Operational criteria for selecting a cDNA microarray data normalization algorithm

C. ARGYROPOULOS[1], A.A. CHATZIIOANNOU[2], G. NIKIFORIDIS[1], A. MOUSTAKAS[3], G. KOLLIAS[2] and V. AIDINIS[2]

[1]Laboratory of Medical Physics, Medical School, University of Patras, 26110 Patras; [2]Institute of Immunology, Biomedical Sciences Research Center 'Alexander Fleming', 16672 Athens, Greece; [3]Ludwig Institute for Cancer Research, Biomedical Center, SE-75124 Uppsala, Sweden

Received September 6, 2005; Accepted October 3, 2005

Abstract. Microarray technology allows gene expression profiling at a global level. Many algorithms for the normalization of raw microarray data have been proposed, but no attempt has yet been made to propose operationally verifiable criteria for their comparative evaluation, which is necessary for the selection of the most appropriate method for a given dataset. This study develops a set of operational criteria for assessing the impact of various normalization algorithms in terms of accuracy (bias), precision (variance) and over-fitting (information reduction). The use of these criteria is illustrated by applying the three most widely used algorithms (global median normalization, spiked-in based normalization and lowess) on a specifically designed, multiply-controlled dataset.

Introduction

cDNA microarray technologies are hybridization-based methods that enable the simultaneous profiling (quantification of expression) of thousands of genes. Emerging and evolving computational methods aim at a more precise analysis of rapidly accumulating microarray data. A prerequisite to any form of microarray analysis is the process of data normalization, which is defined as a transformation of the data that address the random and systemic signal variability, and is intrinsic to every microarray experiment. This variability stems from a number of sources, including chip-to-chip manufacturing differences; unsteady laboratory sample preparation, hybridization and washing protocols; imprecise signal measurements coming from the scanner; and subtle gene-to-gene differences in hybridization efficiency (1). Given the documented impact of normalization on subsequent steps in analysis (2), the proliferation of research on normalization methods (1,3-14), claiming superiority-equivalency over alternative methods is more than justified.

Irrespective of the specific methodology employed, a normalization method is essentially a tripartite process: first, a subset of genes from the targets spotted is selected; second, the expression ratios are fed into a mathematical functional formalism (either parametric or non-parametric); and last, the estimated functional is applied back to the raw data in order to generate normalized measurements. The various proposed formalisms differ in: a) selection process of the gene subset; and b) the specific functional-estimation process employed. However, it is interesting to note that comparisons (when made) usually refer within and not across groups of possible normalization strategies, and methods are normally compared based on how 'straight' scatterplots appear after normalization. Explicitly defined, criteria for comparisons have only recently been utilized (2,8,9,15), but a comprehensive framework that could be used to compare normalization algorithms and practical repercussions of selecting one method over another is still lacking.

In this context, we introduce two non-mutually exclusive views of normalization, namely the calibration and quantitative measurement method perspectives. These perspectives suggest measures of accuracy and precision that can gauge bias and variance reduction and thus derive operationally definite criteria for the comparison of normalization strategies by applying the same graphical and statistical tools used in method agreement clinical research studies (16). Illustration of the use of these criteria is exemplified by comparing the three most widely used normalization strategies: global median, spike-based control and lowess. We then examine the issue of over-fitting, a neglected area in normalization algorithm research, and propose the utilization of theoretic measures to examine information reduction. The proposed framework is operationally definite (and hence verifiable) and could be used not only to compare novel normalization algorithms, but also provide a checklist for the researcher who has read the relevant literature and must choose an algorithm to use for his or her dataset. It is anticipated that the application of the described criteria to the normalization

Correspondence to: Dr V. Aidinis, Institute of Immunology, Biomedical Sciences Research Center 'Alexander Fleming', 16672 Athens, Greece
E-mail: v.aidinis@fleming.gr

Key words: cDNA microarray, normalization method, algorithm, spiked-in normalization, lowess, global median

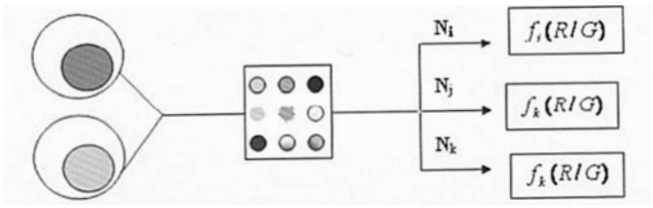

Figure 1. Normalization as measurement. A cDNA microarray is conceptualized as a quantitative measurement method to compare gene expression in two or more biological systems (cells are dyed with red and green fluorescent dye, scanned, then subsequently overlapped and compared). The raw data can be fed into a variety of existing or hypothesized normalization algorithms (Ni, Nk and Nj), which in turn produce different estimations of the unknown gene expression ratio.

of a given microarray dataset would allow for the selection of the most appropriate strategy.

Materials and methods

Arthritic mouse microarray dataset. Tg197 transgenic mice overexpress the gene for the human tumor necrosis factor (hTNF) cytokine and spontaneously develop a severe form of rheumatic disease similar to human rheumatoid arthritis (RA) within 3-4 weeks after birth (17). The arthritic mouse microarray dataset (AMMD) was designed with the specific objectives of: a) understanding global changes in gene profile in the joints of the Tg197 animals as they progress from the normal to diseased phenotype; b) describing clinicopathological and molecular correlates of the disease model; c) discovering downstream targets of TNF signaling that could form the basis of novel drugs against RA; and d) identifying specific disease markers that could be utilized in everyday clinical practice.

In addition and implicit to the AMMD data generation and analysis process was the use of a rigorously controlled experimental and statistical strategy that reflects the strengths and limitations of a decade of microarray research. The dataset incorporates multiple levels of controls: a) spiked-in controls from exogenous, *in vitro* transcribed, bacterial genes at known and varying concentrations (30, 150, 300 1500 and 3000 pg/μl, 5 spots per grid, and 324 spots per array slide); b) empty spots (3 per grid) as negative controls and estimation of background; c) a common reference RNA sample consisting of equal amounts (1:1) of pooled RNA from all diseased, transgenic animals and wild-type controls at equimolar concentrations; d) triplicate hybridizations at every experimental point; and e) three self-self reference sample hybridizations. The MIAME compliant (18) dataset has been submitted to ArrayExpress database (http://www.ebi.ac.uk/arrayexpress/), reference ID pending, and will be publicly released upon publication of the biological interpretation of the results (data not shown). In the current study, only the aspects relevant to comparative normalization algorithm evaluation will be detailed.

Total RNA samples were isolated through disease progression from the joints of arthritic transgenic mice at 2-week intervals (from 2 to 12 weeks) from healthy wild-type littermate animals (a pool aged to the corresponding weeks), and transgenic mice under the prophylactic or therapeutic administration of a disease neutralizing antibody (α-TNF). Each RNA sample consisted of equimolar amounts of total RNA isolated from two male and two female littermate mice of the selected age. All samples were fluorescent-labeled with direct incorporation of Cy3 (reference sample) or Cy5 (experimental samples) and hybridized to Sanger mouse 15K (Mver1.1.1) cDNA glass microarray slides (19) essentially as described by Sanger (http://www.sanger.ac.uk/Projects/Microarrays). Hybridized slides were scanned with the confocal ScanArray Express scanner (Packard Biosciences) utilizing ScanArray software and quantified with the QuantArray software (both Packard Biosciences).

Assessing agreement between normalization methods. The field of quantitative method comparisons offers a general framework and tools for comparative normalization strategy analysis (20). Normalization algorithms can be examined as quantitative (measurement) methods, given the quantitative nature of both the experimental and algorithmic parts of the relation. The experiment measures the biological system in question, and the normalization algorithm makes calculations based on features present in the raw data in an attempt to infer relations present in the biological system (Fig. 1).

Although correlation coefficients are usually quoted as a measure of method agreement and repeatability, their use is fraught with methodological problems (16,21). In particular: i) correlation coefficients measure the strength of the relation between two techniques, but not their agreement. Perfect agreement (i.e. clustering of measurements along the 45˚ line in a scatterplot) is not synonymous with perfect correlation (i.e. tight clustering of repeat measurements along any line; ii) change of scale influences agreement between the two methods, but it does not affect the correlation; iii) correlation depends on the range of the measurement scale with wider ranges associated with higher correlation coefficients; and iv) tests of significance applied to correlation coefficients may show that the two methods are related, but do not indicate agreement. High correlation coefficients can be associated with poor agreement. This is most obvious when considering the impact of normalization on microarray data (Fig. 1) displaying how normalization results in the obvious improvement in agreement of numeric measures of gene expression, without any noticeable effect on the correlation coefficient (22). Points 1-4 are graphically illustrated in Fig. 2 with a simulation study corresponding to a hypothetical self-self hybridization experiment. Uncorrected scale and location bias (Fig. 2b and c) does not affect the correlation coefficient compared to optimally normalized values (Fig. 2a), even though the actual measurements are influenced considerably. On the other hand, high- or low-pass filtering of optimally normalized values (Fig. 2d and e) leads to a reduction of the correlation coefficient even in the case of optimally normalized experiments. Taken together, the figures demonstrate that it will be futile to compare normalization methods, which in general affect scale and location of the original data, on the basis of correlations between coefficients.

Bland-Altman (MA) plots are better suited for this kind of analysis since they are robust with respect to shifts in scale, location and range. They can be utilized not only to assess the limits of quantitative agreement between two methods, i.e. microarrays and immunoassays, for tumor markers (23)

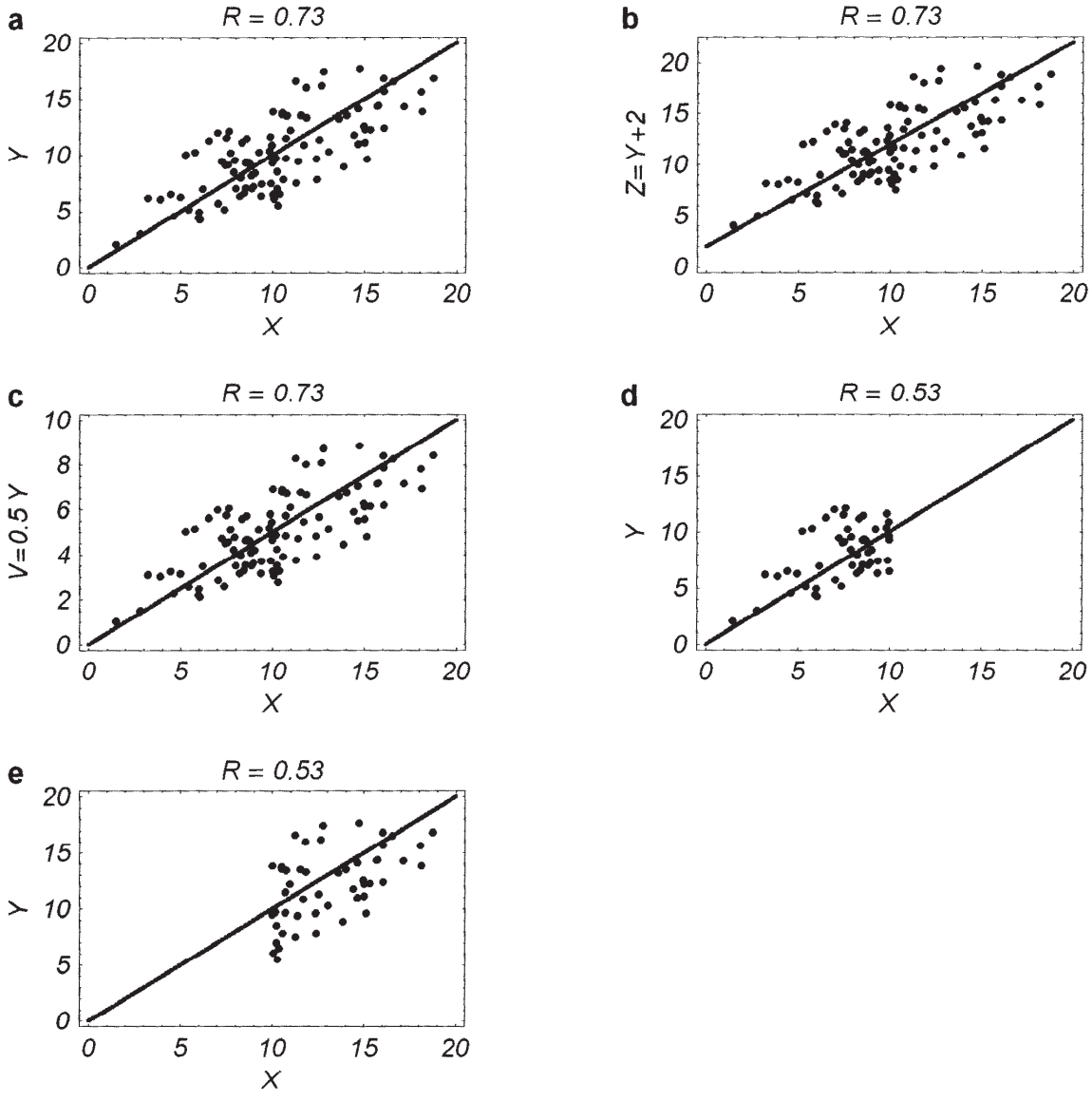

Figure 2. Quantitative method comparison using the correlation coefficient is misleading. A total of 100 points from a normal distribution with mean μ=10 and standard deviation σ=3 were independently drawn in order to simulate a two self-self hybridization experiment. Independently distributed noise in the [-3,3] was added to each data point, and two datasets X and Y were thus generated. Correlation coefficient of X and Y is R=0.73 (a). The same correlation coefficient is insensitive to location and scale bias (datasets Z and V) (b and c, respectively). Low- (d) and high- (e) pass filtering of the original datasets X and Y reduce the coefficient correlation.

or even quality of life indices (24), but also derive relationships of bias versus variance and provide necessary corrections by regressing differences against averages. To use the MA plots, the following steps are required: i) normalize the internal validation data subset, upon which the performance of the compared algorithms is tested, by using both algorithms; ii) calculate normalized expression ratios $\log(R/G)_{i,1}$, $\log(R/G)_{i,2}$ and average signals $[\log(R)+\log(G)/2]_{i,1}$, $[\log(R)+\log(G)/2]_{i,2}$ for every probe *i*, channel (R,G) and normalization method *1,2*; iii) plot the average difference of normalized values against the average normalized expression; iv) if no trend is evident from the graph, find the 95% limits of agreement (25) between the two techniques using confidence intervals based on the familiar paired t-test formula:

$$\bar{\delta} \pm 1.96 \frac{\sqrt{\frac{\sum_i (\delta_i - \bar{\delta})^2}{N-1}}}{\sqrt{N}}$$

v) if trends are present, then the analysis can proceed using errors-in-variable regression using any of the currently available formalisms (i.e. orthogonal least squares, method of moments, and non-parametric methods) (21,25).

Within this article, the symbolism log stands for base 2 logarithms. Algorithms that are found to be in agreement across the intensity range can be classified together, thus aiding the researcher in making comparisons between and within groups of algorithms. The MA plot analysis of the hypothetical dataset of Fig. 2 is shown in Fig. 3. Normalization methods that cannot correct for location and scale bias are associated with scatterplots that cluster away from the x axis (Fig. 3b and c). Range restriction of an optimal normalization algorithm (Fig. 3a) to either high or low values does not affect the MA plot (Fig. 3d and e), thus avoiding the difficulties associated with correlation coefficients.

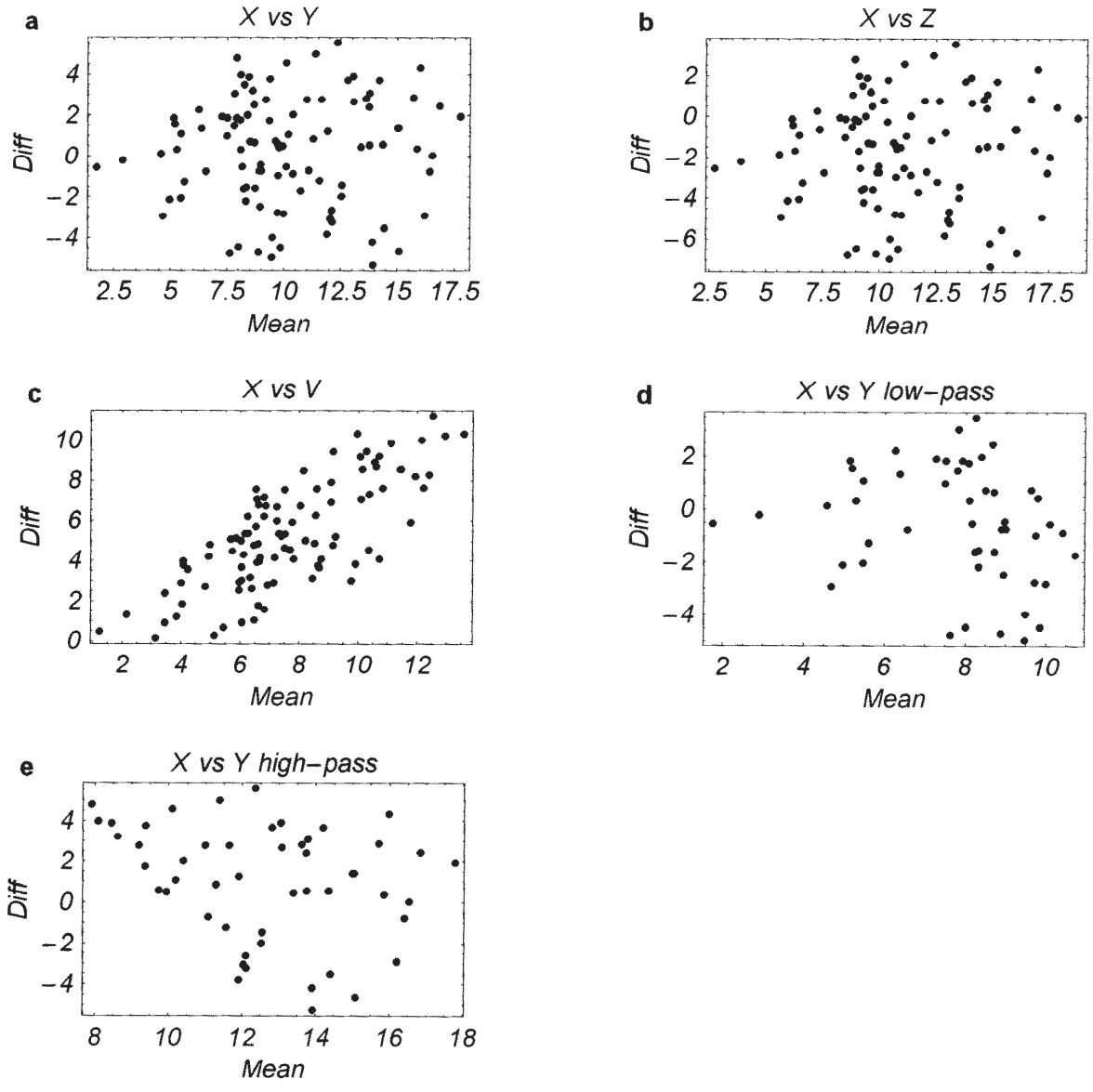

Figure 3. Assessing relative bias and variance with Bland-Altman (MA) plots. An MA plot demonstrates the difference of repeated measures versus their average. The MA plot of an optimally normalized self-self hybridization experiment features points clustered around zero (a). Uncorrected location and scale bias, dataset Z-X and V-Y (b and c, respectively) manifest as deviation from 0. Low- and high-pass versions of unbiased normalization algorithms (d and e, respectively) are not penalized simply because they shrink the dynamic range. The same datasets from Fig. 2 were used.

Normalization as calibration. Calibration is fundamental to achieving consistent measurements and usually involves establishing a relationship between an instrument response and one or more reference values. The calibration problem consists of both estimating and validating the functional relationship. If one foregoes the possibility of *a priori* estimating the expression of tens of genes, then estimation tools from the calibration theory are not applicable (26). However, it is possible to post-hoc validate a proposed normalization strategy if the design strategy has included a number of experiments that can be used for internal validation (27). In a microarray experiment, the internal validation subset can be formed from commonly utilized controls as exogenous spiked-in genes and self-self hybridizations. Estimates of accuracy (i.e. proximity of estimates of the method compared to the hypothesized true values) and precision (i.e. reproducibility of results in subsequent repetitions) can easily be computed from the performance of the normalization method on these subsets. The precision can be characterized by measures of dispersion in the distribution of repetitive measures, and accuracy demands the use of reference internal validation subsets.

Validation using spiked-in controls. Of all possible gene expression measures, the channel ratio was selected and used for the purpose of this study (28).

If we define the measurement as the channel log ratio:log $(R/G)_i$ and take the collection of spiked-in control genes as a reference, then we can use the following definitions of bias and variance as proxies for accuracy and precision, respectively:

$$Bias_i = \sqrt{\frac{1}{n}\sum_j \sum_k \left(\log\left(\frac{R}{G}\right)_{i,j,k} - True\ \log\ Ratio\right)^2}$$

$$Variance_i = \frac{1}{n-1}\sum_j \sum_k \left(\log\left(\frac{R}{G}\right)_{i,j,k} - \left\langle \log\left(\frac{R}{G}\right)_i \right\rangle\right)^2$$

Table I. Comparison of normalization strategies.

Gene Subset	Method	Functional	Effect
All	Global Median	$\log\left(\frac{R}{G}\right)_i \rightarrow \log\left(\frac{R}{G}\right)_i - k_{all}$ $k_{all} = median(\log\left(\frac{R}{G}\right)_i)_{i=1}^{all}$	Center distribution of expression ratio for all genes around zero
Controls (spike – in, housekeeping)	Median	$\log\left(\frac{R}{G}\right)_i \rightarrow \log\left(\frac{R}{G}\right)_i - k_{controls}$ $k_{controls} = median(\log\left(\frac{R}{G}\right)_i)_{i=1}^{controls}$	Center distribution of expression ratio for control genes around zero
All (within grid)	Local weighted regression	$\log\left(\frac{R}{G}\right)_i \rightarrow \log\left(\frac{R}{G}\right)_i - f(A_i)$ $f(A_i) = lowess(MA)_i$	Center distribution of M vs. A values around zero throughout intensity range

Global median normalization, which scales data to have a median expression ratio of 0 is the most common method employed thus far. Another potential strategy is to scale data so a subset (usually spiked-in controls) of genes has a median expression ratio of 0. Lowess is a non-parametric strategy that normalizes the genes located in a local neighborhood of a MA scatterplot to a mean log expression ratio of 0.

In these formulas, the subscript j represents all hybridizations, k is all within-slide replicates, i is the subscript for the i^{th} control gene, *n* is the product of hybridizations x within-slide replicates, < > is the expectation (or averaging) operator, and true log ratio is equal to the ratio of concentrations of spiked-in controls used in the labeling reactions for the two channels.

Since spiked-in controls were introduced at the same concentration in the hybridization reactions, the true log ratio is equal to 0 and the formula for the bias reduces to the root mean square error (RMSE):

$$Bias_i = RMSE_i = \sqrt{\frac{1}{n}\sum_j \sum_k \left(\log\left(\frac{R}{G}\right)_{i,j,k}\right)^2}$$

If a number of controls are introduced (at varying concentrations), then one could examine how bias and variance change across the intensity range. A normalization method A would be preferred over method B if, in addition to small bias and variance, it was associated with a constant performance throughout the mRNA concentration range. An easily computed statistical measure of constancy is provided by the coefficient of variation of replicate spiked-in spots, which should remain constant across the range of mRNA concentration for a preferred method.

Confidence intervals under the normal error model can also be constructed to calculate the limits of agreement between the two normalization techniques (25), using the familiar t-test distribution by averaging all spiked-in controls, hybridizations and array replicates.

Validation using self-self hybridizations. In the same spirit, replicate self-self hybridizations could be performed and provide additional evidence for or against a particular normalization method. In essence, they allow for the calibration of normalization methods by revealing inconsistencies across the signal intensity range. Whereas spiked-in controls spotted in known concentrations/ratios allow such a comparison, the small number of included genes limits its statistical power. Assuming that a signal is monotonically dependent on mRNA concentration, self-self hybridizations afford a broader view of the concentration-response performance of the experiment-normalization combination compared to spiked-in controls.

Avoiding over-fitting: normalization as data compression. It is fairly obvious that any normalization method will lead to a reduction of expression ratio variability compared to the un-normalized values. The variability of the latter is not only due to technical factors, but also the inherent biological variability of the systems examined. An optimal algorithm should at least partially correct for systematic errors (hence, reducing technical variability), but at the same time not over-normalize the gene expression values. If this happens, potential biological differences are suppressed to the point

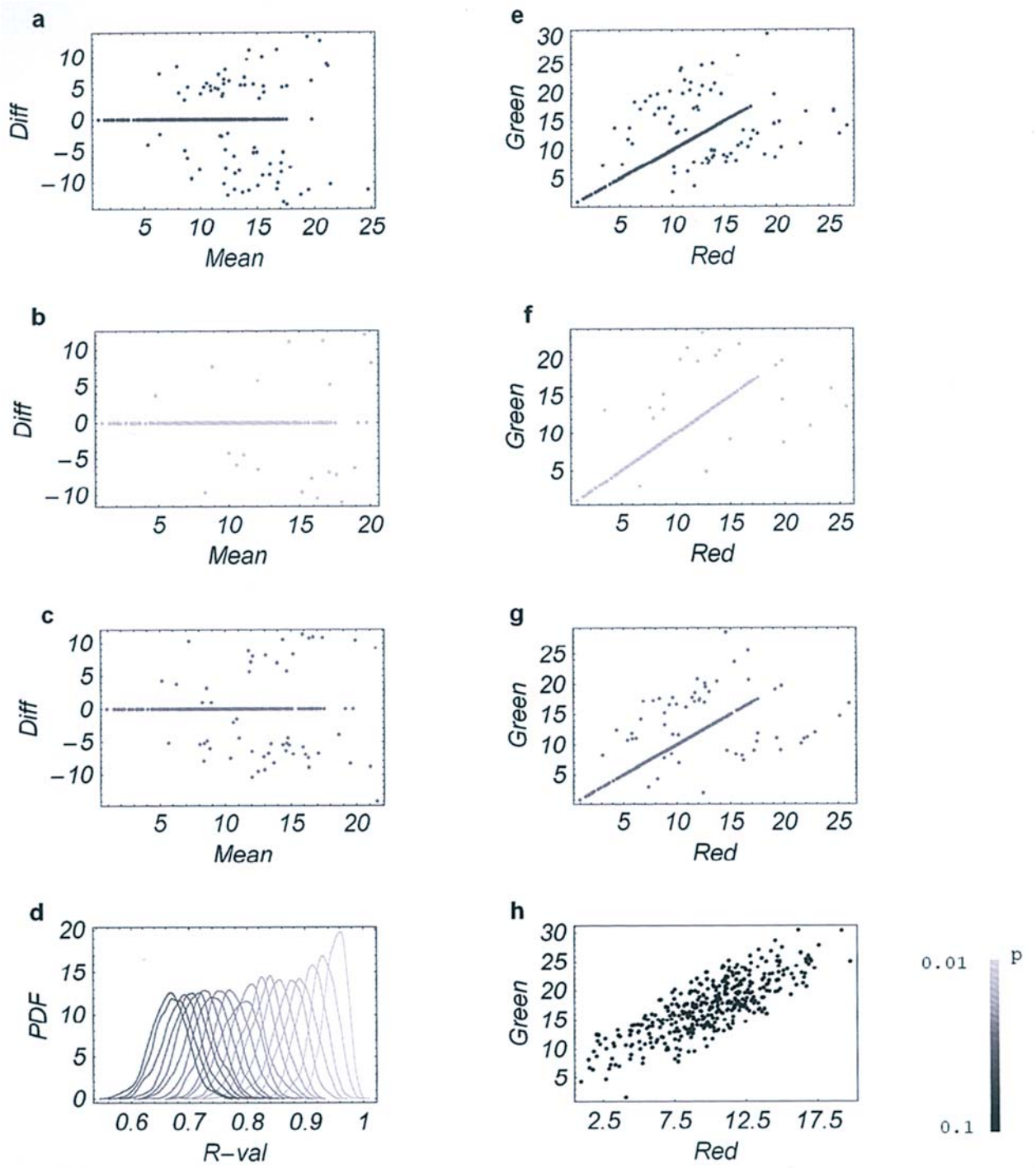

Figure 4. The ChannelFlip algorithm. The algorithm randomly assigns a log expression ratio of 0 to (1-p)% of genes, flips p/2% of the raw expression ratio data, while leaving the rest unchanged. Three transformations of the original dataset (h) are shown both as scatterplots (b, d and f) and MA plots (a, c and e). The graphs are color coded according to the value of *p*, which is shown in the scale (g). R, correlation coefficient; PDF, probability density function; R&G, red and green 'channels'; diff, R-G; mean, (R+G)/2.

that no conclusions about differential effects can be reached.

A global view of technical variability reduction is afforded by the internal validation subset of self-self hybridizations; in this case, total variability is due to technical factors only. Assessing the impact of normalization on biological variability is accomplished by looking at the remainder of the dataset. Reduction of variability can be quantified with theoretic criteria. A less variable distribution of expression-ratio amounts to data compression and hence entropy reduction of the distribution. The entropy of a distribution over a partition X of log-expression ratio range is given by:

$$H = -\sum_{i \in X} p_i \times \log(p_i)$$

The tendency of an algorithm to over-normalize will also be reflected in the divergence of distributions of the two sets

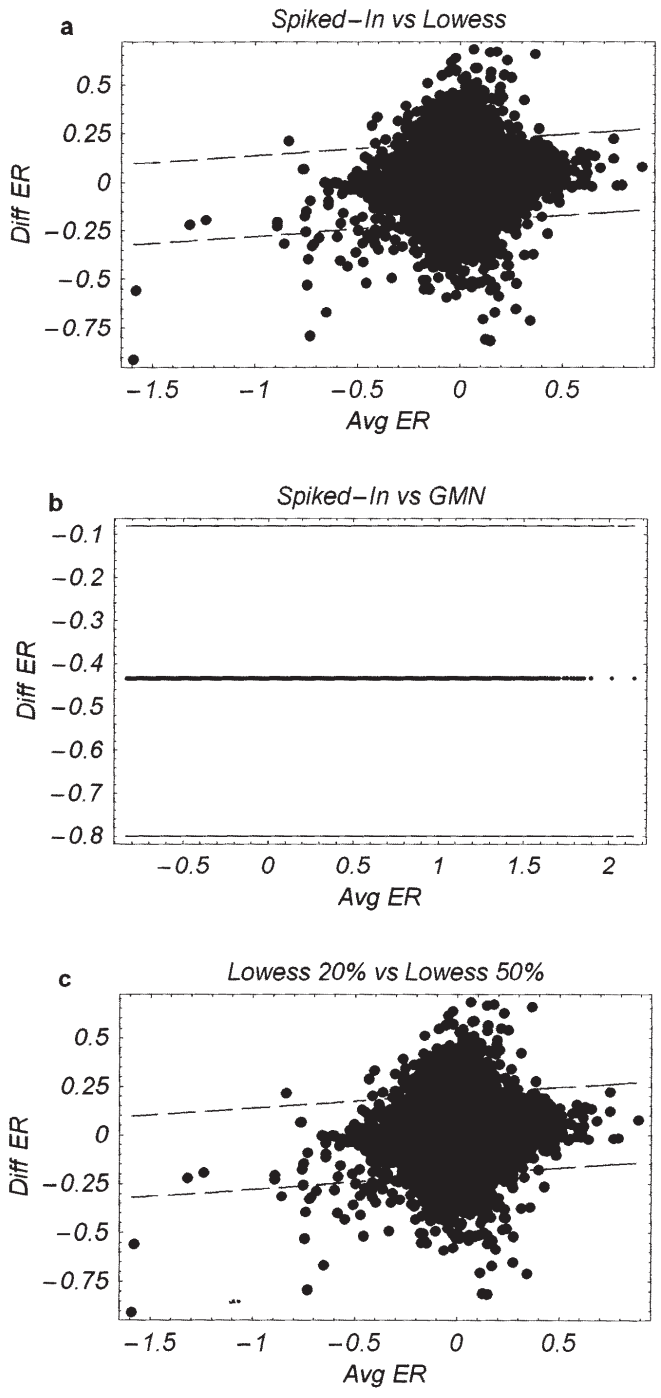

Figure 5. Evaluating normalization agreement using Bland-Altman (MA) plots. The difference of the normalized log-expression ratio (ER) is plotted against the average normalized ER for repeated measures (i.e. application of two normalization algorithms on the self-self hybridization subset).

(self-self, non-self-self) after normalization. If the algorithm tends to over-normalize, divergence measures will decrease compared to the corresponding value before normalization. The Kullback-Leibler (KL) I divergence (or relative entropy) is a well-established measure of the distance between two distributions P, Q and can be used to provide a glimpse of the 'over-normalization' potential of each method (29).

The relative entropy of self-self over non-self-self hybridizations should decrease with the degree of over-normalization as differences between the data subsets are eliminated. In essence, the divergence measures allow us to quantify the reduction of variability beyond the component attributable to technical factors; a normalization algorithm A would be preferred over B if it maintains the divergence between the two subsets more than B.

Algorithms. The employed normalization strategies are given in Table I. These algorithms account for the majority of experimental work with microarrays in the published literature. The first method, global median normalization (GMN) calibrates log-expression ratios to a median of 0, and is the most commonly employed method. Alternatively, one could normalize the log-expression ratio of spiked-in controls (SBN) to a median of 0, and use this constant to normalize all other values. The third method is based on the lowess smoother (30), originally presented by Cleveland *et al* (31,32). A lowess smoother performs locally weighted polynomial and usually linear fittings, and is parameterized by the size of the local neighborhood (as a percentage of the dataset); between 20% and 50% of points are normally included in the local neighborhood, allowing the smoother to accommodate a wide variety of functional relationships between the predictor and response variables. To examine the dependency of lowess-based normalization on the size of the neighborhood, we also compared realizations of lowess using different values of the control parameter (32).

As a case study of over-normalization, we devised an algorithm called ChannelFlip. To each gene, the algorithm randomly: a) assigns a log-expression ratio of 0 (probability 1-p); b) reverses raw log-expression ratio with probability p/2; or c) leaves it unchanged (p/2). ChannelFlip assumes that the majority of genes should have an expression ratio that clusters along the 45˚ line in the scatterplot of R vs. G values, an assumption that is implicitly made in the 'real-world' strategies of Table I. The use of the algorithm is illustrated in Fig. 4 with a simulation; in the series, a typical dataset consisted of 400 points from a normal (10, 3) distribution corresponding to a self-self hybridization experiment. Subsequently, uniformly distributed noise in the (-3, 3) interval was added independently to the two channels. Finally, the 'red' channel was scaled and rotated relative to the first (Fig. 4h). Using an ensemble of such datasets, a researcher without access to the internal workings of the algorithm found the 'optimal' value of the control parameter *p* using maximization of the average intra-channel correlation coefficient (estimated from the empiric distributions of R values) as a criterion (Fig. 4g). Three normalizations of the original dataset in Fig. 4h are shown both as scatterplots (Fig. 4b, d and f) and as MA plots (Fig. 4a, c and e). The graphs are color coded according to the value of *p,* which is shown on the scale (Fig. 4g). It is evident that such an algorithm will shrink the variance of the dataset considerably, depending on the value of *p* (in the limit p=0, ChannelFlip effectively normalizes all genes to an expression ratio of 1). Due to its over-normalization nature, ChannelFlip is insensitive to both location and scale measurement bias for a wide range of the control parameter *p*. Such a method would almost certainly give excellent results in terms of accuracy and precision criteria, when assessed in situations where the

Table II. Average log expression ratio of control genes.

Concentration	20%	25%	30%	35%	40%	45%	50%	GMN	SBN
30 pg/μl	-0.128	-0.128	-0.125	-0.125	-0.126	-0.128	-0.127	0.555	0.105
150 pg/μl	-0.106	-0.108	-0.111	-0.112	-0.114	-0.116	-0.118	0.365	0.189
300 pg/μl	-0.066	-0.059	-0.053	-0.049	-0.045	-0.044	-0.044	0.167	0.004
1500 pg/μl	-0.092	-0.083	-0.078	-0.076	-0.074	-0.073	-0.073	0.313	0.041
3000 pg/μl	-0.118	-0.122	-0.125	-0.127	-0.128	-0.130	-0.132	0.527	0.206
Average	-0.102	-0.100	-0.099	-0.098	-0.098	-0.098	-0.099	0.385	0.109

The average estimated log-expression ratio of control genes as a function of spiked mRNA concentration. Lowess proves to be the most successful in estimating the true expression ratio of 0, and spiked-in based normalization ranked second. SBN, spiked-in based normalization; GMN, global median normalization. A 20-50% lowess normalization with different choices of local neighborhood were used with the smoother.

Table III. Root mean square error (RMSE) performance of normalization algorithms.

Concentration	20%	25%	30%	35%	40%	45%	50%	GMN	SBN
30 pg/μl	0.073	0.073	0.075	0.075	0.075	0.075	0.076	0.257	0.181
150 pg/μl	0.080	0.080	0.079	0.079	0.079	0.078	0.079	0.213	0.232
300 pg/μl	0.058	0.057	0.056	0.055	0.055	0.055	0.054	0.034	0.156
1500 pg/μl	0.060	0.060	0.058	0.057	0.057	0.056	0.055	0.115	0.161
3000 pg/μl	0.085	0.086	0.087	0.088	0.088	0.088	0.089	0.254	0.266
Average	0.071	0.071	0.071	0.071	0.071	0.071	0.071	0.175	0.199

RMSE is the square root of the average deviation and hence estimation bias of the true expression ratio, and spiked-in based normalization ranked second. SBN, spiked-in based normalization; GMN, global median normalization. A 20-50% lowess normalization with different choices of local neighborhood were used with the smoother.

majority of genes are not differentially expressed (i.e. dye swap experiments, massive arrays, etc.), but would obviously be of limited or no value in detecting differentially expressed genes. Note that maximization of the correlation coefficient is not a sensitive criterion of intra-channel bias removal (worse and best cases shown in Fig. 4a and c, respectively, differ by a magnitude of the control parameter p, whereas the correlation coefficient R is only 0.25 higher in the latter).

Implementation. All tested normalization algorithms were developed as notebooks and packages in the Computer Algebra system Mathematica (http://www.wri.com/). Standard vendor supplied packages were used in the construction of the lowess smoother, statistical analysis functions and graph generation. Testing was done in Mathematica versions 4.0 and 4.2 for Windows 2000 Professional and XP, respectively, running on single processor, Pentium IV machines. Since no version-, operating system- and processor-specific libraries were used, the scripts should be portable to any system running Mathematica.

Results

Assessing limits of quantitative agreement. The first step in evaluating normalization algorithms is to establish the relation between normalized ratios obtained by any two techniques in repeated measures i.e. replicates of the same data subset using MA plots. If results obtained by the two methods do not differ widely, then these two methods can be used interchangeably or substituted one for the other in subsequent analysis. For microarrays that generate a multivariate measurement, the construction of an MA plot can be performed in two steps. First, the same array is normalized with both methods and the results of difference vs. average expression ratio are graphed on a per array basis (i.e. to understand dependencies on the array level).

Construction of the composite MA plot is a superposition of graphs obtained in the first step. Fig. 5 represents the method agreement analysis for the competing techniques of Table I; spike-in based normalization (SBN) was considered the 'gold-standard' technique for pair-wise comparisons, and the analysis was carried out on the self-self hybridization subset. The graphs suggest a linear shift-of-scale relationship between normalized measurements obtained with global median normalization (GMN) and SBN; the average difference constant across the average log ratio range is -0.43 with a 95% agreement limit of -0.08 to -0.8 in log scale. This translates to ratios obtained with one technique being from 57.5% to 94.6% compared to the other. The simple relationship between the two methods confirms that both methods essentially 'estimate' the same component of the (unknown) microarray measurement error model, which is hardly surprising given the global nature of both SBN and GMN.

Table IV. Standard deviation of normalized log expression ratio of control genes.

Concentration	20%	25%	30%	35%	40%	45%	50%	GMN	SBN
30 pg/μl	0.594	0.595	0.595	0.596	0.597	0.598	0.599	0.865	0.628
150 pg/μl	0.585	0.586	0.587	0.587	0.588	0.589	0.590	0.765	0.701
300 pg/μl	0.537	0.535	0.534	0.532	0.530	0.529	0.528	0.497	0.570
1500 pg/μl	0.558	0.555	0.552	0.551	0.550	0.548	0.547	0.678	0.587
3000 pg/μl	0.596	0.599	0.602	0.602	0.604	0.605	0.607	0.852	0.719
Average	0.574	0.574	0.574	0.574	0.574	0.574	0.574	0.731	0.641

From all the methods employed, lowess is associated with the smallest standard deviation throughout the intensity range. Spiked-in and global median normalization did much worse compared to the individual lowess realizations. SBN, spiked based normalization; GMN, global median normalization. A 20-50% lowess normalization with different choices of local neighborhood were used with the smoother.

Table V. Coefficient of variation (CV) of normalized log expression ratio.

Concentration	20%	25%	30%	35%	40%	45%	50%	GMN	SBN
30 pg/μl	5.531	5.430	5.265	5.219	5.138	5.062	5.024	2.095	3.715
150 pg/μl	8.092	9.102	10.07	10.83	11.66	11.96	11.90	2.974	136.3
300 pg/μl	6.091	6.676	7.055	7.237	7.386	7.511	7.504	2.163	14.34
1500 pg/μl	4.466	4.668	4.759	4.773	4.748	4.683	4.731	1.559	6.009
3000 pg/μl	5.043	4.925	4.799	4.760	4.713	4.666	4.598	1.616	3.488
Average	5.845	6.160	6.390	6.563	6.729	6.776	6.750	2.082	32.784

The CV of the normalized log-expression ratio of control genes as a function of spiked mRNA concentration. Global median normalization has the lowest CV, but also the highest measures of bias and variance, suggesting a linear relationship between the accuracy and precision of the method throughout the intensity range. SBN, spiked based normalization; GMN, global median normalization. A 20-50% lowess normalization with different choices of local neighborhood were used with the smoother.

Although the methods target the same variance component, they can only be used interchangeably when the identification of differentially expressed genes is based on statistical methods and not on intensity thresholds (i.e. absence of replicates).

No linear relation is evident between the results normalized with lowess and SBN; the relation appears to be non-linear, involving both a shift and change in scale. A preliminary analysis suggested that this relationship could be modeled with a 4th degree polynomial, and thus the limits of agreement are established graphically. The fairly complicated nature of this relationship is anticipated considering the different nature of each normalization method. One measures a global component of the underlying measurement error model, whereas the other estimates a global and local (intensity-dependent) component. In general, these two methods cannot be used interchangeably, and in fact there is no simple rule-of-thumb to predict the functional relation between expression ratios estimated by one technique given the results of the other.

It is not evident if changing the size of the local neighborhood for the lowess smoother will produce results in quantitative agreement for most intents and purposes. Fig. 5c shows the method agreement evaluation for two different values of the local neighborhood (i.e. 20% and 50%). The 95% limit of agreement between expression ratios is fairly constant throughout the intensity range and relatively narrow (-0.15 to 0.25 in log scale). This range corresponds to a 16% change in expression ratio, which would have been observed by switching the normalization strategy. Most criteria for significant fold changes in gene expression would attribute a 16% variation to noise, and the two methods could therefore be used interchangeably for such a purpose.

Bias-variance performance of normalization strategies. The next step in the analysis of existing normalization strategies referred to bias and variance assessment (Table I). Results for the estimated expression ratio (i.e. normalized value), RMSE, variance and coefficient of variation are summarized in Tables II-V. To avoid an overly optimistic assessment of SBN (by definition, the normalized expression ratios of the spiked-in genes generated by this method are centered on 0), we resorted to a holdout re-sampling strategy. Briefly, the 324 control spots present in every array were randomly partitioned into a learning (n=216) and test (n=108) subset. The normalizing constant was calculated from the learning subset, but measures of accuracy and precision were estimated from the test subset. Partitioning to learning and test subsets and estimation of the normalization constant were repeated 1000 times for each array, and the results for all repetitions and arrays were used to construct Tables II-V. The relative size of the learning and test subsets and the

Table VI. Expression ratio entropies for raw and normalized data.

Normalization algorithm	Hybridization entropy	
	Self-self	Non-self-self
Raw data	1.63	2.27
GMN	1.18	1.53
SBN	1.25	1.43
Lowess (50%)	1.04	1.09
Lowess (45%)	1.04	1.09
Lowess (40%)	1.04	1.09
Lowess (35%)	1.04	1.09
Lowess (30%)	1.04	1.08
Lowess (25%)	1.03	1.08
Lowess (20%)	1.03	1.08
ChannelFlip (p=0. 05)	0.61	0.66
ChannelFlip (p=0.02)	0.26	0.27
ChannelFlip (p=0.01)	0.13	0.15

Entropies of expression ratio distributions for raw data are considerably higher compared to normalized data. Normalization always resulted in a greater reduction of the self-self hybridization expression ratio entropy, compared to the non-self-self subset. The greatest reduction is seen with the over-fitting algorithm ChannelFlip, as expected. Regarding lowess, there was no obvious effect of the control parameter (size of neighborhood) on the entropies of the two subsets. To calculate the entropies, the expression ratio scale was partitioned into bins with a 0.5 length in the interval -5 to 5; two additional bins for values <-5 and >5 were also utilized.

Table VII. Relative entropy (Kullback-Leibler divergence) for raw and normalized data and over-fitting.

Normalization algorithm	Relative entropy
Raw data	0.3185
GMN	0.0545
SBN	0.1384
Lowess (50%)	0.0043
Lowess (45%)	0.0041
Lowess (40%)	0.0041
Lowess (35%)	0.0041
Lowess (30%)	0.0042
Lowess (25%)	0.0042
Lowess (20%)	0.0044
ChannelFlip (p=0.05)	0.0175
ChannelFlip (p=0.02)	0.0074
ChannelFlip (p=0.01)	0.0036

The relative entropy of the expression ratio of self-self vs. non-self-self hybridizations was calculated for raw and normalized data. The distance is higher for raw data, and decreases with normalization. The greatest reduction is seen with the over-fitting algorithm ChannelFlip as expected, whereas a spiked-based normalization (SBN) strategy in the 'real-world' normalization strategies was associated with lesser reduction in the divergence. To calculate the entropies, the expression ratio scale was partitioned into bins with a 0.5 length in the interval -5 to 5; two additional bins for values <-5 and >5 were also utilized.

number of repetitions were based on calculations of the expected asymptotic bootstrap error for the first and second moments of the empirical distribution of generated samples. For the rest of the normalization methods, which do not use the spiked-in subset to estimate the normalizing functional no re-sampling strategy was employed (Table I).

Tabulated results demonstrate that the methods are successful to a variable degree in estimating the true expression ratio. Performance is substantially better for the lowess family of normalization methods, followed by spiked-in based normalization (SBN) (mean expression ratio of -0.100 vs. 0.109, true log-expression ratio of 0; Table II). On average, global median normalization (GMN) is associated with the largest bias (estimated mean expression rate of 0.385) in accordance with previous findings (8,30). RMSE metrics of bias confirm the superiority of lowess to the other two methods (Table III); compared to GMN and SBN, it manifests an RMSE that is 2.4 and 2.8 times lower, respectively. It is evident from the data that all currently available methods are characterized by substantial variability at the experimental level (i.e. taking into account all hybridizations). The overall experimental standard deviation (calculated from spiked-in genes present in all arrays) is 4-5 times the size of the mean expression ratio even for lowess-normalized arrays (Table V). Turning our attention to the various realizations of lowess, we find that the choice of control parameter (i.e. size of local neighborhood used in the local fitting) does not have a large impact on the performance of the algorithm in terms of bias and variance. The average bias varies from -0.102 to -0.099 as the size of the neighborhood is increased from 20% to 50% (Table II). The average RMSE, standard deviation and coefficient of variation are essentially the same for all realizations of lowess (Tables III-V). These findings essentially corroborate the method agreement analysis of Fig. 3, which pointed out that the actual size of the neighborhood will have little or no impact on subsequent steps of analysis for the specific dataset. It must be noted that the constant coefficient of variation of global median normalization (Table V) should be interpreted with caution in light of the low accuracy and precision of the method, which decrease proportionally throughout the mRNA concentration range.

Over-fitting quantification. To quantify the over-normalization potential of existing strategies, we examined theoretic measures (entropy and relative entropy) of the log-ratio distributions before and after normalization. Tables VI and VII summarize the effects of the various algorithms on the entropies and relative entropies of the data subsets measured in bits. To calculate the respective entropies, the expression ratio values (-5, 5) were partitioned into groups (bins), each with a length of 0.5 (log-expression scale). For the remaining values belonging to the range (-inf, -5) and (5, inf), two additional bins were provided. In each group, according to the range of values it possessed, the respective expression

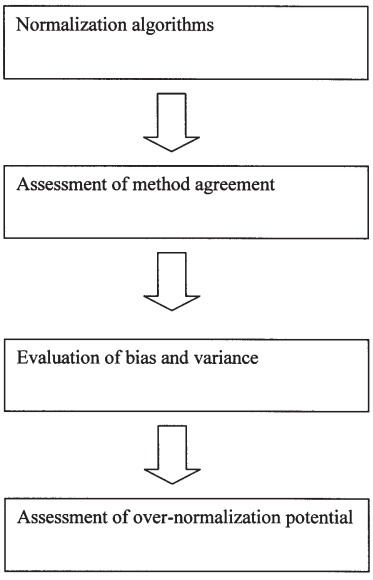

Figure 6. Summary of operational criteria for normalization algorithm comparison. The figure depicts a suggestive flow diagram for comparison among normalization strategies, starting from an initial classification of algorithms using quantitative method comparison criteria. Evaluation of normalization algorithms for a given dataset is best accomplished by the utilization of repeated bias and variance (or accuracy and precision) measurements for typical microarray datasets. Subsequently, theoretic measures can be applied to evaluate algorithms in terms of over-fitting. The algorithm that combines the best overall performance in terms of accuracy and precision and the least over-fitting is optimal for the task at hand.

ratios were assigned; according to the number of values falling within the range of the bin, its relative frequency was calculated to serve as a measure of probability of the group. These values were used for the calculation of both the entropies and relative entropies as described in the relevant paragraph. It is evident that unnormalized data distributions are characterized by the largest entropy measures, consistent with the highly variable log-expression ratios. The self-self hybridization subset has a lower entropy (1.63 vs. 2.27 or 0.64 bits less) compared to the non-self-self subset (Table VI); the latter consists of a number of hybridizations in different physiological states, and hence the excess entropy is a semi-quantitative measure of the magnitude of biological compared to technical variability. The non-zero value of 0.3185 bits for the relative entropy functional confirms the distance between the two distributions in probability distribution (Table VII). Normalization, which reduces variability, is associated with entropy reduction of both self-self and non-self-self hybridization subsets (Tables VI and VII; rows 2-12) irrespective of the specific algorithm employed, even though the value of the observed reduction varied considerably among algorithms.

For the dataset employed, the greatest entropy reduction is effected by the lowess group of algorithms, an effect that appears not to depend on the specific value of the control parameter (i.e. size of local neighborhood) used in the smoother or specific subset (self-self vs. non-self-self). Application of the lowess will result in the reduction of entropies of the expression ratio to roughly 1 bit (Table VI), whereas the GMN algorithm (the most commonly used method in the reporting of microarray research findings) reduces variability to a smaller extent in accordance with reported findings (13). SBN performance was intermediate between GMN and the various realizations of lowess and resulted in entropies of 1.25 and 1.43 for the two subsets. The over-normalizing ChannelFlip resulted in an impressive entropy loss, an effect monotonically decreasing with the value of the control parameter. The limiting case (p=0) would lead to degenerate expression distributions consisting of a single point, yielding expected entropy (and relative entropy) metrics of 0.

After normalization, the less variable self-self hybridization subset still exhibits lower entropy metrics compared to the non self-self subset, but the divergence between the two different subsets is reduced to a variable degree for the various algorithms. Of the competing strategies in Table I, lowess exhibits the greatest reduction in relative entropy (≈0.04 bits; Table VII, rows 4-10). This reduction is of the same magnitude as that imposed by the ChannelFlip algorithm, a case-study of an over-normalization method. Even GMN is not devoid of this 'variance smoothing' effect, although it is of less magnitude; of the methods employed, spiked-in based normalization best preserved the KL divergence, and hence has the least 'over-normalization' potential for the dataset employed.

Discussion

Microarray expression analysis offers an opportunity to generate functional data on a genome-wide scale and should consequently provide much needed data for the biological interpretation of genes and their functions. Applications of microarray technology to oncology have attempted to identify molecular signatures that affect patient outcomes for a variety of solid tumors, e.g. breast (33-36), colon (37,38), hepatocellular (39-41), prostate (42-46), ovarian (22,47,48) and gastric (49-51) cancer and hematologic malignancies, such as ALL and lymphomas (22,52-54). Potential applications of microarray expression profiling in oncology include the identification of signal transduction and transcription factor pathways involved in oncogenesis, optimization of treatment for individual patients, prognostication in individual cases and novel solution to diagnostic problems. Many investigators have used microarray technology to dissect transcriptional profiles that correlate with well-defined features of disease, such as cytogenetic profiles, histological subtypes or prognostically defined patient cohorts. An important, and one of the first, microarray applications in oncology has been the development of new therapeutic agents; in this context, the deployment of microarray-based programs can have a significant impact on all major steps inherent in the development of pharmaceuticals, including but not limited to new target identification, elucidation of the mechanism of action and the establishment of *in vitro* and animal models (55-57).

The power of microarray analysis lies in its capability to simultaneously distinguish and quantify thousands to tens-of-thousands array elements (genes). In the near future, analysis of the complete human transcriptome will likely be possible. The capability for meaningful analysis is predicated on the success of normalization procedures to transform raw expression data into inferences about individual genes or

Table VIII. Ranking of normalization methods tested on the specific dataset according to their performance with respect to the criteria proposed in this study.

Normalization	Average log expression ratio	Bias (RMSE)	Variance	Relative entropy	Total
GMN	3	3	3	2	3
SBN	2[b]	2	2	1	1
Lowess[a]	1	1	1	3	2

Summarizing view of the ranking of the normalization methods according to the results presented in Tables II-IV and VII. The ranking in the last column presents the ranking of algorithms according to the combination of optimal criterion associated with the least over-normalization potential, smaller bias (greater accuracy) and variance (greater precision). [a]As the choice of control parameter (i.e. size of local neighborhood used in the local fitting) does not have a large impact on the performance of the lowess algorithm in terms of bias and variance, we consider its various realizations as instances of the same method. Therefore, lowess is always ranked according to its best performance. [b]Performance of the SBN method was only marginally inferior to that of lowess, yielding estimates of nearly the same range of values.

groups thereof. Despite the importance of normalization, there are no consensus adjustment procedures, thus leaving the microarray experimentalist to ponder the practical repercussions of selecting one normalization method-algorithm over the other. He or she may wonder whether the results generated from a particular form of analysis are sensitive to the normalization step employed and, if so, the quantitative nature of this dependency. Hence, there is a need for a framework or procedure to aid in the comparative evaluation of normalization procedures, which was the imperative for the present study. The proposed framework is operationally definite (and hence verifiable) and could be used to provide a checklist for the researcher who has read the relevant literature and must choose an algorithm to use for his or her dataset. A stepwise approach (Fig. 6) is advocated. First, establish the limits of agreement among the methods employed (Fig. 5), and subsequently calculate measures of accuracy and precision based on two internal validation datasets using spiked-in controls and self-self hybridizations (Tables II-V). Algorithms found to have the smallest bias/variance are assessed in terms of over-normalization potential (i.e. excessive entropy reduction in self-self vs. non-self-self hybridization subsets) by comparing their performance to that of an over-normalization algorithm such as ChannelFlip (Tables VI and VII). The algorithm associated with the smallest over-normalization potential, smaller bias (greater accuracy) and variance (greater precision) is optimal for the dataset at hand. The combination of these criteria with this hierarchy provides a framework for the assessment of the overall performance of normalization algorithms. Table VIII presents a summarized ranking of the tested normalization methods, both for each of the proposed criteria and their overall performance.

Our tri-partite approach finds theoretical justification in three different research areas, namely the fields of quantitative method comparison, regression-calibration and information theory. Application of existing mathematical and graphical tools from these three areas requires the inclusion of internal validation datasets (i.e. repeated measures in statistical parlance), such as self-self hybridizations, spiked-in controls and reference sample designs, which are becoming increasingly common (58,59).

The use of the tri-partite framework is illustrated in a specially designed microarray dataset, normalized with three different methods that together account for the majority of published experimental microarray work. The first step, i.e. method agreement, unsurprisingly revealed that the spiked-in based normalization (SBN) is related to global-median normalization, namely that the addition of a constant in log-space defines a transformation from one method to the other. However, this is a qualitative effect since results obtained by the two techniques cannot be used interchangeably for subsequent analysis (i.e. assessment of significant fold change), whereas lowess normalization is non-linearly related to any method. An interesting finding was the insensitivity of lowess to the specific value of the control parameter. Although we cannot rule out a dataset-specific effect (the common reference sample used in this study precluded the observation of widely varying expression ratios), it is noteworthy that other researchers have made a similar observation (30,60). A large number of such spots consisting of <1% of all spots present on the array surface, and an explicit concentration-dependent relation among spiked-in controls, a dense sampling of technical variability factors including spatial effects, was made possible; hence, the estimation of the normalizing constant is not only feasible, but also gives normalized ratios with a smaller bias and variance than what would have been obtained otherwise. Information reduction metrics reveal that the performance of lowess comes at a price: the relative entropy of expression ratio distributions of the self-self and non-self-self experiments is of the same magnitude as that effected by a devised over-normalization method (ChannelFlip). In other words, lowess has the potential to reduce the biological variability component, and thus complicate forms of analysis that depend on variability measures (i.e. variance ratio comparisons).

Assessing the impact of different values of the control parameter of lowess is best done by method agreement (i.e. MA) plots and theoretic measures between the different realizations of lowess. Such plots are also of value when contemplating the use of other non-parametric normalization strategies (5,6,8,11,13,22,61,62). The multivariate nature of microarray measurements, and the complicated assumptions of statistical models in which these methods rely, render the

comparative evaluation of these methods, in principle, very difficult if not impossible.

The multi-step nature of microarray technology imparts a stochastic character to the quantitative behavior of the measurement process, which coupled to the inherent stochastisticity of the biological systems interrogated, call for a case-based approach to comparative evaluation of microarray normalization strategies using dataset-specific features. In many situations, the researcher will find that there is no single best normalization algorithm for all possible experiments; rather, there are classes of equivalent normalization strategies, each taking advantage of different characteristics of the dataset in which they are deployed. In fact, recent gene expression profiling research programs in malignant mesothelioma used a combination of normalization strategies to identify and experimentally validate differentially regulated control genes (35). If no 'one size fits all' normalization algorithm exists, then the experimentalist must select the 'best' algorithm for the dataset at hand by evaluating alternatives based on their strength/weakness profile. We feel that the selection process is greatly facilitated by the tri-partite process (Fig. 6) proposed in this study, since it relies on simple graphical/statistical measures based on a sound theoretical background.

Acknowledgements

The microarray consortium is funded by the Wellcome Trust, Cancer Research UK and the Ludwig Institute of Cancer Research. We would like to thank the staff of the Sanger Institute Microarray Facility (http://www.sanger.ac.uk/Projects/Microarrays/) for supplying arrays, lab protocols, and technical advice (David Vetrie, Cordelia Langford, Adam Whittaker, and Neil Sutton), Quantarray/GeneSpring datafiles and all data analysis and databases relating to elements on the arrays (Kate Rice, Rob Andrews, Adam Butler, and Harish Chudasama). The Mouse 15K cDNA clone set was obtained from the Laboratory of Genetics NIH/NIA-IRP, Baltimore, MD, USA. All cDNA clone resequencing was performed by Team 56 at the Sanger Institute.

References

1. Tseng GC, Oh MK, Rohlin L, Liao JC and Wong WH: Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. Nucleic Acids Res 29: 2549-2557, 2001.
2. Hoffmann R, Seidl T and Dugas M: Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. Genome Biol 3: RESEARCH0033, 2002.
3. Bilban M, Buehler LK, Head S, Desoye G and Quaranta V: Normalizing DNA microarray data. Curr Issues Mol Biol 4: 57-64, 2002.
4. Cheadle C, Vawter MP, Freed WJ and Becker KG: Analysis of microarray data using z score transformation. J Mol Diagn 5: 73-81, 2003.
5. Chen YJ, Kodell R, Sistare F, Thompson KL, Morris S and Chen JJ: Normalization methods for analysis of microarray gene-expression data. J Biopharm Stat 13: 57-74, 2003.
6. Colantuoni C, Henry G, Zeger S and Pevsner J: SNOMAD (standardization and normalization of microarray data): web-accessible gene expression data analysis. Bioinformatics 18: 1540-1541, 2002.
7. Durbin BP, Hardin JS, Hawkins DM and Rocke DM: A variance-stabilizing transformation for gene-expression microarray data. Bioinformatics 18: S105-S110, 2002.
8. Edwards D: Non-linear normalization and background correction in one-channel cDNA microarray studies. Bioinformatics 19: 825-833, 2003.
9. Kepler TB, Crosby L and Morgan KT: Normalization and analysis of DNA microarray data by self-consistency and local regression. Genome Biol 3: RESEARCH0037, 2002.
10. Quackenbush J: Computational analysis of microarray data. Nat Rev Genet 2: 418-427, 2001.
11. Wang Y, Lu J, Lee R, Gu Z and Clarke R: Iterative normalization of cDNA microarray data. IEEE Trans Inf Technol Biomed 6: 29-37, 2002.
12. Workman C, Jensen LJ, Jarmer H, *et al*: A new non-linear normalization method for reducing variability in DNA microarray experiments. Genome Biol 3: RESEARCH0048, 2002.
13. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J and Speed TP: Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 30: e15, 2002.
14. Zien A, Aigner T, Zimmer R and Lengauer T: Centralization: a new method for the normalization of gene expression data. Bioinformatics 17: S323-S331, 2001.
15. Bolstad BM, Irizarry RA, Astrand M and Speed TP: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19: 185-193, 2003.
16. Altman DG and Bland JM: Measurement in medicine: the analysis of method comparison studies. Statistician 32: 307-317, 1983.
17. Keffer J, Probert L, Cazlaris H, Georgopoulos S, Kaslaris E, Kioussis D and Kollias G: Transgenic mice expressing human tumour necrosis factor: a predictive genetic model of arthritis. EMBO J 10: 4025-4031, 1991.
18. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Stewart J, Schulze-Kremer S, Taylor R, Vilo J and Vingron M: Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 29: 365-371, 2001.
19. Kargul GJ, Dudekula DB, Qian Y, Lim MK, Jaradat SA, Tanaka TS, Carter MG and Ko MS: Verification and initial annotation of the NIA mouse 15K cDNA clone set. Nat Genet 28: 17-18, 2001.
20. Bland JM and Altman DG: Statistical methods for assessing agreement between two methods of clinical measurement. Lancet i: 307-310, 1986.
21. Magari RT: A statistical approach for hematology comparison studies. Lab Hematol 4: 199-203, 1998.
22. Adib TR, Henderson S, Perrett C, Hewitt D, Bourmpoulia D, Ledermann J and Boshoff C. Predicting biomarkers for ovarian cancer using gene-expression microarrays. Br J Cancer 90: 686-692, 2004.
23. Stern P, Bartos V, Vavrova J, Bezdickova D, Pechova M, Uhrova J, Friedecky B, Sprongl L, Zima T and Palicka V: Comparability of eight immunoassay procedures for the determination of CA 15-3 and related markers. Clin Chem Lab Med 41: 1087-1094, 2003.
24. Schwenk W, Neudecker J, Haase O, Raue W, Strohm T and Muller JM: Comparison of EORTC quality of life core questionnaire (EORTC-QLQ-C30) and gastrointestinal quality of life index (GIQLI) in patients undergoing elective colorectal cancer resection. Int J Colorectal Dis 19: 554-560, 2004.
25. Bland JM and Altman DG: Measuring agreement in method comparison studies. Stat Methods Med Res 8: 135-160, 1999.
26. Sundberg R: Multivariate calibration - direct and indirect regression methodology. Scand J Statist 26: 161-207, 1999.
27. Thurston SA, Spiegelman D and Ruppert D: Equivalence of regression calibration methods in main study/external validation study designs. J Stat Planning Inference 113: 527-539, 2003.
28. Tsodikov A, Szabo A and Jones D: Adjustments and measures of differential expression for microarray data. Bioinformatics 18: 251-260, 2002.
29. Kullback S: Information Theory And Statistics. Dover Publications, Mineola, NY, 1997.
30. Dudoit S, Yang HY, Callow MJ and Speed TP: Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments. Stat Sin 12: 111-139, 2002.
31. Cleveland WS: Robust locally weighted regression and smoothing scatterplots. JASA 74: 829-836, 1979.

32. Cleveland WS and Devlin SJ: Locally weighted regression: an approach to regression by local fitting. JASA 83: 596-610, 1988.
33. Callagy G, Pharoah P, Chin SF, Sangan T, Daigo Y, Jackson L and Caldas C: Identification and validation of prognostic markers in breast cancer with the complementary use of array-CGH and tissue microarrays. J Pathol 205: 388-396, 2005.
34. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO and Botstein D: Molecular portraits of human breast tumours. Nature 406: 747-752, 2000.
35. Singhal S, Wiewrodt R, Malden LD, Amin KM, Matzie K, Friedberg J, Kucharczuk JC, Litzky LA, Johnson SW, Kaiser LR and Albelda SM: Gene expression profiling of malignant mesothelioma. Clin Cancer Res 9: 3080-3097, 2003.
36. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, Demeter J, Perou CM, Lonning PE, Brown PO, Borresen-Dale AL and Botstein D: Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci USA 100: 8418-8423, 2003.
37. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D and Levine AJ: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 96: 6745-6750, 1999.
38. Mariadason JM, Arango D, Shi Q, Wilson AJ, Corner GA, Nicholas C, Aranes MJ,Lesser M, Schwartz EL and Augenlicht LH: Gene expression profiling-based prediction of response of colon carcinoma cells to 5-fluorouracil and camptothecin. Cancer Res 63: 8791-8812, 2003.
39. Shi YY, Wang HC, Yin YH, Sun WS, Li Y, Zhang CQ, Wang Y, Wang S and Chen WF: Identification and analysis of tumour-associated antigens in hepatocellular carcinoma. Br J Cancer 92: 929-934, 2005.
40. Yang LY, Wang W, Peng JX, Yang JQ and Huang GW: Differentially expressed genes between solitary large hepatocellular carcinoma and nodular hepatocellular carcinoma. World J Gastroenterol 10: 3569-3573, 2004.
41. Zhang LH and Ji JF: Molecular profiling of hepatocellular carcinomas by cDNA microarray. World J Gastroenterol 11: 463-468, 2005.
42. Foley R, Hollywood D and Lawler M: Molecular pathology of prostate cancer: the key to identifying new biomarkers of disease. Endocr Relat Cancer 11: 477-488, 2004.
43. Halvorsen OJ, Oyan AM, Bo TH, Olsen S, Rostad K, Haukaas SA, Bakke AM,Marzolf B, Dimitrov K, Stordrange L, Lin B, Jonassen I, Hood L, Akslen LA and Kalland KH: Gene expression profiles in prostate cancer: association with patient subgroups and tumour differentiation. Int J Oncol 26: 329-336, 2005.
44. Li Y, Hussain M, Sarkar SH, Eliason J, Li R, Quinn DI, Henshall SM and Sutherland RL: Molecular markers of prostate cancer outcome. Eur J Cancer 41: 858-887, 2005.
45. Li Y, Hussain M, Sarkar SH, Eliason J, Li R and Sarkar FH: Gene expression profiling revealed novel mechanism of action of Taxotere and Furtulon in prostate cancer cells. BMC Cancer 5: 7, 2005.
46. Trojan L, Schaaf A, Steidler A, Haak M, Thalmann G, Knoll T, Gretz N, Alken P and Michel MS: Identification of metastasis-associated genes in prostate cancer by genetic profiling of human prostate cancer cell lines. Anticancer Res 25: 183-191, 2005.
47. Bayani J, Brenton JD, Macgregor PF, Beheshti B, Albert M, Nallainathan D, Karaskova J, Rosen B, Murphy J, Laframboise S, Zanke B and Squire JA: Parallel analysis of sporadic primary ovarian carcinomas by spectral karyotyping, comparative genomic hybridization, and expression microarrays. Cancer Res 62: 3466-3476, 2002.
48. De Cecco L, Marchionni L, Gariboldi M, Reid JF, Lagonigro MS, Caramuta S, Ferrario C, Bussani E, Mezzanzanica D, Turatti F, Delia D, Daidone MG, Oggionni M, Bertuletti N, Ditto A, Raspagliesi F, Pilotti S, Pierotti MA, Canevari S and Schneider C: Gene expression profiling of advanced ovarian cancer: characterization of a molecular signature involving fibroblast growth factor 2. Oncogene 23: 8171-8183, 2004.
49. Haraguchi N, Inoue H, Mimori K, Tanaka F, Utsunomiya T, Yoshikawa K and Mori M: Analysis of gastric cancer with cDNA microarray. Cancer Chemother Pharmacol 54: S21-24, 2004.
50. Terashima M, Maesawa C, Oyama K, Ohtani S, Akiyama Y, Ogasawara S, Takagane A, Saito K, Masuda T, Kanzaki N, Matsuyama S, Hoshino Y, Kogure M, Gotoh M, Shirane M and Mori K: Gene expression profiles in human gastric cancer: expression of maspin correlates with lymph node metastasis. Br J Cancer 92: 1130-1136, 2005.
51. Weiss MM, Kuipers EJ, Postma C, Snijders AM, Pinkel D, Meuwissen SG, Albertson D and Meijer GA: Genomic alterations in primary gastric adenocarcinomas correlate with clinicopathological characteristics and survival. Cell Oncol 26: 307-317, 2004.
52. Burczynski ME, Oestreicher JL, Cahilly MJ, Mounts DP, Whitley MZ, Speicher LA and Trepicchio WL: Clinical pharmacogenomics and transcriptional profiling in early phase oncology clinical trials. Curr Mol Med 5: 83-102, 2005.
53. Virtaneva K, Wright FA, Tanner SM, Yuan B, Lemon WJ, Caligiuri MA, Bloomfield CD, de La Chapelle A and Krahe R: Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. Proc Natl Acad Sci USA 98: 1124-1129, 2001.
54. Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz R, Behm FG, Raimondi SC, Relling MV, Patel A, Cheng C, Campana D, Wilkins D, Zhou X, Li J, Liu H, Pui CH, Evans WE, Naeve C, Wong L and Downing JR: Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell 1: 133-143, 2002.
55. Clarke PA, Te Poele R and Workman P: Gene expression microarray technologies in the development of new therapeutic agents. Eur J Cancer 40: 2560-2591, 2004.
56. Clarke PA, George ML, Easdale S, Cunningham D, Swift RI, Hill ME, Tait DM and Workman P: Molecular pharmacology of cancer therapy in human colorectal cancer by gene expression profiling. Cancer Res 63: 6855-6863, 2004.
57. Sausville EA and Holbeck SL: Transcription profiling of gene expression in drug discovery and development: the NCI experience. Eur J Cancer 40: 2544-2549, 2004.
58. Dobbin K, Shih JH and Simon R: Statistical design of reverse dye microarrays. Bioinformatics 19: 803-810, 2003.
59. Dobbin K and Simon R: Comparison of microarray designs for class comparison and class discovery. Bioinformatics 18: 1438-1445, 2002.
60. Callow MJ, Dudoit S, Gong EL, Speed TP and Rubin EM: Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. Genome Res 10: 2022-2029, 2000.
61. Bilban M, Buehler LK, Head S, Desoye G and Quaranta V: Normalizing DNA microarray data. Curr Issues Mol Biol 4: 57-64, 2002.
62. Colantuoni C, Henry G, Zeger S and Pevsner J: Local mean normalization of microarray element signal intensities across an array surface: quality control and correction of spatially systematic artifacts. Biotechniques 32: 1316-1320, 2002.