



## Analysis of high density expression microarrays with signed-rank call algorithms

W.-m. Liu\*, R. Mei, X. Di, T. B. Ryder, E. Hubbell, S. Dee, T. A. Webster, C. A. Harrington†, M.-h. Ho, J. Baid and S. P. Smeekeens

Applied Research and Product Development, Affymetrix, Inc., 3380 Central Expressway, Santa Clara, CA 95051, USA

Received on November 12, 2001; revised on March 4, 2002; accepted on May 21, 2002

### ABSTRACT

**Motivation:** We consider the detection of expressed genes and the comparison of them in different experiments with the high-density oligonucleotide microarrays. The results are summarized as the detection calls and comparison calls, and they should be robust against data outliers over a wide target concentration range. It is also helpful to provide parameters that can be adjusted by the user to balance specificity and sensitivity under various experimental conditions.

**Results:** We present rank-based algorithms for making detection and comparison calls on expression microarrays. The detection call algorithm utilizes the discrimination scores. The comparison call algorithm utilizes intensity differences. Both algorithms are based on Wilcoxon's signed-rank test. Several parameters in the algorithms can be adjusted by the user to alter levels of specificity and sensitivity. The algorithms were developed and analyzed using spiked-in genes arrayed in a Latin square format. In the call process,  $p$ -values are calculated to give a confidence level for the pertinent hypotheses. For comparison calls made between two arrays, two primary normalization factors are defined. To overcome the difficulty that constant normalization factors do not fit all probe sets, we perturb these primary normalization factors and make increasing or decreasing calls only if all resulting  $p$ -values fall within a defined critical region. Our algorithms also automatically handle scanner saturation.

**Availability:** These algorithms are available commercially as part of the MAS 5.0 software package.

**Contact:** wei-min.liu@affymetrix.com

### INTRODUCTION

High density oligonucleotide microarrays are powerful tools to study gene expression, genotypes, and gene

mutations (Fodor *et al.*, 1993; Lockhart *et al.*, 1996; Mei *et al.*, 2000; Li and Wong, 2001). Analysis of the large amounts of data produced by these microarrays, however, requires robust computer algorithms. While heuristic approaches have been used very effectively in the past, it is usually difficult to adjust parameters with these methods to allow the user to alter levels of sensitivity and specificity often essential to obtain optimal biological results. Statistical methods have also been used to analyze microarray expression data. For example, Callow *et al.* (2000) have successfully applied  $p$ -values obtained from the  $t$ -test to study expression profiling in high-density lipoprotein (HDL) deficient mice. However, the  $t$ -test is a parametric method based on the assumption of normal distribution of data, making this approach sensitive to outliers. In contrast, nonparametric methods are usually less sensitive to outliers. Chen *et al.* (1997) applied the nonparametric Mann–Whitney test for image segmentation of cDNA arrays. Jin *et al.* (2001) also used the Mann–Whitney test on fold-change results of multiple microarray experiments. We have reported previously on nonparametric statistical tests for detection calls (also known as absolute calls) for limited data sets (Liu *et al.*, 2001). In this paper, we further develop these concepts using Wilcoxon's signed-rank test (Wilcoxon, 1945; Hollander and Wolfe, 1999) for making both detection and comparison calls. In addition, since a single constant normalization factor does not fit all genes on the microarrays, we define and perturb two normalization factors and make increasing or decreasing calls only if all of the  $p$ -values obtained from this approach fall within a defined critical region. To provide a highly controlled data set for the development and testing of these algorithms, we used a series of genes spiked-in at known concentrations and arranged in a Latin square format (Box *et al.*, 1978). These algorithms, as well as those providing robust estimation of expression values described in the accompanying paper by Hubbell *et al.*

\*To whom correspondence should be addressed.

†Current address: Mail Code CR145, Oregon Health Sciences University, Portland, OR 97201

(2002) are contained in the new Affymetrix software package, MAS 5.0.

## EXPERIMENTAL DESIGN AND METHODS

We cloned 112 yeast genes and 14 human genes using the TOPO TA cloning kit (Invitrogen). They were labeled and fragmented with the method described by Wodicka *et al.* (1997). Each of the labeled genes were pooled into groups and diluted to concentrations of 0, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512 and 1024 pM. In every microarray experiment, 14 groups of genes in 14 different concentrations were hybridized to the microarray in the presence of a complex background of expressed human genome (30 Mb) and several control genes. For this Latin square design, we perform 14 groups of experiments. The concentrations of the 14 *in vitro* transcript (IVT) groups in the first experiments are 0, 0.25, 0.5, ..., 1024 pM, their concentrations in the second experiments are 0.25, 0.5, ..., 1024, 0 pM, and so on. To reduce the saturation effect, we use low voltage of photomultiplier tube (PMT).

## ALGORITHMS

### Detection calls

Detection calls are used to determine whether the transcript of a gene is detected (present) or undetected (absent). On high-density expression microarrays, a gene is usually interrogated using probes that either perfectly match the sequence in a segment of the target gene (PM probes), or contain a single mismatched nucleotide in the middle position of the corresponding perfectly matched probe (MM probes). The intensity difference of perfect and mismatch probe cells are usually used to make detection calls.

Several different methods have been used to make detection calls (Lockhart *et al.*, 1996; Liu *et al.*, 2001). We begin by defining the discrimination score that can be calculated directly from raw intensity data. The discrimination score of the  $i$ th probe pair in a unit (also known as a probe set) is  $R_i = (PM_i - MM_i)/(PM_i + MM_i)$ , where  $PM_i$  is the intensity of the  $i$ th perfect match cell, and  $MM_i$  is the intensity of the  $i$ th mismatch cell. We see that the discrimination score is a relative measure, therefore if the whole microarray intensity is rescaled by a constant, the discrimination score remains unchanged. We remark in passing that probe pairs of a probe set may be at distant locations on a microarray to minimize the influence of a local defect of a microarray, but the perfect match cell and mismatch cell of a probe pair are always next to each other. Therefore, in the difference of their intensities, i.e. in the numerator of the above ratio, the location-dependent background intensities can be canceled.

We use the one-sided Wilcoxon's signed-rank test to ob-

tain a  $p$ -value for the null hypothesis  $H_0 : \text{median}(R_i) = \tau$  versus the alternative hypothesis  $H_1 : \text{median}(R_i) > \tau$ . Here,  $\tau$  is defined as a small nonnegative number that can be altered by the user to balance experimental specificity and sensitivity. Choosing the proper threshold value of  $\tau$  is important since we observe that the discrimination score of a probe pair can be a small positive number even if the corresponding gene is not present in the hybridization solution. While the origin of this phenomenon is worth additional study, we do not consider it further here. To reduce false detected calls, we chose the default value  $\tau = 0.015$  because it falls between the medians of discrimination scores for transcripts with concentrations of 0 and 0.25 pM for both the human and yeast data.

We next set two significance levels called  $\alpha_1$  and  $\alpha_2$  that serve as the cutoffs of  $p$ -values for detection calls. Specifically, if we let  $\alpha_1$  and  $\alpha_2$  be two small positive numbers such that  $0 < \alpha_1 < \alpha_2 < 0.5$ , we make detected calls for  $p < \alpha_1$ , undetected calls for  $p \geq \alpha_2$ , and marginally detected calls for  $\alpha_1 \leq p < \alpha_2$ . The significance levels,  $\alpha_1$  and  $\alpha_2$ , can also be altered by the user to adjust sensitivity and specificity. In MAS 5.0, the default values are  $\alpha_1 = 0.04$  and  $\alpha_2 = 0.06$  for 15 to 20 probe pairs per probe set. These values in combination with the default value of  $\tau$  result in fewer false detected calls than MAS 4 for the data sets that we studied.

As an example of the performance of MAS 4 and MAS 5.0 in making present calls, and the effect that changing the adjustable parameters has on this performance, we begin with the default setting for parameters  $\tau$  and  $\alpha$ . Under these conditions, MAS 5 gives no false detected calls, while MAS 4 yields 2% of false detected calls. Both algorithms also call genes present with 100% accuracy at target concentrations at and above 4 pM. In contrast, MAS 5 gives fewer present calls than MAS 4.0 at target concentrations between 0.25 and 2 pM. If we maintain  $\tau = 0.015$  and adjust the parameters to  $\alpha_1 = 0.15$  and  $\alpha_2 = 0.17$ , MAS 5.0 yields the same false positive rate as MAS 4.0, namely, 2%. However, MAS 5.0 now calls 100% of the genes present at or above 2 pM, as well as making higher percentages of present calls than MAS 4.0 at concentrations between 0.25 to 2 pM.

In scanning microarray images, saturation can occur when the pixel brightness exceeds the response range of the scanner. The cell intensity used in affymetrix microarrays is the 75th percentile of intensities of the inner pixels in a cell. Thus, if some pixels in a cell are saturated, the cell intensity can be very close to the maximal brightness of the scanner. For example, if the maximal brightness of a scanner is 46 109, we can consider a cell to be saturated if its intensity is larger than or equal to 46 000. Saturated probe pairs cause troubles in using discrimination scores. For instance, if both PM and MM cells are saturated, the discrimination score is zero,

but it also means that the target binds strongly to both PM and MM probes. To deal with this problem, we set forth a series of rules in our algorithm. If one or more mismatch cells in a unit are saturated, we exclude them from further computation. If only the perfect match cell is saturated, and the mismatch cell is not, we still include the probe pair in the detection call computation. If all mismatch cells in a unit are saturated, we make a detected call. To understand this handling, we consider the opposite situation. That is, if a target is absent, the cross hybridization may cause high intensity or saturation of a mismatch cell. In this case, however, it is very unlikely that the cross hybridization can make all mismatch probes saturated.

### Comparison calls

In many applications, we need to evaluate gene expression changes in two different experiments, e.g. stimulated versus unstimulated cells, or normal versus cancerous tissues. This can be done with the comparison call algorithm using two microarrays. Typically, one microarray is designated the baseline and the other the experimental. We define the possible results to be that the studied genes are found to be increasing, marginally increasing, marginally decreasing, decreasing, or exhibit no change at all.

Our comparison call algorithm includes the following steps. We first exclude saturated probe pairs and calculate the quantities on the baseline and experimental arrays. We then calculate the two primary normalization factors for the baseline and experimental array pair, perturb the primary normalization factors and apply them to form the quantities of three Wilcoxon's signed-rank tests for every probe set. Next, we calculate the  $p$ -values of the three Wilcoxon's signed-rank tests, and form the critical  $p$ -value from the three  $p$ -values. Based on two significance levels that are either constant or intensity dependent, we make the comparison calls.

Since saturated intensities can provide inaccurate information, we exclude a probe pair from further calculation of our comparison call algorithm if its PM or MM cell is saturated on either array. If all probe pairs of a probe set are so excluded, then there is no comparison call can be made and this information is output as no call.

To make comparison calls, we use the differences between PM and MM intensities, as well as the differences between PM intensities and background. When the false increasing or decreasing rates are adjusted to be the same, using both of these quantities yields higher rates of true increasing or decreasing calls than using one of them alone.

The background observed with microarrays consists of fluorescence intensity resulting from a variety of factors, including nonspecific binding of labeled target, stain, and incidentally fluorescent species. The background can also vary at different locations within a microarray.

Currently, we divide the microarray into  $N \times N$  zones (default  $N = 4$ ), where the background in a zone is the average of lowest 2% of probe set cell intensities within that zone. To avoid the discontinuity of background on the boundary of zones, we now use a smoothing method in MAS 5. If we assume the background of a zone is at its center, we can denote the background of zone  $(i, j)$  by  $B(i, j)$ , and the coordinates of the center of the zone by  $(X(i, j), Y(i, j))$  ( $i, j = 1, \dots, N$ ). If the coordinates of a cell are  $(x, y)$ , we calculate the background at this cell as a weighted average:  $b(x, y) = \sum_{i=1}^N \sum_{j=1}^N w(i, j) B(i, j) / \sum_{i=1}^N \sum_{j=1}^N w(i, j)$ , where  $w(i, j) = 1 / [(x - X(i, j))^2 + (y - Y(i, j))^2 + 100]$ . Other interpolation methods such as bilinear interpolation work as well. But the method we implemented is smoother.

If there are  $n$  unsaturated probe pairs in a unit, we use two  $n$ -dimensional vectors  $u^{(e)}$  and  $v^{(e)}$  for the 'experimental' array and two  $n$ -dimensional vectors  $u^{(b)}$  and  $v^{(b)}$  for the 'baseline' array. The superscripts  $(e)$  and  $(b)$  denote quantities in the experimental and baseline arrays, respectively. The components of these vectors are  $u_i^{(e)} = PM_i^{(e)} - MM_i^{(e)}$ ,  $v_i^{(e)} = PM_i^{(e)} - B_i^{(e)}$  and  $u_i^{(b)} = PM_i^{(b)} - MM_i^{(b)}$ ,  $v_i^{(b)} = PM_i^{(b)} - B_i^{(b)}$  ( $i = 1, \dots, n$ ).  $B_i$  is the background calculated at the  $i$ th perfect match cell with the formula for  $b(x, y)$ . We omit the coordinates  $(x, y)$  for clarity.

In order to bring the averages of those vectors from the two microarrays to the same level, we use two primary normalization factors: one for  $(PM - MM)$  and the other for  $PM - B$ . There are many different ways to calculate normalization factors, including the use of the intensities on the whole array, or the intensities of control genes. For a general-purpose expression algorithm, we use trimmed means of average quantities of all units. Users also have the option to select specific units for normalization. We use  $m_i = \text{mean}(PM_{ij} - MM_{ij}, j = 1, \dots, n_i)$  and  $s_i = \text{std}(PM_{ij} - MM_{ij}, j = 1, \dots, n_i)$  to denote the sample mean and sample standard deviation of the intensity differences between PM and MM cells in the  $i$ th probe set, where  $n_i$  is the number of unsaturated probe pairs in that probe set. Moreover, we let  $m'_i$  be the average of  $PM_{ij} - MM_{ij}$  within  $[m_i - 3s_i, m_i + 3s_i]$ . We denote the trimmed means of  $m'_i$  between the second and 98th percentiles for the experimental and baseline arrays as  $T^{(e)}$  and  $T^{(b)}$ , respectively, and define the primary normalization factor for  $(PM - MM)$  as  $f = T^{(b)} / T^{(e)}$ . While in most cases, the trimmed means are positive, they can be negative in the rare case where very few genes are present in the hybridization solution. Under those conditions, we replace the negative  $m'_i$  by zero and repeat the above procedure, and the trimmed means should be nonnegative. If one of them is zero, then all units on an array have nonpositive averages, and hence the quality of

this microarray or the assay should be further examined. However, we never experienced this situation and always obtained two positive trimmed means and can form the normalization factor  $f$  without difficulty.

To obtain the primary normalization factor for the differences between PM intensities and the background, we calculate the sample mean  $\mu_i = \text{mean}(PM_{ij}, j = 1, \dots, n_i)$  and sample standard deviation  $\sigma_i = \text{std}(PM_{ij}, j = 1, \dots, n_i)$ . We denote the average of  $PM_{ij}$  in the interval  $[\mu_i - 3\sigma_i, \mu_i + 3\sigma_i]$  by  $\mu'_i$ . Furthermore, we let  $P^{(e)}$  and  $P^{(b)}$  be the respective trimmed means of  $\mu'_i$  between the 2nd and 98th percentiles in the experimental and baseline arrays, and define the normalization factor for  $(PM - B)$  as  $g = P^{(b)}/P^{(e)}$ . We also attempted to use  $PM_{ij} - B_{ij}$  for the normalization factor computation. However, we found that its performance is not as good as using  $PM_{ij}$  only (data not shown).

It is unlikely that two constant primary normalization factors can fit all units on a microarray, and if used, could lead to incorrect comparison calls. To overcome this difficulty, we perturb the primary normalization factors up and down. Specifically, for  $(PM - MM)$ , we use three normalization factors  $f_1 = f * d$ ,  $f_2 = f$ , and  $f_3 = f/d$ ; and for  $(PM - B)$  we use  $g_1 = g * d$ ,  $g_2 = g$ , and  $g_3 = g/d$ . The perturbation coefficient  $d$  is defined as a number larger than or equal to 1. In MAS 5, the default value of  $d$  is 1.1. An example of the influence of changing this value is shown below.

To combine the two vectors of  $PM - MM$  and  $PM - B$  and to make use of the perturbed normalization factors, we form three  $(2n)$ -dimensional vectors  $V_k (k = 1, 2, 3)$  with components  $V_{ki} = f_k \mathbf{u}_i^{(e)} - \mathbf{u}_i^{(b)}$  and  $V_{k,n+i} = c(g_k \mathbf{v}_i^{(e)} - \mathbf{v}_i^{(b)})$  ( $i = 1, \dots, n$ ), where the constant  $c$  is set to a default value 0.2. Then we calculate the  $p$ -values,  $p_k$ , for three one-sided signed-rank tests of the null hypotheses  $H_0^{(k)} : \text{median}(V_{ki}, i = 1, \dots, 2n) = 0$  versus the alternative hypotheses  $H_1^{(k)} : \text{median}(V_{ki}, i = 1, \dots, 2n) > 0$  ( $k = 1, 2, 3$ ). Note that we do not remove outliers at this stage since the nonparametric rank-based method is naturally less sensitive to outliers. To make comparison calls, we set two significance levels  $\gamma_1$  and  $\gamma_2$  that satisfy the condition  $0 < \gamma_1 < \gamma_2 < 0.5$ . If  $p_k < \gamma_1$  is true for  $k = 1, 2, 3$ , we make an increasing call. Similarly, if  $p_k > 1 - \gamma_1$  is true for  $k = 1, 2, 3$ , we make a decreasing call. If we cannot make an increasing call, but  $p_k < \gamma_2$  is true for  $k = 1, 2, 3$ , we make a marginally increasing call. Similarly, if we cannot make a decreasing call, but  $p_k > 1 - \gamma_2$  is true for  $k = 1, 2, 3$ , we make a marginally decreasing call. In the absence of these conditions, we make a no change call. The default values used in MAS 5 are  $\gamma_1 = 0.0025$  and  $\gamma_2 = 0.003$ .

To avoid reporting three  $p$ -values for every probe set, we report only the critical  $p$ -value. It is defined as  $p_c =$

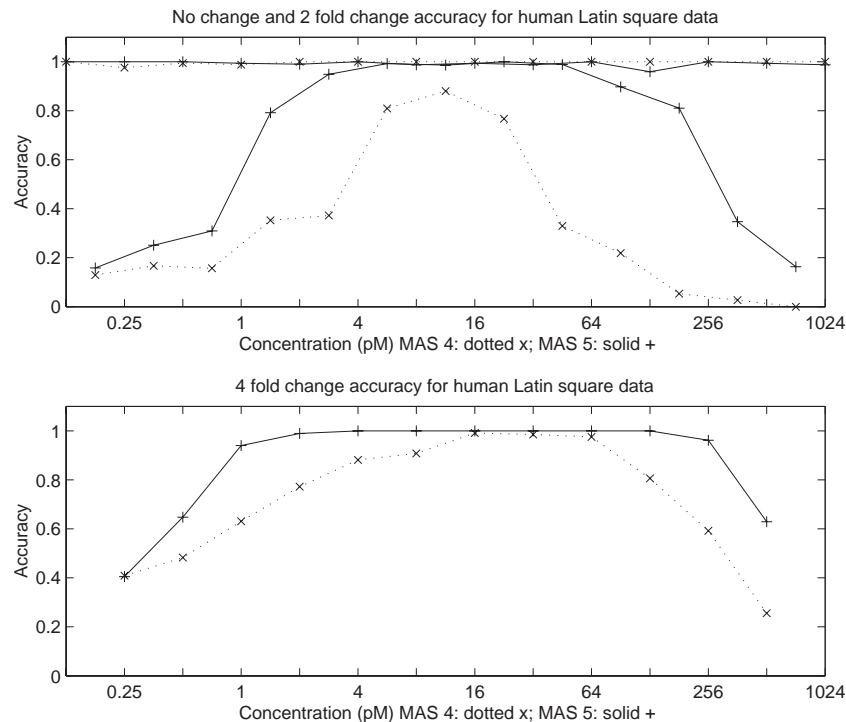
$\max(p_1, p_2, p_3)$  in the case where  $p_1 < 0.5$ ,  $p_2 < 0.5$  and  $p_3 < 0.5$ , and as  $p_c = \min(p_1, p_2, p_3)$  in the case where  $p_1 > 0.5$ ,  $p_2 > 0.5$  and  $p_3 > 0.5$ . In all other cases, we define  $p_c = 0.5$ . With these critical  $p$ -values, users can choose different significance levels at which to mine data without repeated running of software.

Note that we combine  $PM - MM$  and  $PM - B$  to form the  $(2n)$ -dimensional vectors  $V_k$ . While we may also consider  $PM - MM$  and  $PM - B$  separately and then combine their  $p$ -values to make calls, we found during the development of our algorithm that the  $(2n)$ -dimensional vectors  $V_k$  often give higher accuracy, especially when a primary normalization factor is not very accurate. The parameter  $c$  is used to combine and adjust the contributions of the vectors  $PM - MM$  and  $PM - B$ . Its default value  $c = 0.2$  is chosen to obtain robust results using the Latin square data. We tried  $c = 1, 0.5, 0.2, 0.1$ , and  $0.05$ . The value  $c = 0.2$  gives the best results and the differences are small.

We now give an example to explain the role of the perturbation coefficient,  $d$ . For the probe set named 1024\_at with the same concentration 256 pM on both the human genome experimental and baseline arrays, we obtain a  $p$ -value 0.000467 without perturbation. Using this  $p$ -value and the default significance levels  $\gamma_1 = 0.0025$  and  $\gamma_2 = 0.003$ , we would make an erroneous increase call for a target that is known to be present at the same concentration on the two arrays. In setting the perturbation coefficient to  $d = 1.1$ , however, we obtain two additional  $p$ -values, 0.0776 and  $3 \times 10^{-6}$ . The critical  $p$ -value, 0.0776 (the  $p$ -value that is closest to 0.5), yields the correct call of no change.

The default parameters in our algorithm were chosen to give small errors of no change calls. For example, the overall error rate of no change calls for the 12 626 genes on the Hg\_u95 arrays was found to be 0.83% using the replicates in the human data set of 59 microarrays from three different lots. The rate was 0.7% for the human genes spiked in the Latin square experiments. However, the tradeoff for keeping the no change call error rate low is that it increases the error rates of true increasing or decreasing calls. We can balance the tradeoff by reducing the perturbation coefficient  $d$ . Namely, when  $d$  is reduced from 1.1 to 1.08, the no change error rate increases to 1.11% for all 12 626 genes, and 1.26% for the genes spiked in the Latin square experiments. However, the accuracy of two-fold increasing calls are increased from 72.5% to 80.7% for the concentration change from 1 to 2 pM.

Figure 1 shows the accuracy of comparison calls of MAS 4 and MAS 5 for the human Latin square data at various concentrations. We use the default parameters for both algorithms. Marginally increasing and marginally decreasing calls are counted as no change calls. From this, we see that both algorithms have similar accuracy for no



**Fig. 1.** The accuracy of comparison calls of MAS 4 and MAS 5 for the human Latin square data at various concentrations. The default parameters are used for both algorithms. Marginal calls are counted as no change calls. The results of MAS 4 are represented by the dotted curves with marks  $\times$ . The results of MAS 5 are represented by the solid curves with marks  $+$ . (a) No change and generalized two-fold change comparison calls. No change curves are close to the horizontal line  $y = 1$ . The results of comparisons of 0 pM versus 0.25 pM are plotted in the middle of these two values. (b) Generalized four-fold change comparison calls. The results of comparisons of 0 pM versus 0.5 pM are plotted at the middle of these two values.

change comparisons, and MAS 5 demonstrates improved accuracy for two-fold and four-fold change comparisons.

## DISCUSSION

We have described here the algorithms based on Wilcoxon's signed-rank test to make detection and comparison calls on expression microarrays. The accompanying paper Hubbell *et al.* (2002) describes the computation of microarray signals based on Tukey's biweight estimation. These algorithms are based on robust statistical methods, and they together comprise the expression analysis of our newly released microarray software package, MAS 5.0.

There have been recent discussions about the value of approaches that do not use mismatch cells. In both a previous study (Liu *et al.*, 2001) and the accompanying paper by Hubbell *et al.* (2002), we show that using the signal from mismatch cells helps raise the sensitivity of detection calls at low target concentrations. We also notice that using mismatch cells improves the sensitivity of comparison calls at certain concentrations.

Although theoretically the data for signed-rank test should be independent, real data may not be completely independent. To assess this, we used the nonparametric Spearman's test on the discrimination scores of two groups of 12 replicates of human data and found that only 7% of probe pairs have  $p$ -values in Spearman's test below the significance level of 0.01. This indicates that the discrimination scores of most probe pairs can be considered as independent.

Another theoretical requirement of the signed-rank test is that the distribution of data under the null hypothesis be symmetric around the constant on the right hand side of the testing equality or inequality (i.e.,  $\tau$  for detection call algorithm and 0 for comparison call algorithm). This is why we need to introduce a nonzero  $\tau$  instead of using  $\tau = 0$  for detection calls, and why we need to perturb the normalization factors for comparison calls. We can find an ideal value  $\tau'$  such that the discrimination scores in a probe set are symmetric around  $\tau'$  when the target is absent. Of course, this ideal value  $\tau'$  varies from probe set to probe set. For a constant  $\tau$ , there is no guarantee that the distribution of data is symmetric around it when

the target is absent. Consequently, to reduce the false detected calls, we choose a default  $\tau$  value that is a little higher than the ideal  $\tau'$  for many probe sets. Thus, the  $p$ -values calculated in our algorithm are exact only under the condition that the distribution of discrimination scores is symmetric around  $\tau$  when the target is absent. Similarly, for comparison calls, we can find two ideal normalization factors  $f'$  and  $g'$  such that the quantities used in the signed-rank test are symmetrically distributed around 0 when the target has the same concentrations in the baseline and the experimental arrays. The primary normalization factors usually do not satisfy this condition. Therefore, we perturb them. When the ideal normalization factors are in the range of perturbation, it is reasonable to use the three  $p$ -values to make comparison calls.

Further analysis of our methods for detection calls and comparison calls indicates that they are robust. For example, we can multiply a uniformly distributed random factor in the interval [0.9, 1.1] to the intensities of perfect match cells of our yeast data. In doing so, we find that the average difference of absolute call errors is 0.97%, and the average difference of no change comparison call errors is 0.22%. Similarly, the average difference of two-fold comparison call errors is 3.34%, and the average difference of four-fold comparison call errors is 1.2%.

Our algorithms provide the user with parameters that can be systematically adjusted to alter the stringency of calls. For instance, increasing the threshold  $\tau$  of detection calls can reduce the false detected rate; increasing perturbation coefficient  $d$  can reduce the errors of no change calls. Thus, they are very useful for comparisons of different microarray designs, probe selection rules, process controls and scanners. We can adjust the threshold  $\tau$  so that the false detected rates of two treatments are the same, and then compare their true detected rates at different concentrations. We can also adjust the perturbation coefficient  $d$  so that the error rates of no change calls of two treatments are the same, and then compare their accuracy of two-fold change or other increasing or decreasing calls.

The default parameters are set for 15 to 20 probe pairs per probe set. For fewer probe pairs, we suggest increasing the significance levels. For example, for 11 probe pairs per probe set, we recommend the default values  $\alpha_1 = 0.05$ ,  $\alpha_2 = 0.065$ ,  $\gamma_1 = 0.0045$  and  $\gamma_2 = 0.006$ .

Here we comment on a theoretical benefit of the signed-rank test in comparison with the MAS 4 heuristic call algorithms that include the sign test. For detection calls, the sign test applied to 11 probe pairs can only produce 12 different results, while the signed-rank test can give  $2^{11} = 2048$  different results. For our comparison calls with 11 probe pairs, the signed-rank test can yield  $2^{22} = 4194304$  different results. Therefore, for the reduced number of probe pairs, it is relatively easy to adjust our algorithm to balance sensitivity and specificity.

Hubbell *et al.* (2002) applied a logarithmic transformation to intensities and to differences between PM intensities and contrast levels. Using a logarithmic transformation helps obtaining linear responses while the intensities vary significantly. For the detection call algorithm, the most difficult part is in the low-intensity range. For the comparison call algorithm, the most challenging part is to compare similar intensities whose differences are not significantly large. Therefore, we do not use the logarithmic transformation in the call algorithms.

Li and Wong (2001) consider the probe effect using an additive error model. Hubbell *et al.* (2002) use a multiplicative error model of the probe effect to extract expression values. Our focus here is to present robust call algorithms for both stand alone and matched pairs of expression microarrays. Any algorithm that gives expression values can be used to give calls if a threshold of the values can be reasonably set. For example, it is possible to use the log ratios proposed in Hubbell *et al.* (2002) to make comparison calls, which can be done conveniently with a threshold. The call algorithms described here allow users to analyze their data in a different way. Moreover, adjustment of only one threshold parameter is often difficult when attempting to balance the sensitivity and specificity. For example, Li and Wong (2001) provide several methods to make comparison calls, where one of them uses both the threshold of fold change and the threshold of difference of expression indices. In the future, we plan to modify our call algorithms to utilize probe sequence effects.

## CONCLUSION

The algorithms presented here provide  $p$ -values for both detection and comparison calls. Since the signed-rank test is a nonparametric test method, the results are robust. Using discrimination scores helps to reach a reasonable balance of sensitivity and specificity for detection calls. Using the perturbation of normalization factors improves the accuracy of comparison calls. Since the parameters of our algorithms are easy to adjust, our algorithms are also useful for comparisons of different microarray designs, probe selection rules, process controls and hardware.

## ACKNOWLEDGMENTS

We thank Tarif Awad, Dan Bartell, Jon Campman, Simon Cawley, Alex Cheung, Fred Christians, John Chung, Glenn Deng, Helin Dong, Mark Durst, Joy Fang, Luis Jevons, Michael Jordan, Paul Kaplan, Gang Lu, Garry Myada, Jacques Retief, Vivian Reyes, Mei-Mei Shen, Conrad Sheppy, Dan Shulda, David Smith, John Sowatsky, Gene Tanimoto, Kai Wu, Geoffrey Yang and Steve Zanki for helpful discussion, and/or providing data. We also thank the referees for their comments and suggestions.

## REFERENCES

- Box, G.P.E., Hunter, J.S. and Hunter, W.G. (1978) *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*. Wiley, New York.
- Callow, M.J., Dudoit, S., Gong, E.L., Speed, T.P. and Rubin, E.M. (2000) Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Res.*, **10**, 2022–2029.
- Chen, Y., Dougherty, E.R. and Bittner, M.L. (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Optics*, **2**, 364–374.
- Fodor, S.P.A., Rava, R.P., Huang, X.C., Pease, A.C., Holmes, C.P. and Adams, C.L. (1993) Multiplexed biochemical assays with biological chips. *Nature*, **364**, 555–556.
- Hollander, M. and Wolfe, D.A. (1999) *Nonparametric Statistical Methods*, Second edition, Wiley, New York, pp. 36–49.
- Hubbell, E., Liu, W.-m. and Mei, R. (2002) Robust estimators for expression analysis. *Bioinformatics*, this issue.
- Jin, H., Yang, R., Awad, T.A., Wang, F., Li, W., Williams, S.-P., Ogasawara, A., Shimada, B., Williams, P.M., de Feo, G. and Paoni, N.F. (2001) Effects of early angiotensin-converting enzyme inhibition on cardiac gene expression after acute myocardial infarction. *Circulation*, **103**, 736–742.
- 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.
- Liu, W.M., Mei, R., Bartell, D.M., Di, X., Webster, T.A. and Ryder, T. (2001) Rank-based algorithms for analysis of microarrays. In Bittner, M.L., Chen, Y., Dorsel, A.N. and Dougherty, E.R. (eds), *Microarrays: Optical Technology and Informatics, Proceedings SPIE*. **4266**, pp. 56–67.
- Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. and Brown, E.L. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.
- Mei, R., Galipeau, P.C., Prass, C., Berno, A., Ghandour, G., Patil, N., Wolff, R.K., Chee, M., Reid, B. and Lockhart, D.J. (2000) Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. *Genome Res.*, **10**, 1126–1137.
- Wilcoxon, F. (1945) Individual comparisons by ranking methods. *Biometrics*, **1**, 80–83.
- Wodicka, L., Dong, H., Mittmann, M., Ho, M.-H. and Lockhart, D.J. (1997) Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat. Biotechnol.*, **15**, 1359–1372.