# Centralization: a new method for the normalization of gene expression data

Alexander Zien, Thomas Aigner, Ralf Zimmer and Thomas Lengauer

SCAI - Institute for Algorithms and Scientific Computing, GMD - German National Research Center for Information Technology, Schloss Birlinghoven, Sankt Augustin, 53754, Germany and Department of Pathology, University of Erlangen-Nürnberg, Krankenhausstr. 8-10, Erlangen, 91054, Germany

Received on February 5, 2001; revised and accepted on March 30, 2001

## ABSTRACT

Microarrays measure values that are approximately proportional to the numbers of copies of different mRNA molecules in samples. Due to technical difficulties, the constant of proportionality between the measured intensities and the numbers of mRNA copies per cell is unknown and may vary for different arrays. Usually, the data are normalized (i.e., array-wise multiplied by appropriate factors) in order to compensate for this effect and to enable informative comparisons between different experiments. Centralization is a new two-step method for the computation of such normalization factors that is both biologically better motivated and more robust than standard approaches. First, for each pair of arrays the quotient of the constants of proportionality is estimated. Second, from the resulting matrix of pairwise quotients an optimally consistent scaling of the samples is computed.

**Contact:** Alexander.Zien@gmd.de

## INTRODUCTION

Microarrays, like any other technology for measuring RNA expression levels, are subject to errors that arise from various sources. According to their nature, these errors counteract one of the two main attributes of accuracy: *precision* or *correctness*. If a source of error is modeled as a random variable, it is the variance that hampers precision, while a non-zero expectation value impedes correctness by introducing a bias.

**Precision** is dealt with in a variety of ways. First of all, much work is devoted to the development and manufacturing of more precise hardware, including robots and sensors. Computational contributions to increasing the precision of measurements include oligonucleotide selection strategies; precise image analysis; determination and appropriate computational treatment of background intensity; compensation for non-linearity of signal intensity; and more. Generally, the highest noise to signal ratios are observed for low intensities. An important way to improve the significance of measurement values is to perform replicates of experiments, see e.g. (Lee *et al.*, 2000).

**Correctness** is closely related to the question of the interpretation of the data: how can signal intensities be translated into levels of mRNA within the cells? This fundamental challenge equally applies to arrays, Northern blotting experiments and quantitative PCR (qPCR). The relation of measured intensities to the cellular amounts of mRNA molecules is obscured by multiplicative noise (i.e., errors proportional to the measured value) that leads to systematic inconsistencies. The most important sources of such problems are listed in Table 1. Some of these influences lead to systematic biases with respect to genes or spots, see e.g. Li & Wong (2001), while others lead to different scalings for the individual arrays or samples. Many expression analyses are not disturbed by gene-specific multiplicative errors, since they are easily eliminated by taking ratios between different samples.

This paper deals with the proper treatment of array-dependent incorrectness. Obviously, it is important to control this type of error, since it may otherwise lead to false conclusions about the regulation of individual genes in different cellular conditions (as measured from different samples). In fact, normalization methods have been standard procedures in mRNA biology from its very beginning. Also, improvements to the technology are suggested, like furnishing arrays with control spots that can help to compensate for spot variability (Schuchhardt *et al.*, 2000). Most of the error-introducing variables might potentially be controllable with the help of internal standards, which, however, are not yet readily available for laboratory use (Ke *et al.*, 2000; Vu *et al.*, 2000). However, this approach is laborious and costly, and it still might be impossible to account for all variables at the same time.

There are three practical approaches to normalization that are presently in common use: the total RNA approach, the housekeeping gene approach, and the globalization

**Table 1.** Overview of contributions to multiplicative gene expression measurement errors. Errors that depend on the sample or the experimental protocol are called array- or sample-dependent, since they affect all spots/genes of an array equally. Errors depending on sequence (either the expressed mRNA or the spotted cDNA/oligonucleotide) are called gene-dependent, for short; they vary in magnitude on the same array. Additive errors are assumed to be negligible or to be corrected for by subtraction of background intensity by the image analysis software. When the true interest is protein levels, additional gene-dependent effects enter: translation efficiency, efficiency of post-translational modifications, transport efficiency and average protein life span.

| variable | depends on | | | remedy possible |
|---|---|---|---|---|
| | gene / spotted sequence | sample | experimental protocol | |
| number of cells in sample | - | ++ | - | in some cases (cell counting) |
| RNA isolation efficiency | ? | + | ++ | principally yes (e.g. internal standards) |
| RT / labelling efficiency | + | - | ++ | in part (e.g. internal standards) |
| spot size and density | ++ | - | - | yes (e.g. multiple standardized spotting) |
| hybridization / washing efficiency | ++ | - | ++ | yes (two-channel measurements) |
| exposure time; detection sensitivity | - | - | ++ | trivial |

**Table 2.** Overview over common normalization methods in comparison to the proposed method, centralization.

| method | assumption | scaling of expression levels |
|---|---|---|
| **total RNA** | Constant expression of total RNA (or, almost equivalently, ribosomal RNA). | Use fixed amount of total RNA for measurements. |
| **housekeeping** | Housekeeping genes are constitutively expressed (i.e. at constant level). | Divide by intensity of housekeeping genes. |
| **globalization** | The total number of mRNA molecules per cell is constant. | Divide by sum (or mean) of all intensities. |
| **ANOVA** | Different models are possible. Errors are always normally distributed. | Leads to globalization or related procedure. |
| **centralization** | Regulation is well-behaved (e.g., most genes are not significantly regulated OR about equal numbers of genes are up- and downregulated). | Find most probable consistent pairwise scaling based on central tendency of expression ratios. |

approach. Each approach is based on an assumption about cellular gene expression. In particular, in each case some population of RNA molecules is assumed to be present at a constant level in all investigated cells. Therefore, this population can serve as a biological internal standard. An overview over the discussed normalization methods is given in Table 2.

The **total RNA approach** rests on the assumption that, at every time point, each cell carries the same amount of total RNA. More than 90% of the total RNA is 18S and 28S ribosomal RNA (rRNA), which was believed by some people to be constitutively expressed, even when the amount of mRNA varies. Consequently, by using a fixed amount of total RNA for measurements, the first problem mentioned in Table 1 would be circumvented. However, it is erroneous to assume that total RNA levels or rRNA levels are constant (Suzuki *et al.*, 2000). By now, it is well known that different cell types and cells in different conditions produce different amounts of total (and ribosomal) RNA, ranging from less than 2mg to more than 100mg total RNA per $10^9$ cells.

The **housekeeping gene approach** assumes that the expression of housekeeping genes, e.g. GAPDH or $\beta$-actin, is not significantly regulated. This approach has been used in molecular biology for over two decades now (in Northern blotting and PCR experiments etc.). However, it becomes more and more clear that this assumption is wrong (Suzuki *et al.*, 2000; Velculescu *et al.*, 1999; Goldsworthy *et al.*, 1993), although regulation of these genes appears to be low compared to other genes. In fact, for defined cell types analyzed in rather comparable cell states (e.g. isolated cells of a certain type with and without stimulation) it might still be a suitable method (in particular for techniques which do not allow to determine gene expression levels for a high number of different genes in parallel, such as qPCR, Northern blotting and RNAse-protection assays). Overall, however, if one is interested in smaller changes in gene expression levels (less than ten-fold) or in comparing rather different probes, which is the case in many applications, the housekeeping gene approach no longer can be considered appropriate.

The **globalization method** is the most commonly used normalization heuristic in large-scale gene expression biotechnology: for each array, all measured values are divided by their sum (or average). Such a kind of protocol is, among others, implemented in the array analysis program (AtlasImage 1.101 by ClonTech, Germany) that was used for generation and primary evaluation of the data that are discussed in the results. A similar approach of estimating total RNA was suggested for Northern Blotting experiments by probing with a poly-dT probe for total mRNA (Goldsworthy *et al.*, 1993). The globalization method implicitly rests on the assumption that the amount of mRNA per cell is constant. This assumption is theoretically questionable for several reasons. First, adding the intensities of different genes is not per se meaningful since they occur at different scales due to gene-dependent multiplicative errors (c.f. Table 1). Second, and of higher practical importance, often the sum of all expression signals is dominated by the strongest signals (Velculescu *et al.* (1999); see also results). But strongly expressed genes are most likely to be regulated as they represent the major expression products of specialized cells, e. g. immunoglobulin chains for plasma cells, hemoglobin for erythroblasts, etc. Finally, as long as not the whole genome is covered by an array, the question remains whether the set of screened genes forms a representative sample that resembles the behavior of the entire transcriptome.

Recently, **ANOVA** (analysis of variance) was introduced to the field of expression data analysis (Zolotukhin & Lange, 2000; Kerr *et al.*, 2001). ANOVA is a general purpose data analysis technique that leaves certain degrees of freedom for this specific application. The expression measurement can be modeled differently, for example with respect to the potential sources of error. For example, different models are necessary for one- and two-channel arrays. The application of ANOVA to single-channel measurements as described in (Zolotukhin & Lange, 2000) recovers the usual globalization method. Saturated ANOVA models of the type described in (Kerr *et al.*, 2001) lead to a kind of globalization of log-intensities.

Given the limitations of the available solutions, a more sophisticated but still feasible normalization method is of high priority. An appropriate normalization method should fulfill the following criteria: (i) it should not be based on single or few pre-defined genes (as often no specific gene can be named that is expressed at a constant absolute number of mRNA copies per cell), rather it should take into account that a large and variable set of genes can be significantly regulated; (ii) it should only be based on fairly reliable measurements (to ensure statistical soundness); (iii) it should be invariant with respect to the absolute values of detected signals. In this paper, we describe a new algorithm for computing normalization factors that satisfies these requirements.

The proposed method — called **centralization** — rests on the weak assumption that the regulation of gene expression is *well-behaved*. By this we mean that either of two conditions is fulfilled: (i) most genes are not or only moderately regulated; or (ii) approximately equal numbers of genes are up-regulated as are downregulated. In most situations, this seems to be the case, with cell activation being a notable exception. Given this assumption, the central tendency of the expression level ratios of two measurements is a good estimate of the their relative scaling, i.e. the quotient of the experiment-dependent multiplicative error. After computing probability distributions for the pairwise scaling for every pair of measurements, we employ a maximum likelihood approach to find the most probable consistent scaling vector. These factors can then be applied to enable a meaningful comparison of all the measurements among each other.

## METHODS

For the description of our procedure we will use the following notation and conventions. We consider the measurement of $n$ biological samples $K = \{k_1, \ldots, k_n\}$. For clarity we assume that each sample $k \in K$ is measured once and with a separate array. Generalizations to other settings are straightforward. E.g., it is common to perform multiple measurements (on multiple arrays) of the same sample in order to reduce the measurement error. Such replicate measurements should first be brought to scale by normalizing. Here, stronger assumptions may hold, since the samples are identical. Then, averages can be computed and subsequently be treated as a single virtual array with reduced variances of the associated error terms.

For simplicity we also assume that for each sample the same set $G = \{g_1, \ldots, g_{|G|}\}$ of $|G|$ different genes is measured. Multiple measurements of a gene on the same array can most easily be modeled by a single average intensity value and an appropriately reduced associated variance. If different clones or oligonucleotides are used for populating the redundant spots, it is again advantageous to rescale the measured values before averaging. Proper treatment of the case that some genes are measured for some but not all samples will be explicitly described below.

Let $l^*_{g,k}$ denote the true expression level of gene $g$ in sample $k$, i.e. $l^*_{g,k}$ is the (average) number of mRNA copies per cell in the sample. Let $m^*_{g,k}$ denote the measured value, i.e. a number proportional to the signal intensity measured at the spot for $g$ on the array for sample $k$. In this paper, we assume that the main sources of measurement error can be modeled by three independent terms and a small residual $e^{res}$ as follows:

$$m^*_{g,k} = b_{g,k} + c_k d_g l^*_{g,k} + e^{res} \tag{1}$$

Here, $b_{g,k}$ is the so-called background noise. We assume that reasonable estimates of $b_{g,k}$ are supplied by the image analysis software. The background noise can, e.g., be estimated globally ($b_{g,k} := b$), per array ($b_{g,k} := b_k$) by measuring blind spots on the array, or individually for each spot (gene) on each array ($b_{g,k}$), by measuring the signal on the area surrounding the hybridization spots. $c_k d_g$ quantifies the constant of proportionality between the the measured intensity value and true number of mRNA copies per cell. We assume that this multiplicative error can be separated into one part that depends only on the array ($c_k$), and a second part that depends only on the gene ($d_g$). Of course, there may be some residual error $e^{res}$ left over, which may depend on all of $(g, k, l^*_{g,k})$. However, since the most important known sources of error are explicitly modeled by the other terms, $e^{res}$ can be expected to be relatively small.

## Pairwise scalings

Let $k_i, k_j \in S$ be two samples, with unknown constants of proportionality $c_i, c_j$ (short for $c_{k_i}, c_{k_j}$). Thus, the true quotient of the constants of proportionality of the two samples is $q^*_{i,j} := \frac{c_i}{c_j}$. Let $G_{i,j}$ be the set of genes that are considered to be expressed and reliably measured (in the linear part of the dynamic range) in both samples. The other genes are excluded, since ratios of values that are dominated by background noise (as well as saturated intensities) are incorrectly biased towards one. In order to estimate $q^*_{i,j}$, we will use the set of quotients

$$Q_{i,j} := \left\{ q_g \mid q_g := \frac{m^*_{g,k_i} - b_{g,k_i}}{m^*_{g,k_j} - b_{g,k_j}}, g \in G_{i,j} \right\}.$$

The idea is to regard each of the ratios $q_g \in Q_{i,j}$ as an estimate of $q^*_{i,j}$; a related approach was also suggested by (Chen *et al.*, 1997). These estimates are subject to two sources of noise: first, the residual error and, second, the amount of gene expression regulation. We assume both errors to be unbiased. This is relatively safe for the residual error, since the main sources of bias are explicitly taken into account by the other error terms. For the biology, the absence of bias comes back to our main assumption that gene regulation is 'well-behaved'. In particular, the two alternative conditions (see introduction) map well to two properties of the distribution of the values $q_g$.

**Condition (i)**: If the number and extent of upregulation of genes in $G_{i,j}$ in $k_i$ in comparison to $k_j$ is similar to number and extent of downregulation, the distribution is symmetric. At least in the case of the absence of any residual error $e^{res}$, the median of $Q_{i,j}$ is a perfect estimate of $q^*_{i,j}$. This holds even if the average fold change is asymmetric with respect to up- and downregulation. **Condition (ii)**: If more than half of the genes in $G_{i,j}$ are expressed at the same levels in both samples $k_i$ and $k_j$, the distribution of ratios is sharply peaked around $q^*_{i,j}$ and thus unimodal. Again, the median is a perfect estimate of $q^*_{i,j}$, at least for $e^{res} = 0$.

In general, any measure of central tendency of the values in $Q_{i,j}$, including median, mean, trimmed mean and weighted mean, may yield a sensible estimate $\hat{q}_{i,j}$ for $q^*_{i,j}$. However, care must be taken whenever values are averaged: in order to keep the symmetry of up- and downregulation, the arithmetic mean should be computed in the space of log ratios. Recently, this approach was also suggested by (Beißbarth *et al.*, 2000). Alternatively, the slope obtained by (possibly weighted) linear regression of the two sets of intensities may be used. Even a clustering of the (log) ratios can be considered.

The median of $Q_{i,j}$ is a particularly trustworthy estimate $\hat{q}_{i,j}$, since it is very robust with respect to outliers. In order to also obtain a robust estimation of the standard deviation, we apply the following iterative strategy: First, we compute empirical median and standard deviation $\hat{\sigma}_{i,j}$ of all log ratios, using the median as the estimate of the expected value $\hat{\mu}_{i,j}$. Second, we mark as outliers all values $q_g$ that show improbably high deviations with respect to the imposed normal distribution, using $0.5/|Q_{i,j}|$ as a threshold probability. As long as the set of outliers is changed and does not exceed 10% of $G_{i,j}$, we re-estimate $\hat{\mu}_{i,j}$ and $\hat{\sigma}_{i,j}$ on the set of non-outliers (instead of the entire set $Q_{i,j}$) and repeat the second step. Otherwise, we retain the latest estimates of mean and standard deviation. This procedure is guaranteed to provide symmetrical estimates, satisfying $\hat{\mu}_{i,j} = -\hat{\mu}_{j,i}$ and $\hat{\sigma}_{i,j} = \hat{\sigma}_{j,i}$. Later, the implied normal distribution $\mathcal{N}(\hat{\mu}_{i,j}, \hat{\sigma}^2_{i,j})$ will be used in order to estimate the probability of $q^*_{i,j}$ to take on a certain value.

Both estimating the mean by the median, estimating the standard deviation, and parameterizing a normal distribution to the log ratios only make sense if the values are approximately normal distributed. Most important, the empirical density function should be approximately unimodal and symmetric (thus the distribution function should be approximately sigmoid). In an effort to justify the log normal distribution theoretically, we follow a similar line of reasoning as (Chen *et al.*, 1997). They consider the stochastic fluctuations of the number of mRNA copies of each gene, that arise solely from randomness in molecular motions. In (Chen *et al.*, 1997), this is modelled by a normal distribution that is truncated to confine non-vanishing probability to positive numbers. This appears to lead to an asymmetric distribution of log ratios for the comparison of two probes in the same regulatory state, contrasting the symmetry of the situation.

Here, we assume that the number of mRNA copies of each species is Poisson distributed. The Poisson distribution arises naturally when events are counted that occur independently and with constant probability. However, the process of transcription in cells most probably has some amount of memory, e.g. due to mechanisms of re-initiation. But this situation may be seen as independent production events of batches of transcripts of typical size. Also, for an approximately fixed number of mRNA copies and if saturation is avoided, the number of hybridizations to the corresponding spot may be argued to follow a Poisson distribution. Finally, in contrast to the normal distribution, the Poisson assumption naturally assigns

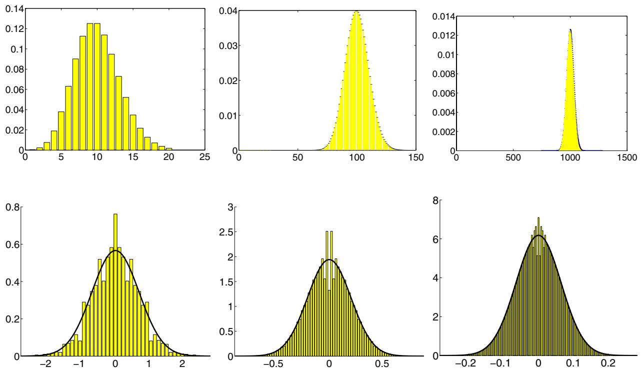

**Fig. 1.** Densities of Poisson variables (upper row) and distributions of log ratios of two such variables with superimposed normal distribution (lower row), shown for three different parametrizations: left, λ = 10; middle, λ = 100; right, λ = 1000. Y-axes: probability density. X-axes of upper figures: realization of Poisson variable (event count); of lower figures: base-two logarithms of ratios of two independent Poisson variables with equal expectation values λ.

positive probability to positive integer numbers (representing counts of events) only. Figure 1 shows example distributions of Poisson variables and the corresponding log ratios. Although the support of the log ratios is a set of measure zero, their distributions can be proved to converge against a normal distribution with zero mean and $2\lambda^{-1}$ variance for $\lambda \to \infty$ (Rudolf Grübel and Niklas von Öhsen, personal communication). Figure 1 shows this convergence to be extremely rapid and accurate for reasonable values of λ. Therefore, the log ratio densities of unregulated (Poisson distributed) genes will behave like a superposition of concentric normal densities, thereby directly leading to unimodality and symmetry.

By visual inspection of a large number of examples, we judge these conditions to be met to a sufficient extent, lending support to our basic assumption. Examples of both empirical values and the estimated parameterization of the normal distribution are shown in Figure 2. Even if the values are not perfectly normally distributed or if the fit is not accurate, the estimated mean and variance will supply a reasonable estimate $\hat{q}_{i,j} := \exp(\hat{\mu}_{i,j})$ of the relative scaling $q^*_{i,j}$ and a useful quantification $\sigma^2$ of the degree of confidence in that estimation.

## Most consistent sample scaling factors

We assume that by some means, e.g. in the fashion described above, we are able to obtain reasonable estimates $\hat{q}_{i,j}$ of the true quotients $q^*_{i,j}$ for any two sample measurements $k_i, k_j \in K$. Now the task is to determine a set of values $\{s_k | k \in K\}$, called *scaling factors*, such that

$$s_k c_k \approx \kappa \text{ for all } k \in K \qquad (2)$$

for some constant (but possibly unknown) $\kappa$. Given such values, we can make the measured expression level values $l_{g,k}$ mutually comparable between different samples $k$ by rescaling them accordingly via

$$(l_{g,k} - b_{g,k}) \to s_k(l_{g,k} - b_{g,k}),$$

because the resulting values are all subject to the same multiplier $\kappa$. I.e., while we still do not know the true number of mRNA molecules for a gene in the sampled cells, we obtain consistent multiples of this number.

Applying a maximum likelihood approach, we seek scaling factors that are most probable, given the estimated probabilities of pairwise scalings $\hat{q}_{i,j}$. Under the assumption of complete independence of all estimated probabilities $\mathcal{N}(\hat{\mu}_{i,j}, \hat{\sigma}^2_{i,j})$, the probability of a scaling vector $s$ can be written as:

$$P(s) = \prod_{i,j=1}^{n} P\left(q^*_{i,j} = \frac{s_i}{s_j}\right) \qquad (3)$$

This assumption is obviously violated, since the $\frac{1}{2}(n^2 - n)$ different pairs of parameters $(\hat{\mu}_{i,j}, \hat{\sigma}_{i,j})$ are computed from only $n$ independent sets of values. On the other hand, the dependencies among the values are symmetrical, thus we can expect to obtain a close to optimal solution when we maximize the product probability. Let $\hat{s} = \arg\max_s P(s)$. By inserting the log-normal distribution into Eq. 3, we get

$$\hat{s} = \arg\min_s \sum_{i,j=1}^{n} \left( \frac{\log \frac{s_i}{s_j} - \hat{\mu}_{i,j}}{\hat{\sigma}_{i,j}} \right)^2$$

A minimum can be found by setting all partial derivatives to zero, which is easier in the space of log ratios. Substituting $t_l := \log s_l$, we get

$$0 = \frac{\delta}{\delta t_k} \sum_{i,j=1}^{n} \left( \frac{t_i - t_j - \hat{\mu}_{i,j}}{\hat{\sigma}_{i,j}} \right)^2 = 4 \sum_{l \neq k} \frac{t_k - t_l - \hat{\mu}_{k,l}}{\hat{\sigma}^2_{k,l}}$$

for each $k = 1, \ldots, n$. Fortunately, the resulting equality system is linear in $t$ and can thus easily be solved by standard methods in $\mathcal{O}(n^3)$ time. However, the corresponding matrix has rank $n - 1$, since the value of $\kappa$ in Eq. 2 leaves one degree of freedom to the solution. In order to restore full rank, we can simply add any constraint on the absolute values of $t$ to any row of the matrix. Here, we choose $\sum_{k=1}^{n} t_k = 0$, keeping the normalized expression levels as close to the raw intensities as possible.

## RESULTS

At the moment, it is virtually impossible to finally validate any method for normalizing gene expression values since,

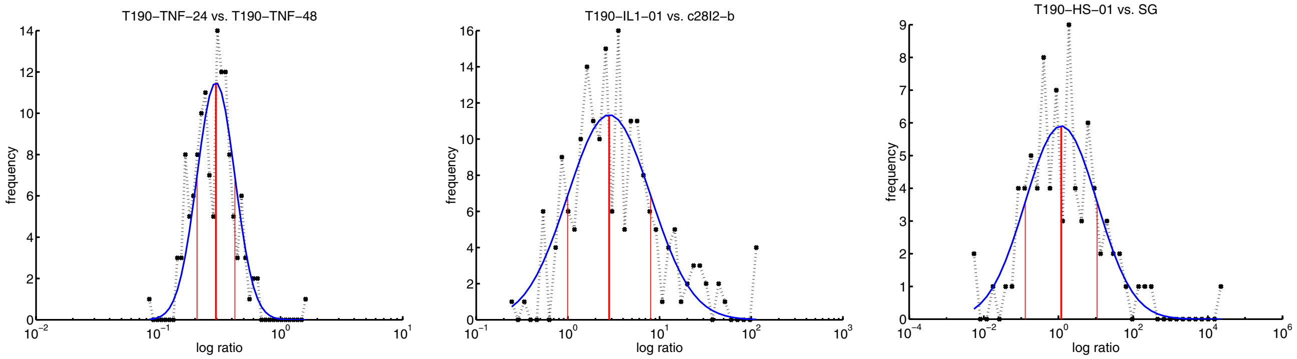

**Fig. 2.** Empirical distributions of $q_g$ for the example pairwise comparisons of samples and superimposed normal distributions. From left to right, showing the cases with the lowest, a typical (median), and the highest estimated variance observed on our data (see results). Estimated mean and standard deviation are indicated by vertical lines.

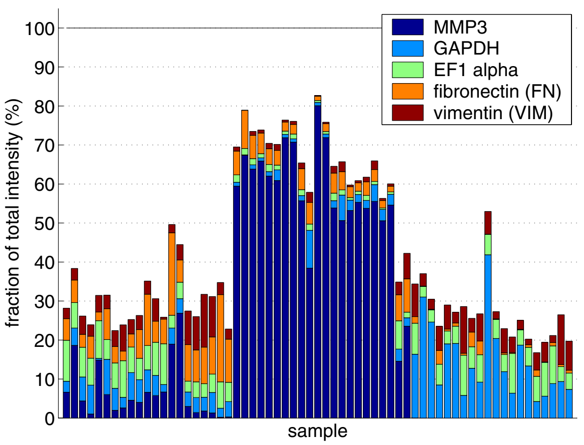

**Fig. 3.** Intensity fractions of those five genes that contribute most to the signal intensity. Left to right: 21 in vivo, 21 cultured cell, and 21 cell line samples.

currently, there is no practical way of obtaining the true numbers of mRNA molecules within the cell by direct measurement. However, we provide experimental evidence which supports centralization by demonstrating its superior robustness on a real life gene expression data set. The data are generated from a total of 89 gene expression measurements, several of which are replicate hybridizations of the same sample, each comprising the same set of 1185 genes as represented on the ClonTech Cancer 1.2 chip. Of the 63 samples, 21 are in vivo samples (Aigner *et al.*, 2001), 21 are human isolated primary chondrocytes cultured under various conditions (McKenna et al., in preparation) and the remaining 21 are different chondrocytic cell lines (transformed or neoplastic; Aigner et al., in preparation).

Figure 3 shows that the contributions of individual genes to the observed overall signal intensity differ dramatically between the different types of samples. In the in vivo samples, the examined five genes account for about 30% of the overall intensity, with quite large individual variations. In most of the cultured cell samples, the single gene MMP3 is responsible for more than 50% of the signal. In the cell lines, MMP3 does not constitute a significant fraction at all, whereas GAPDH accounts for the largest part of the overall signal.

Various analyses of the centralized data yield results that are both biologically plausible and consistent with previous observations, as will be detailed elsewhere. In the following sections, we try to show that alternative normalization methods are less appropriate for the data.

## Housekeeping

As can be seen in Figures 4 and 5, the so-called housekeeping genes are not synchronously regulated. This disproves the assumption that all housekeeping genes are expressed at a constant level. While the choice of a single housekeeping gene from the given set appears to be rather arbitrary, one could argue for a kind of housekeeping globalization. The assumption would be that the sum of all housekeeping genes remains constant within cells. However, most arguments against globalization (see introduction) also apply to this approach. Furthermore, a traditional procedure in this spirit would require a suitable set of housekeeping genes to be known in advance. The observed high variations suggest that no such fixed set exists. Thus, it would be most appealing to base normalization on a flexibly determined set of 'housekeeping' genes that appear to be unregulated as judged from the measurements. In fact, centralization can be interpreted to implement exactly this strategy.

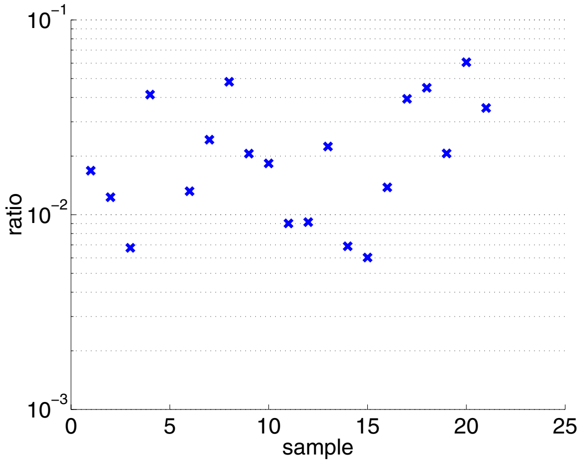

**Fig. 4.** Ratio of expression levels of the two most prominent housekeeping genes, $\beta$-actin over GAPDH, for the 21 in vivo samples. The range of relative regulation is more than 10-fold. Note that these ratios are independent from normalization.

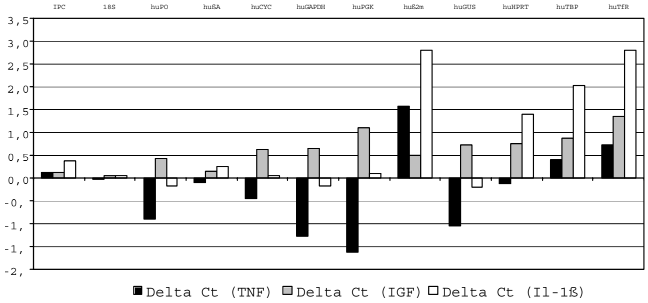

**Fig. 5.** Demonstration of the variation in expression levels of a panel of housekeeping genes (Perkin Elmers ‘Human endogenous control plate’) in three stimulation experiments as measured with the high-precision TAQMAN device. IPC: internal positive control; 18S: 18S rRNA; PO: acidic ribosomal protein; $\beta$A: $\beta$-actin; CYC: cyclophilin; GAPDH: GAPDH; PGK: phophoglycerokinase; $\beta$2m: $\beta$2-microglobulin; GUS: $\beta$-glucoronidase; HPRT: hypoxanthine ribosyl transferase; TBP: transcription factor IID, TATA binding protein; TfR: transferrin receptor. Y-axis: difference in number of PCR rounds, corresponding to $\log_2$ ratio of expression.

## Globalization

The basic assumption of globalization, constant amount of mRNA in cells, can not be disproved from gene expression data, in principle. However, we can show that its practical value is limited. In particular, the assumption that the 1185 genes assayed by the ClonTech arrays are representative in this context is shown to be very questionable.

For both globalization and centralization, we compute normalization factors for all samples based on 200 random subsets of genes. We repeat this for three different sizes of random subsets: one half, a quarter, and one eighth of the 698 genes that are observed at least one in vivo sample. For centralization, we exclude intensities below 10 and above 1000, since they appear to be unreliable. The computed scales are then set in relation to the normalization factors computed with the same method based on all 698 genes. As can be seen in Figure 6, for globalization the scalings depend heavily on the particular subset, indicating low reliability of the method. In contrast, the centralization strategy is much more stable with respect to gene selection. This suggests that the resulting scalings should lead to more reproducible and biologically meaningful results.

For our data, globalization is particularly sensitive to whether MMP3 is included in the set of examined genes, as is manifest in the bifurcation in Figure 6. Thus, at least for chondrocytes, globalization seems completely inappropriate for the comparison of in vivo probes to cultured cells (e.g. serving as disease models) which is a pharmaceutically relevant application of gene expression profiling. In order to evaluate globalization in a simpler setting, we repeated the above experiment on the in vivo samples only; see Figure 6. Again, centralization shows superior reliability.

## DISCUSSION

Since a gold standard for the evaluation of normalization is unavailable, we demonstrated the robustness of centralization and the weaknesses of common competing methods on a real life gene expression data set. But still we believe that the strongest argument for centralization is that the approach is biologically well motivated.

First, there is supporting evidence that the traditional methods to normalize RNA-values contradict the biological rationale to a large extent, as explained in the introduction and substantiated by our results. All of these approaches are based on rather strong biological hypotheses which do not withstand scrutiny. However, as long as no direct measurements are possible, some assumption on biological properties is needed for calculation. Giving this, we have chosen a biological assumption which is substantially weaker, and based our method on the sole assumption that the regulation of genes in cells is well-behaved. This way, we are finally coming back to a kind of housekeeping approach, but not in terms of a fixed pre-defined set of ‘housekeeping’-genes. Instead, we employ a dynamically determined set of genes with sound evidence from the experimental data to behave in a largely non-regulated manner. Also, we eliminate distortions from gene-specific multiplicative errors by relying on ratios of intensities for the same gene.

One problem for all computational normalization methods remains the possibility of cell activation, im-

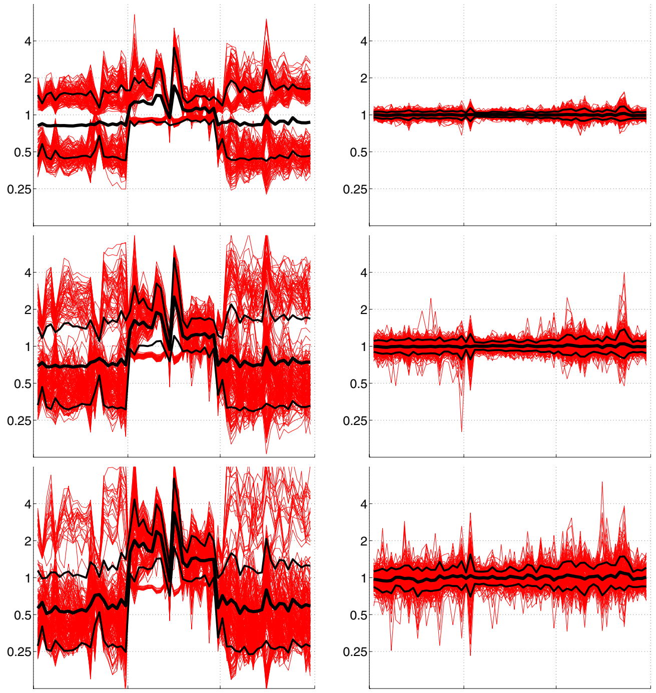

**Fig. 6.** Variability of the results of globalization (left) and centralization (discarding genes with intensity below 10 or above 1000; right) on differently sized random subsets of genes. Rows: top, samples of size 349 genes (half of the in vivo expressed genes); middle, samples of size 174 genes (quarter); bottom, samples of size 87 genes (eighth). Each position on the x-axis corresponds to one of the 63 measured conditions. The y-axis is in logarithmic scale in order to symmetrically represent up- and down-scaling. Each thin gray line represents the scaling computed from one random gene sample. All values are shown relative to the scaling computed from all 698 genes that have a positive intensity value in at least one in vivo measurement, thus unifying this scaling with the horizontal line at scaling one. The fat black line indicates the mean relative scaling over all 200 random samples, the thinner black lines bound the one standard deviation area.

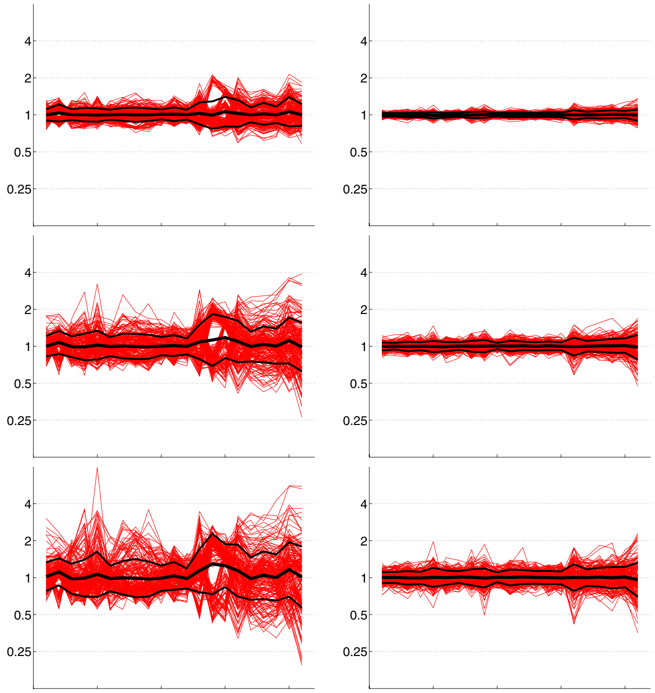

**Fig. 7.** Same sceme as Figure 6, but restricted to the in vivo samples (the leftmost 21 samples of Figure 6).

plying the amplification of most cellular gene products (including housekeeping genes and ribosomal RNA). A proportional increase of the expression of all genes cannot be distinguished from an upscaling of all intensities due to any multiplicative error that takes effect on all spots of an array. However, though such activation is likely to occur to a certain degree, this appears to be limited within the body. It is known that cells requiring high expression capacity in the body form polykaryons (e.g. syncytiotrophoblast, osteoclasts, giant cells, striated muscle cells) or gain polyploidy (hepatocytes, karyocytes). In fact, there seems to be a biological limitation of single karyons with diploid genomes to support the expressional machinery, limiting the systematic error introduced by activation of cells into any normalization procedure. On the other hand, if a cell doubles the expression of all its genes, it should still behave similar to two cells at the original expression level, rendering complete activation irrelevant for most biological questions (e.g., the response to drug treatment). Similarly, if all genes but one are upregulated, it may be equally or even more useful to say that the one gene is downregulated.

As a last point of discussion, we want to indicate that the popular two-channel measurements do not eliminate the need for proper normalization. There, the idea is to measure on the same array each sample simultaneously together with a reference sample, using two different fluorescent dyes. The intensity ratios of each probe against the corresponding reference are invariant with respect to some, but not all multiplicative errors. In particular, if the number of cells from which the mRNA was isolated or the extraction or RT efficiency is unknown, the need for normalization remains. To demonstrate this point in practice, we applied normalization to the background-subtracted intensity values of the well-studied yeast diauxic shift experiment by (DeRisi *et al.*, 1997). The results, shown in Figure 8, show that computational normalization might also be of value for two-channel expression measurement technology. For these yeast data, the underlying assumption of globalization seems to be justified, since virtually all genes

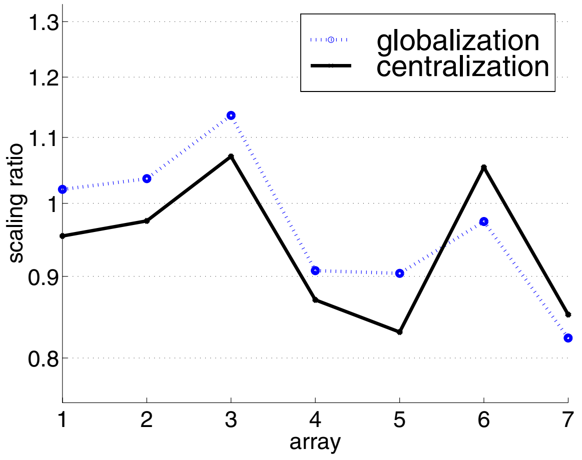

**Fig. 8.** Ratios of normalization factors between red and green channel for each of the 7 two-channel arrays of the yeast diauxic shift time course (DeRisi *et al.*, 1997) as computed with centralization and globalization. In this case the two methods agree quite well, but normalization appears to be necessary.

are assayed and cell specialization can be excluded.

In conclusion, we find that centralization reproduces the results of other normalization methods, where the respective underlying assumptions are justified, and yields more robust estimates otherwise. With its power to assay a huge number of genes at the same time, array technology appears to be superior to any technology focussing on single genes in terms of biologically reasonable normalization, despite the fact that other technologies provide – at least at the moment – a higher precision. If our assumption holds then, array technology, providing more diverse gene data, should allow for more accurate normalization. This effect may even be expected to even compensate for the reduced precision of array experiments. We envision that centralization might prove most useful for two application scenarios. The first is the comparison of cells in significantly different states, including healthy with diseased cells in vivo, or in vitro models of diseases with cells in their native environment. Second, the robustness of centralization will be most useful for the normalization of low gene number assays of gene expression, which can be expected to become popular diagnostic tools in a few years.

## ACKNOWLEDGMENTS

This work was supported by the BMBF (grants 0311615 and 01GG9824). We are grateful to Pia M. Gebhard, U. Erlangen, for technical help and discussions. We thank Niklas von Öhsen, GMD.SCAI, and Rudolf Grübel, U. Hannover, for discussing and proving the Poisson ratio claim. We thank Kathleen Kerr, Jackson Lab, for details on her ANOVA method.

## REFERENCES

Aigner, T., Zien, A., Gehrsitz, A., Gebhard, P. M. & McKenna, L. (2001). Anabolic and Catabolic Gene Expression Pattern Analysis in Normal Versus Osteoarthritic Chondrocytes using cDNA-array Technology. Arthritis Rheum, in press.

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

Chen, Y., Dougherty, E. R. & Bittner, M. L. (1997). Ratio-Based Decisions and the Quantitative Analysis of cDNA Microarray Images. *J Biomed Op*, **2**, 364–374.

DeRisi, J. L., Iyer, V. R. & Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–685.

Goldsworthy, S. M., Goldsworthy, T. L., Sprankle, C. S. & Butterworth, B. E. (1993). Variation in expression of genes used for normalization of Northern blots after induction of cell proliferation. *Cell Prolif*, **26**, 511–518.

Ke, L. D., Chen, Z. & Yung, W. K. (2000). A reliability test of standard-based quantitative PCR: exogenous vs endogenous standards. *Mol Cell Probes*, **14**, 127–135.

Kerr, M. K., Martin, M. & Churchill, G. (2001). Analysis of variance for gene expression microarray data. *J Comput Biol*, **7**, 819–837.

Lee, M. L., Kuo, F. C., Whitmore, G. A. & Sklar, J. (2000). Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proc Natl Acad Sci USA*, **97**, 9834–9839.

Li, C. & Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc Natl Acad Sci USA*, **98**, 31–36.

Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H. & Herzel, H. (2000). Normalization strategies for cDNA microarrays. *Nucleic Acids Res*, **28**, E47.

Suzuki, T., Higgins, P. J. & Crawford, D. R. (2000). Control selection for RNA quantitation. *Biotechniques*, **29**, 332–337.

Velculescu, V. E., Madden, S. L., L, L. Z. *et al.* (1999). Analysis of human transcriptomes [letter]. *Nat Genet*, **23**, 387–388.

Vu, H. L., Troubetzkoy, S., Nguyen, H. H., Russell, M. W. & Mestecky, J. (2000). A method for quantification of absolute amounts of nucleic acids by (RT)-PCR and a new mathematical model for data analysis. *Nucleic Acids Res*, **28**, E18.

Zolotukhin, I. & Lange, J. (2000). Application of Analysis of Variance Schemes to Expression Data. In *Proc German Conf Bioinf*. Logos Verlag, Berlin, Germany, pp. 159–166.