

Quality Weighted Mean and T-test in Microarray Analysis Lead to Improved Accuracy in Gene Expression Measurements and Reduced Type I and II Errors in Differential Expression Detection

Shouguo Gao¹⁺, Shuang Jia²⁺, Martin Hessner² and Xujing Wang^{1*}

¹Department of Physics & the Comprehensive Diabetes Center, University of Alabama at Birmingham, 1300 University Blvd, Birmingham, AL 35294, USA

²The Max McGee National Research Center for Juvenile Diabetes & the Human and Molecular Genetics Center, The Medical College of Wisconsin and Children's Hospital of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226, USA

*Corresponding author: Xujing Wang, Tel: 001-205-934-8186;

Fax: 001-205-934-8042; E-mail: xujingw@uab.edu

+The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

Received December 09, 2008; Accepted December 22, 2008; Published December 26, 2008

Citation: Shouguo G, Shuang J, Martin H, Xujing W (2008) Quality Weighted Mean and T-test in Microarray Analysis Lead to Improved Accuracy in Gene Expression Measurements and Reduced Type I and II Errors in Differential Expression Detection. J Comput Sci Syst Biol 1: 041-049. doi:10.4172/jcsb.1000003

Copyright: © 2008 Shouguo G, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Previously we have reported a microarray image processing and data analysis package *Matarray*, where quality scores are defined for every spot that reflect the reliability and variability of the data acquired from each spot. In this article we present a new development in *Matarray*, where the quality scores are incorporated as weights in the statistical evaluation and data mining of microarray data. With this approach filtering of poor quality data is automatically achieved through the reduction in their weights, thereby eliminating the need to manually flag or remove bad data points, as well as the problem of missing values. More significantly, utilizing a set of control clones spiked in at known input ratios ranging from 1:30 to 30:1, we find that the quality-weighted statistics leads to more accurate gene expression measurements and more sensitive detection of their changes with significantly lower type II error rates. Further, we have applied the quality-weighted clustering to a time-course microarray data set, and find that the new algorithm improves grouping accuracy. In summary, incorporating quantitative quality measure of microarray data as weight in complex data analysis leads to improved reliability and convenience. In addition it provides a practical way to deal with the missing value issue in establishing automatic statistical tests.

Keywords: Microarray; Quality score; Weighted algorithms; Accurate expression measurement

Availability: *Matarray* is freely available from xujingw@uab.edu; or sjia@mcw.edu

Introduction

Microarray technology, which allows genome-level profiling of gene expression changes, has become a widely used genetic tool. However, the technology is prone to noise,

necessitating data filtering (Wang et al., 2001). Often data from poor-quality spots, such as spots with insufficient resolution from noise, or insufficient immobilized probe

material (Hessner et al., 2003) are removed from further analysis, since it is impossible to derive reliable measurements from them. Data filtering creates the problem of missing values (Troyanskaya et al., 2001), which makes combining data from replicates, and the down stream statistical evaluation and data mining cumbersome.

It has been argued that measurement reliability-weighted methods can improve performance of the significant analysis of microarrays (Hughes et al., 2000; Fan et al., 2004). Fan, et al used a LOWESS method to obtain a weight which indicated the reliability of the measured log ratio in an array, then applied both weighted and ordinary t-test to examine the effect of MIF treatment on genes (Fan et al., 2004). Compared with ordinary t-test, the weighted t-test had a better ability to assess the effect of genes, and could identify more significant genes with smaller p -values. However, the weight defined in their paper may not fully capture the inherent variability in microarray measurements, since they only considered the variability brought by low intensity and print-tip effects, and their method was dependent on the number and characteristics of available replicated clones on array. A limited effort has also been made to improve clustering performance by incorporating the error variance information calculated from replicates. Intuitively, gene expression levels that show larger variations over repeated measurements should be assigned lower measurement reliability. For example, Yeung et al systematically evaluated several clustering algorithms that incorporated variability calculated from repeated measurements as weights, and showed that algorithms yield more accurate and more stable clusters (Yeung et al., 2003). However, these methods are all based on a variance calculation from replicates, which requires that an adequate number of replicate hybridizations be made in order to derive a reliable estimation of the variance. This would not always be practical because of the high cost of microarray experiments. In addition, such approaches are not sensitive to quality issues that affect all replicates equally.

Previously we have reported a microarray data acquisition and analysis software package *Matarray* (Wang et al., 2001; Wang et al., 2003). It processes microarray images and acquires gene expression measurements from every spot. A composite quality score q_{com} is defined for every spot according to the signal to noise ratio, spot size and variation, global and local noise distribution, and saturation for detection. If a pre-hybridization third dye (TD) image is also available, as is the case with our three color microarray platform (Hessner et al., 2003; Hessner et al., 2003) then a composite quality score q_{TD} will also be defined similarly according to the informa-

tion from the TD image. In either case, a final quality score q_f will be determined by $q_f = q_{com}$, or $q_f = q_{com} \times q_{TD}$ if TD image is available (Wang and Hessner 2005). Through numerous studies we have demonstrated that our quality metrics capture well the reliability of the data acquired, in the sense that gene expression measurements derived from spots with higher scores are much more accurate and less variable than those derived from spots with lower scores (Wang et al., 2001; Hessner et al., 2003; Wang et al., 2003; Wang and Hessner, 2005). We have also demonstrated the significance of having a quantitative measure of data quality for every spot, through the efficient data filtering and normalization procedures that they led to (Wang et al., 2001; Wang et al., 2003). In this paper, we present a new application to the statistical evaluation of microarray measurements, where the quality scores are utilized to define weights W_o . Using a set of control clones that were spiked in at known input ratios, we show that W_o -weighted mean and W_o -weighted t -test lead to more accurate gene expression measurements and more sensitive detection of gene expression changes. In addition, we introduce a quality-weighted clustering algorithm through the definition of W_o -weighted distance metric. We apply it to a large-scale time series microarray experiments and show that it allows more accurate discrimination of groupings of experimental conditions. In these algorithms filtering of poor-quality data is automatically achieved through their diminishing weights. There is no need to manually flag or remove them explicitly from the data matrix. Therefore the cumbersome data missing value problem is avoided.

Materials and Methods

Microarray Dataset and Processing

Data from 3 different microarray experiments were utilized to validate our quality-weighted algorithm: (1) Profiling of BioBreeding (BB) rat thymus. Gene expressions were compared between the thymus of diabetes prone DR lyp/lyp (referred to as DP) and diabetic resistant DR+/+ (referred to as DR) BB rats (Hessner et al., 2004) at day 40. This analysis utilized 4 animals from each strain, and 4 replicate hybridizations were performed for each animal pair, with 2 hybridizations reverse labeled to control for dye bias. During each hybridization, the labeling reactions of total thymus RNA were spiked with 4 *Arabidopsis in vitro* transcripts (cellulose synthase, chlorophyll a/b binding protein, ribulose-1,5-bisphosphate and triosphosphate isomerase) at known input ratios (30:1, 10:1, 5:1, and 1:1, respectively). Each of our rat arrays possessed 18, 20, 18 and 20 replicate spots corresponding to the 4 *Arabidopsis* clones respectively, giving rise to a total of 1216 data points.

These clones enabled an evaluation of our methods through the comparison of measured output ratios to the known input ratios. (2) Profiling of BB rat liver. Gene expressions in liver were compared between day 65 BB-DR rats and Wistar-Furth (WF) rats. In this experiment, 4 animals from each strain were sacrificed and equal amounts of purified total RNA from the animals of the same strain were pooled. The two pools were then compared in 6 replicate hybridizations, with 3 of them reverse labeled. The transcript abundance of 24 genes that exhibited differential expression (DE) were also measured using quantitative real time RT-PCR, which is generally considered a more quantitative platform than microarrays (Chuaqui et al., 2002). (3) Time course profiling of apoptosis progression in pancreatic islet β cells. Cells from a rat β cell line RIN-m5F were treated with a protein kinase C inhibitor staurosporine (Sanchez-Margalet et al., 1993) at a high dose of 1 μ M, and a low dose of 1nM for 2, 4, and 6 hours, and were compared for differential gene expressions. At each time point, 6 replicate hybridizations were performed, with 3 of them reverse labeled, totaling 18 hybridizations. Cell apoptosis status were confirmed using Annexin V/PI double staining method as described in (Wang et al., 2002), and apoptosis progression under high dose treatment along time was evident. At 2hr for example, the Annexin positive cells was about 20%. At 6 hr after drug treatment, the apoptosis progression has been established with at least 40% cells. In the low dose treatment, the apoptosis rate at any time point is not significantly different from the control sample at t=0.

All experiments were carried out using in-house rat cDNA microarrays that were fabricated using our three-color microarray platform (Hessner, 2003; Hessner et al., 2003). All hybridized arrays were processed using *Matarray*, and an overall quality score Q_f were defined for data from each microarray spot which reflect its quality (Wang et al., 2001; Wang et al., 2003; Wang et al., 2003). Briefly, from cyanine images, non-redundant factors affecting data quality were identified, individual quality scores as well as a composite score q_{com} was determined (Wang et al., 2001). From the prehybridization third dye image, a quality score q_{TD} was calculated similarly to measure the impact of imperfections in array fabrication on hybridization data quality (Hessner, 2003; Hessner et al., 2003; Wang et al., 2006). Together, a final overall assessment of data quality was given by $Q_f = q_{com} \times q_{TD}$ (Wang et al., 2006). Data quality and characteristics were evaluated utilizing the ratio- Q_f plot. In this analysis only spots with $Q_f > 0$ were retained for further analysis, and this comprised of more than 95% of the data.

Quality Weighted Mean and T-test

In statistics, it is known that utilizing the inverse error variance as weight in significance test can improve performance. Unfortunately, true error variance of microarray data is unknown in practice. It can be estimated from adequate number of replicates; however this would increase experimental cost (Tjaden, 2006). We have optimized our quality score definitions such that error variance monotonically decreases as quality score increases (Wang et al., 2001; Wang and Hessner, 2006). So improvement in statistical tests can be expected when the quality scores are incorporated as weights. Assuming that g_{ij} is the expression measurement for gene i in target sample j , Q_{ij} is the corresponding quality score, and there are N samples, we define the weighted mean and weighted standard error (SE) by:

$$\bar{g}_i^w = \frac{\sum_{j=1}^N g_{ij} W_{Q_{ij}}}{\sum_{j=1}^N W_{Q_{ij}}}$$

$$SE_i^w = \sqrt{\frac{\sum_{j=1}^N (g_{ij} - \bar{g}_i^w)^2 W_{Q_{ij}}}{(N-1) \sum_{j=1}^N W_{Q_{ij}}}} \quad (1)$$

Where weight is defined as $W_{Q_{ij}} = Q_{ij}$. Replacing the mean and SE in $t_i^n = \frac{\bar{g}_i - g_{i0}}{SE_i}$ by their weighted counterparts, we define the quality-weighted t -test by:

$$t_i^w = \frac{\bar{g}_i^w - g_{i0}}{SE_i^w} \quad (2)$$

where g_{i0} is the expected value of g_{ij} , The two-sample weighted t -test can be defined accordingly. The row-mean of $\bar{Q}_i = \sum_{j=1}^N Q_{ij} / N$ are used to indicate the overall quality for each gene. If $W_{ij}=0$ (ie $Q_{ij} = 0$) for all j , or all j but one, then the arithmetic mean will be calculated for g_{ij} and the p -value will be set to 1. In *Matarray* normally all data points with $\bar{Q}_i = 0$ are filtered (Wang et al., 2003; Wang and Hessner, 2005). Here in this new approach, data filtering is built-in and the contribution from bad data points is automatically minimized through reduction in their weights, eliminating the need of physical removal of substandard data. Furthermore, it automatically gives the best data the highest weights; and therefore has the potential of generating more sensitive and accurate measurements.

As our sample size is not big (Allison et al., 2006), we have also implemented penalized weighted t -test (Comander, et al., 2004) to reduce false positives resulted from coincidental small SE:

$$t_i^p = \frac{\bar{g}_i^w - g_{i0}}{SE_i^w + s_0} \quad (3)$$

where s_0 is a small constant. In this work, we choose s_0 to be the 75th percentile of SE in expression measurements for all genes. When sample size is small, penalized t -test usually performs better than normal t -test (Comander, et al., 2004).

Quality-weighted Clustering

We define quality-weighted similarity measures that weight expression values with quality scores such that contributions from low quality data points are reduced. Using average-linkage hierarchical clustering of samples as an example, for each pair of sample a and sample b , we define the weighted distance metric by:

$$d_{ab} = \sqrt{\frac{\sum_{i=1}^n (g_{ia} - g_{ib})^2 (W_{Q_{ia}} + W_{Q_{ib}})}{\sum_{i=1}^n (W_{Q_{ia}} + W_{Q_{ib}})}} \quad (4)$$

Where n is the number of genes used for cluster analysis. The distance metric for gene pairs, or other types of similarity measurements can be defined similarly. In this study after calculating distance matrix of all pairs of samples, average-linkage hierarchical clustering algorithm was applied to cluster samples in data set 3.

Implementation

All algorithms are implemented in our in-house software *Matarray* (Wang et al., 2001; Wang et al., 2003). It is freely available, with documentation, examples and a tutorial.

Results

Spiked-in Control Clones and RT-PCR Demonstrate that Weighted Mean is More Accurate

We have found that the weighted mean generates more accurate gene expression measurements than do the arithmetic means. In figure 1A, comparison between the measured and the actual input ratios are given for the *Arabidopsis* control clones, where a highly linear relationship ($R^2 > 0.99$; $p < 0.0001$) is observed, with the exception of the last data point (spiked-in ratio 30:1, Cy5: Cy3). Overall, the weighted means exhibited less compression in measurements than the arithmetic means (slope=0.888 versus 0.824, $p < 0.01$). The random sampling method also proved that weighted means possess significantly higher slope (data not shown). A closer examination of the spots contributing to the last data point revealed significant intensity saturation in one dye channel (Cy5), which led to under-estimation of the fold difference between the two dye channels.

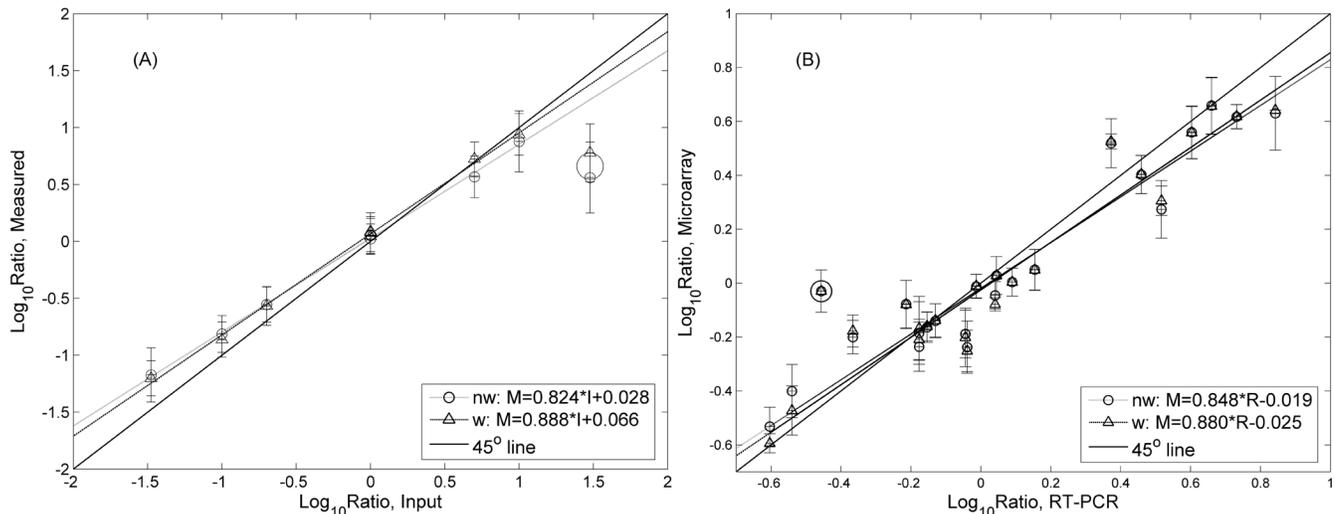


Figure 1: W_Q -weighted statistics leads to improved accuracy and sensitivity over the non-weighted approach. In both (A) and (B) symbols represent the measure data, and lines are the linear regressions. (A) The measured ratios (M) are compared with the actual input ratios (I) for the spiked-in *Arabidopsis* clones. Both weighted mean (w) and non-weighted mean (nw) ratio measurements exhibit good agreement with the input ratios, with $R^2 > 0.99$; $p < 0.0001$. There is less compression in the weighted mean ratios, as the slope of the linear regression is closer to 1. The last data point (circled) corresponding to a 30:1 input ratio is excluded in the linear regression. (B) Measurements by microarrays (M) are compared with those by RT-PCR (R) for 24 genes in the rat liver experiment. Highly linear relationships are observed between the two platforms ($R^2 > 0.96$; $p < 0.0001$). One data point (circled) where $W_Q = 0$ for all replicates is excluded in the linear regression.

In figure 1B the microarray measurements were compared to the RT-PCR results for the 24 genes in the rat liver experiments, and again an overall good agreement was observed. After removal of one gene (circled) where $Q_f = 0$ in all replicates, a highly linear relationship ($R^2 > 0.96$; $p < 0.0001$) was observed for the remaining 23 genes. Again the weighted means exhibited an improved agreement with RT-PCR over the non-weighted means (slope=0.880 versus 0.848 for the arithmetic means, $p \approx 0.32$).

The weighted mean ratios agreed better than ordinary arithmetic means with the true input ratios, as well as with the ratio values measured by RT-PCR. In addition, reduction of ratio compression should help to improve sensitivity in identifying significant genes, since severe ratio compression can push some truly significant genes into background noise. The following subsection shows that this is indeed the case.

Weighted T -test is more Sensitive in Detecting Differential Expressions

To evaluate the performance of the weighted t -test, we compared the p -values derived using both weighted and non-weighted tests. We found that in general the weighted t -test allowed more genes to be detected with significant p -values. Using the rat thymus data set as an example, we have

plotted in figure 2A the p -values defined by W_Q -weighted t -test against those by normal t -test for all genes. Spots corresponding to the *Arabidopsis* clones were not included. The weighted t -test predicted more genes to be differentially expressed between the DP and DR rats. For example, the genes in the lower right quadrants are significant at $p = 0.1$ according to the weighted t -test, but not according to the non-weighted t -test. On the contrary, only a few genes were identified by non-weighted test, but have been missed by weighted t -test, see upper left part of figure 2A.

To further verify that this is due to more sensitive detection rather than to a higher false positive rate, we again turned to the *Arabidopsis* control clones. Each of our rat arrays possessed 76 spots corresponding to the 4 *Arabidopsis* clones. Therefore, this experiment generated a total of 152 *Arabidopsis* data points in each sample comparison from the two directions of labeling. 40 of them corresponding to the clone spiked in at 1:1 ratio served as (non-DE) negative controls. The remaining 112 corresponding to an input ratio that was significantly different from 1 served as (DE) positive controls. The results were summarized in table 1. We found that the type II error (false positive rates) were comparable between the weighted and non-weighted t -tests. Specifically, at $p = 0.01$, 5 out of the 40 negative controls were significant according to non-weighted t -test and 7 according to the weighted test. On the other

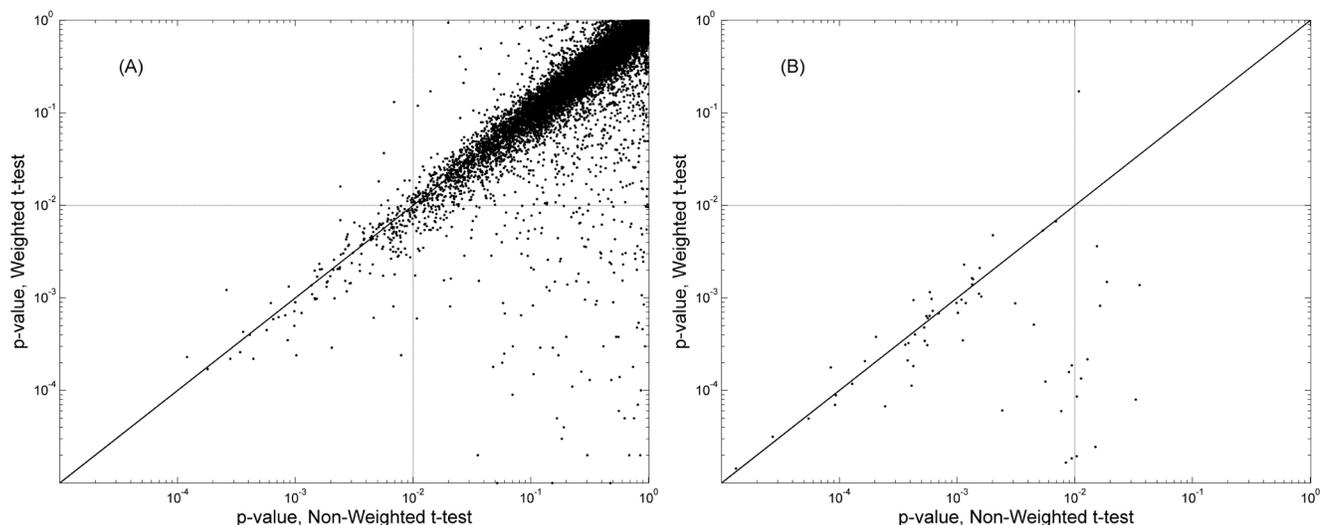


Figure 2: The weighted t -test leads to identification of more genes with significant changes. (A) The p -values determined using weighted and non-weighted t -tests are compared for all genes in the rat thymus data set. More genes have significant p -values according to the weighted t -test (lower right quadrant). (B) The p -values are compared for all spiked-in *Arabidopsis* spots corresponding to the input ratios of 30:1, 10:1 and 5:1 (excluding those corresponding to 1:1). The measured ratios of these spots should all be significantly different from 1. The weighted t -test is able to detect more of them as evidenced by the more significant p -values.

	<i>t</i> -test	weighted <i>t</i> -test	weighted penalized <i>t</i> -test
Type I error	18	1	2
Type II error	5	7	4

Table 1: Comparison of type I errors and type II errors produced by different algorithms at $p = 0.01$. The number of available positive and negative controls is 81 and 40 respectively.

hand, the type I error rate was significantly reduced in the weighted approach (Figure 2B). 81 of 112 positive controls had $\bar{Q}_i > 0$, and the weighted *t*-test was able to detect all but one (Type I error rate: 1.2%). In contrast, non-weighted *t*-test missed 18, leading to a type I error rate of 22.2%, significantly higher than that of the weighted test ($p < 0.0001$). This result indicates indirectly that those data points in the lower right quadrant of figure 2A are likely to be true positives. Since the microarray technology is often utilized as an explorative tool to be followed by conformational measures, more sensitive detection is highly desirable.

To further reduce the false positive rate we have introduced penalized weighted *t*-test given by equation 3. In figure 3A, the Gaussian fitting to the *p*-value distribution calculated according to the weighted penalized and non-weighted *t*-tests were plotted, for the positive and negative control clones of the rat thymus analysis. The vertical line

corresponds to the user-specified *p*-value cutoff which is used to balance the tradeoff between type I and type II error probabilities. This figure clearly demonstrates that both type I and type II error rate were significantly reduced within commonly used threshold *p*-values for significance [0.05-0.001]. Therefore, in comparison to the non-weighted *t*-test, the power of weighted method to detect gene expression changes was also significant enhanced. On the other hand, penalized *t*-test was expected to lower the type II error rate, indeed, at $p = 0.01$, the type II error according to the 40 negative controls has reduced from 7 to 4. The type I error rate in the penalized approach was 2.5% (2 out of 81), not significantly different from that of the non-penalized weighted *t*-test ($p > 0.5$, χ^2 -test). Therefore introducing penalized test further reduced type II error rate as expected without compromising sensitivity of detection.

We have then used receiver operating characteristic

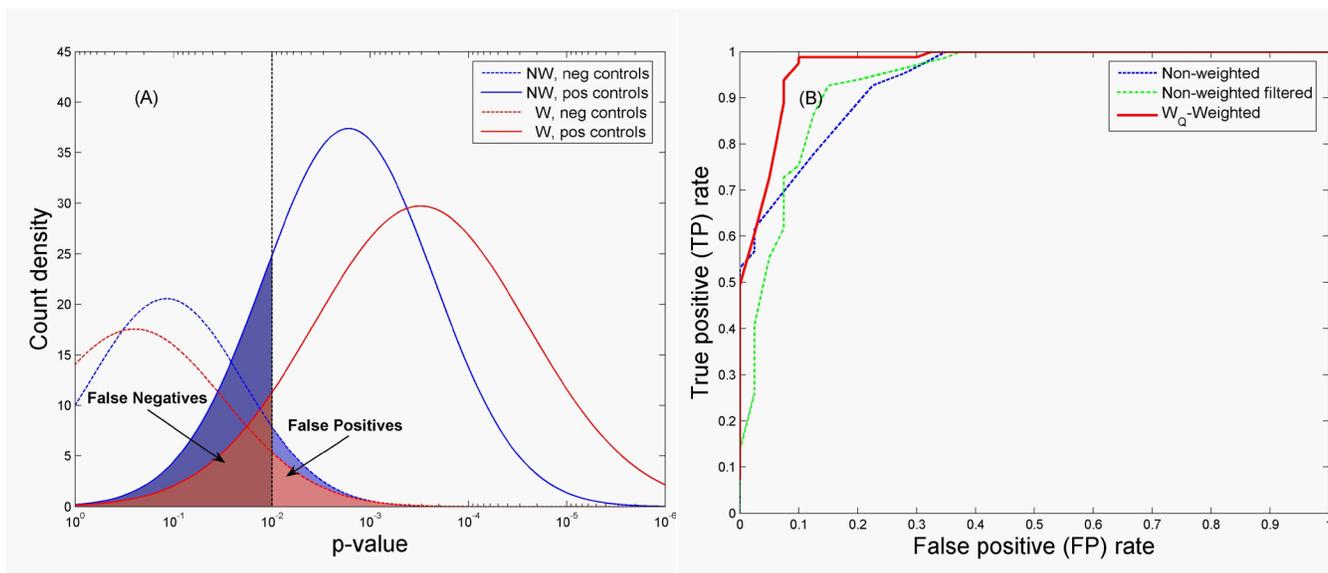


Figure 3: Comparison between the weighted penalized *t*-test and normal *t*-test demonstrates the advantage of the former. In an experiment that profiled rat thymus, we have 40 data points corresponding to a control clone spiked in at known Cy3:Cy5 input ratios of 1:1, which serve as negative controls, and 112 data points corresponding to spike-in ratios of 5:1, 10:1 and 30:1, which serve as positive controls. (A) The relative frequencies are plotted against the *p*-values derived using the penalized weighted (W) and non-weighted (NW) *t*-tests. The areas under positive controls curves on the left of vertical line correspond to false negatives, and the areas under negative controls curves on the right of vertical line correspond to false positives. (B) ROC plot showing true positive rate against false positive rate.

(ROC) curves to quantitatively compare the results between penalized weighted and non-weighted *t*-test. In figure 3B, true positive rate was plotted against false positive rate for the Arabidopsis control clone data set. This underlines the obvious fact that weighted method can lead to higher true positive rate at the same false positive level. AUC (Area Under Curve) of ROC is a standard performance measure of algorithm, the AUC for non-weighted curve was 0.94 and AUC for W_q -weighted curve was 0.97, so our weighted algorithm can more sensitively identify significant genes without increasing false positive rates.

For comparison we also added non-weighted results calculated by normal *t*-test for the data in which bad points were manually flagged and removed. Interestingly, manually flagging method did not significantly improve over non-weighted no filtering method. Specifically at stringent type I error rates, manually filtering significantly lowered the sensitivity to detect differentially expressed genes. After carefully checking the data, we found that this was primarily due to the decrease of available replicates, almost all data points on the left end of the curve (FP rate<0.1) were calculated from only 2-4 replicates retained after manual filtering. The number of replicates for these data points was below 5, the number found to be the minimally required in order to achieve reliable statistical inference (Allison et al., 2006). In summary, we have found that the W_q -weighted statistics allows more accurate and sensitive detection of gene expression changes. It allows efficient filtering of poor quality data, and is more convenient than the

manual flagging method.

Weighted Clustering Yields More Accurate Grouping

We have found that the weighted clustering generally leads to more sensitive detection of groupings among samples. Figure 4A-4B show the result of the average-linkage hierarchical clustering of samples from data set 3, which profiled gene expression changes in pancreatic islet β cells during apoptosis progression (single- and complete-linkage hierarchical clustering algorithm give very similar results).

Samples collected at the same time should be close to each other in the dendrograms. It is also reasonable to expect that experiments using the same labeling method to cluster together if all other conditions are identical. Therefore we expect that at the top level of the hierarchy, there are three clusters, each correspond to one of the three time points; at the next level, samples form 6 subgroups named by collection time and labeling methods: 2H-f (2H-forward labeling), 2H-r (2H-reverse labeling), 4H-f, 4H-r, 6H-f and 6H-r. Clearly, the weighted algorithm can discriminate the groupings among the samples much better than the non-weighted algorithm. Specifically, (1) weighted method could distinguish samples at 2h perfectly, and three forward labeling and reverse labeling experiments have been strictly divided into two subgroups; (2) for the samples at 4h, weighted algorithm put two experiments with forward labeling into one group with very high similarity, whilst the non-weighted

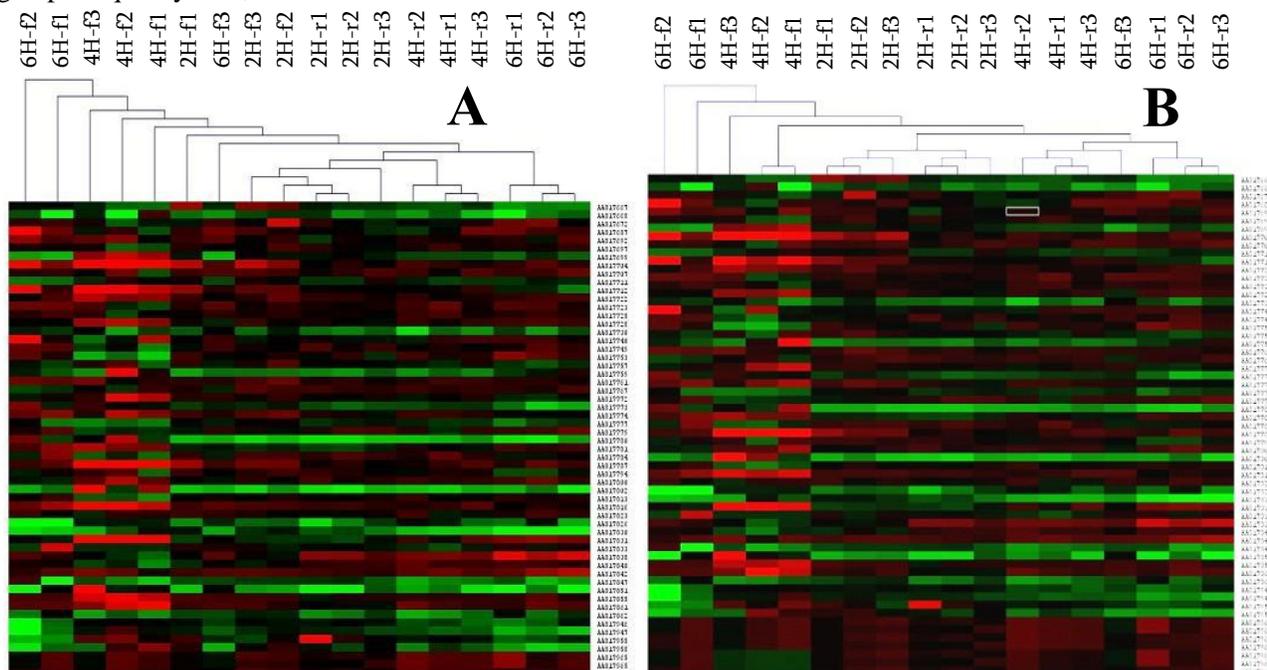


Figure 4: Weighted clustering can better recover sample relationships. Presented are the dendrograms of hierarchical clustering of the 18 arrays in data set 3 using non-weighted (A) and weighted (B) algorithms. On top are the hybridizations named by sample collection time and labeling methods (forward or reverse), the last number represents replicate index.

algorithm failed to group them correctly. On the other hand, all the sensible findings obtained by non-weighted algorithm have also been found by weighted approach, such as 6H-r group and 4H-r group.

Discussion

In this work we have extended our quality score definitions to the statistical evaluation of microarray data by introducing the W_q -weighted mean and W_q -weighted t -test. We have shown that the new approach leads to improved accuracy in gene expression measurements, and more sensitive detection of expression changes. Recently, we have further investigated the impact of such improvement on the biological interpretation of the data. We examined the results from ontological analysis of DE genes defined by the weighted or non-weighted tests, using OntoExpress (Draghici et al., 2003) and EASE (Hosack et al., 2003). We have found that the weighted t -test led to annotations with more focused, logical biological themes (data not shown). Our quality score weighted approach can be further extended to other statistical models, such as mixture models (Kauermann and Eilers, 2004; Newton et al., 2004). We have shown that weighted clustering algorithms incorporating quality scores performed better to group samples. Similarly, it can be applied to cluster genes of similar variation patterns, and will also likely lead to improved performance in identifying meaningful relationships between gene groups, so that more biological information can be extracted from microarray data.

Data from microarray experiments are usually in the form of large matrices of gene expression measurements or log ratios between the target samples and controls. Normally, each row corresponds to a gene and each column corresponds to a condition. Data filtering to remove low-quality elements, which is necessary in microarrays, results in missing values in the matrices. It is difficult to set up automatic statistical tests where the gene expression matrix is incomplete, and the sample size varies from gene to gene due to missing values. Many pattern finding methods, including principal component analysis and singular value decomposition need complete data sets. Clustering methods such as hierarchical clustering (Eisen et al., 1998) can handle missing values by ignoring them when calculating cluster distance, however, doing so can lead to spurious results (Oba et al., 2003). To deal with the missing values in a dataset, the most straightforward approach is to simply remove the whole row or column that contains missing values. This will not be practical for large data sets that profile multiple conditions, as there often too many genes possess missing values (Ouyang et al., 2004). Other

methods include replacing the missing values with zeros or row means. But they can often lead to high deviations from true values (Troyanskaya et al., 2001; Oba et al., 2003). More sophisticated imputation approaches that utilize the information from the whole data set to estimate the missing values have also been proposed, examples include methods that utilize measurements from other genes that have similar or correlated expression patterns (Troyanskaya et al., 2001; Bo et al., 2004); utilizing the principle components of the gene expression matrix (Troyanskaya et al., 2001; Oba et al., 2003); and model based approaches such as Gaussian mixture (Ouyang et al., 2004) and Bayesian (Oba et al., 2003; Zhou et al., 2003) models. These approaches often require high number of replicates. In addition, the performance of different algorithms varies, and the accuracy and robustness of the estimation often depend on data characteristics, including data size, data quality, correlation between data from different conditions, and experimental designs. There is no single algorithm that has been deemed the best under all conditions. These issues have added to the complexity of the already challenging microarray data analysis. In addition of improved data quality, our approach eliminates the need to manually flag or remove bad data points, and hence the missing value problem is avoided. The convenience will be more significant for large data sets where a great number of genes can be affected.

Acknowledgement

This work is supported in part by National Institute of Biomedical Imaging and Bioengineering under grant No. 1R01EB001421, the National Institute of Diabetes and Digestive and Kidney Diseases under grant No. R01DK080100, and the National Institute of Allergy and Infectious Diseases under grant No. 9R56AI078713-05A1. We thank Lisa Meyer for performing the experiments.

References

1. Allison DB, Cui X, Page GP, Sabripour M (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet* 7: 55-65. » [CrossRef](#) » [PubMed](#) » [Google Scholar](#)
2. Bo TH, Dysvik B, Jonassen I (2004) LSImpute: accurate estimation of missing values in microarray data with least squares methods. *Nucleic Acids Res* 32: e34. » [CrossRef](#) » [PubMed](#) » [Google Scholar](#)
3. Chuaqui RF, Bonner RF, Best CJ, Gillespie JW, Flaig MJ, et al. (2002) Post-analysis follow-up and validation of microarray experiments. *Nat Genet* 32: 509-14. » [CrossRef](#) » [PubMed](#) » [Google Scholar](#)
4. Comander J, Natarajan S, Gimbrone MA, Jr, García-

- Cardeña G (2004) Improving the statistical detection of regulated genes from microarray data using intensity-based variance estimation. *BMC Genomics* 5: 17. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
5. Draghici S, Khatri P, Shah A, Tainsky MA (2003) Assessing the functional bias of commercial microarrays using the onto-compare database. *Biotechniques* 55-61. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
6. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 95: 14863-8. » [Pubmed](#) » [Google Scholar](#)
7. Fan J, Tam P, Woude GV, Ren Y (2004) Normalization and analysis of cDNA microarrays using within-array replications applied to neuroblastoma cell response to a cytokine. *Proc Natl Acad Sci USA* 101: 1135-40. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
8. Hessner M, Wang X, Hulse K, Meyer L, Wu Y, et al. (2003) Three color cDNA microarrays: quantitative assessment through the use of Fluorescein-Labeled Probes. *Nucl Acids Res* 31: e14. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
9. Hessner MJ, Wang X, Khan S, Meyer L, Schlicht M, et al. (2003) Use of a three-color cDNA microarray platform to measure and control support-bound probe for improved data quality and reproducibility. *Nucl Acids Res* 31: e60. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
10. Hessner MJ, Wang X, Meyer L, Geoffrey R, Jia S, et al. (2004) Involvement of eotaxin, eosinophils, and pancreatic predisposition in development of type 1 diabetes mellitus in the BioBreeding rat. *J Immunol* 173: 6993-7002. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
11. Hosack DA, Dennis G, Jr, Sherman BT, Lane HC, Lempicki RA, et al. (2003) Identifying biological themes within lists of genes with EASE. *Genome Biology* 4: p4. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
12. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, et al. (2000) Functional discovery via a compendium of expression profiles. *Cell* 102: 109-26. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
13. Kauermann G, Eilers P (2004). Modeling microarray data using a threshold mixture model. *Biometrics* 60: 376-87. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
14. Newton MA, Noueiry A, Sarkar D, Ahlquist P (2004) Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* 5: 155-76. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
15. Oba S, Sato MA, Takemasa I, Monden M, Matsubara K, et al. (2003) A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics* 19: 2088-96. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
16. Ouyang M, Welsh WJ, Georgopoulos P (2004) Gaussian mixture clustering and imputation of microarray data. *Bioinformatics* 20: 917-23. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
17. Sanchez MV, Lucas M, Solano F, Goberna R (1993) Sensitivity of insulin-secreting RIN m5F cells to undergoing apoptosis by the protein kinase C inhibitor staurosporine. *Exp Cell Res* 209: 160-3. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
18. Tjaden B (2006) An approach for clustering gene expression data with error information. *BMC Bioinformatics* 7: 17. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
19. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, et al. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics* 17: 520-5. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
20. Wang X, Becker FF, Gascoyne PR (2002) Membrane dielectric changes indicate induced apoptosis in HL-60 cells more sensitively than surface phosphatidylserine expression or DNA fragmentation. *Biochim Biophys Acta* 1564: 412-20. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
21. Wang X, Ghosh S, Guo SW (2001) Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Research* 29: E75-82. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
22. Wang X, Hessner MJ (2006). Quantitative quality control of microarray experiments: toward accurate gene expression measurements. *Gene expression profiling by microarrays – clinical implications*. H. W. K., Cambridge.
23. Wang X, Hessner MJ, Wu Y, Pati N, Ghosh S (2003) Quantitative quality control in microarray experiments and the application in data filtering, normalization and false positive rate prediction. *Bioinformatics* 19: 1341-1347. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
24. Wang X, Jia S, Meyer L, Xiang B, Chen LY, et al. (2006) Accurate gene expression measurements by cDNA microarrays utilizing TDAV. *BMC Bioinformatics* 7:378. » [Google Scholar](#)
25. Wang X, Jiang N, Feng X, Xie Y, Tonellato PJ, et al. (2003) A Novel Approach For High Quality Microarray Processing Using Third-Dye Array Visualization Technology. *IEEE Trans. NanoBioscience* 2: 193-201. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
26. Yeung KY, Medvedovic M, Bumgarner RE (2003) Clustering gene-expression data with repeated measurements. *Genome Biol* 4: R34. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
27. Zhou X, Wang X, Dougherty ER (2003) Missing-value estimation using linear and non-linear regression with Bayesian gene selection. *Bioinformatics* 19: 2302-7. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)