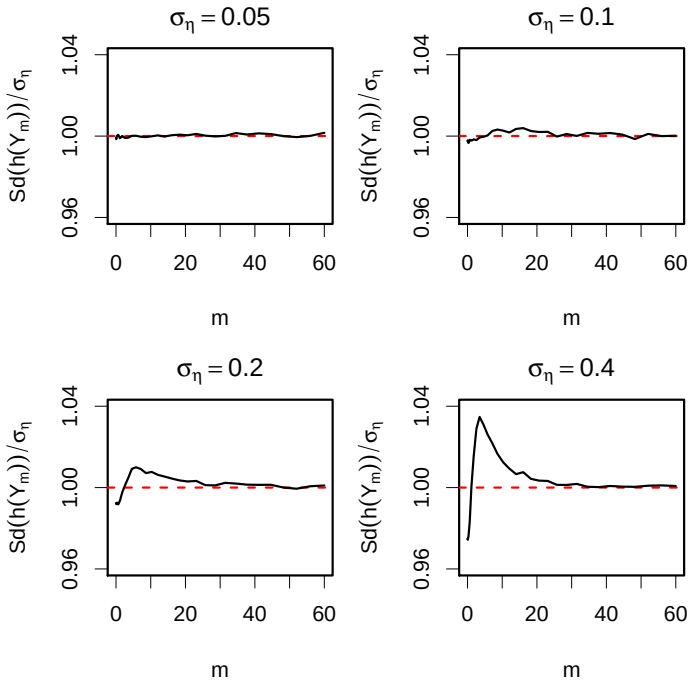

Figure 2: Validation of the approximation used in Section 3. For a family of random variables distributed according to the error model (26), the plots show the standard deviation of the transformed values $h(Y_m) = \mathrm{arsinh}(cY_m)$, divided by the asymptotic value σ_η that is obtained for $m \to \infty$.

An example for the effect of the variance stabilizing transformation (10) on real data is shown in Fig. 3. RNA from biopsies of adjacent parts of a kidney tumor was labeled with red and green dyes, respectively, and hybridized to a cDNA microarray. Fluorescence intensities were measured with a laser scanner. Per-spot summary intensity values were determined with the ArrayVision software (Imaging Research Inc., St. Catharines, Ontario, Canada). A spot's intensity was ob-

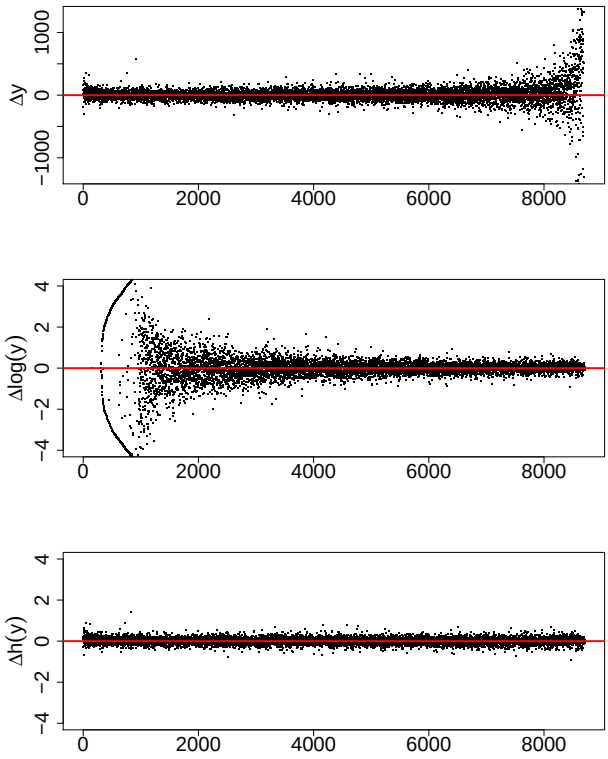

Figure 3: Three different transformations applied to data from a cDNA microarray. The top panel shows the difference $\Delta y_k = y_{k2} - y_{k1}$ between background-subtracted and calibrated red and green intensities on the y-axis versus the rank of their sum $y_{k1} + y_{k2}$ on the x-axis. Similarly, the middle plot shows the log-ratio $\Delta \log(y_k) = \log(y_{k2}) - \log(y_{k1})$ versus the rank of the sum $\log(y_{k1}) + \log(y_{k2})$ and the bottom plot shows $\Delta h(y_k) = h(y_{k2}) - h(y_{k1})$ versus the rank of $h(y_{k1}) + h(y_{k2})$. Plotting against the rank distributes the data evenly along the x-axis and thus facilitates the visualization of variance heterogeneity.

tained by subtracting the median brightness of surrounding pixels from the median of those within. The expression levels of almost all genes in the two samples are expected to be unchanged. Thus, the vertical width of the band of points represents the scale of the distribution of differences. For the untransformed intensities (upper plot), the width increases to the right, in accordance with the error model described in Section 2. For log-ratios (middle plot), the differences are largest at the left end, for measurements of low intensity. The far left region of the plot corresponds to spots in which one or both of the intensities were smaller than 1, in which case the logarithm was replaced by the value 0. While the log-ratios have an interquartile range (IQR) of about 0.2 in the high-intensity regime, the IQR is as large as 0.95 for the points with ranks 1500 to 2000. A constant IQR of about 0.2 along the whole range of intensities is observed for the generalized log-ratio Δh (bottom plot). Note that this value is the same as that in the high-intensity end of the log-ratio plot, in agreement with the asymptotic relationship (14).
The marginal distribution of Δh is compared a against a normal distribution in the quantile-quantile (QQ) plot Fig. 4. While the distribution is roughly normal in the center and approximately symmetric, its tails are heavier than normal. This observation, which we have made on many data sets, motivates the use of normal theory and the robust modification of Section 4.

5.2 Simulation of data

To verify the computational feasibility of the estimator constructed in Section 4 and to investigate its behavior for different sample sizes, we ran simulation studies. Simulated data were generated according to model (11) with $\mathcal{L}_\varepsilon = N(0, c^2)$.
Values for the parameters μ_{ki} were generated as follows. First, following [13], for each gene k a value μ_k was drawn according to

$$\mu_k = \operatorname{arsinh}(m_k), \qquad 1/m_k \sim \Gamma(1,1). \tag{27}$$

The density of the reciprocal $\Gamma(1,1)$ distribution is shown in Fig. 5. To model the mixture of non-differentially and differentially expressed genes, indicators $p_k \in \{0,1\}$ were generated with $P[p_k = 1] = p_{\text{diff}}$. For each gene with $p_k = 1$ and for each sample $i \geq 2$ a factor $s_{ki} \in \{-1, 1\}$ was drawn with $P[s_{ki} = 1] = p_{\text{up}}$ and an amplitude z_{ki} was drawn from the uniform distribution $U(0, z_{\max})$. Thus, p_k indicates whether or not gene k shows differential expression; s_{ki}, whether it is up- or down-regulated in sample i compared to sample 1; z_{ki}, by how much. These were combined to obtain

$$\begin{aligned} \mu_{k1} &= \mu_k \\ \mu_{ki} &= \mu_k + p_k s_{ki} z_{ki}, \qquad i \geq 2. \end{aligned} \tag{28}$$

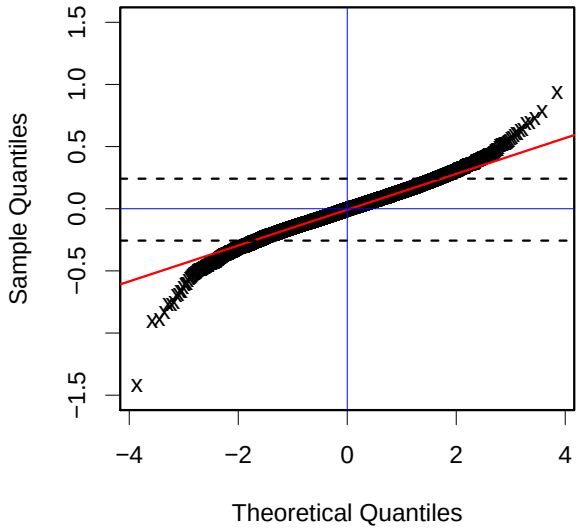

Figure 4: Normal QQ-plot of the distribution of Δh, using the same data as in Fig. 3. The dashed lines correspond to the 5% and 95% percentiles of the sample distribution, respectively.

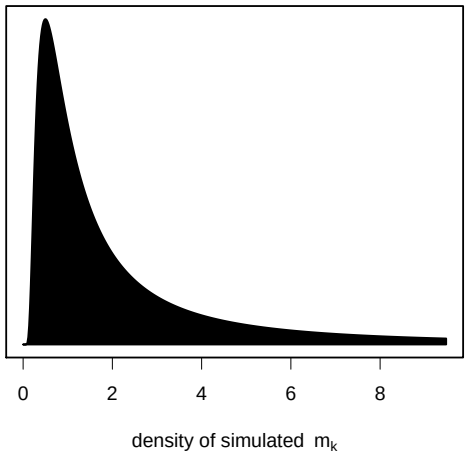

Figure 5: Density of the reciprocal Γ distribution, $1/m_k \sim \Gamma(1,1)$. At the right end, the plot is cut off at the 90%-quantile of the distribution.

Values for the calibration parameters a_i and b_i were generated through

$$\begin{aligned} a_1 &= 0, & a_2, \ldots, a_d &\overset{\text{iid}}{\sim} U(-\Delta a, \Delta a) \\ b_1 &= 1, & b_2, \ldots, b_d &\overset{\text{iid}}{\sim} LN(0,1), \end{aligned}$$

where $\Delta a = 0.95$ roughly corresponds to the peak of the distribution shown in Fig. 5, $U(-\Delta a, \Delta a)$ is the uniform distribution on the interval $[-\Delta a, \Delta a]$, and $LN(0,1)$ is the log-normal distribution that corresponds to the standard normal distribution. This yielded simulated data $y_{ki} = a_i + b_i \sinh(\mu_{ki} + \varepsilon_{ki})$.
The matrix y_{ki} of simulated data was presented to the software implementation in the Bioconductor [14] package `vsn`. It returns the estimated transformations $\hat{h}_1, \ldots, \hat{h}_d$, parameterized by $\hat{a}_1, \ldots, \hat{a}_d$ and $\hat{b}_1, \ldots, \hat{b}_d$ (see Eqn. (10)), as well as the matrix of transformed data $\hat{h}_{ki} = \operatorname{arsinh}((y_{ki} - \hat{a}_i)/\hat{b}_i)$. Generalized log-ratios were calculated as

$$\Delta\hat{h}_{ki} = \hat{h}_{ki} - \hat{h}_{k1} \tag{29}$$

and compared to the true values

$$\Delta h_{ki} = h_{ki} - h_{k1}, \tag{30}$$

simulation		A	B	C	D
number of probes	n	$384, \ldots, 69120$	9216	9216	9216
number of samples	d	2	$2, \ldots, 64$	2	2
proportion of differentially expressed genes	p_{diff}	0	0	$0, \ldots, 0.6$	0.2
proportion of up-regulated genes	p_{up}	-	-	$0.5, 1$	$0, \ldots, 1$
amplitude of differential expression	$z_{\max}$	-	-	2	2
trimming quantile	q_{lts}	$0.5, 0.75, 1$			
asymptotic coefficient of variation	c	0.2			

Table 1: Simulation parameters.

with $h_{ki} = \mu_{ki} + \varepsilon_{ki}$, by means of the root mean squared deviation

$$\delta = \sqrt{\frac{1}{|\kappa|(d-1)} \sum_{i=2}^{d} \sum_{k \in \kappa} (\Delta \hat{h}_{ki} - \Delta h_{ki})^2}, \tag{31}$$

where κ is the set of k for which $p_k = 0$.
For a given set of simulation parameters, this procedure was repeated multiple times, resulting in a simulation distribution of the root mean squared error δ. This was used to obtain the error bars shown in Figs. 6, 8, and 9. The error bars are centered at the mean and extend by twice the standard error of the mean in each direction.

5.3 Simulation results

Four series of simulations were performed to investigate the influence of the number of probes, the number of samples, the proportion of differentially expressed genes, and the choice of the trimming quantile q_{lts}. The parameter settings are summarized in Table 1.
The dependence of the estimation error δ on the number of probes n and the number of samples d was investigated in simulation series A and B. The results are shown in Fig. 6. In the left plot, the number of probes n varies from 384 to 69120. From the plot, a scaling of the root mean squared error approximately as

$$\delta \propto \frac{1}{\sqrt{n}} \tag{32}$$

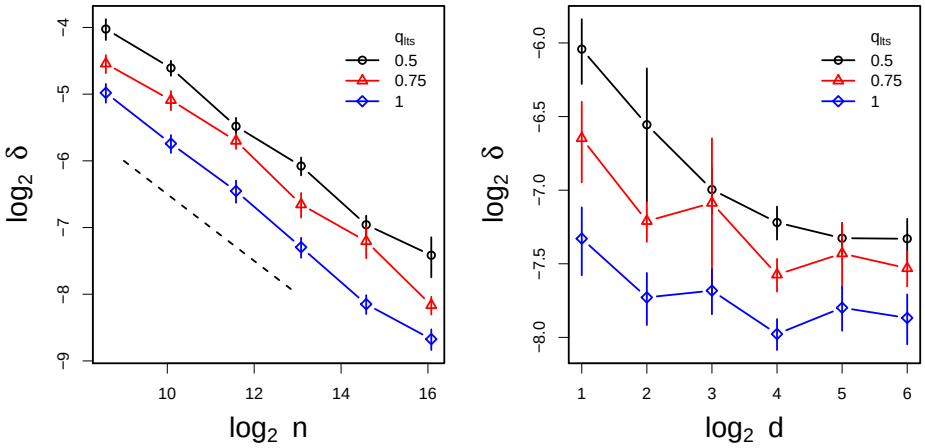

Figure 6: Dependence of the estimation error δ on the number of probes n (left panel) and the number of samples d (right panel), for three different choices of the trimming quantile q_{lts}. The dashed line in the left panel has a slope of $-1/2$ and corresponds to a scaling $\delta \propto 1/\sqrt{n}$.

can be observed. In the right plot, d varies from 2 to 64, again with three different values of q_{lts}. While δ does slightly decrease with d, the decrease is much slower than that with n and does not show an obvious scaling such as (32). The difference between the two plots may be explained by the fact that the number of parameters of the transformations (10) is $2d$. Thus, the number of data points per parameter remains constant when d is increased, but increases proportionally when n is increased.

The dependence of the required computation time on the parameters n, d and q_{lts} is shown in Fig. 7. The plots indicate a scaling approximately as

$$t_{\text{CPU}} \propto n \times d. \tag{33}$$

The computation times were measured with the Bioconductor package vsn version 1.0.3 and R version 1.6.1 on a DEC Alpha EV68 processor at 1 GHz. On this system, the proportionality factor in (33) is about 2 ms.

The effect of the presence of differentially expressed genes on the estimation error δ was investigated in simulation series C and is shown in Fig. 8. With $q_{\text{lts}} = 1$, i. e., without use of the robust trimming criterion, δ becomes large even in the presence of only a few differentially expressed genes. As p_{diff} increases, the estimation error remains smaller with trimming at the median ($q_{\text{lts}} = 0.5$) than at the 75%

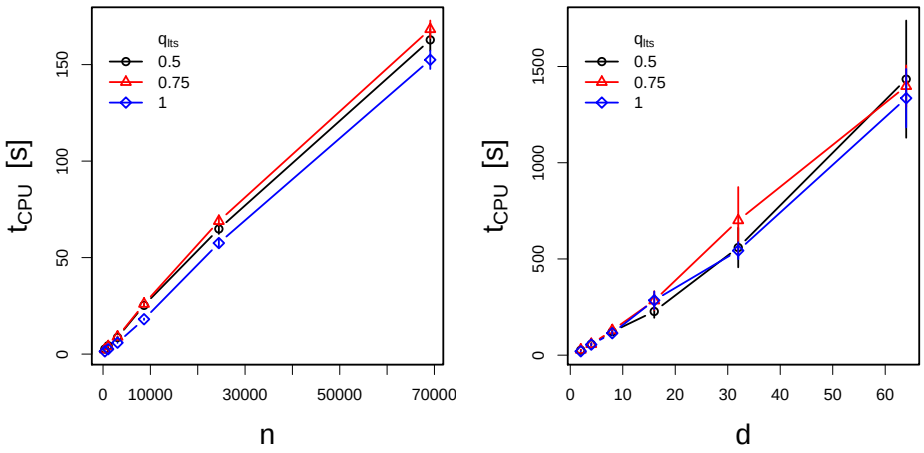

Figure 7: Dependence of the computation time t_{CPU} on the parameters n, d, and q_{lts}. Parameter values were as in Fig. 6.

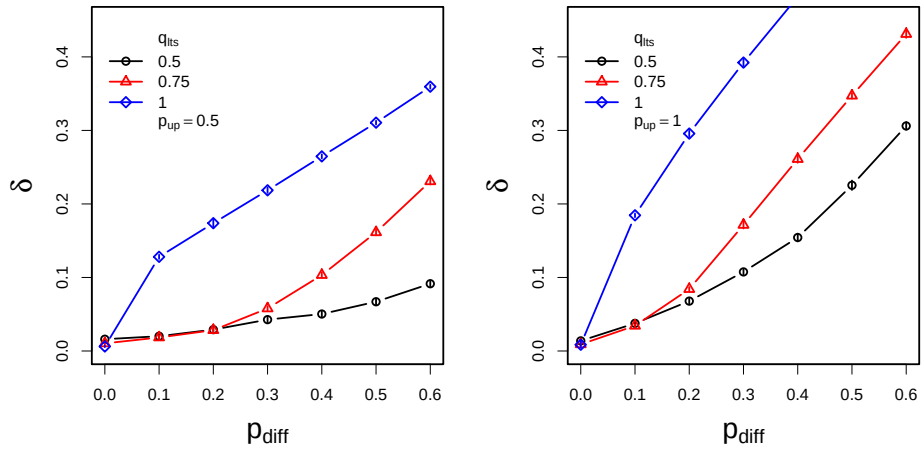

Figure 8: Estimation error δ for different mixture proportions p_{diff} of differentially expressed genes and three different choices of the trimming quantile q_{lts}. In the left plot, half of the differentially expressed genes are up-regulated while the other half is down-regulated. In the right plot, all of the differentially expressed genes are up-regulated.

quantile ($q_{\text{lts}} = 0.75$). Asymmetric situations ($p_{\text{up}} = 1$, right panel) are worse than symmetric ones ($p_{\text{up}} = 0.5$, left panel), but still can be handled reasonably as long as the proportion of differentially expressed genes is not too large.

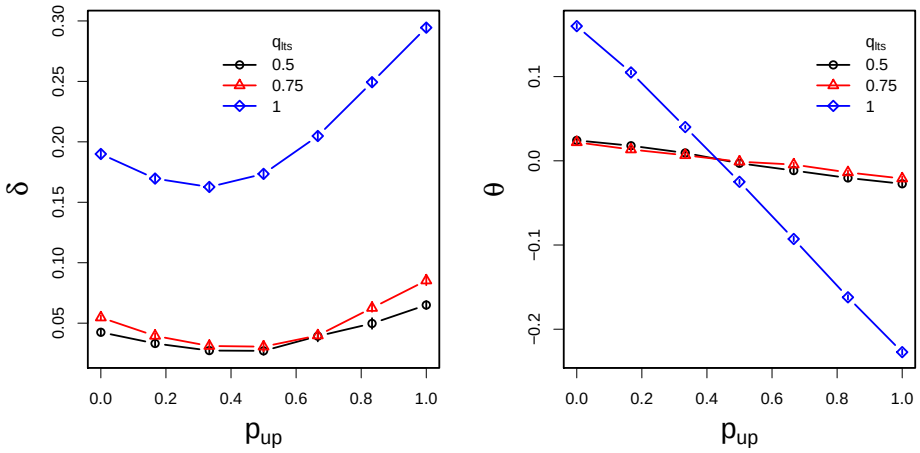

Figure 9: Estimation error δ and bias θ for different values of the asymmetry parameter p_{up}. Among the set of differentially expressed genes, p_{up} is the fraction of up-regulated and $1 - p_{\text{up}}$ the fraction of down-regulated genes.

Another look at the influence of asymmetry between up- and down-regulated genes is shown in Fig. 9, which was obtained from simulation series D. Here, p_{diff} was fixed at 0.2. The right panel shows the average bias

$$\theta = \frac{1}{|\kappa|} \sum_{k \in \kappa} \left(\Delta \hat{h}_{k2} - \Delta h_{k2} \right). \tag{34}$$

The robust procedures ($q_{\text{lts}} = 0.5$ and 0.75) perform much better than the unrobust one ($q_{\text{lts}} = 1$). Bias θ and error δ are smallest when the fractions of up- and down-regulated genes are about the same. An interesting feature of Fig. 9 is the asymmetry of the plots about the vertical line $p_{\text{up}} = 0.5$. The presence of up-regulated genes has a somewhat stronger effect on δ and θ than that of down-regulated ones. This appears to be related to the skewness of the distribution of μ_k (see Eqn. (27)), and to the curvature of the function (10).

6 Discussion

Models are never correct, but they may be useful. There is a trade-off between the wish to describe as much detail as possible and the problem of overfitting. Model (17) has two parameters per sample, an offset a_i and a normalization factor b_i, and one parameter for the whole experiment, the standard deviation σ_ε. In addition, we assume that for each non-differentially expressed gene k the center of the distribution of intensities is parameterized by μ_k for all samples i. Many extensions and variations of the model are possible. Instead of offsets and normalization factors that are common for all measurements from a sample, print-tip-specific or plate-specific parameters may be used [15]. Systematic differences between different production batches of arrays or reagents could be modeled by allowing for different values of μ_k in the different batches. Furthermore, we have assumed that the standard deviations of the additive noise terms for different samples are related to each other via $\mathrm{Sd}(\nu_{ki}) = \text{const}$. We find this to be an acceptable approximation for the data we have encountered, but other relationships, such as $\mathrm{Sd}(\bar{\nu}_{ki}) = \text{const}$. or $\mathrm{Sd}(\nu_{ki}) = \lambda_i$ with further parameters λ_i could also be appropriate.

Error modeling and calibration depend on the particularities of the technologies used. For this article, we have simply considered the probe intensities as given, setting aside important questions such as how the labeling and detection are realized and how the probe intensities are obtained from the fluorescence images. Whether or not the intensities measured from an experiment accord to the assumptions laid out in Section 2 has to be verified case by case; however, we have generally found good agreement with data from two-color spotted cDNA arrays, from radioactive nylon membranes, and from Affymetrix genechips. An advantage of the modeling approach compared to a heuristic, algorithm-oriented approach is that it provides criteria for quality control: by explicitly stating the assumptions made on the data, insufficient data quality can be detected by statistical tools such as residual analysis.

On cDNA arrays, typically each probe is sensitive for a distinct gene transcript. The calibrated and transformed intensities may be directly used as a measure for the abundance of transcripts in the samples. On Affymetrix genechips, multiple oligonucleotides of potentially different specificity and sensitivity probe for the same transcript. There, we recommend to use our method on the individual probe intensities. Approaches to the question of how these then can be combined into per-gene summary values are described in the references [16, 17, 18].

The computation time consumed by our software implementation is generally too large for interactive use. For example, with $n = 40000$ probes, $d = 100$ samples and a proportionality factor of 2 ms in Eqn. (33), $t_{\mathrm{CPU}} \approx 2\text{h}$. The time-critical part of the computations is the iterative likelihood optimization. There are two ways to

reduce the time: First, the proportionality factor could be reduced by implementing code in C instead of R. Second, sublinear scaling can be achieved instead of Eqn. (33) by not using the data from all probes, but only from a (quasi-)random subset.

From the point of view of the user, two important issues are interpretability and bottom-line performance. In these respects, the method presented here needs to be compared against established approaches that are based on the logarithmic transformation and the calculation of (log-)ratios. A comparison on real data showed higher selectivity and sensitivity for the idenfication of differentially expressed genes [3]. The (log)-ratio has, at first sight, the advantage that it can be simply and intuitively interpreted in terms of "fold change". However, the value of the log-ratio is highly variable or may be undefined when either the numerator or the denominator are close to zero. Since many microarray data sets include genes that are not or only weakly expressed in some of the conditions of interest, the significance of fold changes can be difficult to assess for a large and potentially important part of the data. Furthermore, many authors have noted a need for non-linear normalization transformations to be applied in conjunction with the log-transformation [15, 18, 19, 20]. These have the goal of removing an intensity dependent bias from the log-ratios. They are often implemented through a scatterplot smoother or a local regression estimator. The result of that has no longer a simple interpretation in terms of fold changes.

The approach presented in this paper offers a rational and practicable solution to these problems. Through the criterion of variance stabilization, we arrive at a transformation that corresponds to the logarithm when the intensity is well above background, but has a smaller slope for intensities close to zero. Thus, the generalized log-ratio Δh coincides with the usual log-ratio $\Delta \log$ when the latter is meaningful, but is shrunk towards zero when the numerator or denominator are small. Non-linear calibration transformations are often motivated by the curvilinear appearance of the scatterplot on the log-log scale. This may be caused, for example, by differences in the overall background between different arrays or color channels. In that case, these differences may also be modeled by the family of transformations (10), which allow different offsets a_i for different arrays or colors i. An advantage of this over calibration by local regression is that it does not depend on the choice of a smoothing bandwidth, and that the offsets a_i and scaling factors b_i are easier to interprete.

While our approach approximately removes the intensity-dependence of the variance, this does not necessarily mean that the variance of the data for all genes is the same. There may still be technical or biological reasons for the variance of one gene being different from that of another, or even from itself under a differ-

ent biological condition. For instance, the tightness of regulatory control could be different for a highly networked transcription factor than for a protein with mainly structural function. However, whether or not such gene- or condition-specific variances play a role, in any case the removal of the intensity-dependence should be advantageous for subsequent analyses.

Acknowledgements

We thank Günther Sawitzki and Dirk Buschmann for fruitful discussions, and Robert Gentleman and Jeff Gentry for useful suggestions and practical advice on the software implementation of the Bioconductor package.

References

[1] Blythe P. Durbin, Johanna S. Hardin, Douglas M. Hawkins, and David M. Rocke. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, 18 Suppl. 1:S105–S110, 2002. ISMB 2002. 1

[2] Peter Munson. A ”consistency” test for determining the significance of gene expression changes on replicate samples and two convenient variance-stabilizing transformations. Genelogic Workshop on Low Level Analysis of Affymetrix Genechip data, http://stat-www.berkeley.edu/users/terry/zarray/Affy/GL_Workshop/genelogic2001.html, 2001. 1

[3] Wolfgang Huber, Anja von Heydebreck, Holger Sültmann, Annemarie Poustka, and Martin Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl. 1:S96–S104, 2002. ISMB 2002. 1, 3, 19

[4] T. Ideker, V. Thorsson, A.F. Siegel, and L.E. Hood. Testing for differentially expressed genes by maximum-likelihood analysis of microarray data. *Journal of Computational Biology*, 7:805–818, 2000. 1

[5] David M. Rocke and Blythe Durbin. A model for measurement error for gene expression analysis. *Journal of Computational Biology*, 8:557–569, 2001. 1, 3

[6] Wolfgang Huber. Vignette: Robust calibration and variance stabilization with vsn, 2002. The bioconductor project, http://www.bioconductor.org. 1

[7] Tim Beißbarth, Kurt Fellenberg, Benedikt Brors, Rose Arribas-Prat, Judith Maria Boer, Nicole C. Hauser, Marcel Scheideler, Jörg D. Hoheisel, Günther Schütz, Annemarie Poustka, and Martin Vingron. Processing and quality control of DNA array hybridization data. *Bioinformatics*, 16:1014–1022, 2000. 3

[8] David M. Rocke and Blythe Durbin. Approximate variance-stabilizing transformations for gene-expression microarray data. *Bioinformatics*, in press, 2002. 5

[9] Xiangqin Cui, M. Kathleen Kerr, and Gary A. Churchill. Data transformations for cDNA microarray data. Technical report, The Jackson Laboratory, http://www.jax.org/research/churchill, 2002. 5

[10] Peter J. Rousseeuw and Annick M. Leroy. *Robust Regression and Outlier Detection*. John Wiley and Sons, 1987. 5, 7

[11] Susan A. Murphy and A. W. van der Vaart. On profile likelihood. *Journal of the American Statistical Association*, 95:449–465, 2000. 6

[12] Peter J. Rousseeuw and Katrien van Driessen. Computing LTS regression for large data sets. Technical report, Antwerp Group on Robust & Applied Statistics, 1999. http://win-www.uia.ac.be/u/statis/abstract/Comlts99.htm. 8

[13] M.A. Newton, C.M. Kendziorski, C.S. Richmond, F.R. Blattner, and K.W. Tsui. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, 8(1):37–52, 2001. 11

[14] Ross Ihaka and Robert Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996. http://www.bioconductor.org. 13

[15] Sandrine Dudoit, Yee Hwa Yang, Terence P. Speed, and Matthew J. Callow. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12:111–139, 2002. 18, 19

[16] Affymetrix, Santa Clara, CA. *Microarray Suite Version 5.0, User's Guide*. 18

[17] Cheng Li and Wing Hung Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *PNAS*, 98:31–36, 2001. 18

[18] Rafael A. Irizarry, Bridget Hobbs, Francois Collin, Yasmin D. Beazer-Barclay, Kristen J. Antonellis, Uwe Scherf, and Terence P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 2003. Accepted for publication. http://biosun01.biostat.jhsph.edu/∼ririzarr/papers/index.html. 18, 19

[19] Eric E. Schadt, Cheng Li, Byron Ellis, and Wing Hung Wong. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry*, Supplement 37:120–125, 2001. 19

[20] Thomas B. Kepler, Lynn Crosby, and Kevin T. Morgan. Normalization and analysis of DNA microarray data by self-consistency and local regression. *Genome Biology*, 3(7):research0037.1–0037.12, 2002. 19