BIOINFORMATICS

Vol. 20 no. 11 2004, pages 1772–1784
doi:10.1093/bioinformatics/bth170

Normalization of single-channel DNA array data by principal component analysis

Radka Stoyanova[1], Troy D. Querec[2,†], Truman R. Brown[3] and Christos Patriotis[2,*]

[1]Division of Population Science and [2]Department of Medical Oncology, Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia, PA 19111-2497 and [3]Hatch Center for MR Research, Columbia University, 710 W. 168th St., New York, NY 10032, USA

Received on July 9, 2003; revised on January 8, 2004; accepted on February 8, 2004
Advance Access publication March 22, 2004

ABSTRACT

Motivation: Detailed comparison and analysis of the output of DNA gene expression arrays from multiple samples require global normalization of the measured individual gene intensities from the different hybridizations. This is needed for accounting for variations in array preparation and sample hybridization conditions.

Results: Here, we present a simple, robust and accurate procedure for the global normalization of datasets generated with single-channel DNA arrays based on principal component analysis. The procedure makes minimal assumptions about the data and performs well in cases where other standard procedures produced biased estimates. It is also insensitive to data transformation, filtering (thresholding) and pre-screening.

Contact: Christos.Patriotis@fccc.edu

INTRODUCTION

The development of high-density DNA arrays (oligonucleotide and cDNA) has revolutionized our ability to characterize biological processes and samples genetically by monitoring the relative expression of thousands of genes simultaneously (Bowtell, 1999; Debouck and Goodfellow, 1999; Duggan *et al*., 1999; Lander, 1999). To meet the challenges for interpretation of this complex data, sophisticated software packages have become available for analysis of the gene expression profiles, such as ScanAnalyze (Eisen and Brown, 1999), ArrayExplorer (Patriotis *et al*., 2001) and ImaGene (Biodiscovery, Inc.). An important, but still unresolved, issue is associated with the normalization of the relative expression of genes across a series of microarray experiments. In order to compare the results from multiple samples, which is the ultimate goal of these studies, it is obligatory that the individual array datasets be normalized to correct for the inherent experimental differences. The critical element in this process is the discrimination of the interesting, biological variation from the obscuring variation, which is related to the experimental conditions (Hartemink *et al*., 2001). This is why the initial attempts towards normalization of array datasets relied on the concept that a group of genes could be identified *a priori* and serve as 'housekeeping' genes, assuming that their expression will reflect directly the obscuring experimental variation. As discussed in detail below, if such a subset of genes could be identified reliably, then well-defined normalization factors could be estimated to within the accuracy inherent in the measurements. Unfortunately, as shown by others (Butte *et al*., 2001; Selvey *et al*., 2001) and by us in this report, this simple concept works only in very limited cases. (Here and in the rest of the paper, we will refer to the *a priori* specified housekeeping genes as 'designated' in order to distinguish them from those determined to be the 'true' housekeeping genes. The latter represent the subset of genes whose expression is invariant to the particular biological and/or experimental variables in the multiple microarray experiments being compared.)

The realization that in most of the cases the 'designated' housekeeping genes cannot be used for reliable normalization has spurred the development of alternative approaches for normalization. The majority of these approaches determine normalization factors on the basis of averages over the behavior of the entire set of genes measured (Schuchhardt *et al*., 2000). Typically, these methods utilize the mean or median of the array intensities (Quackenbush, 2001) and linear (Golub *et al*., 1999) or orthogonal regression (Sapir and Churchill, 2000). A variety of non-linear techniques were also proposed (Schadt *et al*., 2000, 2001; Li and Wong, 2001; Bolstad *et al*., 2003).

There is also a series of methods that identify a subset of genes in the data that can be assumed as housekeeping (Zien *et al*., 2001; Kepler *et al*., 2002). All these approaches perform

*To whom correspondence should be addressed.

†Present address: Emory University, GDBBS, 1462 Clifton Road, Dental Bldg, Suite 314, Atlanta, GA 30322, USA.

satisfactorily when the following two assumptions about the data are met:

(1) the majority of the genes (in the fitting segment for the non-linear approaches, or overall) are not affected by the experimental variables, i.e. they can all be regarded as housekeeping genes; and

(2) the subset of differentially expressed genes are 'activated' symmetrically, i.e. the overall intensity change of up- and down-regulated genes is similar.

Here we present a novel normalization approach that performs satisfactorily even when the conditions above are not met, which is the most commonly observed scenario. In contrast to the methods requiring the selection of a baseline array, this method analyses the entire dataset simultaneously, and, as such, it is considered a complete data method (Bolstad *et al*., 2003). The goal of the technique is to determine in a multi-array experiment if there is a subset of genes whose expression may be considered unaffected by the 'interesting' (biological) sources of variation and if there are such, to identify this set of specific, 'data-driven' housekeeping genes and use them for normalization. Briefly, if the results from each array measurement are represented in a multi-dimensional vector space where each axis is a different sample, then the entire experiment can be represented as a series of points corresponding to the strength of each gene's expression in each sample measured. If a set of genes with an unchanged relative expression is present, their intensity levels will represent points along a straight line through the origin. We present a principal component analysis (PCA)-based method for identifying such a line, if one exists. The factors determined from the expression of these genes can be used to normalize the gene expression in the individual array datasets.

MATERIALS AND METHODS

Theory

Consider a gene expression dataset consisting of m arrays with n genes each. Let **D** be the data matrix containing in its rows the measured expression levels, and let g_{ij} be the measured expression level of the i-th gene in the j-th array ($i = 1, \ldots, n$, $j = 1, \ldots, m$). We seek to identify a subset, **S**, of s genes ($s \leq n$) whose expression remains constant over the experimental conditions of the study. Mathematically, for the genes in S the following equations hold:

$$q_j g_{ij} = c_i \quad \text{or} \quad g_{ij} = c_i / q_j,$$

where q_j is the j-th normalization constant and c_i is the true concentration of the i-th gene, which is constant across the samples. If we plot the points g_{ij} in an m-dimensional space, we can see that they lie along a line through the origin, which has projections along the axes of $\{1/q_j\}$. If we can find such a line, we will have identified our desired relative normalization constants (relative since unless at least one of the c_is is known, it is impossible to normalize the data absolutely).

We now turn to the problem of identifying the genes in **S**. The obvious method is to calculate the densities in the cloud of n data points in the m-dimensional data space, which represent the directions of n gene levels in the m observations. In reality, this is difficult because there are approximately N^{m-1} directions for examining if each orientation is divided into N segments. In order to reduce the dimensions of the space that needs to be examined, we use PCA to identify the directions along which the principal variations of the genetic expressions lie in the original m-dimensional space. We project the data points onto the first two of these directions and examine their angular distribution to determine if a line through the origin is present. Note that the original line in the full space need not lie in this plane as its projection into the plane will also be a line through the origin.

PCA is used commonly for reducing the dimensionality of complex data (Anderson, 1971) and has been used previously in the analysis of microarray data from time-course experiments (Alter *et al*., 2000, 2003), for normalization of gene expression ratios obtained from two different microchips of two-channel arrays (Nielsen *et al*., 2002) and for partitioning large-sample microarray-based gene expression profiles (Peterson, 2003). It is also an inseparable part for exploration of large genomic datasets (Misra *et al*., 2002). Previously, we have applied the PCA technique for removing 'unwanted' variation in multi-spectral datasets (Stoyanova and Brown, 2002).

Briefly, PCA identifies the directions of the largest variations in the data via the principal components (PCs), and represents the data in a coordinate system defined by the PCs ($\vec{P}_1, \vec{P}_2, \ldots$), as follows:

$$\mathbf{D} = R_1 \vec{P}_1 + R_2 \vec{P}_2 + R_3 \vec{P}_3 + \cdots + R_m \vec{P}_m, \qquad (1)$$

where $\vec{P}_j$ ($1 \times m$) and R_j ($n \times 1$) are row and column matrices; R_j contain the projections of the data along the PCs ($j = 1, \ldots, m$), generally called scores. Below, some of the relevant properties of the PCs are listed.

(1) $\vec{P}_j$ are eigenvectors of the data-covariance matrix (calculated around the origin, rather than around the mean) and are orthonormal, i.e.

$$\vec{P}_i \cdot \vec{P}_j = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j. \end{cases}$$

(2) The PCs are ordered by the decreasing amount of variation in the data they explain. Let $\Lambda_1, \Lambda_2, \ldots, \Lambda_m$ be the eigenvalues of the covariance matrix ($\Lambda_1 > \Lambda_2 > \cdots > \Lambda_m$). Each PC explains a portion of the total variance of **D**, proportional to its corresponding eigenvalue.

(3) The magnitude of R_j is proportional to its corresponding eigenvalue, Λ_j.

(4) **D** can be represented sufficiently with fewer than m PCs [Equation (1)]. PCA provides a representation of the data in a lower-dimensional space of significant variables.

(5) The PCs are a linear combination of the original data. The coefficients of this linear combination (R_i) are typically referred to as loadings and represent the projections of the PCs along the axes of the original m-dimensional space.

(6) The PCs minimize the squared distances of the variables (gene-expression levels) and themselves.

From the last three properties, it follows that the loadings of the first PC may serve as normalization coefficients of the arrays. In many cases, when the assumptions (1) and (2) (see Introduction) are met, as discussed in detail below, PCA can provide directly the normalization coefficients sought. In other cases, we can use the first two PCs to detect linear behavior in a subset of genes **S** ($s \leq n$) that are the 'true' housekeeping genes. PCA applied only to the genes in **S** will identify the appropriate normalization line in the entire m-dimensional data space. Its projections can then be used as normalization factors.

The procedure [dubbed PCA(line)] tests automatically for the existence of and detects the group of genes, which are distributed 'tightly' along a line in the plane defined by the first two PCs. We chose this plane because by definition it contains the largest variations in the expression levels. Although the actual straight line of the desired normalization may not lie completely in this plane, its projection in the plane is also a straight line and will serve to identify the desired set of genes. To identify such a line, we divide the part of the plane that contains all the points into small angular segments and determine the number of data points (genes) in each segment. The segment(s) containing the data-driven housekeeping genes will contain a disproportionally large density of points. This procedure is described below and given in detail in Appendix 1.

Initially, we assume **S** is an empty set (**S** $\equiv \emptyset$). In the plane defined by $\vec{P}_1$ and $\vec{P}_2$, we partition the angle through the origin defined by the genes with maximal and minimal components on $\vec{P}_2$ in p equal angular segments. Let s_k ($k = 1, \ldots, p$) be the subset of genes in **D**, that belong to the k-th segment ($s_1 \cup s_2 \cup \cdots \cup s_p = \mathbf{D}$). We recommend that p be set initially to contain on average at least 10 genes per segment. Let θ_k be the angular densities defined as the number of genes in each segment, s_k, and $M(\theta_k)$ and $V(\theta_k)$ be, respectively, the sample mean and variance of θ_k. Then, the density of the k-th segment is considered to be significant if

$$\theta_k > M(\theta_k) + \mu\sqrt{V(\theta_k)}, \qquad (2)$$

where μ is a parameter indicating the number of standard deviations above the mean that is required for significance. If a normal distribution of θ_k is assumed, then $\mu = 1.96$ will correspond to a one-sided test with a type-I error of 2.5%. However, in most cases, due to different procedures for microarray image quantification as well as the specific prefiltering of the data, the distribution of θ_k is unknown. In cases where a normal distribution of θ_k cannot be assumed, it is recommended that their histogram be examined and μ be set appropriately. For added stringency of the test, the genes in segment s_k are assumed to be housekeeping genes only if θ_{k+1} of the neighbouring segment s_{k+1} is also tested significant. Then the genes in the two segments are merged in **S**, i.e. **S** $\equiv s_k \cup s_{k+1}$. If the angular density of the genes of further contiguous segments is detected to be significant, then these genes are added to **S**. After all segments are tested, PCA is applied to **S** and the reciprocal values of the loadings of the resultant first PC are used as normalization coefficients.

If the procedure failed to identify at least two significant contiguous segments, then either all the genes in the data can be assumed to be housekeeping (**S** $\equiv$ **D**), or, in the extreme situation, the housekeeping genes are either too few to be detected or not existent (**S** $\equiv \emptyset$). In the first case, the loadings of the first PC from the initial PCA of **D** are the true normalization coefficients and can be used for direct normalization. There is not very much to be done in the second case—the PCA-derived normalization would be as erroneous as the ones produced by any other linear technique. Let λ_1 be the fraction (in per cent) of the first eigenvalue, Λ_1, from the total variance in the data. In this case, a low λ_1 (in our experience $<60\%$) will be indicative of a lack of normalizing genes.

Biological samples (datasets)

Human ovarian surface epithelial cell lines Microarray datasets obtained from experiments with RNA of human ovarian surface epithelial (HOSE) cells were analyzed using Atlas 1.2 Human arrays (ClonTech). The details of array preparation and data extraction are described elsewhere (Patriotis *et al.*, 2001). Briefly, the HOSE cells were derived from a short-term primary cell culture obtained from one of the ovaries of an individual predisposed to ovarian cancer. The short-term HOSE cell culture was transduced with a Cytomegalovirus-based vector expressing the Simian Virus-40 large T-antigen. As a result, the *in vitro* lifespan of the cells, while still 'mortal' (118M), was considerably extended, leading to the spontaneous outgrowth of an 'immortal'/non-transformed cell line (118Im). Following multiple passages in culture, the 118Im cell line gave rise spontaneously to cells that acquired anchorage-independent growth characteristics and, ultimately, the potential to grow tumours *in vivo* when inoculated in nude mice (118NuTu) (Frolov, A. *et al.*, unpublished data). In the first experiment, the cDNA probes were derived from total RNA purified from 118M, 118Im and 118NuTu. In the second experiment, microarray data were obtained from 118NuTu cells treated for different lengths of time (0, 24, 48 and 72 h) with the synthetic retinoic acid derivative Fenretinide (4-HPR) (Moon *et al.*, 1979).

Lymphoma data (LD)

The dataset was constructed from the supplementary datasets of Golub *et al*. (1999). The microarray measurements were performed with RNA of samples obtained from bone marrow and peripheral blood from patients with acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML) at the time of diagnosis using high-density oligonucleotide Affymetrix arrays. In the paper referred to, the data were normalized by pair-wise linear regression (LR) between the first sample (baseline) and the rest of the samples in the dataset. Only genes with satisfactory quality (marked with 'P' in the datasets provided) in each pair were considered for the regression. The normalized datasets, as well as the normalization factors, are supplied at http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi. The data used here were non-processed and 'non-normalized', and the combined datasets resulted in a data matrix containing 72 arrays and 7129 genes.

Simulated data

The values in the simulated datasets were chosen to be realistically probable, based on our experience with data obtained with the Atlas 1.2 CLONTECH arrays (Patriotis *et al*., 2001). The number of genes was set to 500, in agreement with our observation that between 30 and 50% of the genes are expressed in any of the samples investigated in our lab. In the first array, the expression levels, g_{i1} [in arbitrary units (a.u.)], were simulated using the relation $g_{i1} = 2^u$, where u is uniformly distributed between 1 and 16.

In all simulated datasets of pairs of arrays a multiplication factor of 1.2 was applied to the second array, equivalent to $q_1 = 1$ and $q_2 = 1.2$. Gene intensities were assumed to be background-corrected, and (unless noted otherwise) signals with intensities less than 200 were zeroed (thresholded).

'Noise' data

The sources of noise in microarray datasets are multiple and complex, and they contribute simultaneously with variable amounts to the total variance in the data. Generally, the total noise contribution to the measured signal represents a variable mixture of the contribution of two components: one is independent of gene intensity and affects the expression of all genes equally, and the other is gene-dependent and increases with the magnitude of the gene expression. To investigate the contribution of noise to the process of normalization, we simulated two pairs of replicate arrays, as described above. Random noise was added to each array. In the first set, the noise was gene independent (N_1)—uniformly distributed random noise between -2500 and 2500—and in the second set, a gene-dependent (N_2), uniformly distributed noise whose magnitude was $\pm 10\%$ of the gene intensities. Formally,

$$N_1 = -2500 + 5000\,u$$
$$N_2 = \frac{g_{i1,2}}{10}(2u - 1) \qquad u = U(0,1). \tag{3}$$

'Signal' dataset 1

'Signal' dataset 1 (SD1) contained two pairs of simulated arrays. The first pair satisfied conditions (1) and (2) (see Introduction) by choosing a substantial number of the genes to be housekeeping (250) and the number and magnitude of change of up- and down-regulated genes to be equal. The second pair was constructed to illustrate a scenario where these assumptions are not met: the housekeeping genes (150) were not a majority, and more genes were 'up-regulated' (200) than 'down-regulated' (150) (the details about the simulated up- and down-regulation are given in Appendix 2). Two independent sets of random noise were added to each array, generated as the sum of half of both gene-dependent and -independent noise [Equation (3)], i.e. $\frac{1}{2}(N_1 + N_2)$.

'Signal' dataset 2

'Signal' dataset 2 (SD2) contained eight arrays with 500 genes each. The first array in SD2 was generated randomly, as described above. The gene expression levels of the remaining seven arrays were generated with the idea of recreating a scenario where progressive changes occur in the studied samples (e.g. time-response to treatment or undergoing a process of immortalization and malignant transformation). The details of simulation parameters for up- and down-regulation are given in Appendix 3. The arrays were multiplied with coefficients generated at random between 0.3 and 3. Finally, random noise, generated as described for SD1, was added to each array.

RESULTS

Housekeeping genes in HOSE cells

Figure 1(a) depicts the correlation plot of the 'designated' housekeeping genes in the first experiment with HOSE cells: 118M on the x-axis, and on the y-axis 118Im (black series) and 118NuTu (gray series). The expression of these genes is well correlated ($R^2 = 0.96$), and, in this case, they can be used for normalization of the data. Figure 1(b) depicts the correlation plot of the expression of the same set of housekeeping genes in the 118NuTu, untreated (0 h, x-axis) and treated with 4-HPR for 24, 48 and 72 h (y-axis; black circles, gray triangles and shaded squares, respectively). In this case, the correlation between the expression of the 'designated' housekeeping genes is quite poor ($R^2 = 0.43$, 0.81 and 0.85, respectively). From these data, it is clear that the expression profiles of the 'designated' housekeeping genes are changed non-uniformly in the cells in response to the drug treatment.

'Noise' data

Figure 2(a) and (b) (left panels) depict the correlation between the data in the two pairs of simulated arrays in this dataset together with the linear trendline through the origin. Note that the regression coefficient in both cases is very close to the true value of the multiplication factor 1.2. The fit is slightly tighter

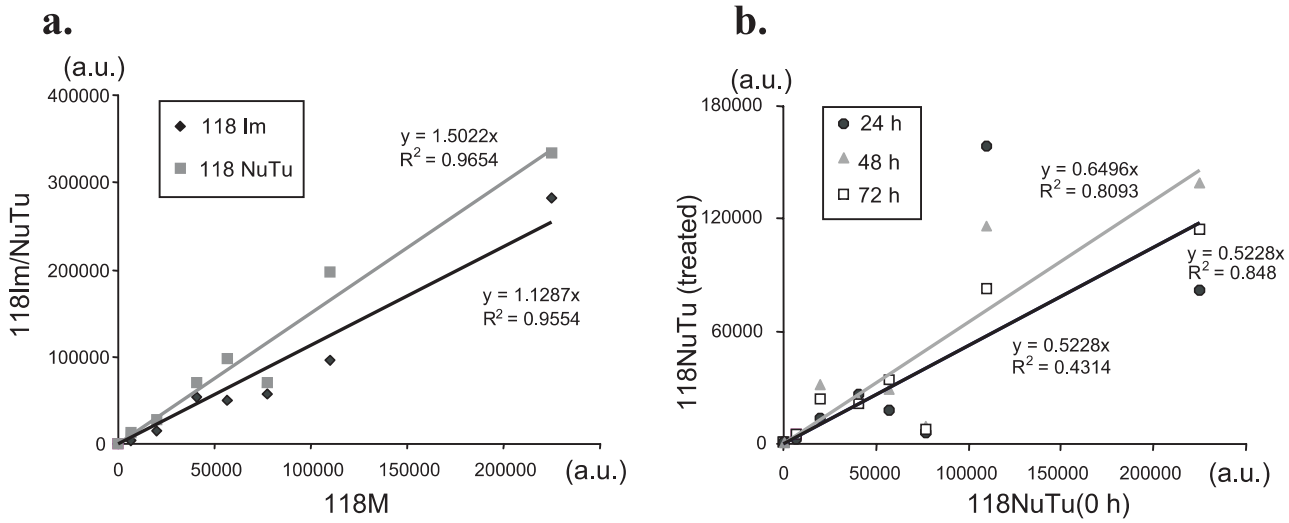

Fig. 1. Correlation plots of the intensities of the 'designated' housekeeping genes in two microarray experiments. (**a**) HOSE cell lines at different stages of malignancy, on the x-axis 118M, and on the y-axis, 118Im (black) and 118NuTu (gray). Regression lines are indicated in black and gray, respectively; (**b**) 118NuTu cell line following treatment with Fenretinide, on the x-axis at 0 h and on the y-axis after 24 (black circles), 48 (gray triangles) and 72 h (squares) of treatment. Regression lines are indicated in black solid, black dashed and gray, respectively (note that the black solid and black dashed regression lines are overlapping).

for the second dataset ($R^2 = 0.986$ versus $R^2 = 0.992$), which reflects the smaller contribution of the noise in the overall gene intensities. Figure 2(c) (left panel) depicts the correlation between two replicate array datasets obtained from 118M. The genes depicted by gray squares represent the 'designated' housekeeping genes. On the right panels in Figure 2 the correlation of the logarithmic transforms of the data from the left panels are presented (due to the restriction of the logarithmic function to only positive numbers, for this comparison, only genes that are expressed simultaneously in the two arrays are used). Comparison of the graphs of simulated [Fig. 2(a) and (b)] and real [Fig. 2(c)] noise indicates the similarity in the overall distributions, although the real data have a greater variance.

'Signal' dataset SD1

The graphs of the two pairs of arrays in this dataset, together with the regression line through the origin, are presented in Figure 3. The housekeeping genes are marked in green. In the case of the first pair [Fig. 3(a)], it is clear that the regression line is along the line of normalization and, therefore, all the above reference normalization methods will perform well. Obviously, this is not the case with the second dataset [Fig. 3(b)], and we applied the PCA (line) procedure for determining the subset of housekeeping genes.

After thresholding, 296 genes were found with non-zero intensities simultaneously in both arrays (132 up-regulated, 88 down-regulated and 76 housekeeping). PCA was applied to this set ($\lambda_1 = 96\%$). The representation of the data along the first two PCs is shown in Figure 4(a) [note that the first PC, $\vec{P}_1$, is along the regression line of this rotated version of Fig. 3(b)]. The procedure for automatic detection of the housekeeping genes is schematically illustrated in Figure 4(b). The angle encompassing all data points (between 1.069 and 2.438 radians) was divided into 50 segments. The histogram of the angular densities θ_k ($k = 1, 2, \ldots, 50$) is presented in Figure 4(c) [$M(\theta_k) = 5.92$ and $\sqrt{V(\theta_k)} = 5.18$]. For $\mu = 1.96$, three contiguous segments, starting at $p = 22$, contained points with a significantly higher density [Equation (2)]. A total of 63 points (subset **S**) from these segments were extracted. These genes (orange points), together with the original set of housekeeping genes (in green), are presented in Figure 4(d). The collinearity between the identified genes and the housekeeping genes is apparent. Thirty-two of the genes in **S** belong to the original set of 76 housekeeping genes in the analyzed data, indicating that the procedure recovered successfully a substantial fraction of them (32/76, or >40%). Moreover, the procedure detected an additional 31 genes whose expression changes in accordance with a housekeeping gene behavior. PCA was applied to the data in **S** ($\lambda_1 = 99\%$), and the first PC loading factors were $q_1 = 0.635$ and $q_2 = 0.773$, corresponding to a relative normalization factor of 1.217.

Simulated dataset SD2

PCA was applied to 205 genes with non-zero intensities in all eight arrays (88 up-regulated, 52 down-regulated and 64 housekeeping) ($\lambda_1 = 96\%$). The points in the $\vec{P}_1$ and $\vec{P}_2$ plane were within 1.079 and 1.938 radians. As in the case of SD1, the densities of points in 50 segments were calculated ($M(\theta_k) = 4.08$ and $\sqrt{V(\theta_k)} = 5.21$). For $\mu = 1.96$, three

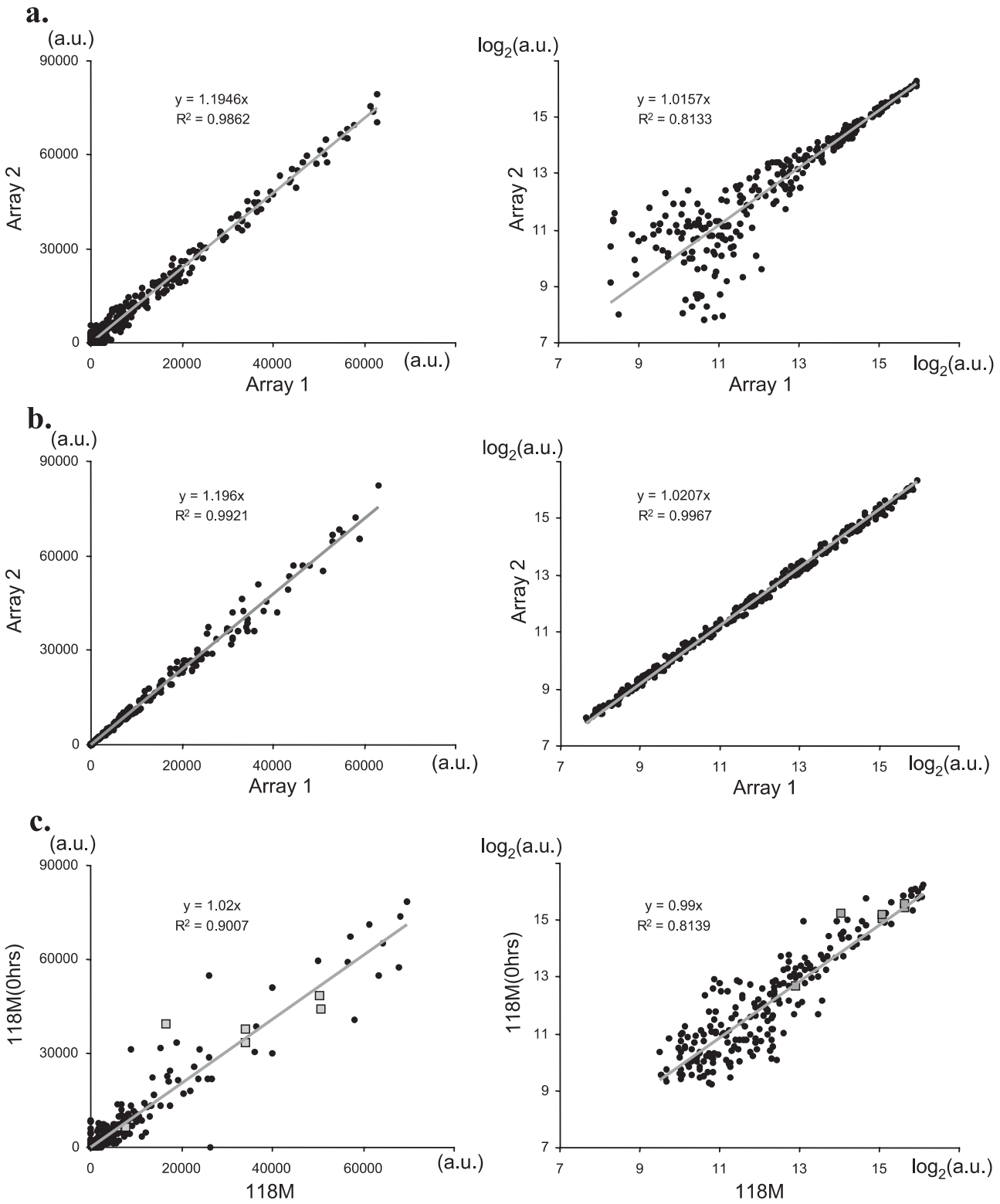

Fig. 2. Correlation plots of gene intensities in replicate arrays, displayed on untransformed (left panels) and logarithmic scales (right panels) with indicated LR line (gray): (**a**) simulated data, containing gene-independent noise; (**b**) simulated data, containing gene intensity-dependent noise; (**c**) two replicate arrays of 118M cell line. The genes shown in gray squares represent the designated housekeeping genes included in the arrays by the manufacturer.

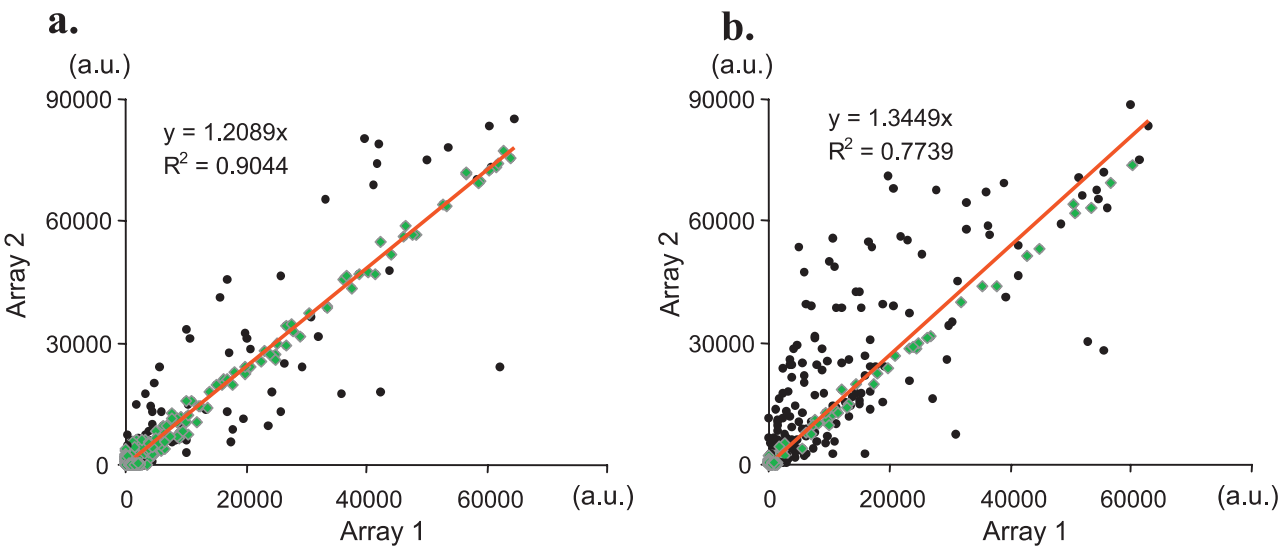

Fig. 3. Correlation plots of gene intensities of two simulated array datasets (SD1) with indicated housekeeping genes (green squares) and indicated LR line (orange): (**a**) 'symmetric' case, where the majority of the genes are housekeeping and the number and magnitude of up- and down-regulated genes is similar; (**b**) the housekeeping genes are of a relatively smaller number, and the up-regulated genes dominate the distribution.

contiguous segments containing a total of 64 points (subset **S**) contained a significant number of points. The majority of the points in **S** belonged to the original set of housekeeping genes analyzed (44, or 69%), and the remaining 20 were split between the 12 up-regulated and eight down-regulated genes. PCA was applied to the data in **S** ($\lambda_1 = 99\%$), and the normalization coefficients $q_j (j = 1, \ldots, 8)$ were calculated as the loadings of the first PC.

We compared the accuracy of the PCA(line)-estimated normalization factors with the ones estimated by LR and mean (MEAN). We scaled all normalization factors so that their sum was equal to 1, and the correlation between the true values (x-axis) and the estimated values (y-axis) are presented in Figure 5(a). Although the overall correlation between the true and estimated normalization factors is quite good [$R^2 = 0.9964$, 0.9862 and 0.9726 for PCA(line), LR and MEAN estimates, respectively], it is clear that PCA(line) provides the best estimates. We also calculated the error for each individual array, defined as the percentage difference of the estimated from the true normalization factor, and the minimum, maximum and average error values are presented in Figure 5(b). This analysis indicated that the error of the PCA(line)-derived estimates is on average lower by a factor of 2 and 3 as compared with the ones derived by LR and MEAN, respectively.

We further investigated the effect of data thresholding on the PCA(line) procedure. We re-analyzed SD2 by applying PCA to all 500 genes in the dataset. Since some of the scores along $\vec{P}_2$ were negative, the data points spanned the entire plane (between 0.03 and 6.27 radians). In this case, we set $p = 200$ and $\mu = 4$. Two consecutive segments [Fig. 5(c)], containing a total of 77 genes, were determined to have significant angular densities. The overwhelming majority of genes (55) in this set belonged to the original set of housekeeping genes. The housekeeping gene sets derived by PCA (line) on thresholded and unfiltered data were strongly overlapping—all but four were identical to the 64 housekeeping genes determined with the thresholded data. Finally, the PCA-determined normalization factors in this case were virtually identical to the ones determined with the thresholded data.

Lymphoma Data

PCA was applied to all 7129 genes in the dataset ($\lambda_1 = 88.31\%$). All loadings of $\vec{P}_1$ were scaled by the first one, resulting in a normalization factor of 1 for the first array. Figure 6(a) depicts the comparison between LR- and PCA-derived (yellow circles) values. The high correlation ($R^2 = 0.99$) between the two series is apparent. Further, we applied the PCA(line) procedure. Three contiguous segments (from a total of 200), containing 1095 genes, were above the threshold [$M(\theta_k) = 35.64$, $\sqrt{V(\theta_k)} = 72.21$, $\mu = 4$]. PCA was applied to the intensities of the genes in **S** ($\lambda_1 = 93.85\%$) and the loadings of $\vec{P}_1$ rescaled appropriately and compared with the LR results [Fig. 6(a), black circles]. While showing an overall good agreement with the LR-derived results ($R^2 = 0.92$), they also indicate, in some individual cases, substantial differences with the PCA(line)-estimated values. The average absolute value of the relative difference between LR- and PCA-derived factors was 7.52%, with a range of 0.07–30.84% in the case of array #65 [Fig. 6(a), marked with an arrow]. We then examined the correlation of the intensities of the genes marked with 'P' (those of satisfactory quality) in arrays # 1

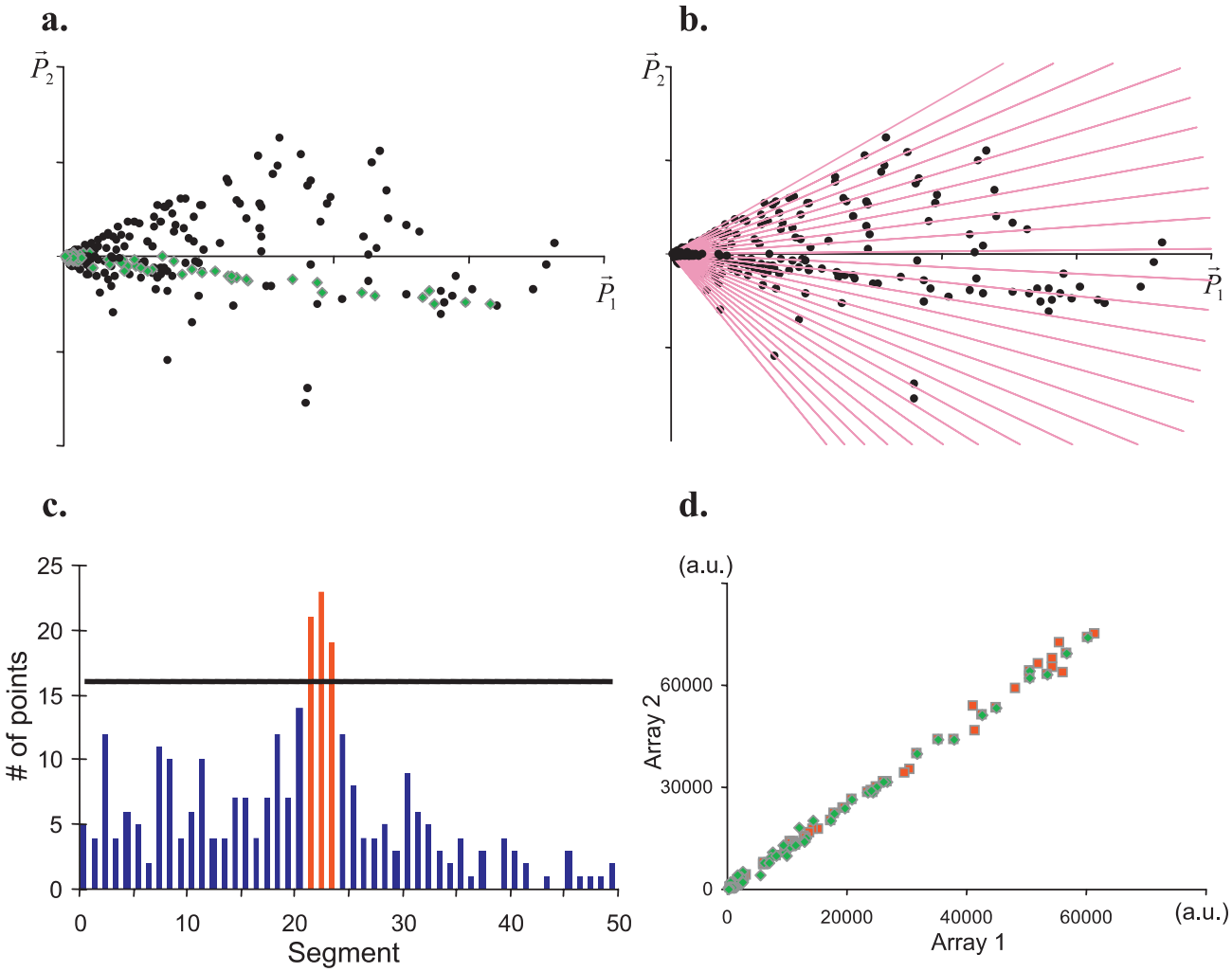

Fig. 4. (**a**) The data from Figure 3b, presented in the PC-plane; (**b**) schematic illustration of segmentation of the part of the PC-plane containing the data; (**c**) histogram of the angular densities of the segments; (**d**) 'true' (green) and PCA(line)-detected housekeeping genes (orange).

and # 65 [Fig. 6(b)]. The normalization lines [represented in orange and blue, respectively, for LR and PCA(line)] indicate that in the case of LR, a handful of strongly expressed genes are driving the normalization. A similar graph was obtained with arrays #1 and #58, which also showed a large difference between the two normalization procedures.

To determine how the number of segments in the plane impacts the estimated normalization coefficients, we ran the procedure with $p = 100$, 300, 400 and 500. In all cases, the procedure extracted essentially the same subset of normalizing housekeeping genes. The number of genes for each p was 1410, 1192, 1092 and 1162, respectively. We estimated a (5×5) correlation matrix of the derived normalization factors for each value of p. All coefficients in the correlation matrix were greater than 0.994, indicating the high degree of reproducibility between the derived normalization factors for different numbers of segments (p). We also estimated the coefficient of variation (COV) between the five series of estimates. The average COV for the 72 normalization factors was 1.71%.

DISCUSSION

Normalization of gene intensities in multi-array experiments is crucial for the ultimate biological interpretation to be meaningful (Hoffmann *et al.*, 2002). Only after proper normalization can changes in expression of a given gene amongst the studied samples in the experiment be characterized quantitatively. Conversely, erroneous (or no) normalization may lead to inaccurate estimation of the changes in gene expression including wrong conclusions with regard to their up- or down-regulation. While optimal normalization is still a subject of discussion, individual investigators are faced daily with many questions about the analysis of these complex

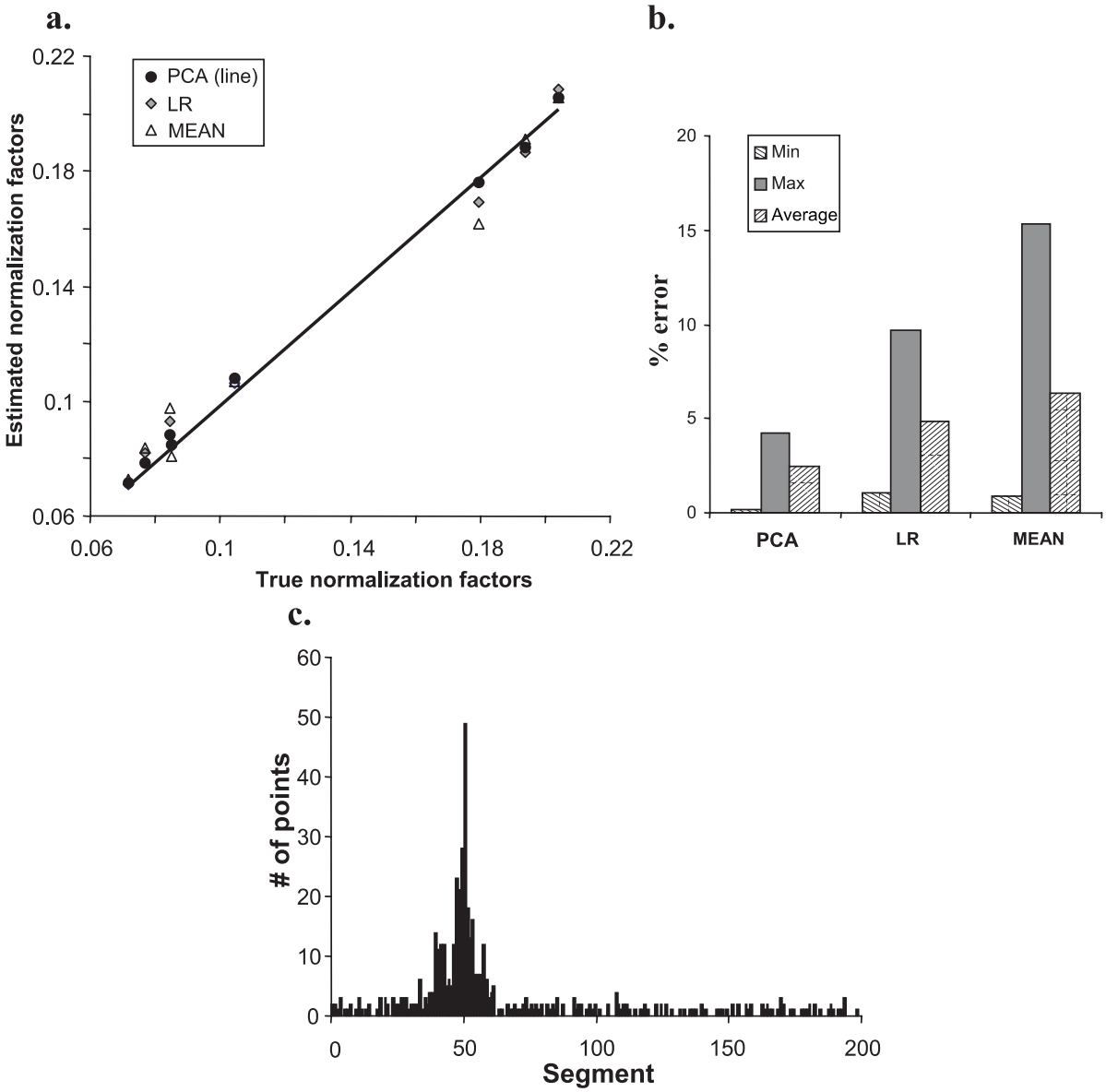

Fig. 5. (**a**) Relation of 'true' normalization factors and factors estimated via PCA(line), LR and MEAN in a simulated dataset containing eight arrays. The black line indicates the line of identity; (**b**) ranges (minimum and maximum) and average of the absolute values of relative errors of estimation of the normalization factors in the three estimates; (**c**) histogram of the angular densities of the segments in the PCA(line) for unfiltered data.

data. For example, should the array data be logarithmically transformed prior to normalization; should low intensity spots be discarded, and, if so, what is the right cut-off limit for this operation; should the mean or median intensity of the arrays be used for normalization; or alternatively, do 'designated' housekeeping genes play reliably their assigned role?

In this report, we address all these questions and present a simple procedure for normalization of datasets generated with single-channel arrays based on PCA. The procedure makes minimal assumptions about the data and does not require any pre-processing, pre-screening or filtering of the data.

The need for alternative normalization techniques arose with the realization that genes assumed as housekeeping and 'designated' by the manufacturers as such on arrays are not reliable for accurate data normalization. In the first experiment with HOSE cells, investigating a set of three cell lines with close genetic origin, the 'designated' housekeeping genes change in a coordinated fashion, and it is likely that they fulfill their role as normalizing genes. This result is anticipated

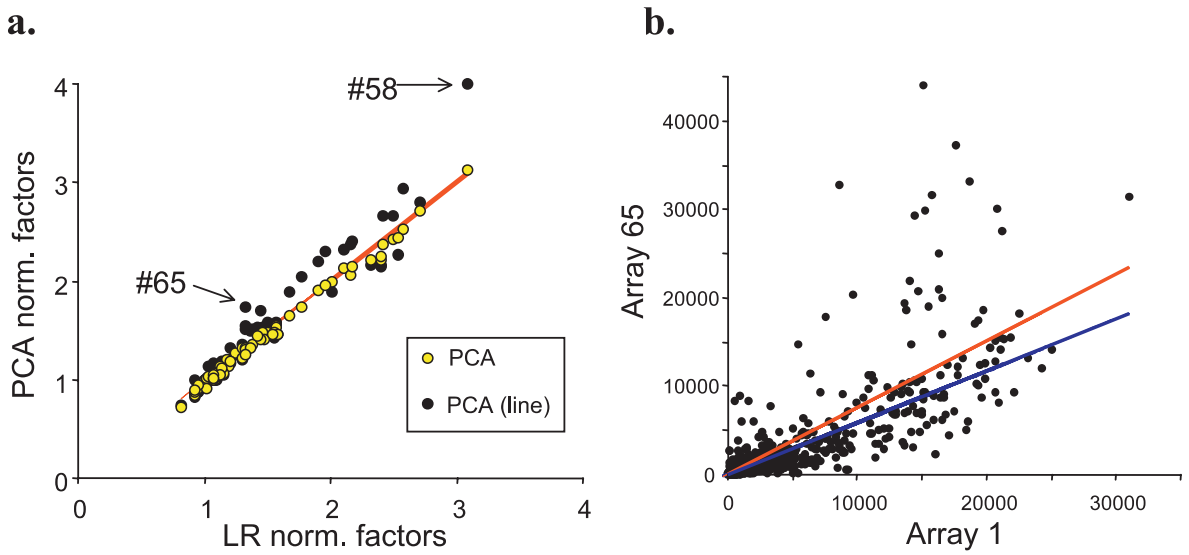

Fig. 6. (**a**) Correlation between LR- estimated (x-axis) and PCA- or PCA(line)-estimated (yellow series and black series, respectively) normalization factors for the LD. The orange line indicates the identity line. The arrows point at arrays with a large relative difference; (**b**) correlation plots of intensities of genes marked with 'P' in arrays #1 and #65. The normalization lines derived by the LR and PCA(line) estimates are indicated in orange and blue, respectively.

since the three cell lines were cultured under standard growth conditions and the observed differences in the global gene expression profiles are related to only a small subset of genes associated with the sequential transition of the cells through the process of malignant transformation. Conversely, in the second experiment, the 'designated' housekeeping genes appear to change differentially in response to treatment with Fenretinide. This is consistent with the dramatic biochemical changes associated with the process of cells undergoing programmed cell death (Querec, T.D. *et al*., manuscript in preparation). The major alterations in the global gene expression profile that precedes and leads to the triggering of apoptosis affect the expression states of most housekeeping genes.

Pre-processing of the data prior to normalization is an important issue. Typical steps include background correction, logarithmic transformation and/or thresholding. We believe that the background should be removed prior to normalization, so that the normalization line goes through the origin. Although we simulated gene intensities, as described in the Materials and methods section, there is no theoretical basis to assume that real data comply with this distribution. Log-transformation has the advantage of transforming the noise distributions approximately to Gaussian. This property can be used for estimating the probabilities of differentially expressed genes (Kerr *et al*., 2000). The PCA-based normalization procedure, however, is based on identifying the genes along the normalization line in the dataset and is invariant to prior transformation. Moreover, based on 'noise'-simulated data, as well as from the HOSE cell replicates, it is apparent that log-transformation may be detrimental to the analysis as it increases the relative contribution of the gene-independent noise in genes expressed at low levels. Because of these adverse effects, and the fact that by estimating the numbers of genes in the segmented plane the PCA(line) procedure allows low-expressed genes to be taken into consideration, we chose to implement our normalization procedure on raw (untransformed) data.

The described procedure is also insensitive with respect to prefiltering (thresholding) of the data, given that the parameter μ [Equation (2)] is adjusted appropriately. In the case of 'thresholded' data, $\mu = 1.96$ will be sufficient to discriminate between the sought housekeeping genes and the rest [Fig. 4(c)]. This μ-value will merely distinguish the 'noise' genes from the signal ones in non-prefiltered data. Thus, a larger μ [as in the case shown in Fig. 5(c)] is required to detect the normalizing genes sought. We therefore strongly recommend exploring the characteristics of the angular histogram of the data before setting the appropriate μ-value.

The only assumption made about the distribution of the intensities of the houseskeeping genes for PCA(line) is that they are distributed along a straight line. This assumption is very sensible for single-channel arrays, unlike the case of the double-channel arrays, where it is known that a non-linear dependence exists between the gene expression levels among the two channels (Yang *et al*., 2002). Furthermore, it has been shown recently that even for these arrays the linear and non-linear normalization methods perform similarly (Park *et al*., 2003). In our experience, most of the non-linear effects are due to improper scanning settings, which, besides the unwanted variations, produce saturated spots also.

We consider the identification of the housekeeping genes with intensities within the linear range, as proposed by the PCA(line) routine, to be a reliable and robust source for normalization.

The linearity is the basis of the stability of the approach with respect to the parameter p—it is sufficient to detect a small subset of S to identify uniquely the normalization line. Conversely, a larger set of genes along this line will not impede the calculation of the normalization parameters. Still, in order to obtain meaningful histograms of the number of genes in each segment, we recommend that p initially be selected to contain on average at least 10 genes per segment. The condition for linearity naturally excludes genes with saturated expression levels and it thus contributes significantly to reducing the interference of these typically large signals in the normalization process.

Conditions (1) and (2) (see Introduction) are instrumental for the successful performance of the referenced normalization procedures. However, in single-channel arrays, such as the Affymetrix platform and radiolabeled filter arrays, it is a common phenomenon that the detected number of up-regulated genes is larger than the number of the down-regulated ones. This is due to the fact that the signals of genes expressed at low levels and undergoing down-regulation are close to or below the background level, and, therefore, their change is either undetected or deemed statistically insignificant. When these conditions hold, as in the case of the simulated data in Figure 3(a), PCA will be successful in determining the normalization factors with the following advantages, as compared with the other referenced techniques:

- It provides an objective measure through the magnitude of the first eigenvalue of how 'tightly' the data are distributed along the first PC.
- It simultaneously determines normalizing coefficients for the entire dataset. A common approach for normalization of multiple experiments is to choose one array as the baseline and to apply normalization (Golub *et al*., 1999). In order to avoid the lack of symmetry of this procedure, the baseline is computed frequently as the average gene expression profile (Tusher *et al*., 2001). This is achieved naturally with PCA as the first PC is an approximation of the 'average' array in the dataset.
- Viewing the entire set of multiple array data simultaneously allows proper down-weighing of the 'noise' genes, which, during individual comparisons, may affect strongly the calculation of the normalization coefficients.

The advantages of PCA are underscored in the LD example, where a single PCA step applied to the entire dataset estimates normalization coefficients that are almost identical to the ones determined by the pair-wise LR procedures, using only well measured genes in each pair [Fig. 6(a)].

The PCA(line) procedure, besides having the above listed general advantages of PCA, can also deal successfully with situations where conditions (1) and (2) do not apply. In the simulated datasets, the PCA(line) results are closest to the true values as judged by the relative mean-square errors from the three procedures tried. Visual inspection of the LR and PCA(line) normalization lines in the graph shown in Figure 6(b) suggests that this is also true for the Affymetrix data. In addition, it eliminates the need for using a baseline array, which, as shown by Bolstad *et al*. (2003), has a clear disadvantage relative to the complete data methods for normalization such as the one proposed here.

In conclusion, the proposed normalization procedure improves significantly the accuracy and precision of the measured gene expression levels. Such procedures will become even more relevant with further refinement and standardization of the microarray technology.

ACKNOWLEDGEMENTS

The authors would like to thank Dr S. Litwin for reviewing the manuscript critically and for his suggestions for improving it further. The authors would also like to thank Dr P. Tamayo for his helpful discussion regarding the Affymetrix array dataset. The work described in this report was supported by funds provided through NIH grant R29-CA73676 to C.P., P01-CA41078 to T.R.B., P50-CA83638 (PI: R. Ozols) and a Guzik Foundation Award to C.P. and R.S. C.P. is a Liz Tilberis Scholar of the OCRF, Inc.

REFERENCES

Alter,O., Brown,P.O. and Botstein,D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci., USA*, **97**, 10101–10106.

Alter,O., Brown,P.O. and Botstein,D. (2003) Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proc. Natl Acad. Sci., USA*, **100**, 3351–3356.

Anderson,T.W. (1971) *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

Bolstad,B.M., Irizarry,R.A., Astrand,M. and Speed,T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Bowtell,D.D. (1999) Options available—from start to finish—for obtaining expression data by microarray. *Nat. Genet.*, **21**, 25–32.

Butte,A.J., Dzau,V.J. and Glueck,S.B. (2001) Further defining housekeeping, or 'maintenance,' genes Focus on 'A compendium of gene expression in normal human tissues'. *Physiol. Genomics*, **7**, 95–96.

Debouck,C. and Goodfellow,P.N. (1999) DNA microarrays in drug discovery and development. *Nat. Genet.*, **21**, 48–50.

Duggan,D.J., Bittner,M., Chen,Y., Meltzer,P. and Trent,J.M. (1999) Expression profiling using cDNA microarrays. *Nat. Genet.*, **21**, 10–14.

Eisen,M.B. and Brown,P.O. (1999) DNA arrays for analysis of gene expression. *Methods Enzymol.*, **303**, 179–205.

Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A., Bloomfield,C.D. and Lander,E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Hartemink,A., Gifford,D., Jaakola,T. and Young,R. (2001) Maximum likelihood estimation of optimal scaling factors for expression array normalization. *Proc. SPIE*, **4266**, 132–140.

Hoffmann,R., Seidl,T. and Dugas,M. (2002) Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. *Genome Biol.*, **3**, RESEARCH0033.

Kepler,T.B., Crosby,L. and Morgan,K.T. (2002) Normalization and analysis of DNA microarray data by self-consistency and local regression. *Genome Biol.*, **3**, RESEARCH0037.

Kerr,M.K., Martin,M. and Churchill,G.A. (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, **7**, 819–837.

Lander,E.S. (1999) Array of hope. *Nat. Genet.*, **21**, 3–4.

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

Misra,J., Schmitt,W., Hwang,D., Hsiao,L.L., Gullans,S. and Stephanopoulos,G. (2002) Interactive exploration of microarray gene expression patterns in a reduced dimensional space. *Genome Res.*, **12**, 1112–1120.

Moon,R.C., Thompson,H.J., Becci,P.J., Grubbs,C.J., Gander,R.J., Newton,D.L., Smith,J.M., Phillips,S.L., Henderson,W.R., Mullen,L.T., Brown,C.C. and Sporn,M.B. (1979) N-(4-hydroxyphenyl)retinamide, a new retinoid for prevention of breast cancer in the rat. *Cancer Res.*, **39**, 1339–1346.

Nielsen,T.O., West,R.B., Linn,S.C., Alter,O., Knowling,M.A., O'Connell,J.X., Zhu,S., Fero,M., Sherlock,G., Pollack,J.R. *et al.* (2002) Molecular characterisation of soft tissue tumours: a gene expression study. *Lancet*, **359**, 1301–1307.

Park,T., Yi,S.G., Kang,S.H., Lee,S., Lee,Y.S. and Simon,R. (2003) Evaluation of normalization methods for microarray Data. *BMC Bioinformatics*, **4**, 33.

Patriotis,P.C., Querec,T.D., Gruver,B.N., Brown,T.R. and Patriotis,C. (2001) ArrayExplorer, a program in visual basic for robust and accurate filter cDNA array analysis. *Biotechniques*, **31**, 862–872.

Peterson,L.E. (2003) Partitioning large-sample microarray-based gene expression profiles using principal components analysis. *Comput. Methods Programs Biomed.*, **70**, 107–119.

Quackenbush,J. (2001) Computational analysis of microarray data. *Nat. Rev. Genet.*, **2**, 418–427.

Sapir,M. and Churchill,G.A. (2000). Published: The Jackson Laboratory **Poster**.

Schadt,E.E., Li,C., Ellis,B. and Wong,W.H. (2001) Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J. Cell Biochem. Suppl.*, **37**(suppl.), 120–125.

Schadt,E.E., Li,C., Su,C. and Wong,W.H. (2000) Analyzing high-density oligonucleotide gene expression array data. *J. Cell Biochem.*, **80**, 192–202.

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

Selvey,S., Thompson,E.W., Matthaei,K., Lea,R.A., Irving,M.G. and Griffiths,L.R. (2001) Beta-actin—an unsuitable internal control for RT–PCR. *Mol. Cell Probes*, **15**, 307–311.

Stoyanova,R. and Brown,T.R. (2002) NMR spectral quantitation by principal component analysis. III. A generalized procedure for determination of lineshape variations. *J. Magn. Reson.*, **154**, 163–175.

Tusher,V.G., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci., USA*, **98**, 5116–5121.

Yang,Y.H., Dudoit,S., Luu,P., Lin,D.M., Peng,V., Ngai,J. and Speed,T.P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.

Zien,A., Aigner,T., Zimmer,R. and Lengauer,T. (2001) Centralization: a new method for the normalization of gene expression data. *Bioinformatics*, **17** (Suppl. 1), S323–S331.

APPENDIX 1: ALGORITHM DESCRIPTION

(1) Construct the data matrix $\mathbf{D}(i, j)$, where

$i = 1, \ldots, n$ (n—total number of genes on each array),

$j = 1, \ldots, m$ (m—total number of arrays in the dataset).

(2) (Optional) thresholding of the data:

(2.1) Set the values in $\mathbf{D}$ smaller than a given value (e.g. 200 a.u. for the Clontech data) to 0.

(2.2) Remove from $\mathbf{D}$ genes with 0 intensities in at least one array, resulting in a new data matrix $\mathbf{D}'(n' \times m)$, where $n' \leq n$.

(3) PCA of $\mathbf{D}$ (here and in the rest of the text $\mathbf{D}$ should be substituted by $\mathbf{D}'$ in the case of thresholding, as well as n by n').

(3.1) Calculate $\mathbf{C}$—the covariance matrix of $\mathbf{D}$:

$$\mathbf{C} = \frac{1}{n-1}\mathbf{D}^{\mathbf{T}}\mathbf{D},$$

where $\mathbf{D}^{\mathbf{T}}$ denotes the transpose matrix of $\mathbf{D}$.

(3.2) Calculate eigenvectors $\mathbf{Q}$ and eigenvalues $\mathbf{\Lambda}$ of the covariance matrix $\mathbf{C}$, i.e.:

$$\mathbf{CQ} = \mathbf{Q\Lambda}$$

The rows in $\mathbf{Q}$ are the PCs $\vec{P}_1, \vec{P}_2, \ldots, \vec{P}_m$.

(3.3) Calculate the scores $R = \mathbf{D}\,\mathrm{P}^{\mathrm{T}}$.

(4) Let R_1^i and R_2^i be the scores of the i-th gene along $\vec{P}_1$ and $\vec{P}_2$.

(4.1) Disregard genes for which $R_2^i = 0$.

(4.2) Calculate the angle $\varphi_i, i = 1, \ldots, n$ (in radians), between $\vec{P}_2$ and the vector with coordinates (R_1^i, R_2^i), as follows:

$$\varphi_i = \begin{cases} 2\pi + \arctan(R_1^i/R_2^i), \\ \quad \text{if } R_1^i \le 0 \text{ and } R_2^i > 0, \\ \arctan(R_1^i/R_2^i) \\ \quad \text{if } R_1^i > 0 \text{ and } R_2^i > 0, \\ \pi + \arctan(R_1^i/R_2^i) \\ \quad \text{if } R_1^i > 0 \text{ and } R_2^i < 0, \end{cases} \qquad i = 1, \ldots, n.$$

(5) Segment the part of the plane defined by the first 2 PCs in p partitions.

(5.1) Determine the segment $\theta = \max(\varphi_i) - \min(\varphi_i)$

(5.2) Determine a step $\delta = \theta/p$

(5.3) Define the subset of genes s_k in each of the p segments, defined as

$$s_k \in [(k-1)\delta \min(\varphi_i), k\delta \min(\varphi_i)],$$
$$k = 1, \ldots, p.$$

(6) Determine the subset of housekeeping genes **S**.

(6.1) Determine the number of genes θ_k in each subset s_k.

(6.2) Estimate the mean $M(\theta_k)$, and variance, $V(\theta_k)$, of the distribution of θ_k.

(6.3) Evaluate if

$$\theta_k > M(\theta_k) + \mu\sqrt{V(\theta_k)}$$

holds for any k. μ is a cut-off parameter, which can be set to 1.96 if a normal distribution of θ_k is assumed [see body of the paper, Equation (2)].

If none of the segments satisfies the condition it means that either none of the genes can serve as a housekeeping gene ($\mathbf{S} \equiv \emptyset$) or all genes in the dataset can be assumed to be housekeeping genes ($\mathbf{S} \equiv \mathbf{D}$). Then the loadings of $\vec{P}_1$ (3.2) may be used as normalizing factors.

(6.4) The expression levels of the genes in each array should be divided by these loadings.

End of the Procedure

(6.5) Let Z denote the set of these segments that satisfy the condition in 6.3. If for a certain q, $\zeta_q \in Z$, then

(6.5.1) If $\zeta_{q+1} \notin Z$, then

(6.5.1.1) If there are no other qs, for which $\zeta_q \in Z$, then proceed as in 6.4.

(6.5.1.2) Conversely, proceed as in 6.5.

(6.5.2) If $\zeta_{q+1} \in Z$, then the genes in these two segments are assumed to be housekeeping genes; $\mathbf{S} \equiv s_q \cup s_{q+1}$. Add to S the genes of any consecutive segments that belong to Z.

(6.5.2.1) Apply PCA (3.2) to the gene expression levels in **S**. The loadings of $\vec{P}_1$ can be used as normalizing factors. The expression levels of the genes in each array should be divided by these loadings.

End of the Procedure

APPENDIX 2: SIMULATED DATASET

Let g_{i1} be the gene intensity of the i-th gene in the first array ($i = 1, 2, \ldots, 500$). The corresponding intensities in the second array in SD1 were generated as follows.

$$\left| \begin{array}{ll} g_{i2} = q_{12} * \min[\alpha_{\text{up}}\, g_{i1}, \beta_{\text{up}}] & i = 1, \ldots, 200, \\ g_{i2} = q_{12} * \max[\alpha_{\text{down}}\, g_{i1}, \beta_{\text{down}}] & i = 201, \ldots, 350, \\ g_{i2} = q_{12} * g_{i1} & i = 351, \ldots, 500, \end{array} \right. \tag{A.1}$$

where $q_{12} = 1.2$, and the αs and βs are random numbers within the following intervals:

$$\alpha_{\text{up}} = (1, 10],$$
$$\beta_{\text{up}} = (g_{i2}, g_{\max}], \quad \text{where } g_{\max} = 80\,000,$$
$$\alpha_{\text{down}} = (0, 1/10],$$
$$\beta_{\text{down}} = (g_{\min}, g_{i2}], \quad \text{where } g_{\min} = 0.$$

APPENDIX 3: SIMULATED DATASET

Let g_{ij} be the gene intensity of the i-th gene in the j-th array ($i = 1, 2, \ldots, 500$; $j = 1, 2, \ldots, 7$). Equation (A.1) describes the generation of the data in SD2 (q_{12} substituted correspondingly with q_{1j}, randomly generated scaling parameters between 0.3 and 3), derived from the intensities of the genes in the first array, where α_{up}^j and α_{down}^j are consistent with a simulated gradual increase in fold of changes between 1.5 and 4.5 with an increment of 0.5, both for up- and down-regulated genes. Formally,

$$\begin{array}{ll} \alpha_{\text{up}}^j = (1, 1 + j * \text{step}], & \\ \alpha_{\text{down}}^j = (0, 1/(1 + j * \text{step})], & j = 1, \ldots, 7 \end{array}$$

where step $= 0.5$.