BMC Bioinformatics

Methodology article

Open Access

Three-parameter lognormal distribution ubiquitously found in cDNA microarray data and its application to parametric data treatment

Tomokazu Konishi*

Address: Faculty of Bioresource Sciences, Akita Prefectural University, Akita 010-0195, Japan

Email: Tomokazu Konishi* - konishi@akita-pu.ac.jp

* Corresponding author

Published: 13 January 2004

BMC Bioinformatics 2004, **5**:5

Received: 19 September 2003

Accepted: 13 January 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/5

© 2004 Konishi; licensee BioMed Central Ltd. This is an Open Access article: verbatim copying and redistribution of this article are permitted in all media for any purpose, provided this notice is preserved along with the article's original URL.

Abstract

Background: To cancel experimental variations, microarray data must be normalized prior to analysis. Where an appropriate model for statistical data distribution is available, a parametric method can normalize a group of data sets that have common distributions. Although such models have been proposed for microarray data, they have not always fit the distribution of real data and thus have been inappropriate for normalization. Consequently, microarray data in most cases have been normalized with non-parametric methods that adjust data in a pair-wise manner. However, data analysis and the integration of resultant knowledge among experiments have been difficult, since such normalization concepts lack a universal standard.

Results: A three-parameter lognormal distribution model was tested on over 300 sets of microarray data. The model treats the hybridization background, which is difficult to identify from images of hybridization, as one of the parameters. A rigorous coincidence of the model to data sets was found, proving the model's appropriateness for microarray data. In fact, a closer fitting to Northern analysis was obtained. The model showed inconsistency only at very strong or weak data intensities. Measurement of *z*-scores as well as calculated ratios was reproducible only among data in the model-consistent intensity range; also, the ratios were independent of signal intensity at the corresponding range.

Conclusion: The model could provide a universal standard for data, simplifying data analysis and knowledge integration. It was deduced that the ranges of inconsistency were caused by experimental errors or additive noise in the data; therefore, excluding the data corresponding to those marginal ranges will prevent misleading analytical conclusions.

Background

Since microarray data contain systematic variations that are derived from various experimental sources, the data should be normalized prior to comparison with other such data. In order to perform such normalization, some stable data characters that represent the data set are found and/or assumed. By making such characters identical, each data set is adjusted to other data sets, to a reference experiment's data, or to a mathematics model. A normalization method is based on ideas or concepts in which elements of data are considered to be the stable characters, and on the design of calculations regarding how data sets are to be adjusted. It is clear that these concepts affect the normalization results; such concepts behind the

normalization are often closely connected with the evaluation of differences in data. Indeed, these concepts should originate from experimental observations and/or biologically appropriate assumptions. As an introduction, it might be helpful to describe the concepts on which previously reported methods for microarray data normalization have been based.

Taking ratios with stable elements of the data is one of the simplest methods by which a set of relative data has been normalized. Candidates for such stable elements can be data for house-keeping gene(s), data for control experiments or the median of a set of data. Within a group of data that share this stable element, the calculated ratio can be compared. Such ratio-based methods have been frequently used in the field of molecular biology, since many of the determination methods of mRNA produce relative values. The relative nature is also expected in microarray data. One of the pioneer works in the statistical treatment of microarray data followed the ratio-based scheme [1], assuming a rigid distribution model for the ratios, and allowing objective decisions by seeking the ratio data that exceeded a cut-off value for their deviation. However, it has become clear that such calculated ratios often fluctuate depending on the signal intensity [2-7]. This unstable character, or intensity-dependent effect of measured log-ratio, disagrees with the original assumption, and becomes problematic in data analyses.

Tseng et al. and Yang et al. [2,3] followed the ratio-based scheme but stabilized the log-ratios by compensating them using the LOWESS technique. In addition, Workman developed a method that used a different calculation technique [4]. Due to the flexibility of their non-linear compensations, the methods can adjust any pair of data sets by resolving the fluctuations. However, even such strong methods could not achieve the assumed stability of the model in regard to log-ratio deviations, by which differences in gene expressions are measured [1]; rather, determined ratios were dependent on signal intensity.

In order to solve the intensity-dependence problem in determined ratios, Huber et al. and Durbin et al. [5,6] produced a new scheme that recognizes the differences in mRNA levels, not in terms of ratios but in terms of the difference values in arsinh functions, which have stability in their statistical behaviors. The adjustment is performed by linear transformation of one of the pair-wise data sets prior to the arsinh conversion, by trial and improvement, evaluated by likelihood analysis in arsinh values [5]. In fact, this method adjusts the data over the entire range of determination. However, since microarray data might have additive noise [5,8], which will affect data especially at lower intensities, the stability of deviations at all signal intensities is still doubtful. Additionally, the processed results are not comparable with those of authentic analyses, such as Northern and/or RNase protection assays, since the arsinh function is incompatible with ratios.

An alternative attempt involves the adjustment of a group of data together at one time. Kerr et al. [9] assumed an ANOVA model, in accordance to which data logarithms are linearly adjusted to have the least differences in relation to each other. Since this method treats data sets simultaneously, it finds the most suitable solution among the data sets. However, this method cannot self-evaluate the appropriateness of the model and the design of the adjusting process. Additionally, this process requires an inordinate amount of calculation and, if a set of data is added afterwards, all the calculations have to be performed from the beginning.

In order to reduce the amount of calculation, it is preferable to adjust each data set in terms of a rigid model. Such a method can normalize data sets one by one, and the normalized data can be compared directly with each other without further adjustment. Additionally, if the model is appropriate, normalization will be highly accurate. In such model-based normalization methods, the simplest assumed stability might be the total amount of mRNA in a tissue, which forms the basis of the normalization of the total intensity [7,10] or the global standardization [11,12]. However, the appropriateness of the model in terms of the stability, not of the total amount of mRNA but of the sum of the determined numerals, is difficult to evaluate; the determined data is not always proportional to the amount of mRNA, since the determined data contains background, the value of which is difficult to estimate exactly [13]. Some alternative methods assume a model for the statistical distribution of data. In many cases, a lognormal distribution would be the optimal model for microarray data, and indeed this distribution has been reported for some data sets [2,14,15]. Additionally, Hoyle et al. [16] have found that microarray data are in agreement with both Benford's law and Zipf's law, and suggested the lognormal model and power law model to be good candidates for assumptions concerning the distribution. However, the real data distributions sometimes do not fit closely to these models [9,16]. Such inappropriateness in a model can be found to skew data histograms or probability plots.

The intensity data of microarray experiments always contain a certain level of background [8], and inadequate estimation of this background can affect the assumed stabilities in ratio-based normalization methods as well as lognormal data distributions. Background has been estimated based on the hybridization image [13], which is then subtracted from intensity data; this estimation technique is based on the supposition that the background

level is consistent between the DNA spot and the surrounding space. However, because surface properties may differ between the DNA spot and the surrounding space, the respective backgrounds can also differ (a possible extreme example of such difference is the antiprobe [17] with dark DNA spots against a bright surrounding area). Indeed, failure in background estimation will originate intensity dependency of calculated log-ratio; such an effect can even be seen in simulation data [18]. Additionally, an under-estimation of background in both data sets will reduce the differences at lower signal intensities; such a phenomenon has been observed in determinations evaluating a microarray's mechanical characteristics [19]. Adding or subtracting a constant to or from a series of numerals affects the logarithm values in a non-linear way, and biased errors in background estimation can affect the distribution of microarray data in the same manner.

In this article, a model-based normalization method that finds the background by calculation is introduced. The method assumes stability in data distribution of each set of single-channel microarray data. The method uses a three-parameter lognormal distribution model; since the image-based local background estimation [13] can generate a constant error deriving from the different surface properties on a DNA chip, it is reasonable to handle the background as an unknown quantity. The three-parameter model was established by introducing the unknown as the threshold parameter to a lognormal distribution model [15]. To maintain the objectivity of data treatment, the parameter is restricted as a common constant within a single-channel data set; this treatment is based on assumptions that the background is mainly provoked by non-specific binding of pigments to DNA spots, and that each DNA spot binds a fixed amount of the pigments. The common constant is found as the value that, when universally subtracted, produces a data distribution that most closely fits the model. The appropriateness of the assumed distribution model is evaluated by means of coincidence between the model and the resulting data distribution in many different microarray experiments. Additionally, a ratio-based treatment of the normalized data is introduced. The stability of signal versus ratio relationship is shown below, as well as a correlation with Northern blot analyses.

Results

Fitness of the three-parameter model to data

In all the cases examined – approximately 300 different samples with at least 16 cDNA stamping type chips, as well as 50 samples with 4 synthetic oligonucleotide chips – a threshold parameter whose subtraction could transform the data distribution to fit the model could be found. As an example, the distribution of open resource data [20] from a study of human fibroblasts [21] is shown in Figure 1. In this case, both the intensity data and the local background-subtracted data clearly show distributions different from that of the model; the histograms are asymmetric and the probability plots are distant from the ideal value (Fig. 1, blue and green). However, subtraction of the threshold parameter transforms the distribution to one closer to the model (Fig. 1, red). The histogram shows the ideal bell shape, and the probability plot at larger than -1.4 shows a linear relationship; this area includes 92% of the original intensity data. Such normality was also confirmed by a chi-square test [22]; at a range from 2.35 to -1.55 standard units, seven out of eight randomly chosen data [21] passed the test at a 5% significance level. The only data set that failed the test was derived from an experiment in which hybridization had apparently not taken place (most likely due to the presence of an air bubble during processing). This parametric nature of data was also observed in experiments that used other DNA chips, including yeast, rice, *A. thaliana*, *B. subtilis*, *C. elegans*, and *E. coli* (some examples are shown in Fig. 2), as well as in commercially available DNA chips such as Atlas Glass Array (Clontech), and synthetic oligonucleotide probe chips such as GeneChip (Affymetrix) and Agilent Oligo microarray (data not shown).

During the threshold parameter estimation process, the parameter became larger than some of the intensity data, producing negative values whose logarithms could not be calculated. In the experiment shown in Figure 1, for example, 1.2% of the data fell into this class, although numbers of such data were variable between experiments. Since microarray data might contain a certain level of additive noise [8], it is highly possible that some of the DNA spots produce signals so faint that the negative noise can mask them completely. Consequently, those data were simply treated as "signals not detected".

Appropriateness of the three-parameter model evaluated from technical reproducibility

Although microarray data was consistent with the distribution model over a wide range of intensities, no data showed perfect consistence. Rather, the probability plots necessarily bent downwards at lower intensities (Figs. 1,2). Such discrepancies could be caused either by additive noise in the measurement system or by the unsuitability of the model. These possibilities were verified through data reproducibility, which was confirmed on a scattergram with repeated hybridization experiments involving the same RNA sample. Since the breakdown of the model occurs at different intensities, each data set is consistent with the model at a different range of signals. Naturally, we cannot expect data reproducibility at the inconsistent ranges where the data does not obey the model. In such ranges, if the inconsistency occurs because of model unsuitability, the normalization may give the data a

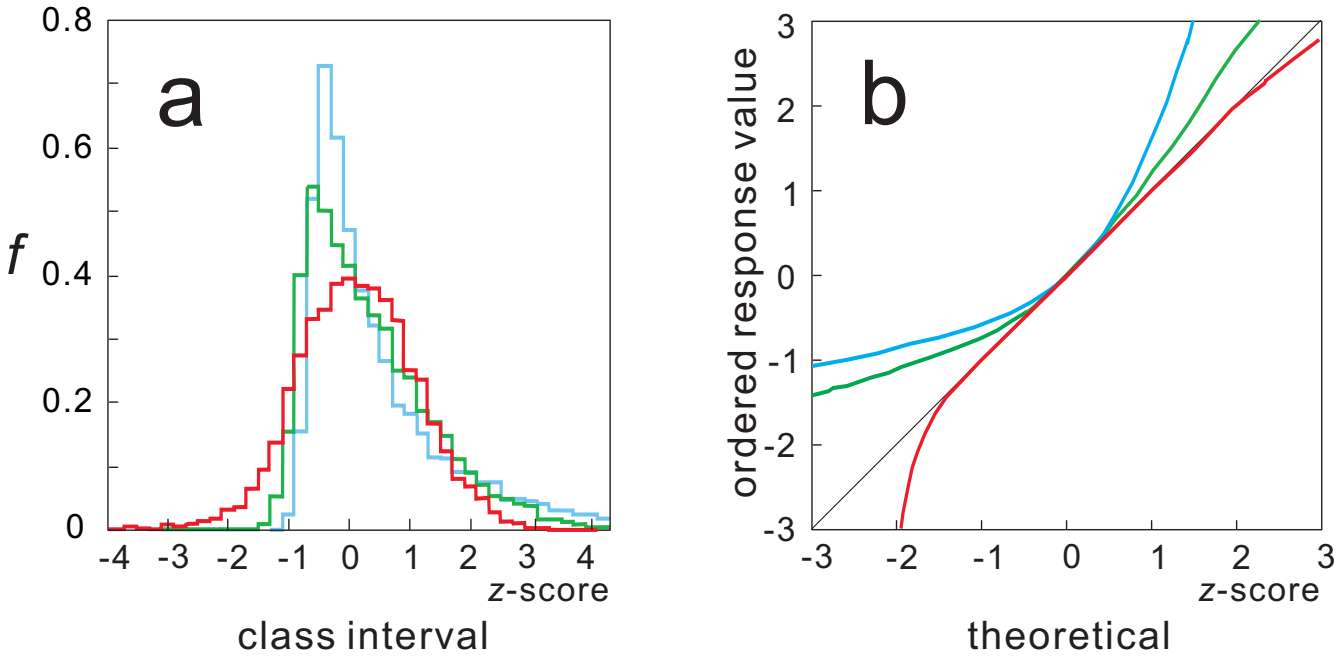

Figure 1
Effects of compensating the threshold parameter γ. (**a** Histograms of the logarithms of human cDNA microarray data [21] (from the Stanford Microarray Database [20], ID 5731). The histograms for the original intensity data (blue) and local background-subtracted [13] data (green) both skew to the left, whereas that for the γ-subtracted data shows the typical shape of a normal distribution (red). *f*, frequency of data. (**b**) Normal probability plots for logarithms of the original intensity data (blue), local background-subtracted data (green), and the γ-subtracted data (red). In the plot, x-axis presents theoretical values for normal order statistic medians, whereas y-axis presents ordered logarithms of data [26]. The γ-subtracted data show reasonable linearity, characteristic of a normal distribution. Definitions of parameters and data treatments are described in Materials and Methods.

biased error; such bias will make a bend(s) in the scattergrams between repeated experiments. Alternatively, if the inconsistency is caused by noise, the noise will affect the reproducibility by dispersing the scattergram.

The reproducibility was verified using open resource data in the Stanford Microarray Database [20]. The reproducibility that should be issued here is not the biological one but that of the hybridization process as well as that between different dyes. Such artificial reproducibility [23] can be measured with sets of repeated hybridization of the same RNA. In the Stanford system, each hybridization experiment contains control RNA; for example, in the time course experiment on serum shock to human fibroblasts [21], a zero-time RNA sample is labeled with Cy3 and hybridized to each DNA chip as the control. Data for those controls were normalized and compared (Fig. 3a and 3b) in order to visualize the reproducibility in hybridization. For another example, in the comparison with A. thaliana tissues, Horvath et al. made dye-swap experiments [24]. Two pairs of hybridizations were compared, using data from the same RNA but differently colored and hybridized separately to DNA chips (Fig. 3c and 3d). If the difference in dyes could cause specific alterations to data, the scattergram would be bent or tilted.

The scattergrams for repeated experiments showed consistent data reproducibility (Fig. 3). Most of the data were plotted within the 1.4-fold difference lines (the method for calculating the ratio is described below) above the breakdown levels of the model. In contrast, in the discrepancy region (red dots), the reproducibility was lost and the data became randomly plotted (Fig. 3), suggesting that the discrepancies between data and the model were caused by noise rather than unsuitability of the model. In some cases, the upper part of the data also bent down below the y = x line in the probability plots (Fig. 2, ID 1593 and 15973). Such a breakdown was typically found in cases in which the signals of the intensity data were relatively large. Such breakdowns may be caused by the saturation of array scanners, which may ruin the signal response. Since the diameters of the DNA spots and also

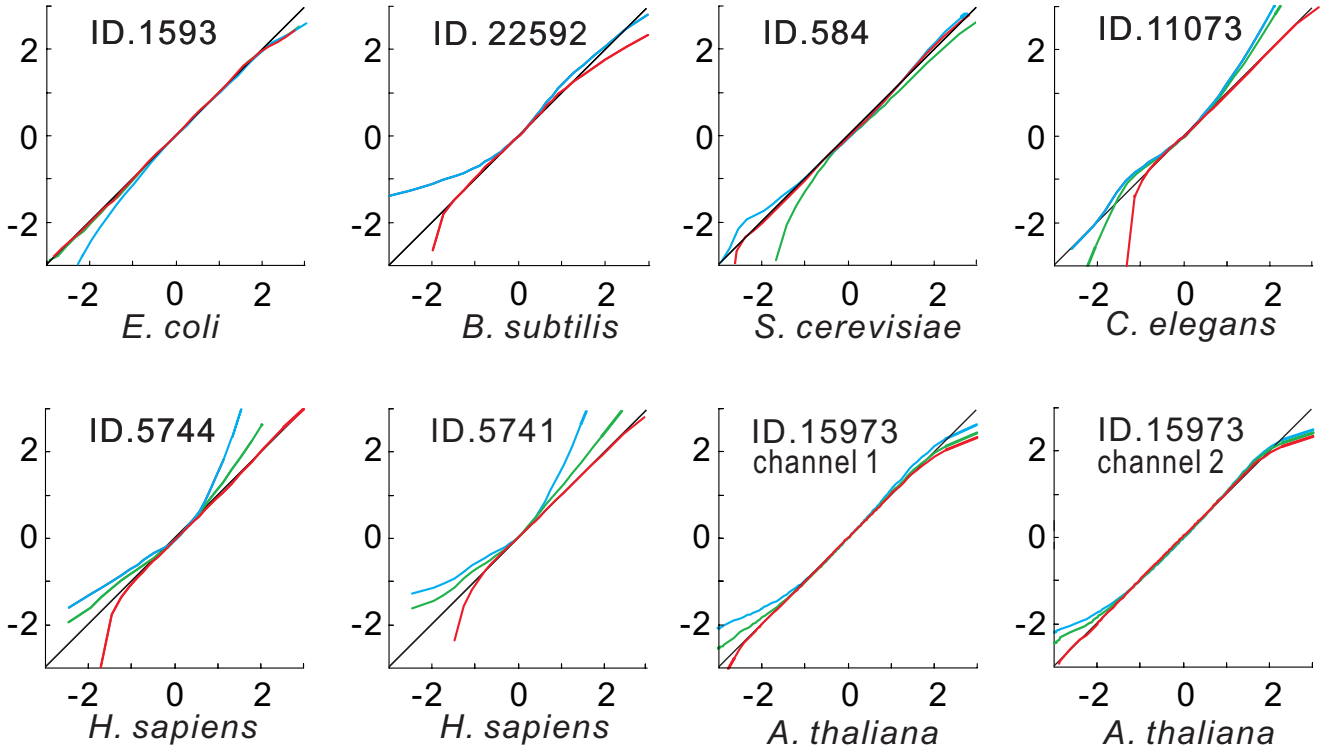

Figure 2

Distributions of data obtained from the StanfordMicroarray Database. [20]. These probability plots [26] show wide ranges of linearity, proving the appropriateness of the three-parameter lognormal model (red) for normalizing microarray data. The plots for the original intensity data are also shown (blue). Local background-subtracted [13] data (green) of ID 22592 could not be normalized since a large proportion of the data (over 25%) was negative; the parameter σ, which is found from interquartile range, for those grids could therefore not be calculated.

the DNA concentrations within each spot are uneven, such saturation will also add a level of noise to the data, rather than just creating distortion in the linear signal response. Actually, the plotted data above the upper limit of linearity showed inconsistent reproducibility (Fig. 3, panel b). There were no bends observed in the scattergrams, showing the appropriateness of the three-parameter model, and contradicting the dye-specific alterations to data (panels c and d).

Stable nature of σ values found from logarithms of γ subtracted microarray data

Values of the shape parameter, σ, were found to be stable within one set of experiments, between different hybridizations and between clone sets. As an example, the values obtained from human cells [21] showed little divergence, and the values remained the same throughout the time-course experiment on serum shock (Table 1). The shape parameter was measured from each grid of data, which were derived from clones spotted with identical pins, at six intervals in the time-course experiment: zero to 24 hr after the serum shock treatment. Each DNA chip consisted of 4 grids, and each grid contained different sets of cDNA clones. If the parameter had an unstable nature, the values among the grids and/or among the time-course intervals would have diverged. However, the averages of the six σ values were almost identical among the grids (Table 1). Additionally, the values were not affected by cellular conditions; during the time course experiment, the standard deviations of the values (Table 1, treatment) remained at the same level as those obtained from control hybridizations, showing that small deviations may occur due to experimental noise but not cellular conditions (Table 1).

Comparison of the normalized data on a ratio basis

The obtained results demonstrate that we can expect lognormal distributions in microarray data. Within the range in which data obey the distribution model, logarithms of the data can be normalized to *z*-scores. How, then, can we evaluate the change of expression levels presented in *z*-scores? It will, of course, be useful if the normalized data can also be compared to results obtained by conventional

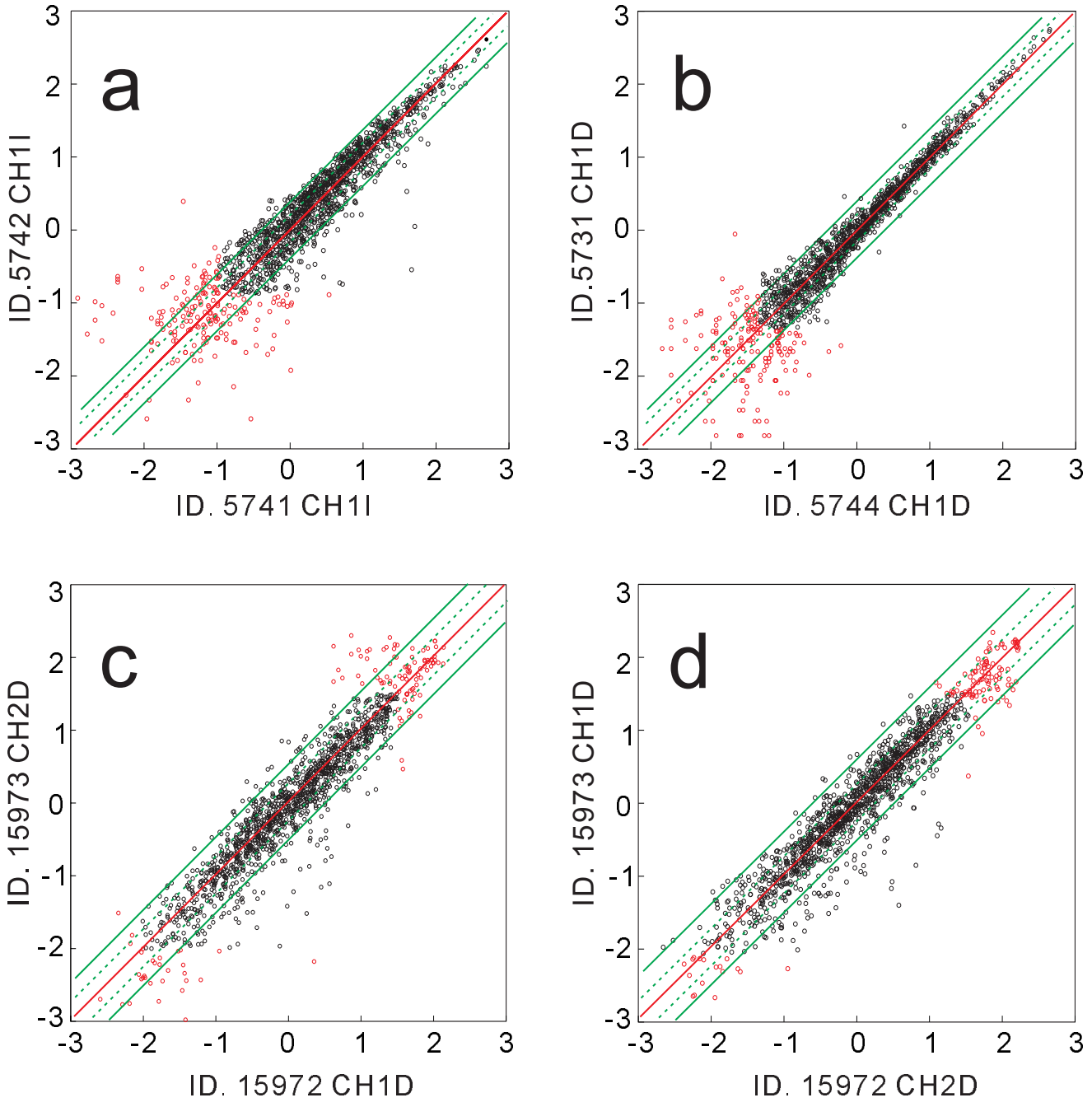

Figure 3

Examples of scattergrams for repeated hybridization data, normalized by means of the three-parameter log-normal distribution model. The scattergrams are of two hybridization experiments with the same control probe, thus demonstrating the technical reproducibility of the experiments. Red colored dots indicate that one or both of the paired data is in the model-inconsistent range. Lines beside the dots show 2-fold (solid green lines) and 1.4-fold (dashed green lines) differences. The number of data used for each plot was 1,500. Data sets were obtained from control hybridizations in Iyer et al. [21] (panels **a** and **b**), and from dye-swap hybridizations in Horvath et al. [24] (panels **c** and **d**).

Table 1: The stability of shape parameter σ values in human fibroblast data.

experiment	σ[a]	SD[b]	n[c]
control (ch1)	0.65	0.08	24
grid 1	0.61	0.08	6
grid 2	0.66	0.06	6
grid 3	0.65	0.09	6
grid 4	0.69	0.08	6
treatment (ch2)	0.65	0.06	24
grid 1	0.69	0.06	6
grid 2	0.68	0.04	6
grid 3	0.65	0.06	6
grid 4	0.60	0.04	6

[a]means of σ ; [b]standard deviations of σ ; [c]number of grids used for calculating the means and standard deviations.

methods, such as Northern analysis. In most conventional analyses, ratios are used to indicate differences in expression levels. Since such analyses do not provide information about the distribution of expression levels, the *z*-scores cannot be calculated. In order to normalize the data, the amounts of total RNA, rRNA, and/or housekeeping genes are commonly used as standards instead. Under such limitations, ratio methods are a convenient choice for evaluating the differences in gene transcript levels, i.e. the number of mRNA molecules transcribed from a gene and accumulated in a cell.

Assuming stability in the population distribution or transcript levels of genes, ratios can be calculated from *z*-scores obtained from microarray experiments according to the following formula. Since background-subtracted microarray data may have a linear relationship to the transcript level, each datum can be expressed as

(datum for *i*th spot at *j*th hybridization) = $a_j b_i x_{j,i}$,

where a_j is a factor that compensates for differences in sensitivities of detection between hybridization experiments, b_i is another sensitivity compensation factor between different DNA spots on a DNA chip, and $x_{j,i}$ is the transcript level of a gene. Consider two sets of background-subtracted data, $a_1 b_i x_{1,i}$ and $a_2 b_i x_{2,i}$ for i = 1...n, in different hybridizations on identical array chips. Since the same normal distribution is assumed for $\log(x_{1,i})$ and $\log(x_{2,i})$, the values for the shape parameter (σ) are the same. According to *z*-normalization of the data, the normalized data will be,

$$Z_{j,i} = \{\log(a_j b_i x_{j,i}) - \mu_j\}/\sigma$$

where μ_j is the observed scale parameter for each hybridization experiment. The difference in the normalized data between $Z_{1,i}$ and $Z_{2,i}$ can be presented as

$$Z_{1,i} - Z_{2,i} = \{\log(a_1 b_i x_{1,i}) - \mu_1\}/\sigma - \log(a_2 b_i x_{2,i}) - \mu_2\}/\sigma$$

$$= \{\log(x_{1,i}/x_{2,i}) + \log(b_i) - \log(b_i) + \log(a_1) - \log(a_2) - (\mu_1 - \mu_2)\}/\sigma. \quad (1)$$

In this formula, both μ_1 and μ_2 are the scale parameters that can be defined as

$$\mu_j = 1/n\sum_{i=1}^{n}\{\log(a_j b_i x_{j,i})\} = \log(a_j) + 1/n\sum_{i=1}^{n}\{\log(b_i)\} + 1/n\sum_{i=1}^{n}\{\log(x_{j,i})\}.$$

According to the stable nature on the distribution of $x_{j,i}$, the average of $\log(x_{j,i})$, $1/n\sum_{i=1}^{n}\{\log(x_{j,i})\}$, will be common between the experiments. Here we can express $\mu_1 - \mu_2$ appearing in formula (1) as

$$\mu_1 - \mu_2 = \log(a_1) + 1/n\sum_{i=1}^{n}\{\log(b_i)\} + 1/n\sum_{i=1}^{n}\{\log(x_{1,i})\} - [\log(a_2) + 1/n\sum_{i=1}^{n}\{\log(b_i)\} + 1/n\sum_{i=1}^{n}\{\log(x_{2,i})\}] = \log(a_1) - \log(a_2).$$

This leads to the difference of normalized data (1)

$$Z_{1,i} - Z_{2,i} = \{\log(x_{1,i}/x_{2,i})\}/\sigma$$

From this formula, the abundance ratio of RNA, $x_{1,i}/x_{2,i}$, can be found from the difference of *z*-scores as

$$x_{1,i}/x_{2,i} = 10\text{^}\{\sigma^*(Z_{1,i} - Z_{2,i})\}.$$

Data comparison with Northern analyses

Comparisons of normalized microarrays with other conventional methods could constitute a rigorous inspection for data treatments, the normalization and the transformation to a ratio-based method. If appropriate, the results will be similar to those obtained using conventional analytical methods. Based on this assertion, obtained results were compared between microarrays and Northern blot analyses in a cold-treatment time course experiment using rice seedlings. Some of the clones issued in the microarray experiment were semi-randomly selected, and ratios were determined by Northern analysis (Materials and Methods). Differences in the transcript levels of each gene were calculated by means of both microarray and Northern analysis, and the logarithms of the calculated ratios were compared between the two methods (Fig. 4). If the results coincided with each other, the scattergram would show positive correlation.

The scattergram showed a close correlation between the results of microarray and Northern blot analyses (panel **a**), suggesting that the proposed treatment of the microarray data provides an appropriate method for analyzing the data. For reference, the same comparison using other normalization methods is shown in panels **b** and **c**, providing more dispersed results with different tendencies. The coincidence shows the appropriateness of the presented data treatments, since the background subtraction is critical to the ratio calculation. If the subtraction were inaccurate, the scattergrams would never coincide.

Stability in signal intensity versus ratio relationships

As mentioned in the Background, ratio-based normalization methods assume that a determined ratio should be independent of signal strength [7]. Such an assumption, however, is not prefigured in data normalized by means of the parametric method, since the method normalizes each single-channel data set and does not have restrictions on the relations between pair-wise data. In order to check the dependency of ratios in parametrically-normalized data, log-ratios between 0 hr and 24 hr after serum treatment [21] were presented on a plot [5] in which the x-axis showed the rank of averages of signal intensity and the y-axis showed the log-ratio (Fig. 5). If the ratio is independent of signal intensity, data would be plotted horizontally, not be tilted or bent. Additionally, the vertical width of the plot would be uniform. It should be recalled that LOWESS is a calculation process that makes the plot horizontal. Since LOWESS cannot make the width uniform, the variance stabilization [5,6] and/or intensity-filtering method [7] are proposed.

In the graph, the data normalized by means of the parametric method were plotted horizontally along the zero-difference line, forming a uniform width above and below the line at the model consistent ranges of data (Fig. 5, black dots). In detail, the moving average (green line) demonstrated the ratio to be independent of the signal intensity. Part of the moving standard deviation (blue line) showed larger divergences of the data at the model-inconsistent range of data (red dots). However, such instability in the ratios was not observed within the model-consistent range of data (black dots). Actually, the stability in the moving deviations was much greater than that achieved by global normalization [10] (Fig. 6a) or by the LOWESS method [3] (Fig. 6b). Additionally, the parametric method is free from the inconsistencies found in a pair of reciprocal calculations in variance stabilization [5] (Figs. 6c and 6d). Furthermore, the log-ratios obeyed a constant distribution that seems to be normal at any model consistent ranges of signal intensity (Fig 5, histograms).

The appropriateness of the parametric method and of the rejection of the data class that does not obey the model were further investigated in dye-swap experiments of a two-colored system [24]. In experiments with high technical reproducibility, the determined ratios of expression levels of the experimental pairs would coincide. Fig 7 shows a rank vs. log-ratio plot for one of the pair hybridizations, after LOWESS treatment. Unlike the case in Fig. 6b, the plot showed a stable vertical width, showing no clear intensity dependence in determined ratios. Hence, the plot gives the impression that LOWESS achieves ideal normalization. Indeed, the other set of data in the pair showed an equally ideal rank vs. log-ratio plot (data not shown). However, in comparing the determined ratios, which showed the technical reproducibility of the experiment, correlation was found at only limited range of the rank (Fig. 7, lower part of the panel). Obviously, with average intensities of large and very small, the two experiments show a poor correlation. Such failure of reproduction, which can frequently be observed with other two-colored experiments, reduces the total reproducibility of the experiments. To make matters worse, the ranges of data that have no reproducibility cannot be predicted from a pair of data sets normalized by the LOWESS method.

In contrast, the same data sets of Fig. 7 were normalized by the parametric method in Fig. 8. The calculated ratios showed independency of signal within the model-consistent area (Fig. 8, plots colored in black; those colored in red are the range of data that does not obey the distribution model; only one set of the pairs' rank vs. log-ratio plots is presented at the upper part of the panel). Within the model-consistent range, the determined ratio showed high reproducibility (Fig. 8, lower part of the panel; plots colored in black). If differences between the dyes created bias in the data, the scattergrams would be bent, leaning

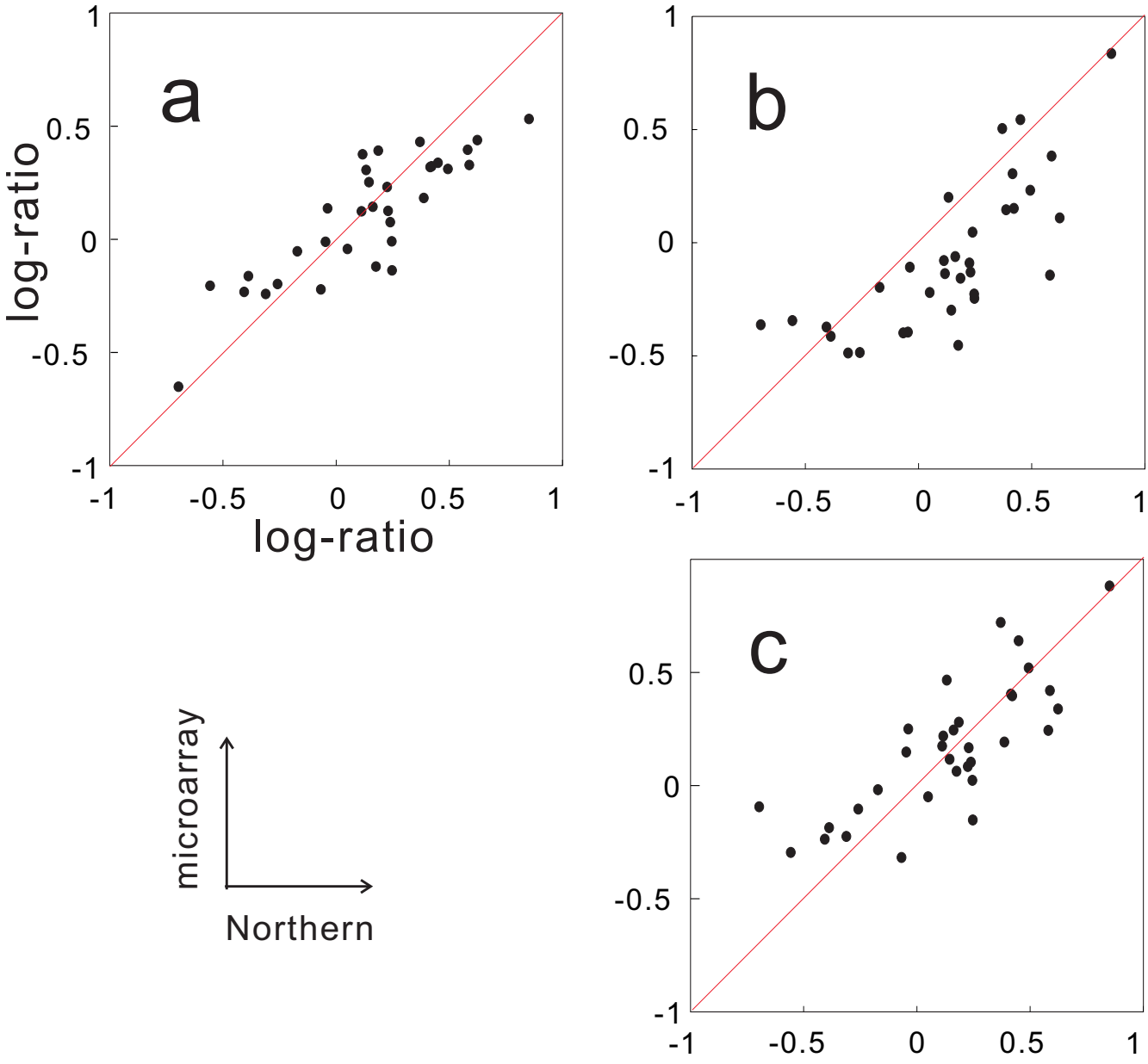

Figure 4

Comparison of the results of microarray and Northern blot analyses. Data were obtained from a time-course experiment for cold treatment of rice seedlings [15]. **a**. In the microarray analysis, intensity data were normalized by means of the three-parameter lognormal distribution model, and transformed to ratio-basis in comparison with control plants. The number of data points was 33, and the correlation coefficient was r= 0.8579. **b**. The same data set as in **a**., but the background of microarray data was subtracted using local background method [13] and then the data were normalized by globalization. The correlation coefficient was 0.7613. **c**. The same as **b**., but normalized using LOWESS method [3]. The correlation coefficient was 0.7845.

or apart from the zero point of the plot. However, the scattergrams suggest that there were no such color-based biases in the data. In contrast, in the model-inconsistent area (plots colored in red), the plot showed a less significant level of correlation, showing the low reproducibility of determined ratios in that area.

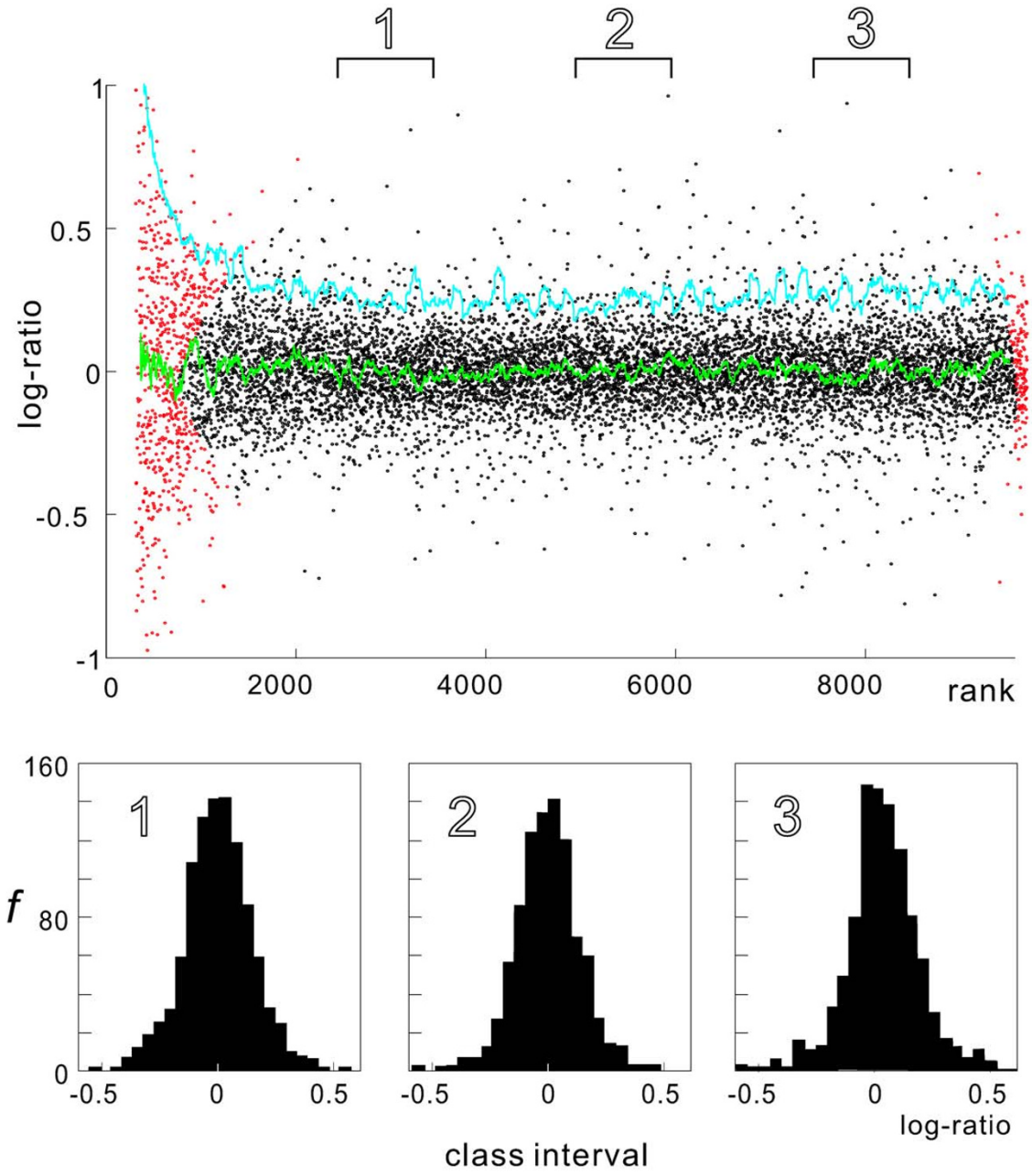

Figure 5

Stability of determined log-ratio against signal intensity. The upper part of the panel is an example of a rank vs. log-ratio plot for data comparison [5]. Each channel of the data was normalized by means of the three-parameter lognormal distribution model prior to the comparison. In the plot, the x-axis shows the ranks of average z-scores, and the y-axis shows the log-ratio that was calculated by means of the difference in z-scores. Intensity dependence of the determined log-ratio can be found by tilting and/or varying the width of the bands formed by the dots (see Results for details), as well as by noting changes in the moving averages (green line) and moving standard deviations (blue line). Red colored dots indicate that one of the paired data is in the model-inconsistent range. In the lower part of the panel, distributions of measurements are presented in histograms at the indicated ranges of ranks. Variations in intensity dependence can be visualized in the particular shape, height, width, and/or the center value of such histograms. Data were obtained from 0 hr and 24 hr after serum shock treatment [21]; the total number of data was 9677, and the experiment ID was 5731 [20]. In the control and treatment experiments, data ranked below -1.4 and above 2.2 were considered inconsistent with the distribution model (red dots). The 300 lowest-intensity data were excluded as "not detected" since they were found to be negative through γ subtraction. The window of the moving average and standard deviation calculation was 200.

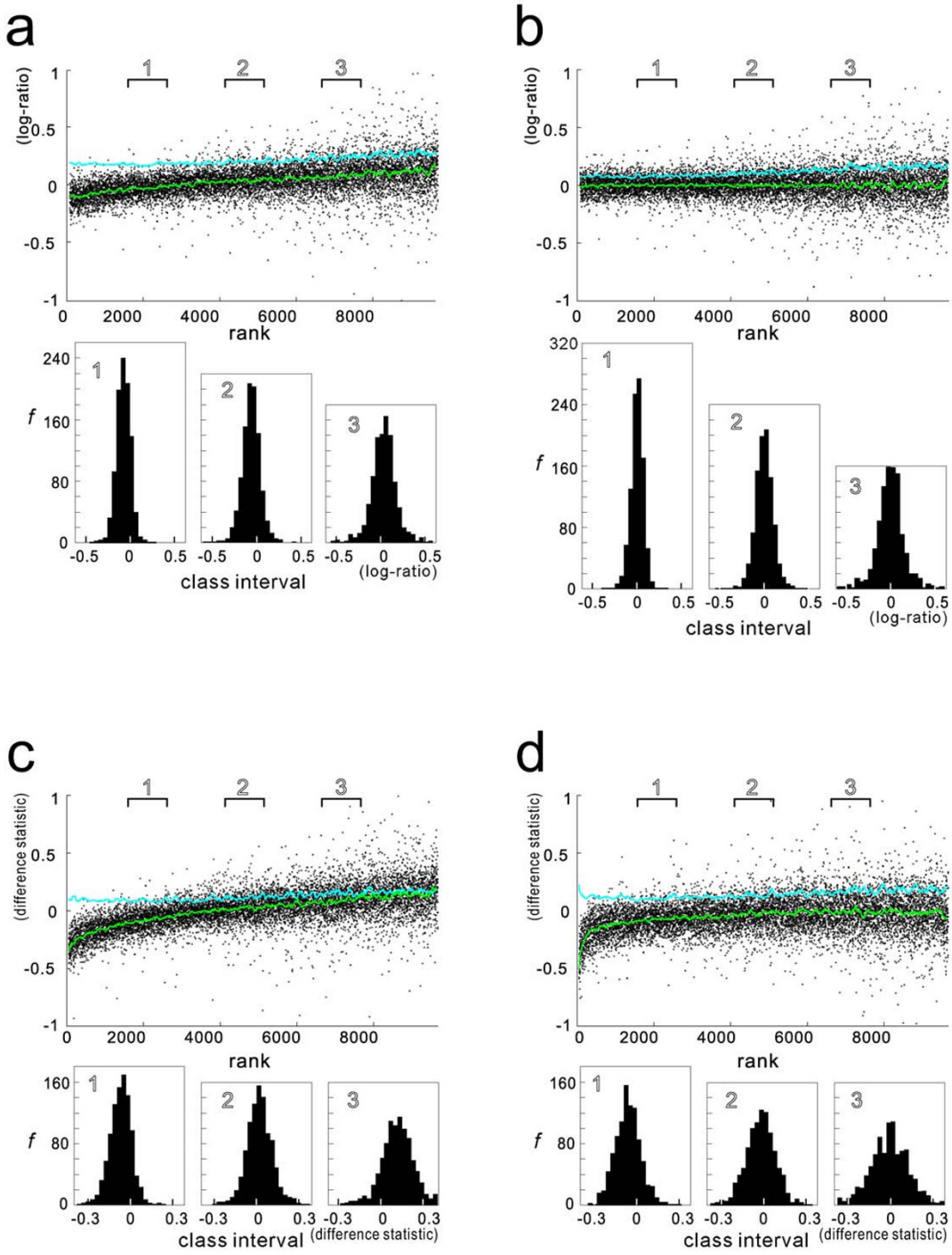

Figure 6

Detection of intensity dependence of measured differences in other normalization methods, using the same data sets shown in Fig. 5. The upper part of the panels are examples of rank vs. log-ratio plots for data comparison [5], and the lower part of the panels are histograms at the indicated ranges of ranks. **a**. Normalized by a globalization. **b**. Normalized by LOWESS method [3]. **c**. Normalized by a variance stabilization method [5]. **d**. The same method as seen in panel c but using the opposite direction of adjustment; the control data were adjusted to the 24 hr data. Since the method transforms only one of the data sets, the resulting normalized data can be different; in this case, the same result will form a mirror image of panel **c**. Green and blue lines show the moving average and moving standard deviation, respectively; the window of the calculation was 200.

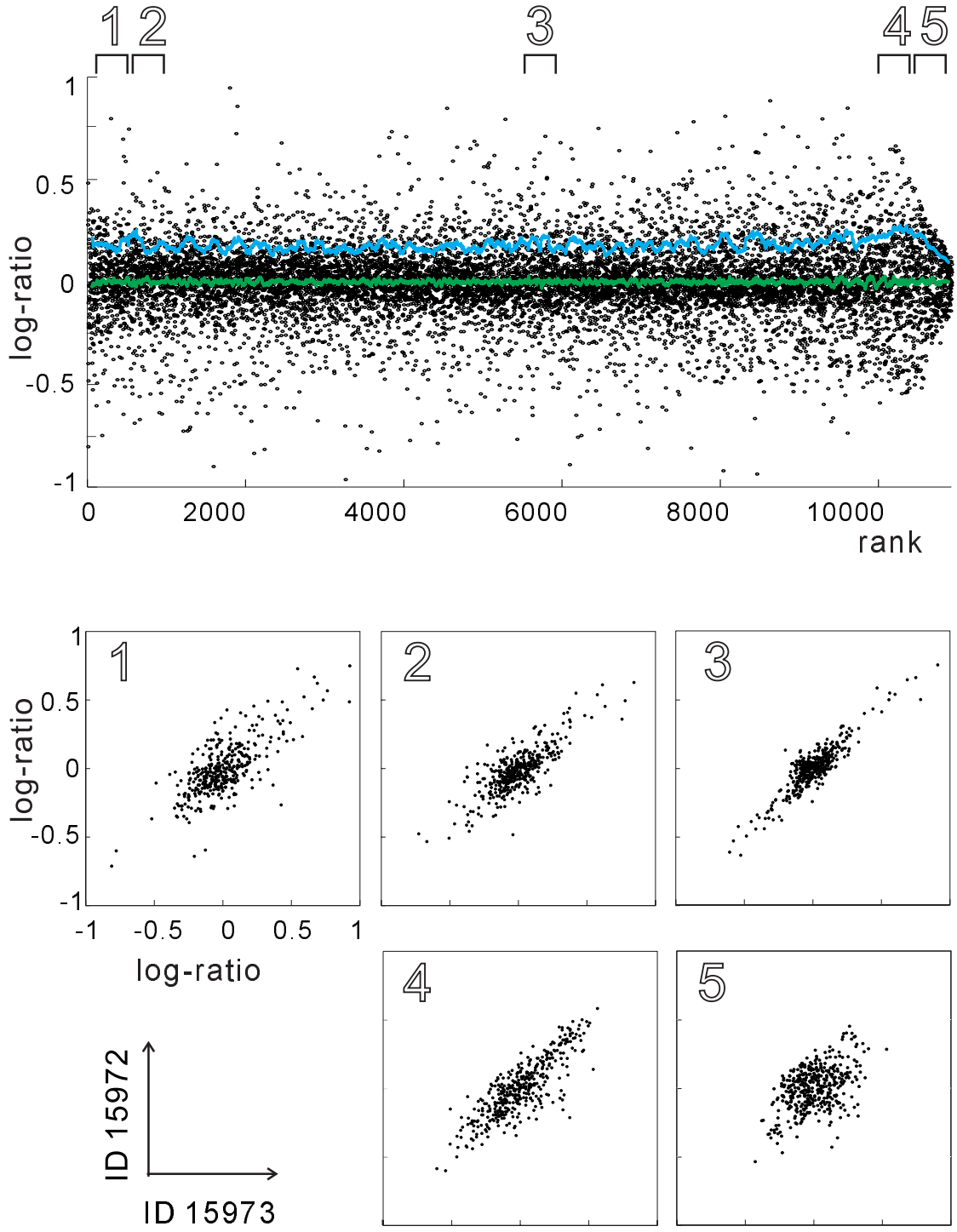

Figure 7

Intensity dependence found in reproducibility of log-ratio measurements. Each channel of data in a pair of dye-swapping hybridizations was normalized by means of LOWESS [3] method. The rank vs. log-ratio plots present intensity dependence in terms of measured log-ratio (upper part of the panels; only one of the pair is shown). Scattergrams (lower part of the panels) present the reproducibility of the log-ratio measurements at the given range of intensity ranks. Data were obtained from dye-swap experiments [24]; the experiment IDs were 15972 and 15973 [20], and each channel was comprised of 11501 data. The rank vs. log-ratio plots presented above are from the ID 15973 data. The green and blue lines show the moving averages and moving standard deviations, respectively; the window of the calculation was 200.

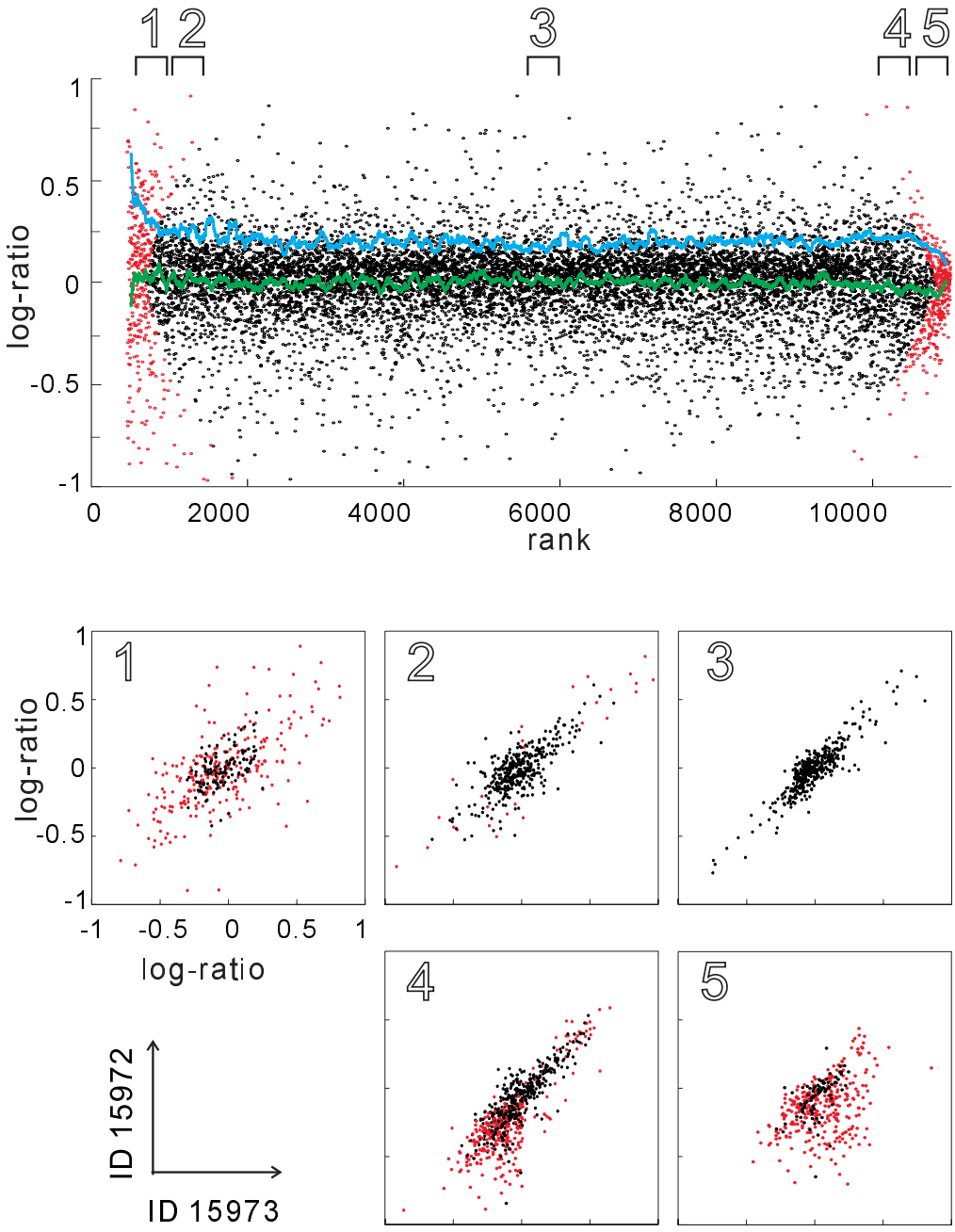

Figure 8

Improved reproducibility achieved by means of predicting the noise-affected class of data by the parametric normalization. The same data sets shown in Fig. 7 were used. Each channel of data in a pair of dye-swapping hybridizations was normalized by means of the parametric method. Red and black dots indicate model-inconsistent and consistent data, respectively. In the control experiment, data ranking below -2.3 and above 1.8 were inconsistent with the distribution model, while those below -2.8 and above 1.9 in the treatment experiment were inconsistent with the distribution model. The 223 lowest-intensity data were excluded as "not detected" since they were found to be negative through γ subtraction. The rank vs. log-ratio plots present intensity dependence in terms of measured log-ratio (upper part of the panels; only one of the pair is shown). Scattergrams (lower part of the panels) present the reproducibility of the log-ratio measurements at the given range of intensity ranks. Data were obtained from dye-swap experiments [24]; the experiment IDs were 15972 and 15973 [20], and each channel was comprised of 11501 data. The rank vs. log-ratio plots presented above are from the ID 15973 data. The green and blue lines show the moving averages and moving standard deviations, respectively; the window of the calculation was 200.

Discussion

The data distributions found in the public resources [20] and rice cDNA microarray [15] demonstrate the appropriateness of the three-parameter lognormal model for microarray data distribution. All the probability plots show wide ranges of coincidence between the normalized data and the model (Fig. 2). Small classes of data, at the largest and smallest intensities, are inconsistent with the model, but these appear to be due to measurement errors rather than the inappropriateness of the method. If this assumption holds true, we can expect lognormal distributions in the transcript levels of genes, i.e. the number of mRNA molecules transcribed from a gene and accumulated in a cell. Since microarrays can be considered as measurements of random samples of the transcript levels, and the population has the same distribution manner with its random samples, the transcript levels must be lognormal. It may be the common nature of cells, since this distribution is found ubiquitously in many experiments on different DNA chips.

The assumed stability in the distribution of transcript levels, which was the basis of the conversion of *z*-scores to ratios, may represent the state of real cells in a sample. Stabilities of the lognormal distribution, the expected distribution of transcript levels, can be observed from those of the two parameters, σ and μ. Since the parameter σ may not be affected by experimental conditions, the value for the population should be the same as that determined from microarray data. The stability of the parameter σ is observed clearly in data (Table 1). Unfortunately, the stability of μ cannot be confirmed from microarray data, which has a relative nature; experimental conditions will affect μ. The relativity is derived not only from the signal detection method, but also from the RNA sample preparation process. However, since σ is stable, changes in μ mean that most of the genes change their expression levels down or up simultaneously. Such synchronous decrease or increase of materials may rarely occur in cells, which otherwise show homeostatic natures.

In a pair of normalized data sets, the ratios calculated by means of the three-parameter model were distributed approximately lognormally and the distribution was found to be stable in relation to the signal intensities (Fig 5). Since each normalization method is based on different assumptions, each of which reflects the criteria used to evaluate the intensity of data or difference between data; different methods can lead to different conclusions. Interestingly, the distribution of log-ratios fortuitously satisfies the assumptions that are used in other normalization methods. For instance, where only a limited number of gene expression levels are changed and they are well-behaved [25], stability of signal versus log-ratio [1-4] and intensity independence of measured log-ratio [1,5-7] can be expected. Since these assumptions have been adopted from a biological point of view, the stable distributions may be an *a priori* characteristic of the differences in transcript levels. Confirming such a characteristic in real data might be another means of verifying the appropriateness of the parametric normalization method: finding the real background as well as the center and deviation of data distribution, and managing the expressional changes. Certainly, such stabilities will be a great value in data mining on a ratio-basis [1], since a fixed threshold value can be used to select affected genes in sets of experiments. Furthermore, such stability will help in designing comparisons of data [23] by reducing the possibility that different designs will lead very different conclusions.

In normalization of microarray data, treatment of data at lower intensities can seriously affect the calculation results. According to the expected lognormality in signal distributions, the range of signal data necessarily becomes quite wide, and this characteristic complicates exact measurements at the lower and higher intensity ends of the signal. Additionally, unlike errors that are caused by signal saturation, which can be resolved simply by re-scanning the DNA chip at lower excitation intensity, the additive noise is difficult to cancel or reduce. Such additive noise will critically damage faint signals. If such tainted data is included in the normalization process, the additive noise can affect the entire data set. In order to avoid such effects, the choice of a robust calculation method will be important. For example, the parameter calculation used in this article uses data only within the interquartile range. Of course, in cases in which the additive noise becomes comparable to the lower quartile, even this method will become noise-sensitive.

The range of data that are inconsistent with the distribution model should be canceled prior to further bioinformatic analyses, since such data may contain additive noise at a level that seriously affects the signal. Generation of such a data class can be simulated using simple calculations, for example, addition of random numbers to an ideal series that are lognormally distributed. Such noise numbers will create a bend in the probability plot for the resulting series (data not shown). Indeed, in the inconsistent signal range of data, determined *z*-scores showed low reproducibility in repeated hybridizations (Fig. 3) and in the calculated ratios in the dye-swap experiment (Fig. 8). The low reproducibility is not derived from the parametric normalization method, since the corresponding range of data normalized by LOWESS also lacks reproducibility (Fig 7). Cancellation of such data classes by the model does not mean sacrificing a range of measurement; rather, it can prevent a waste of labor, which is often initiated by noise in data.

Many experimental errors are possible sources for additive noise; however, in my experience, critical ones that can affect a large part of data are derived from insufficient signal intensity or uneven hybridization. Shortage in the amount and/or inadequate quality of RNA would lead to the former problem. Unfortunately, re-scanning of hybridized chips with higher extinction power rarely changes the signal/noise ratios; it might expand the noise levels as well as the signal level. Unevenness of hybridization might be compensated with various calculation methods. However, such compensations require information for the differences and/or similarities in the unevenness between the background and the signal; for example, if the unevenness occurs on the backgrounds at the same rate on the signals, compensation should be performed before the g subtraction. In contrast, the parametric normalization method does not determine the level of multiplicative noise from a set of data. The noise can occur from variation of DNA amount in each spot, and this would cause errors in the determination of expressional changes. The error can be cancelled in multi-color comparisons within the same chip's data; however, it will appear in inter-chip comparisons of data.

Conclusions

A close fit was found ubiquitously between the three-parameter model and real data. The coincidence was stable across biological treatments of subjects. Such commonness and stability in the manner of distribution can be explained without inconsistency if these features are *a priori* characteristics of a living cell.

Using the distribution model, data were successfully handled parametrically. The calculation methods for the data ratios as well as for normalization were introduced. Some characteristics found in the normalized data and in the results obtained from the analysis showed improved data handling in the following categories:

Advances in data accuracy and reliability. Normalized data and calculated ratios showed high levels of experimental reproducibility. Moreover, it was shown that the normalization method could identify the noise-affected ranges of data intensity, allowing for the exclusion of affected data prior to detailed analysis. Calculated ratios and their determination reproducibility were independent of signal intensity.

Expansion of the groups of experiments and of measurement methods that can compare data. The commonness of data distribution suggests that the model-based method may be applicable to a wide range of experiments. At least, the removal of the need for special reference RNA hybridization means that data comparisons are no longer restricted. Additionally, differences between the normalized data can be translated to a ratio basis. Indeed, the calculated ratios correlated closely with those of Northern analysis. It became possible to compare and integrate the ratio-basis results among experiments and/or with other measurement methods.

As mentioned above, and summarised in Table 2, it is clear that data normalization and comparison based on the three-parameter lognormal distribution model will markedly improve the handling of microarray data.

Methods

Data resources

Data used in this article were obtained from open resources at Stanford University [20] or from experiments using rice seedlings [15].

The lognormal distribution model and estimation of the parameters

The method assumes that the original intensity data, (r_i) for i = 1,2...n, obey a lognormal distribution. The probability density function of the intensity data used was:

$$f(r_i) = (1/\sigma/(2\pi)^{1/2})\exp[\{-\log(r_i-\gamma) + \mu^2/(2\sigma^2)] \text{ for } r_i > \gamma,$$

where σ, μ and γ are the shape, scale and threshold parameters, respectively.

The parameter σ was found through trial and improvement calculation processes; in the trial, the distribution of $\log(r_i-\gamma)$ was checked by normal probability plotting [26], and the value that gave the best fit to the model was selected for γ. The fitness was evaluated by the sum of absolute differences between the model and $\log(r_i-\gamma)$, within the interquartile range of data. The parameter μ was found as the median of $\log(r_i-\gamma)$, and the parameter σ was found from the interquartile range of $\log(r_i-\gamma)$; these are known as robust alternatives for the arithmetic mean and standard deviation, respectively. Parameters μ and σ were found for each data grid, a group of data for DNA spots that were printed by an identical pin in order to avoid divergences caused by pin-based differences [27]. *Z*-normalization was carried out for each datum as

$$Z_{ri} = \{\log(r_i-\gamma)-\mu\}/\sigma.$$

Intensity data (r_i) less than γ were treated as "data not detected", since such data might contain negative noise larger than the signal (see Results).

Northern analyses

RNA samples were obtained from a time course experiment on rice seedlings exposed to cold-stress [15]. During the time-course experiment, 7 clones were randomly selected from those showing a higher magnitude of

Table 2: Comparative table of normalization methods

	3-parameter lognormal method [this work]	LOWESS [2.3]	house-keeping genes	globalization [7]	A N O V A [9]	variance stabilization [5,6]
assumed stable character	statistical data distribution	constant ratio tendency	expression levels of particular genes	sum of signal data	smallest differences in log(data)	smallest differences in arsinh(data)
Can the assumption be verified?	yes [22,26] Figs. 1,2	no	yes[1]	no	no	no
units for expression level	z-score	ratios to a reference	ratios to the stable genes	fraction (ppm)	ratios to a reference	statistical differences to a reference
data transformation for the adjustment	subtraction of a constant	non-linear	no	no	linear on logarithms	linear
numbers of data sets normalized in a calculation	1	2	1	1	all the sets to be compared	2
Can it compare multiple data sets without reference RNA?	yes	no	yes	yes	yes	no
amount of calculation	medium	medium	least	least	vast	large-medium
reproducibility	yes Figs. 3,8	no Fig. 7	nt[3]	no [7]	no[2]	no Fig. 6cd
Is the ratio tendency independent of signal intensity?	yes Figs. 5,8	yes Fig. 6b	nt[3]	no Fig. 6a	no[2]	yes Fig. 6cd
Is the ratio variance independent of signal intensity?	yes Figs. 5,8	no Fig. 6b[7]	nt[3]	no Fig. 6a	no[2]	yes Fig. 6cd
Can it find the level of additive noise in the data?	yes	no	no	no	no	no

[1]A possible method is as follows: three or more gene candidates are chosen for the stable expression. Among the candidates, the ratio of every pair of data for a microarray experiment are calculated. If the expression levels are stable, the ratios are also stable. Such a combination of genes, however, has not been reported. [2]Data are not presented. [3]Since the stable genes could not be found among the data sets used in this article, this method was not tested (also see a review article [11]).

increase or decrease (more than 1.5-fold) from microarray experiments that were normalized with globalization, and 7 clones were selected by a totally random manner. For those clones, northern blotting analyses were performed with the same RNA batch that was used for probing microarray experiments. Radioactivity of detected bands on probed membranes was measured using the BAS system (Fuji). For each band on an image, the signal intensity was detected as the sum of signal values in pixels within the band. The background was estimated based on the average intensities of the electrophoresis lane but excluding the band itself. The relative signal of a band was calculated by subtracting the background from the intensity data. Each signal datum was normalized by creating ratios to the control samples. For 4 clones out of 14 clones, northern analyses could not detect the signals.

Acknowledgements

I would like to thank M. Araki and K. Takahashi for their assistance in microarray experiments; and Drs. S. Youssefian and H. Wabiko for their comments on the manuscript. Some of the data used were from a study that was supported by an MAFF Rice Genome Project grant, MA-2109. A data normalization service based on this method is commercially available at http://www.skylight-biotech.com.

References

1. Chen Y, Dougherty ER, Bittner ML: **Ratio-based decisions and the quantitative analysis of cDNA microarray images.** *J Biomed Optics* 1997, **2:**364-374.
2. Tseng GC, Oh MK, Rohlin L, Liao JC, Wong WH: **Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects.** *Nucleic Acids Res* 2001, **29:**2549-2557.

3. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP.: **Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.** *Nucleic Acids Res* 2002, **30:**e15.
4. Workman C, Jensen L, Jarmer H, Berka R, Gautier L, Nielser H, Saxild HH, Nielsen C, Brunak S, Knudsen S: **A new non-linear normalization method for reducing variability in DNA microarray experiments.** *Genome Biol* 2002, **3:**RESEARCH0048.
5. Huber W, Von Heydebreck A, Sultmann H, Poustka A, Vingron M: **Variance stabilization applied to microarray data calibration and to the quantification of differential expression.** *Bioinformatics* 2002, **18:**S96-104.
6. Durbin BP, Hardin JS, Hawkins DM, Rocke DM: **A variance-stabilizing transformation for gene-expression microarray data.** *Bioinformatics* 2002, **18:**S105-110.
7. Quackenbush J: **Microarray data normalization and transformation.** *Nat Genet* 2002, **32(Suppl):**496-501.
8. Rocke DM, Durbin B: **A model for measurement error for gene expression arrays.** *J Comput Biol* 2001, **8:**557-569.
9. Kerr MK, Martin M, Churchill GA: **Analysis of variance for gene expression microarray data.** *J Comput Biol* 2000, **7:**819-837.
10. Quackenbush J: **Computational analysis of microarray data.** *Nat Rev Genet* 2001, **2:**418-427.
11. Sherlock G: **Analysis of large-scale gene expression data.** *Brief Bioinform* 2001, **2:**350-362.
12. Bilban M, Buehler LK, Head S, Desoye G, Quaranta V: **Normalizing DNA microarray data.** *Curr Issues Mol Biol* 2002, **4:**57-64.
13. Yang YH, Buckley MJ, Speed TP: **Analysis of cDNA microarray images.** *Brief Bioinform* 2001, **2:**341-349.
14. Olshen AB, Jain AN: **Deriving quantitative conclusions from microarray expression data.** *Bioinformatics* 2002, **18:**961-970.
15. Konishi T: **Parametric treatment of cDNA microarray data.** *Genome Informatics* 2002, **13:**280-281 [http://www.jsbi.org/journal/GIW02/GIW02P166.pdf].
16. Hoyle DC, Rattray M, Jupp R, Brass A: **Making sense of microarray data distributions.** *Bioinformatics* 2002, **12:**576-584.
17. Eisen MB, Brown PO: **DNA arrays for analysis of gene expression.** *Methods in Enzymology* 1999, **303:**179-205.
18. Kepler TB, Crosby L, Morgan KT: **Normalization and analysis of DNA microarray data by self-consistency and local regression.** *Genome Biol* 2002, **3:**RESEARCH0037.
19. Yue H, Eastman PS, Wang BB, Minor J, Doctolero MH, Nuttall RL, Stack R, Becker JW, Montgomery JR, Vainer M, Johnston R: **An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression.** *Nucleic Acids Res* 2001, **29:**e41.
20. Gollub J, Ball CA, Binkley G, Demeter J, Finkelstein DB, Hebert JM, Hernandez-Boussard T, Jin H, Kaloper M, Matese JC, Schroeder M, Brown PO, Botstein D, Sherlock G: **The Stanford Microarray Database: data access and quality assessment tools.** *Nucleic Acids Res* 2003, **31:**94-6.
21. Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JCF, Trent JM, Staudt LM, Hudson J Jr, Boguski MS, Lashkari D, Shalon D, Botstein D, Brown PO: **The transcriptional program in the response of human fibroblasts to serum.** *Science* 1999, **283:**83-87.
22. Ichihara K: *Statistics for Biosciense – practical teqhnique and theory Tokyo: Nankodo*; 1990.
23. Churchill GA: **Fundamentals of experimental design for cDNA microarrays.** *Nat Genet* 2002, **32(Suppl):**490-495.
24. Horvath DP, Schaffer R, West M, Wisman E: **Arabidopsis microarrays identify conserved and differentially expressed genes involved in shoot growth and development from distantly related plant species.** *Plant J* 2003, **34:**125-134.
25. Zien A, Aigner T, Zimmer R, Lengauer T: **Centralization: a new method for the normalization of gene expression data.** *Bioinformatics* 2001, **17:**S323-331.
26. **NIST/SEMATECH e-Handbook of Statistical Methods** [http://www.itl.nist.gov/div898/handbook/]
27. Schuchhardt J, Beule D, Malik A, Wolski E, Eickhoff H, Lehrach H, Herzel H: **Normalization strategies for cDNA microarrays.** *Nucleic Acid Res* 2000, **28:**e47.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral