Which to use? - microarray data analysis in input and output data processing

Qingwei Zhang[1*†], Rie Ushijima[2*], Takatoshi Kawai[2], and Hiroshi Tanaka[1]

[1] Department of Bioinformatics, Medical Research Institute, Tokyo Medical and Dental University,
1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8510, Japan
[2] Laboratory of Seeds Finding Technology, Eisai Co., Ltd.
5-1-3 Tokodai, Tsukuba-shi, Ibaraki 300-2635, Japan
[†]Present address: Pharmaceutical Research Laboratory, AJINOMOTO Co., Inc.
1-1 Suzuki-cho, Kawasaki-ku,Kawasaki, 210-8681 JAPAN

** The first two authors contributed equally 1. E-mail: zhancom@tmd.ac.jp, 2. r-ushijima@hhc.eisai.co.jp*

(Received May 31, 2004; accepted June 18, 2004; published online July 13, 2004)

Abstract

Along with the developments and advances in microarray technology, data analysis is becoming an increasingly critical step of the microarray system for unraveling complicated biological mechanisms. As there are various platforms for microarray technology and microarrays are used for different purposes, many methods have been consequently devised for data analyzing. It is not easy, however, to choose the most appropriate method for each situation. This review focuses on the currently available methods for “input” and “output” data processing, including normalization in raw data processing and the use of ontology and meta-analysis in data aggregation. By presenting detailed explanations of both the major established methods and several state-of-the-art approaches, this review aims to provide a brief overview of the trends in microarray data analysis.

Key Words: DNA microarray, GeneChip, normalization, summarization, gene ontology, meta-analysis

Area of Interest: Genome Wide Experimental Data Analysis

1. Introduction

The use of microarray technology has become commonplace today along with the rapidly expanding fields of genomics and proteomics. The ability to monitor genome-wide changes in gene expression levels and to detect sequence changes of tens of thousands of genes simultaneously make it a valuable research and diagnostic tool with a wide range of uses today. Experiments using of microarrays can easily provide overwhelming amounts information, although successful insights into the fundamental mechanisms behind a phenomenon being addressed depend on the quality of the subsequent data analysis and interpretation. Nevertheless, the methodologies for microarray data analysis are still under development and many methods hitherto provided are in need of improvement as well as validation. Generally speaking, microarray data analysis covers a variety of processes beginning with the processing of raw images and ending with data annotation. Here we give only a partial review on the latest developments in microarray data analysis of gene expression data. As the means for extracting desired genes is probably the most rapidly developing issue, which is subject to different demands, we leave it to other papers. We focus here, instead, on the fundamental issues of input and output data processing in microarray data analysis: (1) the normalization methods for cDNA and oligo type microarrays, (2) the use of ontology tools to systematically annotate co-regulated genes, and (3) the meta-analysis to take advantage of analysis results derived from using different array technologies and backgrounds.

2. Normalization of cDNA microarray data

Normalization is a process of removing systematic variations in microarray-derived data. There are many sources of such variations, which may be caused by differences in plates, chips, and dyes; sequence-specific preferences; differences in sample preparation; scanner malfunction; and so on. Considering that various protocols are used today and different systematic features could arise in different types of experiments, it is almost impossible to identify all of the sources of systematic variation with only our current limited knowledge of the possible sources of systematic variation. In addition, not only do the raw data contain both the random and systematic variation from different resources, but also it is possible that the variation in gene expression level might just reflect a natural divergence in living systems. Thus, it might not be practical to exclude all the sources of systematic bias from the raw data. We can see, however, that recently many methods have been developed and others further improved, providing us with a deeper understanding of the data contents and with better procedures for acquiring the actual intensity data.

2.1 Linear models in normalization

Error model might be one of the most commonly used techniques for assuming the intensity of spots in the microarray. The simplest error model is employed by the commonly used global normalization using the mean or median of the spot intensities [1]. If the identical sample is labeled with different fluorescent dyes in a typical two-channel experiment, the model is given by:

$$I_A = aI_B + \varepsilon \qquad (1).$$

where a is a normalization constant, ε is an independent random error, and I_A and I_B stand for the observed expression levels in two channels, respectively, for a given gene. Similarly, in a comparative experiment with the assumption of an equal abundance of intensity over two channels for most genes and assuming no other variables (e.g. dye effect, slide difference, spatial location), the same model of equation 1 is commonly used (Figure 1b). The estimation of the multiplicative constant a is similar to a simple regression analysis; for instance, to estimate the regression coefficient a by using the error model:

$$I_A = aI_B + b + \varepsilon, \qquad (2)$$

except that the intercept b is zero. By the least-squares method, the estimated a is derived as $\overline{I_A} / \overline{I_B}$, e.g., the ratio of the means of the intensities in both channels. Instead of a mean value, the median of the spot intensities is used in order to reduce the influence from outliers. Consequently, one channel intensity is scaled so that: $I'_A = I_{A,}$ $I'_B = aI_B$. Thus the adjusted intensity ratio T_i' becomes: $T_i' = (1/a)I_A / I_B = (1/a)T_i$.

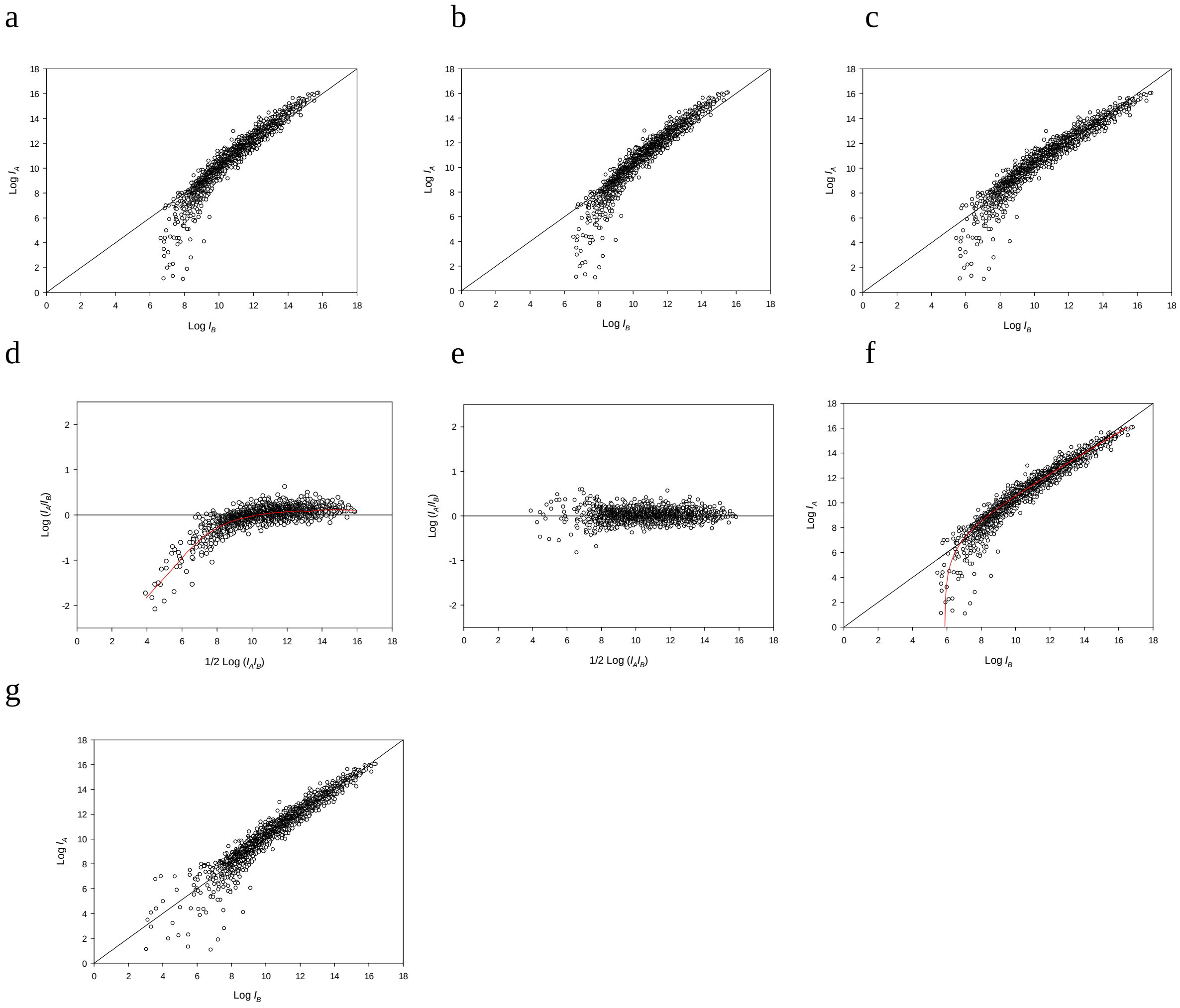

Figure 1. Normalized results obtained by different methods for a raw cDNA microarray data set.

a) The original raw data plotted on the log scale. b) Normalized by average intensity. c) Normalized by a simple linear model of log intensity. d) MA-plot, where the red line indicated the regression by *lowess* function. e) Normalized by using *lowess* function. f) Regression line of the bias model. g) Normalized data by the bias model.

In dealing with the up- and down-regulated genes, the log intensity ratio is more preferable than the intensity ratio for both data displaying and calculation. For example, the base 2 log ratio of 2 times up- and 2 times down-regulated are -1 and 1, respectively, which are symmetrical around zero. In contrast, the corresponding simple ratios for the above cases are non-symmetrical on the axis as 0.5 and 2, respectively. In addition, when displayed on the x-y plot, spots representing down-regulated genes all shrink between zero and 1 on the axis, which is a poor condition for data viewing.

For a log-transformed ($\log I_A$, $\log I_B$) data set, if most genes are assumed to be stably expressed, the orthogonal linear regression is one of the commonly used methods with the linear error model (Figure 1c):

$$\log_2 I_A = a\log_2 I_B + b + \varepsilon \qquad (3).$$

Basically, this model takes the same linear model as in equation 2. It is not difficult to find that; however, the relation for intensity I_A and I_B is non-linear as $I'_B = 2^b I^a_B$ and $I'_A = I_A$ after normalization. Consequently, when most genes are stably expressed, the linear regression of the log-transformed data usually outperforms that of the normal scale. The former one performs both a rotation and shift of the regression line ($\log_2 T_i' = \log_2(I_{Ai}/I_{Bi}{}^a)-b$), whereas the later one only shifts the regression line ($\log_2 T_i' = \log_2 T_i - \log_2 a$).

It is ideal to remove all differentially expressed genes for global normalization; this is why only house-keeping genes or spiked DNAs are used alternatively for regression. Nonetheless, the house-keeping genes might not be stably expressed all the time, and spiked DNAs might not reflect the variance carried by the sample of interest. Wang et al. [2] provided an iterative regression normalization, which is based on the same model as that of orthogonal linear regression. Iterative regression, however, is employed to infer the regression coefficient *a* and intercept *b*.

It is obvious that the above models are too simplistic to account for various sources of systematic error. To overcome the problems of such simplified models, Kerr et al. [3] provided a more general error model represented by:

$$\log_2\left(Y_{ijkg}\right) = \mu + A_i + D_j + V_k + G_g + (AG)_{ig} + (VG)_{kg} + \varepsilon_{ijkg} \qquad (4).$$

This model might seem to be very confusing. However, equation 4 can be considered as an expansion of equation 3 by modifying the error more precisely by the addition of possible factors. In short, μ is the overall average signal; A_i, D_j, V_k, and G_g represent the effect of array, dye, experimental condition, and gene, respectively; and $(AG)_{ig}$ and $(VG)_{kg}$ represent the interaction effect. By employment of experiment design and performance of analysis of variance (ANOVA), each effect can be estimated and the normalized log ratio is given by subtraction of all estimated effects from the log ratio. To infer the confidence intervals for the estimates of effects, a bootstrapping technique was employed. Such ANOVA models are not restricted to cDNA type microarrays, Yang et al. [4] provided a similar model for oligonucleotide microarrays, and they applied the maximum likelihood method for estimation.

With employment of the ANOVA models, one can test the significance of the effects in the error model with detailed quantitative data and identify the sources of variation. Considering that other effects such as plate, spatial, and so on may also exist, a model such as Kerr's is not sufficient enough to account for all sources of variation. One might tend to expand the variables to account for different variation. Thus the ANOVA model seems to be a very convenient parametric model with *a priori* accounting of different sources of variance, but without the assumption of equal expression levels for most genes as used by global normalization. Nonetheless, it shall be noted that the inclusion of additional effects means the loss of degrees of freedom. As a result, estimation of the error variance will be unfeasible with inadequate data and too many variables in the model. In practical application, an ANOVA model with reasonably few variables, for instance:

$$Y_{gk} = \mu + G_g + V_k + (GV)_{gk} + \varepsilon \qquad (5)$$

(where only experimental effect and gene effect are mainly accounted for as influencing the log ratio) provided by Lee at al. [5], might be convenient for calculation and be sufficient to remove the main part of the systematic variation.

2.2 Non-linear models in normalization

In addition to the above linear regression models, many nonlinear models for normalization have also been provided. For experiments using two different fluorescent dyes, the locally weighted regression, also known as *lowess* (or *loess*) smoothing [1], is probably one of the most commonly used methods for normalizing two-color data. The *lowess* smoothing is performed based on a (A, M) plot, which is the mean log intensity (denoted as A) vs. the log intensity ratio (denoted as M) plot introduced by Dudoit et al. [6]. For spots on the (A, M) plot, the M and A are given by:

$$M = \log_2 (I_A / I_B) \text{ and } A = (1/2)(\log_2 I_A + \log_2 I_B), \tag{6}$$

which is equal to a clockwise rotation of the ($\log_2 I_A$, $\log_2 I_B$) plot by 45 degrees followed by rescaling (Figure 1d). The performance of *lowess* smoothing resembles a moving average procedure. On the (A, M) plot, only neighboring points contained in the window are locally regressed to fit a smooth function of A to M. This corresponds to drawing a bias line of M varying smoothly subject to A. Thus the log ratios, or M values are normalized by subtraction of the bias (Figure 1e) as:

$$M' = M - f(A). \tag{7}$$

There is a good reason for transforming the original ($\log_2 I_A$, $\log_2 I_B$) plot by a 45 degree rotation to the (A, M) plot followed by scaling. Under ideal conditions, the A value is the log of the geometric average of intensity I_A and I_B reflecting the intensity abundance; and the M value corresponds to the log ratio. Therefore *lowess* smoothing estimates an intensity dependent bias in the log ratio. However, if *lowess* is performed for the ($\log_2 I_A$, $\log_2 I_B$) plot, the procedure is to fit a smooth function of $\log_2 I_A$ to $\log_2 I_B$, which carries a completely different meaning.

Although *lowess* smoothing seems to be very convenient for normalizing data of various forms on the scatter plot, it shall be noted that the (A, M) plot assumes the same intensity for identical samples, as $\log_2 I_A = \log_2 I_B$ on the ($\log_2 I_A$, $\log_2 I_B$) plot. If, for example, a linear relation exists for $\log_2 I_A$ and $\log_2 I_B$, such as $\log_2 I_A = a \log_2 I_B + b + \varepsilon$, the log ratio M should be calculated as:

$$M = \log_2 (I_A / I'_B) = \sqrt{2} \cos\theta (\log_2 I_A + b) - \sqrt{2} \sin\theta (\log_2 I_B). \tag{8}$$

(where θ denotes the anti-clockwise angle between the I_B axis and the orthogonal regression line) after normalization [7] (Figure 2a). Without considering whether the pre-processing method is right or not remains a problem; it is clear that the (A, M) plot equals a 45 degree rotation followed by scaling only under the condition where θ equals 45 degrees and the intercept is zero. The difference between log ratios derived by using the MA-plot and the linear model (Equation 3) is illustrated in Figure 2b.

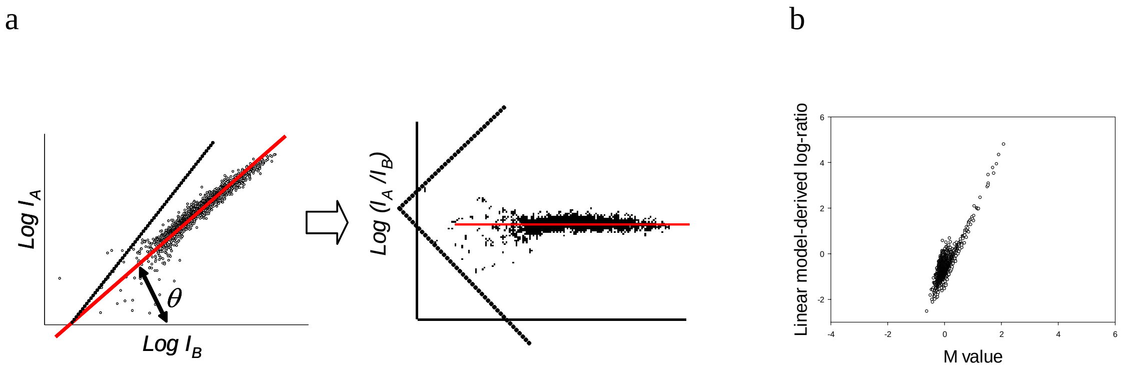

Figure 2. The MA-plot vs. linear model-derived plot of log intensity.
a) The rotation of a linear model-derived plot of log intensity. b) The correlation of log intensity ratio of MA-plot with that derived by rotation of linear model normalized log intensity (the same data of Figure1)

Yang et al. [1] emphasized that spatial biases or print-tip biases can easily be identified and corrected by application of *lowess* normalization when compared with several other global normalization methods. On the other hand, however, *lowess* smoothing is also based on the assumption that most of the genes are equally expressed in both compared samples. Considering that this assumption may not hold true for all data sets, it may be inaccurate to perform *lowess* smoothing by inclusion of all the points. To find a control sample that spans the intensity range with a constant expression level, Yang et al. advocated performing *lowess* smoothing by using a titration series including all the genes present on the microarray. They designated the titration series as the Microarray Sample Pool (MSP). As an alternative to using the MSP, Tseng et al. [8] ranked expression levels and performed iterations to select un-differentially expressed genes for *lowess* smoothing.

As an improved version of the *lowess* smoothing, Yang et al. [9] used dye-swap normalization, which does not need *a priori* information, by assumption of self-consistency of the data sets. For every spotted gene i, assuming a multiple constant for the relation between the intensities of two samples, the dye-swap experiments give the log ratio set as:

$$M_i = \log_2(k_i)(I_A / I_B) = \log_2(I_A / I_B) + c_i \tag{9}$$

and the log ratio for reversely labeled data as:

$$M_i' = \log_2(k_i')(I_B / I_A) = -\log_2(I_A / I_B) + c_i' \tag{10}$$

where c_i and c_i' account for the different properties of the dyes. By approximation and transformation of the equations, we have $c_i = (1/2)(M_i + M_i')$; thus c_i can be estimated from the data plotted in the scatter plot $(1/2)(A + A')$ vs. $(1/2)(M + M')$. By comparing the results between the *lowess* smoothing and dye-swap normalization applied to experimental data sets, Sanchez-Cabo et al. [10] indicated that the dye-swap normalization is more accurate, and is especially efficient in normalizing the genes with low expression levels. In addition, the dye-swap normalization was indicated to show a smaller coefficient of variation than that of *lowess* normalization.

Using the polynomial spline as a similar smoothing procedure, Huang et al. [11] suggested a semi-linear model for log ratios as:

$$Y_{ij} = \phi_{ij}(x_{ij}) + z_i' \beta_i + \varepsilon_{ij} \tag{11}$$

where β_i is the gene effect, z_i' is the covariate, and $\phi_{ij}(x_{ij})$ is a nonparametric component that needs to be estimated from the data.

In a two-channel microarray experiment, scatter plot tilting on low intensity is not unusual; Zhang et al. [12] advocated a simple semi-linear model as:

$$\log_2 I_{Ai} = k \log_2 (I_{Bi} - a) + b + \varepsilon, \tag{12}$$

where a is a constant that is estimated from the data set. The simple subtraction of the constant a from one channel seems to be effective enough instead of smoothing techniques (Figure 1f, g).

Among the nonlinear methods, there are ANOVA-like models that consider mixed effects.

Kepler et al. [13] suggested a model for the log intensity Y_{ijk} as:

$$Y_{ijk} = V_{ij}(\alpha_k) + \alpha_k + \delta_{ik} + \sigma(\alpha_k)\varepsilon_{ijk}. \tag{13}$$

where $V_{ij}(\alpha_k)$ and $\sigma(\alpha_k)$are the intensity dependent normalization constant and variance scaler, respectively; α_k is the mean log intensity and δ_{ik} is the differential effect. Note here that Y_{ijk} is log intensity in contrast to the log intensity ratio denoted in Kerr's model. Similarly, Fan et al. [14] provided a model by application of the (A, M) plot as:

$$M_{gi}. = \beta_r + \gamma_c + f(A_{gi}) + \mu_g + \varepsilon_{gi}. \tag{14}$$

where β and γ correspond to print tip block effects of different rows and columns, respectively; and f is the intensity effect as a function of intensity A.

From the above examples, we have seen some variations in the application of local regression in non-linear normalization methods. There are many other methods not mentioned here. By means of this or that, smoothing techniques or iteration procedures seem to be commonly used in data fitting; however, how to make a better, more accurate model compared with the current models still remains an open question. The baseline is that a so-called gold standard might not exist, and validation studies of different models are inadequate. By efforts to define a standard for the published microarray data by the Microarray Gene Expression Data (MGED) organization, we may expect to see more accurate methods produced by use of sufficient qualitative data sets.

3. Normalization and summarization of Affymetrix GeneChip

The Affymetrix GeneChip measures the gene expression levels with an 11-20 set of oligonucleotide probe pairs comprised of perfect match (PM) and mismatch (MM) probes. PM probes are designed to be perfectly complementary to a subsequence of a particular mRNA, and MM probes, whose central base is changed to the counter one of its PM sequence, are designed to discriminate non-specific hybridization. For this feature of chip design, Affymetrix data analysis consists of a probe set aggregation or summarization to obtain the measure of gene expression in addition to the background correction and data normalization procedures. Normalization can be applied not only onto gene expression measures but also onto probe levels. Currently available useful methods are summarized in Table 1. Affymetrix supplies the Microarray Suite 5.0 (MAS5) algorithm [15], and the other methods are implemented in the BioConductor affy package and its related R packages [16].

Table 1. Currently available useful methods for Affymetrix GeneChip data analysis

method	background correlation	normalization	probe set summarization
MAS5	zone-weighted	linear scaling	Tukey bi-weight on log2(PM-IM)
dCHIP		invariant set	MBEI on linear PM
RMA	distribution modeling	quantiles	median polish on log2(PM)
others		VSN	

3.1 Background correction of GeneChip data

As there is not enough space around the probes of an Affymetrix GeneChip to calculate the background intensity, MAS5 adopts the 2nd percentile of the probe values as the background intensity [15]. To reflect the spatial background drifting, the entire array area is divided into 4 by 4 zones, and the background value of a certain point is calculated from 16 zone backgrounds with a weight function proportional to the distance between that position and each zone centroid. To avoid getting a negative value after subtraction of the position specific background, a small threshold value is preset. This background correction changes the PM and MM log-intensity distribution into a normal-like distribution (Figure 3b).

Irizarry et al. [17] solved the negative value problem by a signal and noise convolution model. They assumed that PM intensity distribution can be modeled by an exponentially distributed signal component S and a normally distributed noise component N. The convolution product of S and N will be an ex-Gaussian function with three parameters, mean μ and standard deviation σ of normally distributed N and decay parameter α of exponentially distributed S. They estimated these parameters by using a density kernel estimation applied to the observed PM intensity values. After the adjustment of noise components, pure positive signal components will be estimated (Figure 3c). This background correction is implemented in the robust multi-array analysis (RMA).

3.2 Normalization of GeneChip data

MAS5 normalization among multi-array experimental datasets is a simple linear scaling not on the probe-level intensity data but on the summarized gene-level intensity data [15]. This simple normalization is not effective on the dataset whose probe-level intensity distribution contains large chip by chip differences.

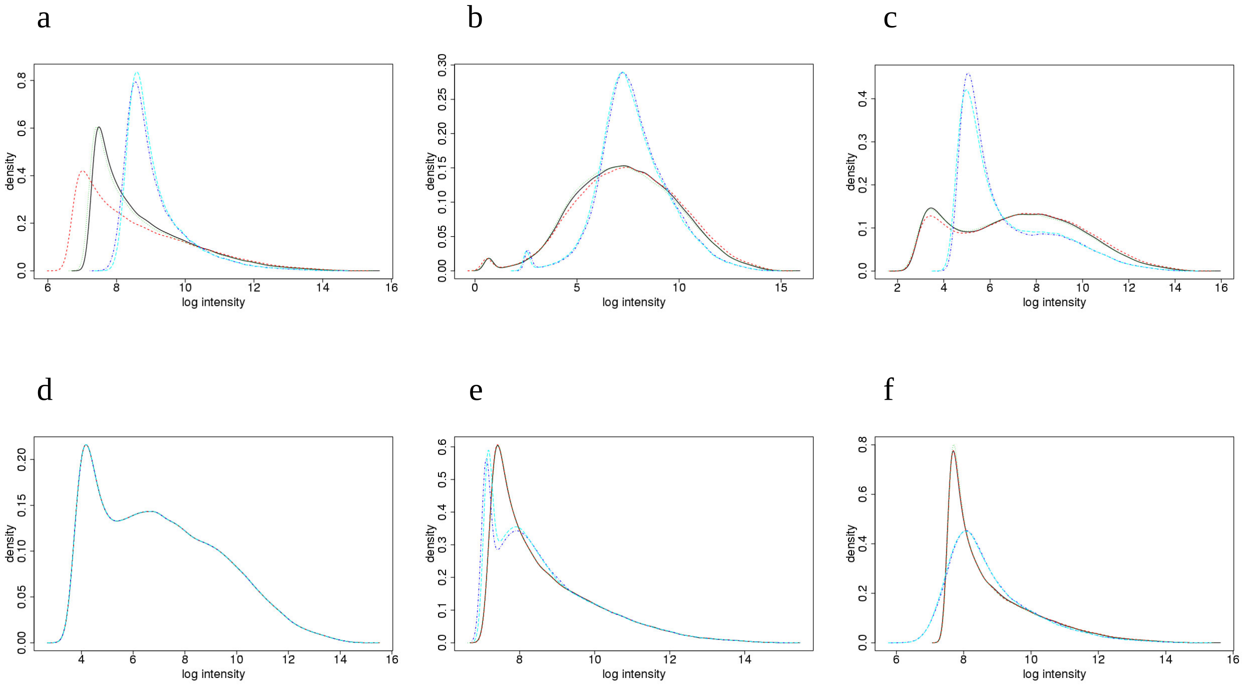

Figure 3. Background correction and normalization results by different methods for five differentially distributed raw GeneChip PM intensity data sets.
a) The original raw PM data plotted on the log scale. b) MAS5 background corrected data. c) RMA background corrected data. d) RMA background correction followed by quantiles normalization. e) dChip normalization (invariant set). f) VSN normalization.

Quantile normalization adopted in RMA imposes the distribution of probe intensities for each array to the averaged distribution by taking the mean values across the same quantile intensities for their normalized values [18]. This normalization makes it possible to compare a set of differentially distributed probe intensity arrays caused by the differences of cell conditions, scanner conditions, chip lots, etc. Because this method forces the value of quantiles to be equal, it is possible that a certain probe could have the same value across all the arrays (Figure 3d). However, in practice, it would not be a problem because the expression measure is calculated from a set of probes.

dChip software [19][20] incorporates "invariant set" normalization [21]. This method does not force all of the quantiles to the reference values at once, rather it forces the selected non-differentially expressed genes (or probes) to have equal values followed by smoothing spline normalization among these points (Figure 3e). Schadt et al. [21] showed that this normalization method kept the expression ratio values between two datasets under consideration unchanged for the biological meaningful ones.

Huber et al. [22][23] introduced a variance stabilization technique, called VSN, to the calibration of probe intensity data. The analysis of replicate microarray data typically shows that the variance of the measured probe intensities increases with their mean value. Huber et al. showed that the affine transformation of the measured probe intensities followed by the arc-hyperbolic-sine transformation stabilizes the variance along the whole intensity range.

$$h_{k,i} = \operatorname{asinh}(a_i + b_i \cdot y_{k,i}), \text{ where } \operatorname{asinh}(x) = \ln(x + \sqrt{x^2 + 1}) \tag{15}$$

$h_{k,i}$ is the variance stabilization transformation of the measured probe intensity $y_{k,i}$, which is the intensity of the kth probe on the ith array. The array-dependant variables a_i and b_i are estimated from the measured intensities by using the maximum likelihood estimator. As the arc-hyperbolic-sine function has the formula of $\ln(x + \mathrm{sqrt}(x^2 + 1))$, this transformation can be applied even onto background subtracted negative intensity values (Figure 3f).

3.3 Summarization of GeneChip

After making the background correction followed by the intensity normalization, we need to describe the final step, intensity summarization. As the Affymetrix GeneChip oligonucleotide arrays are designed such that each gene is represented by a set of several PM and MM probe pairs, it is necessary to summarize a set of probe intensities to an aggregated expression measure. There are several methods to average the probe intensities; each of them is derived to be insensitive to outliers.

A key algorithm used in MAS5 is the one-step Tukey's biweight algorithm [15]. To calculate an average value from a set of measured intensities, this biweight algorithm determines a robust average unaffected by outliers. The one-step biweight algorithm begins by calculating the median value for a dataset as the starting average value. In case of signal value computation, this dataset consists of the $\log(PM - CT)$, where CT is a Change Threshold described later, within a probe set of a given gene. The one-step biweight algorithm shifts the median value to the mean of weighted MAD (mean absolute deviation) value calculated from Tukey's biweight function and the absolute distances between each $\log(PM - CT)$ and the median value. This new average value is the output of one-step Tukey's biweight algorithm. In this process, MAS5 uses CT instead of MM. When $PM > MM$, CT is identical to MM. When $PM \leq MM$, CT is replaced with $PM \times \mathrm{Tb}(MM / PM)$, where Tb is the function of Tukey's biweight.

It is a common feature that the hybridization efficiency profile of a set of probes is conserved throughout multiple arrays. Once a common pattern of hybridization efficiencies is derived, it is quite easy to identify outlier probes. Therefore, the use of multiple arrays together could lead to more robust results. There are two noteworthy methods that can analyze multi array data. One is Li and Wong's product model in the linear scale, and the other is Irizarry et al.'s additive model in log scale.

Li and Wong's method [19] assumed that each probe measure $y_{ij} = PM_{ij} - MM_{ij}$ can be modeled as the product of θ_i (model-based expression index (MBEI)) in array i and probe-sensitivity index ϕ_i for given probe j, accompanied with random error ε_{ij}:

$$y_{ij} = PM_i - MM_{ij} = \theta_i \cdot \phi_j + \varepsilon_{ij} \tag{16}$$

After the model fitting procedure, the MBEI θ will be used for gene expression values. Though the initial model was designed with $PM - MM$, in most cases the PM only model is in good agreement with the original one [20]

Irizarry et al.'s method [24] is referred to as the log scale robust multi-array analysis (RMA). They modeled the probe intensity data as follows.

$$\log_2(PM_{ij}) = e_i + a_j + \varepsilon_{ij} \tag{17}$$

where e_i is an expression measure of array i, a_j is the probe j specific affinity effect, and ε_{ij} is the error term. To fit this model to the measured intensities, RMA uses the median polish algorithm [25]. Before applying this algorithm, make a matrix of PM_{ij} with j probes along the row and i arrays along the column for a PM set of each gene. When the median polish procedure, where the median value of each column and row is swept out from the matrix one after another, is applied to this matrix, it comes to be unchanged. At this point, the row median residuals are probe specific affinity effect offsets a_j, and residuals left in the initial matrix are error term ε_{ij}, and the sums of the grand median of this matrix and each column median residuals retain the individual expression measures e_i.

We reviewed some widely used Affymetrix GeneChip data analysis methods from the viewpoint of the algorithms involved. There is no solution that can be said to be the best recipe. Some method-comparison papers [24][26][27][28] have indicated that model-based methods like RMA and dChip provide better solutions in comparison with MAS5; however, we must make our own prescription with a suitable combination of background correction, normalization, and summarization. Affycomp [29], which is a benchmark platform for measuring Affymetrix GeneChip expression, will be a big help as an example for this purpose.

4. Classification and categorization of co-regulated genes by Gene Ontology Analysis

How can we get useful information from the microarray data that can possibly be used for comparison among the samples? One practical solution for this question is gene ontology analysis. Gene Ontology (GO) [30] is widely accepted as the standard for vocabulary; and it consists of three categories, biological process, molecular function, and cellular component. A gene expression profile experiment usually gives us a list of a large number of co-regulated genes, which sometimes gives us a feeling of being lost in a deep forest. GO categorized gene expression profiles provide such a tool that is able to highlight some pathways hidden by fallen leaves.

McCarrol et al. [31] prevailed in the common transcriptional profile for aging of the nematode and fruit fly by using this approach. They obtained adult-onset expression profiles of these highly diverged animals by Affymetrix oligonucleotide microarrays. It is easy to imagine that most transcriptional changes were specific to worms or to flies; however, after allocating them into nearly individual GO categories, they found that an unexpected shared feature of aging in nematode and fly was the repression of orthologous genes involved in diverse ATP-utilizing molecular transport functions, including primary active transporters, ion transporters, and ABC transporters.

GO analysis tools are collected at the Gene Ontology Consortium [32]. One easy to use tool, called the NetAffx GO Mining Tool [33], provides graphical and interactive views of GO term relationships, and frequency in a queried co-regulated gene list. This tool creates an interactive hierarchical graph image dividing the queried genes into each GO node with the number of counts or percentage against the total number of genes belonging to that node or the number of queried genes. It can reconstruct sub-graphs dynamically and colored as a heat map according to the numbers of counts or percentage. This tool is good at pointing out the "hot spots" from the GO hierarchical graph; however, it is difficult to extract a relationship between gene sets and certain depths of the GO terms, because the sub-graph has several depths of hierarchies and each node is comprised of differing ranges of gene numbers.

To extract GO categories of interest, in practice, it is very useful to have a high-level view of the three ontologies. GO slims are designed for those purposes [32]. GO slims contain small numbers of high-level GO terms, and GO slim assignment emphasizes the specific expression profile patterns of the individual samples at a glance. Bono et al. [34] examined tissue-specific gene expression profiles using RIKEN mouse cDNA microarrays, and analyzed with GO slim terms of molecular function ontologies. From this analysis, they found that the expression profile of placenta has a high proportion of "signal transduction functions" including placental lactogen2, placental growth factor, and prolactin-like proteins.

FatiGO [35] is another GO tool that extracts relevant GO categories for a given gene set with respect to the rest of the genes using Fisher's exact test, which considers the multiple-testing nature of the statistical contrast performed. When a lot of hypotheses are tested simultaneously, the rate of false positives increases and the individual p-values no longer correspond to significant findings. To assess the statistical significance, p-value adjustments are needed. FatiGO returns adjusted p-values based on three different ways of accounting for multiple testing. Those are a step-down minP method and two methods using the false discovery rate (FDR). Step-down minP proposed by Westfall and Young [36] adjusts p-values referring to the result of re-sampling and a permutation test of unadjusted ones. FDR is the expected proportion of false positives among the rejected hypotheses. Benjamini and Hochberg's p-value adjustment [37] applies the m/k coefficient to k-th ordered p-values to control FDR at the specified level, and it works for the case that many genes are differentially expressed. Benjamini and Yekutieli's method [38], which is a modified version of Benjamini and Hochberg's one, works on general dependency cases. With the adjusted p-values, we can recognize significant GO categories under investigation. In this type of GO analysis the number of members belonging to each GO category is important. The result derived from GO analysis without a sufficient number of members is not reliable. Al-Shahrour et al. [35] indicated that GO level 3 constitutes a suitable compromise between the information quality and the number of genes annotated in each category. Figure 4 shows the numbers of human, mouse, and yeast genes annotated in each GO level within FatiGO. In the case of yeast, coverage of the molecular function ontologies is diminished from GO level 3. This is agreeable with the assumption adopted by FatiGO.

It is obvious that genes sharing common biological features can be easily identified by GO analysis. However, quite a number of genes still lack GO annotations at the moment. Thus the high throughput systems, like microarray analysis, are yet in an early stage for aggregating the knowledge from using GO analysis. As the GO analysis has emerged and developed rapidly over the past several years, we can be quite optimistic about the future advancements and use of GO.

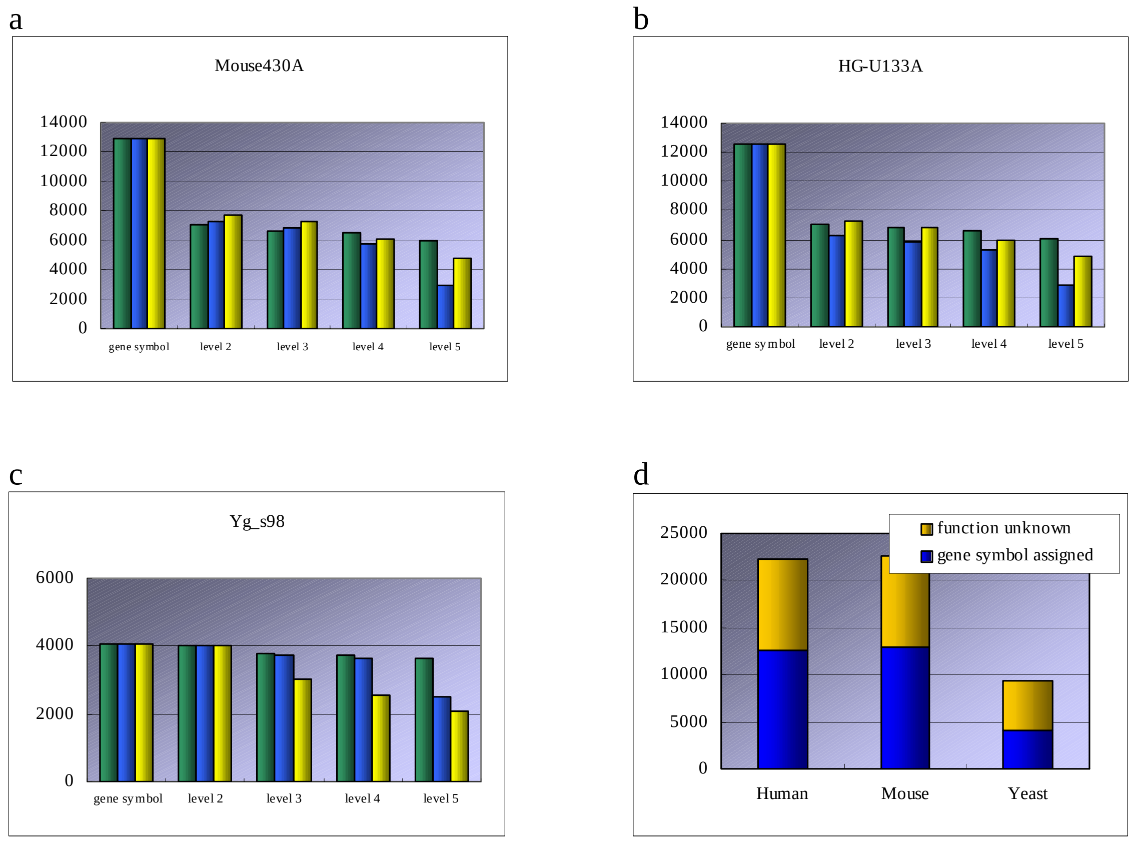

Figure 4. The number of GO assigned genes at each level.
The number of genes that have their own gene symbols and the number of GO assigned genes at each GO level (from 2 to 5) for a) human genome array, b) mouse genome array, and c) yeast genome array. Three bars represent biological process (left), cellular component (middle), and molecular function (right). d) The coverage of gene symbols for these genome arrays.

5. Meta-analysis of microarray expression data

With the upcoming availability of public microarray data repositories, the problem of how to extract, compare, and integrate information from enormous amounts of accumulating microarray data is becoming another important challenging task. Appropriate extraction and integration of the enormous amounts of microarray data not only save time and costs, but also combine the separative information and bring new insights into the underlying mechanisms. Although to a certain extent the congruences in observed expression among different data sets were reported, a considerable inconsistency might still exist among data sets derived from using different analysis techniques, microarray platforms, protocols, or samples [39][40][41].

To identify differentially expressed genes in tissues by using cross-platform data sets of oligonucleotide microarray, Serial Analysis of Gene Expression (SAGE), and Expressed Sequence Tag (EST) human gene expression data, Huminiecki et al. [42] introduced a simple method called Preferential Expression Measure (PEM) for scoring. For SAGE and EST data sets, the PEM equals to log (observed SAGE or EST tag count/expected tag count assumed for a uniform expression), while for oligonucleotide data, the PEM equals to log(observed intensity/ overall intensity mean). Using the PEM value, a comparison of tissue expression over platforms thus becomes possible.

However, without evaluation of the method, whether the method can be generally used or not is unknown.

To integrate the results across independent studies that address a related set of research questions, a set of statistical procedures called meta-analysis was introduced, intending to provide statistically sound results [43]. There are generally two types of meta-analytical statistical methods. One combines the significance, such as *P* values or *Z* scores, while the other combines effect sizes, such as Cohen's *d* statistic or correlation coefficient [43].

Rhodes et al. [44] might be the first to use meta-analytic procedures to integrate microarray data. They introduced a combined statistic *S* using the individual *P* values, where the statistic *S* is defined as:

$$S = (-2) \times \log(P_1) + (-2) \times \log(P_2) + ... + (-2) \times \log(P_n) \qquad (18)$$

here P_i stands for the *P* value of *i*th study in *n* multiple studies. To evaluate the significance over studies, firstly, the summary static *S* was calculated using each P_i derived from random permutation *t* tests [45]. Secondly, the summary static *S* was compared to simulated static *S*, where each P_i was computed by randomly assigning the group labels to the samples in each study. The comparison was performed 100,000 times and the significance is equal to the fraction of randomly simulated *S* greater than or equal to the actual value. Significant genes were finally picked up through a multiple testing correction, the false discovery rate (FDR) adjustment.

Although the combination of individual *P* values enlarges the sample size and is easy for implementation, it focuses on an overall probability instead of on distributions. Therefore, such combined significance tests do not provide an estimate of the magnitude of effects. To tackle the question, the fixed-effects model (FEM) and random-effects model (REM) are commonly used and a general form is given hierarchically as:

$$\begin{aligned} y_i &= \theta_i + \varepsilon_i & \varepsilon_i &\sim N(0, \sigma_i^2) \\ \theta_i &= \mu + \delta_i & \delta_i &\sim N(0, \tau_i^2) \end{aligned} \qquad (19)$$

For FEM, the between-study variance τ^2 is assumed to be zero and the study-specific mean θ_i equals to the overall mean μ. Thus the observed effect size of the *i*th study is only subject to sampling error ε_i alone. For REM, a study specific variance δ_i, whether considered as a constant or a probability distribution, also need to be estimated in addition to overall mean μ. The choice of FEM or RAM is made through the homogeneity test of independent studies.

Concerning the differences between means, Choi et al. [46] applied both FEM and REM to assess the mean difference. The effective size for each study is given by $y_i = (m_A\text{-}m_B)/s_i$ where m_A and m_B stand for the means of two competitive groups, respectively and s_i stands for the estimated pooled standard deviation. Statistically significant genes are chosen by comparing the threshold with the *z* statistic, which is computed as the ratio of estimated μ over its standard error. Under the FEM, a permutation [47] was employed to avoid assuming a normal distribution in calculating the *z* statistic. Under the REM, a Bayesian approach was applied to estimate both ε_i and δ_i by assumption of a *t* distribution for θ_i. Although positive results were demonstrated by database query, further detailed validation might be necessary as various approaches for FEM or REM are already available and there is much room here in modeling strategy.

The application of meta-analysis to microarrays has just begun. The meta-analysis mentioned above only deals with differential expression of the genes and thus depends on the statistical tests for the detection of differential expression. However, meta-analysis can also be expected to be used for integrating the results of other types of procession, such as analysis of variance (ANOVA), clustering, discrimination, and so on. Moreover, versatile statistics such as correlation coefficient, conventional *t*, *F*, and χ^2 may also be used, and microarray specific meta-analytical methods are expected to appear in the near future. With the development of a standard format for microarray data, such as the Minimum Information About a Microarray Experiment (MIAME) provided by the Microarray Gene Expression Data (MGED) Society, more convenient methods for meta-analysis and improvements in analysis results are expected to be seen.

Various numerical techniques are available today to identify groups of genes of potential interest; however, they don't provide information to interpret the biological function and related pathways for the genes. To annotate the function of several genes at a time is not difficult for one expert, whereas annotating over hundreds of genes might be beyond one's knowledge and ability. From accumulated overwhelming information, one possible approach to interpret the unknown function of genes is to link the genes of interest to the information content of published literature.

If we could expand the meaning of meta-analysis which was originally defined by Glass [48] as "….the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings", text mining techniques applied for microarray data interpretation can also be considered as a special variation of meta-analysis in the sense of integrating information.

One commonly performed procedure to retrieve information from the biological literature is keyword indexing. As keywords represent the essential concepts contained in a text, similar keywords are assumed to be found for a specific gene. When each gene is assigned with a keyword list ranked by correlated appearance in the literature, a network of genes with possible relevant function can be constructed. By constructing a network from the co-occurrence of gene symbols in the title or the abstract of an article record in the MEDLINE records, Tor-Kristian et al. [49] showed the effectiveness of such a method in finding functionally related genes by analyzing two publicly available microarray data sets. Similarly, Masys et al. [50] used the hierarchical structure of Medical Subject Headings (MeSH) to retrieve the annotation for genes in the branching tree of terminology.

Focusing on the frequencies of certain terms by using the MEDLINE abstracts, Chaussabel et al. [51] performed clustering analysis based on term occurrences to find functionally relevant genes. Clustering results of the Medline literature database indicated the feasibility of such a unique approach and demonstrated the possibility of combining multivariate analysis for data exploration.

There are other text mining techniques suitable for gene annotation [49]. However, it shall be remembered that the information that can be retrieved from the literature is limited and thus can only provide a superficial assessment. With the improvement in gene nomenclature and synonyms list, text mining methods are ready to produce better results, yet it might take time for statistical techniques to be applied to provide results with better accuracy.

6. Future directions

By the results of substantial studies on the statistically reliable methods such as normalization and data integration, we can be less aware of the differences between data of various backgrounds or diverse platforms today. However, to make the best choice from a group of methods still mainly relies on one's experience as the data are often case sensitive. With the development in microarray technology in the future, consistency and compatibility in data is supposed to relieve the urgent reliance on statistically robust methods. Consequently, less difference among the results of employing various methods will ease automated selection of the method as an embedded function in the microarray equipment.

The ultimate goal of using microarrays is to find the genes responsible for specific function or diseases. Thus, finding the interpretation of microarray results remains a crucial issue. Combining microarray data with other sources of genomic and biomedical information, such as employing gene ontology, are already showing promise. Future studies may focus on establishing systematically more effective way to aid interpretation. However, on the other hand, for complex biological mechanisms, how to keep integrated knowledge from other sources from becoming overwhelming remains tantalizing. Currently, we are much safer from the problem as we have just initiated the attempts to integrate other sources of information for data interpretation. The next several years will be expected to see a boost in this active research area.

We thank the referees for their comments and suggestions.

References

[1] Y. H. Yang, S. Dudoit, P. Luu, D.M. Lin, V. Peng, J. Ngai and T.P. Speed, *Nucleic Acids Res.*, **30**, e15 (2002).
[2] Y. Wang, J. Lu, R. Lee, Z. Gu and R. Clarke, *IEEE Trans. Inf. Technol. Biomed.*, **6**, 29-37 (2002).
[3] M.K. Kerr, M. Martin and G..A. Churchill, *J. Comput. Biol.*, **7**, 819-837 (2000).
[4] Y. Yang, J. Hoh, C. Broger, M. Neeb, J. Edington, K. Lindpaintner and J. Ott, *J. Comput. Biol.*, **10**, 157-169 (2003).
[5] M.L. Lee, F.C. Kuo, G.A. Whitmore and J. Sklar, *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 9834–9839 (2000).
[6] S. Dudoit, Y.H. Yang, M.J. Callow and T.P. Speed, *Stat. Sin.*, **12**, 111-139 (2002).
[7] Q.W. Zhang, N. Ono, Y. Takahara and H. Tanaka, ISMB2002, poster microarrays 10A (2002).
[8] G.C. Tseng, M.K. Oh, L. Rohlin, J.C. Liao and W.H. Wong, *Nucleic Acids Res.*, **29**, 2549-2557 (2001).
[9] Y.H. Yang, S. Dudoit, P. Luu and T.P. Speed, SPIE BIOS 2001 (2001).
[10] F. Sanchez-Cabo, K.H. Cho, P. Butcher, J. Hinds, Z. Trajanoski and O. Wolkenhauer, *Technical Report*, http://www.sbi.uni-rostock.de/publications.htm, (2003).
[11] J. Huang, H.C. Kuo, I. Koroleva, C.H. Zhang and M.B. Soares, Technical Report 321, http://www.stat.uiowa.edu/techrep/, (2003).
[12] Q.W. Zhang, N. Ono, Y. Takahara and H. Tanaka, *Gene*, **324**, 89-96 (2004).
[13] T.B. Kepler, L. Crosby and K.T. Morgan, *Genome Biol.*, **3**, RESEARCH0037 (2002).
[14] J. Fan, P. Tam, G.V. Woude and Y. Ren, *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 1135-1140 (2004).
[15] Statistical Algorithms Description Document, Affymetrix, Inc. Technical documentation - white papers, http://www.affymetrix.com/support/technical/whitepapers.affx (2002).
[16] L. Gautier, L. Cope, B.M. Bolstad, and R.A. Irizarry, *Bioinformatics* **20**, 307-315 (2004).
[17] R.A. Irizarry, B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, R. Scherf and T.P. Speed, *Biostatistics*, **4**, 249-264 (2003).
[18] B.M. Bolstad, R.A. Irizarry, M. Astrand and T.P. Speed, *Bioinformatics*, **19**, 185-193 (2003).
[19] C. Li and W.H. Wong, *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 31-36 (2001).
[20] C. Li and W.H. Wong, *Genome Biol.*, **2**, RESEARCH0032 (2001).
[21] E.E. Schadt, C. Li, B. Ellis and W.H. Wong, *J. Cell. Biochem. Suppl.*, **Suppl. 37**, 120-125(2001).

[22] W. Huber, A. von Heydebreck, H. Sultmann, A. Poustka and M. Vingron, *Bioinformatics*, **18 Suppl 1**, S96-S104 (2002).
[23] W. Huber, A. von Heydebreck, H. Sultmann, A. Poustka and M. Vingron, *Statistical. App. Genet. Mol. Biol.*, **2**, Article 3 (2003).
[24] R.A. Irizarry, B.M. Bolstad, F. Collin, L.M. Cope, B. Hobbs and T.P. Speed, *Nucleic Acids Res.*, **31**, e15 (2003).
[25] J.W. Tukey, *Exploratory Data Analysis*, Chapter 11, Addison-Wesley, MA. (1977).
[26] A.C. James, J.G. Veitch, A.R. Zareh and T. Triche, *Bioinformatics* **20**, 1060-1065 (2004).
[27] D. Rajagopalan, *Bioinformatics* **19**, 1469-1476 (2003).
[28] W.J. Lemon, S. Liyanarachchi and M. You, *Genome Biol.* **4**, R67 (2003).
[29] L.M. Cope, R.A. Irizarry, H.A. Jaffee, Z. Wu and T.P. Speed, *Bioinformtics* **20**, 323-331 (2004).
[30] The Gene Ontology Consortium, *Nat. Genet.*, **25**, 25-29 (2000).
[31] S.A. McCarroll, C.T. Murphy, S. Zou, S.D. Pletcher, C.S. Chin, Y.N. Jan, C. Kenyon, C.I. Bargmann and H.Li, *Nat. Genet.*, **36**, 197-204 (2004).
[32] The Gene Ontology Consortium, *Nucleic Acids Res.*, **32**, D258-D261 (2004).
[33] G. Liu, A.E. Loraine, R. Shigeta, M. Cline, J. Cheng, V. Valmeekam, S. Sun, D. Kulp and M.A. Siani-Rose, *Nucleic Acids Res.*, **31**, 82-86 (2003).
[34] H. Bono, K. Yagi, T. Kasukawa, I. Nikaido, N. Tominaga, R. Miki, Y. Mizuno, Y. Tomaru, H. Goto, H. Nitanda, D. Shimizu, H. Makino, T. Morita, J. Fujiyama, T. Sakai, T. Shimoji, D.A. Hume, Y. Hayashizaki, Y. Okazaki, RIKEN GER Group and GSL Members, *Genome Res.*, **13**, 1318-1323 (2003).
[35] F. Al-Shahrour, R. Diaz-Uriarte and J. Dopazo, *Bioinformatics*, **20**, 578-580 (2004).
[36] P.H. Westfall and S.S. Young, Resampling-based multiple testing: Examples and methods for p-value adjustment, John Wiley & Sons, New York (1993).
[37] Y. Benjamini and Y. Hochberg, *J. Royal Stat. Soc.*, **B57**, 289-300 (1995).
[38] Y. Benjamini and D. Yekutieli, *Ann. Statist.*, **29**, 1165-1188 (2001).
[39] A.T. Rogojina, W.E. Orr, B.K. Song and E.E. Jr. Geisert, *Mol. Vis.*, **9**, 482-496 (2003).
[40] W.P. Kuo, T.K. Jenssen, A.J. Butte, L. Ohno-Machado and I.S. Kohane., *Bioinformatics*, **18**, 405-412 (2002).
[41] Y. Moreau, S. Aerts, B. De Moor, B. De Strooper and M. Dabrowski, *Trends Genet.*, **19**, 570-577 (2003).
[42] L. Huminiecki, A.T. Lloyd and K.H. Wolfe., *BMC Genomics*, **4**, 31 (2003).
[43] L.V. Hedges and I. Olkin, Statistical Methods for Meta-Analysis, Academic Press, New York (1985).
[44] D.R. Rhodes, T.R. Barrette, M.A. Rubin, D. Ghosh and A.M. Chinnaiyan, *Cancer Res.*, **62**, 4427-4433 (2002).
[45] I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, O.P. Kallioniemi, B. Wilfond, A. Borg and J. Trent, *N. Engl. J. Med.*, **344**, 539-548 (2001).
[46] J.K. Choi, U.Yu, S.Kim and O.J. Yoo, *Bioinformatics*, **19 Suppl. 1**, i84-i90 (2003)
[47] V.G. Tusher, R. Tibshirani and G. Chu, *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 5116-5121 (2001).
[48] G. Glass, *Educational Researcher*, **5**, 3-8 (1976).
[49] T.K. Jenssen, A. Laegreid, J. Komorowski and E. Hovig, *Nat. Genet.*, **28**, 21-28 (2001).
[50] D.R. Masys, J.B. Welsh, J. Lynn Fink, M. Gribskov, I. Klacansky and J. Corbeil, *Bioinformatics*, **17**, 319-326 (2001).
[51] D. Chaussabel and A. Sher, *Genome Biol.* **3**, RESEARCH0055 (2002).