BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 11 2010, pages 1431–1436
doi:10.1093/bioinformatics/btq163

Gene expression

Advance Access publication April 16, 2010

Novel statistical framework to identify differentially expressed genes allowing transcriptomic background differences

Zhi-Qiang Ling[1,2,†], Yi Wang[3,†], Kenichi Mukaisho[2], Takanori Hattori[2], Takeshi Tatsuta[4], Ming-Hua Ge[1], Li Jin[3], Wei-Min Mao[1,*] and Hiroyuki Sugihara[2,*]

[1]Zhejiang Cancer Research Institute, Zhejiang Province Cancer Hospital, Banshanqiao Guangji Road 38, Hangzhou 310022, China, [2]Department of Pathology, Shiga University of Medical Science, Otsu 520-2192, Japan, [3]MOE Key Laboratory of Contemporary Anthropology and Center for Evolutionary Biology, Institution of Life Sciences and Institutes of Biomedical Sciences, Fudan University, Shanghai 200433, China and [4]Department of Surgery, Shiga University of Medical Science, Otsu 520-2192, Japan

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Tests of differentially expressed genes (DEGs) from microarray experiments are based on the null hypothesis that genes that are irrelevant to the phenotype/stimulus are expressed equally in the target and control samples. However, this strict hypothesis is not always true, as there can be several transcriptomic background differences between target and control samples, including different cell/tissue types, different cell cycle stages and different biological donors. These differences lead to increased false positives, which have little biological/medical significance.

Result: In this article, we propose a statistical framework to identify DEGs between target and control samples from expression microarray data allowing transcriptomic background differences between these samples by introducing a modified null hypothesis that the gene expression background difference is normally distributed. We use an iterative procedure to perform robust estimation of the null hypothesis and identify DEGs as outliers. We evaluated our method using our own triplicate microarray experiment, followed by validations with reverse transcription–polymerase chain reaction (RT–PCR) and on the MicroArray Quality Control dataset. The evaluations suggest that our technique (i) results in less false positive and false negative results, as measured by the degree of agreement with RT–PCR of the same samples, (ii) can be applied to different microarray platforms and results in better reproducibility as measured by the degree of DEG identification concordance both intra- and inter-platforms and (iii) can be applied efficiently with only a few microarray replicates. Based on these evaluations, we propose that this method not only identifies more reliable and biologically/medically significant DEG, but also reduces the power-cost tradeoff problem in the microarray field.

Availability: Source code and binaries freely available for download at http://comonca.org.cn/fdca/resources/softwares/deg.zip

Contact: maowm1218@hotmail.com; sugihara@belle.shiga-med.ac.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on 14 November 2009; revised on 2 April 2010; accepted on 9 April 2010

1 INTRODUCTION

Microarray technology is widely used for gene expression analysis of biological samples that undergo different treatments or represent different phenotypic groups. The treated sample or the group of interest is called the ‘target’, and the other sample/group is called the ‘control’. One of the most important usages of such an experimental design is to identify genes that are differentially expressed between target and control. This is usually achieved by statistical tests using biological replicates (replicates using different biological cases) and technical replicates (replicates of a single case on multiple arrays) (Allison *et al.*, 2006).

Usually *t*-test or analysis of variance on the equality of mean microarray gene expression measurement between the target group and the control group are employed to identify differentially expressed gene (DEG) (Cui *et al.*, 2003). Fold change (FC) method is also used as a simple and empirical alternative. More sophisticated tests, such as shrinkage methods (Budhraja *et al.*, 2003; Cui *et al.*, 2005; Smyth *et al.*, 2005; Tusher *et al.*, 2001) are available and possess greater power. All of these methods are based on the null hypothesis that genes that are irrelevant to the phenotype/stimulus are expressed equally in the target and control samples. However, this strict hypothesis is not always true, as there can be several transcriptomic background differences between target and control, including different cell/tissue types, different cell cycle stages and different biological donors. These differences lead to increased false positives, which have no biological/medical significance, when traditional statistical tests are employed. For example, when the sample size is large, the testing of gene expression differences between brain tissue samples and blood samples will identify most genes as DEG. Following the traditional framework investigators will also exhaust their financial resources to maximize the sample size in order to gain greater power for the identification of smaller gene expression differences (Allison *et al.*, 2006). However, the small differences they seek may be simply explained by the aforementioned transcriptomic background differences, which are irrelevant to the phenotype or stimulus of interest.

To overcome the problems that can be encountered when using the traditional statistical testing framework, we propose a new approach to identify DEG in the presence of different transcriptomic backgrounds. We extend and broaden the traditional null hypothesis of no difference to normal distributed differences and identify

[*]To whom correspondence should be addressed.

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

DEG as outliers of the null hypothesis distribution. Interestingly, this extended/broadened null hypothesis can be tested easily and efficiently with few replicates. We evaluated our method using our own triplicate microarray experiment, which was further validated by reverse transcription–polymerase chain reaction (RT–PCR) and on the MicroArray Quality Control (MAQC) dataset (Shi *et al.*, 2006). This evaluation suggests that our method (i) results in fewer false positives and false negatives, as measured by the degree of RT–PCR concordance, (ii) can be applied to different microarray platforms and results in better reproducibility as measured by DEG identification concordance both within and between platforms and (iii) can be applied efficiently with few microarray replicates. Based on these evaluations, we hypothesize that the proposed method not only identifies more reliable and biologically/medically significant DEG, but also reduces the power-cost tradeoff problem in the microarray field.

2 MATERIALS AND METHODS

2.1 Cell culture

A human gastric cancer cell line, KATO-III, which was maintained as described previously (Ling *et al.*, 2007), was used in these experiments. Antibiotic and antimycotic drugs were not used in culture in order to avoid their possible effects on gene expression. Cells in the logarithmic growth phase were used. Peripheral blood lymphocytes were cultured by standard methods for 3 days after stimulation with phytohemagglutinin (PHA-P; Sigma, St Louis, MO, USA).

2.2 RNA preparation

Total cellular RNA was extracted from cultured cells using IsoGen (Nippon gene, Toyama, Japan), which is an acid guanidinium thiocyanate–phenol–chloroform method, according to the manufacturer's instructions. The cells were suspended in 400 μl IsoGen lysis buffer at room temperature, mixed with 80 μl chloroform and precipitated in isopropanol with 20 μg glycogen (Roche, Mannheim, Germany). The RNA pellet was resuspended in RNase-free water, and the concentration of RNA was measured by a spectrophotometer (GeneQuant pro, Biochrom Ltd, Cambridge, UK). Five micrograms of RNA were treated with 5 U of RNase-free DNase I (TaKaRa Bio Inc., Otsu, Japan) for 30 min for the removal of DNA contamination. To confirm the absence of residual DNA, the RNA samples were subjected to degenerate oligonucleotide primed–PCR amplification, as described below. As a result, no PCR product was detected in the samples that were examined. After phenol extraction and precipitation in isopropanol, the pellet was resuspended in 10 μl of RNase-free water.

2.3 RNA amplification with T7 RNA polymerase

To initiate first strand cDNA synthesis, 1 μl of 100 pmole/μl oligo(dT)24-T7 primer (TaKaRa) (5′-pGGCCAGTGAATTGTAATACGA CTCACTATAGGGAGGCGGTTTTTTTTTTT TTTTTTTTTTTT-3′) was added to 10 μl of a solution containing 5 μg total RNA. The solution was then incubated at 65°C for 10 min, chilled on ice and equilibrated at room temperature for 10 min. Then, the first and second cDNA strands were synthesized as described previously (Luo *et al.*, 1999; Kan *et al.*, 2001) with a modification of the incubation temperature from 42°C to 37°C during first strand synthesis. After phenol/chloroform extraction, the aqueous layer was collected and purified by isopropanol precipitation two times. The pellet was then resuspended in 8 μl of RNase-free water. The MEGAscript in vitro Transcription Kit (Ambion, Austin, TX, USA) was used for cRNA production with T7 RNA polymerase. After the addition of 2 μl of 10× reaction buffer, 2 μl each of 100 mM adenosine triphosphate, cytidine triphosphate, guanosine triohosphte and uridine triphosphate and 2 μl T7 RNA polymerase to 8 μl of cDNA, the solution was incubated at 37°C for 5 h. Then, 1 μl of RNase-free DNase I was added, and the solution was incubated at 37°C for 15 min. After mixing with 400 μl IsoGen lysis buffer at room temperature, chloroform extraction and isopropanol precipitation were performed. The cRNA pellet was then resuspended in 50 μl of RNase-free water.

2.4 Microarray analysis

The Human Cancer Chip Version 4.0 microarray (IntelliGene, TaKaRa) was used. This array is spotted with 886 cDNA fragments from human genes, including 588 that have been identified as cancer related. A complete list of the genes is available on the TaKaRa web site (http://bio.takara.co.jp/catalog/Catalog_d.asp?C_ID=C1219). cDNA microarray analysis with an array scanner (Affymetrix 428; Affymetrix, Santa Clara, CA, USA) and ImaGene 5.0 software (BioDiscovery, El Segundo, CA, USA) was carried out as per the manufacturer's instructions. Briefly, the probe was T7 amplified and cRNA labeled in the same way as in Ling's protocol (Ling *et al.*, 2007). Hybridization and washes were performed as previously described (Kan *et al.*, 2001) with some modifications. After heat denaturation, the probe was cooled at room temperature. Each post-hybridization wash was performed for 5 min. Expression profiling that incorporated dye swapped, with Cy5-labeled test cDNA and Cy3-labled reference cDNA, was used to check for non-specific reactions. The microarray analyses with normal labeling and dye swapped were each performed in triplicate. We adopted the global normalization by ImaGene. Up- and downregulation in array spots were defined by the T/R>2.0 and <0.5, provided that signal counts of T (Cy3) and R (Cy5) were >500.

2.5 RT–PCR analyses

All significantly DEGs and many randomly selected unchanged genes identified by Takara's build-in fold change method are validated (totally 278 genes) by quantitative RT–PCR with a LightCycler (Roche Applied Science, Mannheim, Germany) using cDNA that was prepared from mRNA for microarray analysis. The primers were designed based upon the sequence described in the Human Organized Whole Genome Database (http://howdy.jst.go.jp). We treat the rest of 886 genes as non-DEGs in subsequent data analysis.

2.6 MAQC dataset preprocessing

The MAQC dataset is a benchmark of several commonly used microarray platforms. Normalized data from five high-throughput platforms mentioned on the MAQC official web site (Yang *et al.*, 2002) were downloaded. These platforms include ABI (Applied Biosystems, 5791 Van Allen Way Carlsbad, CA 92008, USA), AFX (Affymetrix, 3420 Central Expressway Santa Clara, CA 95051, USA), AG1 (Agilent Technologies, one color, 395 Page Mill Road, 94303 Palo Alto, CA, USA), GEH (GE Healthcare, Princeton, NJ 08540, USA) and ILM (Illumina, 9865 Towne Centre Dr, San Diego, CA 92121-1975, USA). The National Cancer Institute platform was not included for two main reasons: (i) a reproducibility that is known to be significantly lower than that of other platforms; and (ii) the observed signal truncation, in which the signal above 3.7e4 is truncated as 3.7e4, is problematic, since many of the signals are above the truncation threshold. For every platform, there were three test sites with five technical replicates each. Missing data and negative values were replaced by the mean value of the within group signal.

2.7 Signal normalization and variance stabilization transformation

There are many issues with signal normalization across multiple arrays (Bolstad *et al.*, 2003; Quackenbush *et al.*, 2002; Yang *et al.*, 2002). In this study, quantile normalization, which forces the signal of every chip to be equal at every quantile point (Bolstad *et al.*, 2003) was selected as it is superior to other known methods (Bolstad *et al.*, 2003).

Plenty of methods have been proposed as the best way to stabilize signal variance when technical replicates are used (Huber *et al.*, 2002; 2003; Motakis *et al.*, 2006). Here, the Box-Cox power transformation, which is based upon the observation of the within group signal-variance plot, was used. The log-scaled signal, sg, and the corresponding within group variance, vg, across genes support a desired linear regression model, $lnv_g = alns_g + b$. As Cox-Box power transformation takes the form of $s'_g = (s_g^k - 1)/k$. And $k = (2-a)/2$ can be determined from the estimated linear regression model. Since k is a tuning parameter that characterizes the signal-variance response of a particular microarray platform, it can also be estimated from published data of the same type of microarray before the experiment. This 'signal' is referred to as the quantile normalized and variance stabilized, transformed signal in the later context.

2.8 Identifying DEG

As previously mentioned, under the null hypothesis, the difference in the gene signal of the g-th gene between target and control samples is modeled as $d_g \sim N(0,\sigma_0^2)$, and the variance-stabilized sample mean signal difference of the g-th gene between target and control samples is modeled as $d'_g \sim N(d_g, \sigma_1^2)$. Such a hierarchical null model has a simple one parameter marginal distribution $d'_g \sim N(0, \sigma_0^2 + \sigma_1^2) = N(0, \sigma^2)$, which eases the corresponding inference and significance testing.

The real distribution of d'_g across genes is a mixture of genes under the null hypothesis (non-DEG) and genes under alternative hypotheses (DEG). To estimate the scale parameter σ of the normal distribution under the null hypothesis in the presence of pollution (DEG), robust estimators have to be employed. When the number of genes is large (>5000), the median absolute deviation (MAD; 37% Gaussian asymptotic efficiency) estimator is used for speed, and, when the number of genes is not large (<5000), the Qn (82% Gaussian asymptotic efficiency) estimator (Kan *et al.*, 2001) is used for efficiency. After that estimation, the Z test is used to calculate the raw significance (*P*-value) of each gene, and the FDR procedure (Benjamini *et al.*, 1995) is performed as a final multiple testing correction.

The distribution of d'_g using our own three replicates is shown in Figure 2. The center peak of the distribution is mainly contributed by the genes under the null hypothesis, while the outliers of the distribution are contributed by DEG. An intuitive description for our method is that the center peak is fitted with a zero-centered normal distribution with outliers (DEG) detected using the estimated normal distribution.

It is worth to mention that the asymptotic statistical power $P \sim 1/\sigma^2 = 1/(\sigma_0^2 + \sigma_1^2) = 1/(\sigma_0^2 + \sigma_{1-1}^2/N)$, where σ_{1-1}^2 is the measuring variance/error when one sample is used and N represents the sample size and $\sim$ means 'in proportion to'. If the term σ_0^2 is not small then increasing/decreasing N does not increase/decrease the power significantly due to the mathematical property of the above sample size power function. Thus, our method is efficient in the case of small sample size as $P \sim 1/(\sigma_0^2 + \sigma_{1-1}^2)$, $N = 1$. We also reduce the power-cost tradeoff problem in microarray field by arguing that large sample size will be a waste as $P \sim 1/\sigma_0^2$, $N = \infty$.

We summarize our procedure as follows:

(1) Quantile normalization of multiple chips signal.

(2) Perform variance stabilization transformation.

(3) Calculate average target-control signal difference.

(4) Robust estimation of the total variance by MAD or Qn.

(5) Access raw significance by Z test.

(6) Perform false discovery rate multiple testing correction.

2.9 The triplicate dataset

This experiment included one pair of samples with three technical replicates. Each replicate contained a two-color microarray and a dye-swapped, two-color microarray. To minimize the potential dye bias problem, we first averaged the signal of the two dyes from each replicate on a log scale. Quantile normalization was then used to normalize the signal of the three replicates.

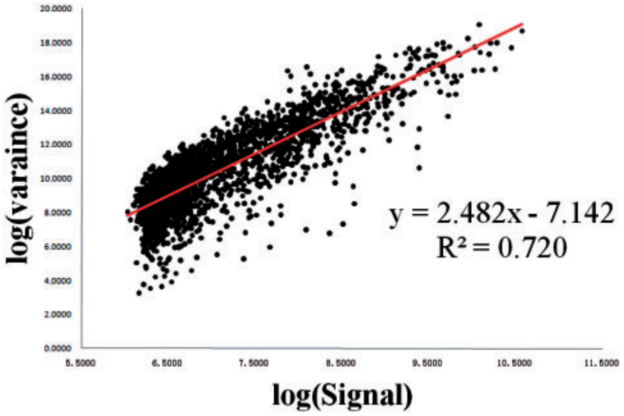

Fig. 1. Signal-variance plot on a log-scale.

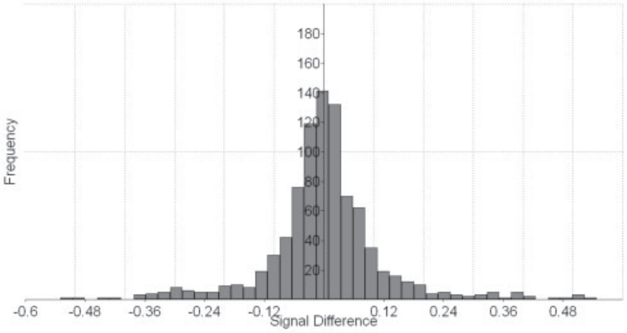

Fig. 2. Distribution of d'_g over genes.

To estimate the parameters of the Cox-Box power transformation (Box *et al.*, 1964) that was mentioned in the method part, a regression of the log-transformed signal variance against the log-transformed signal mean that was estimated from the three technical replicates was performed. The simple linear regression model fit the data well (Fig. 1). Seventy-two percent of the log-scaled estimated variance could be explained by the regression model. The estimated regression model was $lnv_g = 2.482lns_g - 7.142$, thus the Cox-Box transformation that was used was $s'_g = (s_g^{-0.241} - 1)/-0.241$. In addition, we also performed Kolmogorov–Smirnov normality test of the differences of non-DEG signal strength ($N = 800$, P-value $= 0.053$), indicating that our core hypothesis approximately holds.

Our result was first compared with commonly used methods with three replicates, using RT–PCR as the gold standard. Fold change is a widely used method for the identification of DEG. The criterion used here was an at least 2-fold difference between target and control samples, using Takara's build-in procedure which perform background subtraction first and then calculate fold changes. The *t*-test (Microsoft Excel, unequal variance) was performed on both untransformed data and variance-stabilization transformed (VST) data. Each method was tested using two FDR cutoffs, which were 0.05 and 0.1. The results of this comparison are shown in Table 1. The fold change method performed poorly, as demonstrated by a 45.8% false positive rate and 60.5% statistical power. The *t*-test on the original data showed very low power with >50% false positives, while the *t*-test on the variance-stabilized data performed slightly better. Although it had a higher statistical power, more than half of its positive findings were not of sufficient size and were not confirmed by RT–PCR. Our method shows the best balance between false

Table 1. Comparison of DEG detected by different methods using three replicates

Method	True negative	True positive	False positive	False negative	FDR (%)	Power (%)
FDR = 0.05						
FC	756	52	44	34	45.8	60.5
t-test	789	9	11	77	55.0	10.5
VST, *t*-test	755	69	45	17	39.5	80.2
Ours	799	72	1	14	1.4	83.7
FDR = 0.10						
t-test	784	14	16	72	53.3	16.3
VST, *t*-test	692	79	108	7	57.8	91.9
Ours	789	75	11	11	12.8	87.2

Table 2. Comparison of DEG detected by our method using different array pairs

Sample	True negative	True positive	False positive	False negative	FDR (%)	Power (%)
FDR = 0.05						
rep1&2	789	73	11	13	13.1	84.9
rep1&3	799	63	1	23	1.6	73.3
rep2&3	799	59	1	27	1.7	68.6
FDR = 0.10						
rep1&2	783	74	17	12	18.7	86.0
rep1&3	794	74	6	12	7.5	86.0
rep2&3	798	69	2	17	2.8	80.2

Table 3. Comparison of DEG detected by our method using different replicates

Sample	True negative	True positive	False positive	False negative	FDR (%)	Power (%)
FDR = 0.05						
rep1	793	71	7	15	9.0	82.6
rep2	796	63	4	23	6.0	73.3
rep3	800	34	0	52	0.0	39.5
FDR = 0.10						
rep1	782	74	18	12	19.6	86.0
rep2	791	70	9	16	11.4	81.4
rep3	798	51	2	35	3.8	59.3

positives and false negatives, as demonstrated by a false positive rate around the selected FDR cutoff and an acceptable statistical power (>80%).

We also evaluated our method using both two replicates and one replicate (Tables 2 and 3, respectively). When two replicates were used, the statistical power remained around 80%, which is acceptable. When only one replicate was used, the power was around 70%, but the variation was large. These observations suggest that our method can be used when few replicates are performed or are possible. In this situation, one replicate can be tried, two replicates will be acceptable and three replicates will be better (see Supplementary Material for details).

Table 4. Reproducibility (%) of the *t*-test carried by the MAQC consortium

$\alpha = 1\text{e-}3$, FC>2	Test site X				
	ABI	AFX	AG1	GEH	ILM
Test site Y					
ABI	88.0	71.7	80.2	66.7	68.6
AFX	73.5	94.4	88.3	74.9	77.0
AG1	67.5	72.5	89.8	67.0	67.5
GEH	66.2	72.5	78.9	88.8	65.8
ILM	76.5	83.8	89.3	73.9	92.6

Table 5. Reproducibility (%) of our method using $P < 0.3$

$\alpha = 0.3$	Test site X				
	ABI	AFX	AG1	GEH	ILM
Test site Y					
ABI	91.9	77.8	73.3	69.8	75.4
AFX	76.6	94.7	76.6	71.3	81.2
AG1	75.5	80.0	90.7	72.0	79.3
GEH	72.6	75.3	72.8	90.0	75.1
ILM	74.2	81.1	75.7	71.0	92.6

2.10 The MAQC dataset

The MAQC dataset was used to evaluate the portability of our method to other microarray platforms and the reproducibility of DEG identification on both an intra- and inter-platform basis. Reproducibility was defined as the percent overlap of genes on the DEG list for test site Y that were also present on the DEG list for test site X. For the ease of reading, the site-to-site reproducibility was summarized with the platform-to-platform reproducibility by arithmetical averaging corresponding site-to-site pairs.

A comparison of the reproducibility of the *t*-test used by the MAQC paper (Shi *et al.*, 2006) and our method revealed that reproducibility is positively correlated with the size of the DEG list. An intuitive example is that, when all the genes in both platforms are DEG, the reproducibility will reach 100%. To make the comparison fair, the DEG significance threshold of the two methods was adjusted to produce approximately the same number of DEGs. Two comparisons were carried out: (i) DEG identified by the MAQC paper using a *t*-test with $P < 0.001$ and $\text{FC} > 2$, which indicated that $P < 0.3$ should be used with our method to generate a similar number of DEG; and (ii) DEG identified by our method using $P < 0.05$, suggesting that the *t*-test should use $P < 5\text{e-}9$.

The reproducibility of our method and the *t*-test are shown in Tables 4–7. When the size of the DEG list is large (36% of total genes), the reproducibility of both the *t*-test and our method are high (most platform pairs >70%) and our results (average reproducibility = 78.7%) are slightly more reproducible than that of the *t*-test (average reproducibility = 77.4%). When a stricter threshold is applied with a reduced number of DEGs (17% of total genes) the situation changes. Our method results in a significantly more reproducible DEG (average reproducibility = 72.8%) than that with the *t*-test (average reproducibility = 59.5%). In addition, our method demonstrated its robustness of reproducibility against threshold changes. These observations lead to the conclusion that our method is portable among different platforms and results in higher reproducibility than that of the *t*-test intra- and inter-platforms.

Table 6. Reproducibility (%) of t-test using $P < 5\text{e-}9$

$\alpha = 5\text{e-}9$	Test site X				
	ABI	AFX	AG1	GEH	ILM
Test site Y					
ABI	64.0	84.6	71.3	48.7	64.0
AFX	34.3	83.2	56.7	35.6	50.1
AG1	38.3	75.0	72.0	39.6	53.2
GEH	42.0	76.1	63.9	67.3	57.1
ILM	44.0	83.3	67.4	45.4	71.4

Table 7. Reproducibility (%) of our method using $P < 0.05$

$\alpha = 0.05$	Test site X				
	ABI	AFX	AG1	GEH	ILM
Test site Y					
ABI	92.0	72.7	61.1	57.4	69.6
AFX	70.5	93.2	65.7	61.1	77.1
AG1	70.3	77.9	87.7	63.5	76.4
GEH	67.1	73.6	64.5	88.4	72.2
ILM	67.1	76.5	64.1	59.5	89.9

3 DISCUSSION

In this article, we present a statistical framework to identify DEGs from microarray experiments in the presence of transcriptomic background differences between target and control samples. We evaluate its performance by our own triplicate microarray experiment followed by RT–PCR validation and a comparison with the MAQC dataset on multiple microarray platforms. Our method can reduce both false positive and false negative (in biological/medical sense) results and can be applied effectively with fewer replicates compared with other methods. The technique is also portable across different microarray platforms. Its reproducibility is not only higher than the traditional t-test, but is also robust, since it is not greatly affected by the significance cutoff threshold. This method brings new insight to the significance of DEGs and may provide supporting evidence for those DEGs that would be accepted by more biomedical investigators.

The definition of statistical significance depends on how the null hypothesis is defined. Currently, most methods define the null hypotheses as $d_g = 0$ and develop a corresponding testing framework. This null hypothesis, however, is too idealistic and sensitive if its biological meaning is seriously considered. In reality, cells or biological tissues can never be exactly the same, especially after undergoing different treatments: (i) suppose we have highly matched cells lines, and the cell expression network changed as a whole after the treatment. We can image that expression level goes up and down across different genes, but only those leading genes indicated by obvious response to the treatment is of our interest while the rest result from background expression differences; (ii) many experimental designs are not so perfect that they compare different cells at different time points, different tissues and different status (e.g. cancer and nearby tissue). Another example is the benchmark datasets created by MAQC consortium compares human universal RNA against human brain RNA. Slight variations in d_g around zero should be allowed, otherwise the null hypothesis will be always rejected when the sample size is large enough, which would result in many meaningless positive findings. Therefore, in this article, the null hypothesis was loosened to become more realistic and robust, as indicated by $d_g \sim N(0, \sigma_0^2)$.

The marginal distribution of the sample-averaged target-control difference was still normally distributed, even with the addition of the unknown variance σ_0^2. Given that the number of genes in microarray experiment is usually large, the overall variance σ^2can be estimated using the transcriptome-wide distribution of d'_g. The robust estimators MAD or Qn were selected to estimate σ_0^2 in the presence of DEG pollution. Although MAD is easy and quick to calculate, it has a low Gaussian efficiency (37%). On the other hand, the Qn calculation is computationally slow but has a high Gaussian efficiency (82%) (Croux *et al.*, 1992). Both MAD and Qn have the maximum 50% breakdown point, at which their maximum robustness is indicated.

Our method assumes that the mean target-control difference of the g-th gene is distributed as $d'_g \sim N(d_g, \sigma_1^2)$. In order to ensure that this assumption can be approximately held, the signal should be properly normalized and transformed. In this article, quantile normalization and the Box-Cox power transformation were selected. Extensive comparison between these two methods and other methods was not performed, since they work well enough on our dataset (Fig. 1).

The RT–PCR significance cutoff was set to a 2-fold change, which is consistent with that used by the fold change methods. This cutoff is arbitrary but generally acceptable by biomedical investigators. The high degree of concordance of our results with those obtained in the RT–PCR analyses suggests that the DEGs identified by high-throughput microarray experiments using our method can be trusted and applied to other situations with little risk.

Our findings demonstrate that the reproducibility/portability of our method is higher than that of t-test. This is because t-test requires estimation of individual gene variance which is inaccurate and unstable in small sample size, thus its reproducibility is limited. On the contrary, our method estimates one common variance (contributed by both background variance and measuring variance) across thousands of genes and thus is more stable and reproducible.

Funding: Natural Science Foundation of Zhejiang Province, China (No. Y2080749 and No. Y2091110; partial); Zhejiang Province Science and Technology Fund for excellent returnee (No. 2008004); Science and Technology General Project of Zhejiang Province (No.2009C33143); Grants-in-Aid for Scientific Research (15590301, 16590276 and 17590300) from Japanese Society for the Promotion of Science.

Conflict of Interest: none declared.

REFERENCES

Allison,D.B. *et al.* (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.*, **7**, 55–65.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, **57**, 289–300.

Bolstad,B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Box,G.E.P. and Cox,D.R. (1964) An analysis of transformations. *J. Roy. Stat. Soc. Ser. B*, **26**, 211–252.

Budhraja,V. *et al.* (2003) Incorporation of gene-specific variability improves expression analysis using high-density DNA microarrays. *BMC Biol.*, **1**, 1.

Croux,C. and Rousseeuw,P.J. (1992) *Time-Efficient Algorithms for Two Highly Robust Estimators of Scale*. Physica, Heidelberg.

Cui,X. and Churchill,G.A. (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.*, **4**, 210.

Cui,X. *et al.* (2005) Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, **6**, 59–75.

Huber,W. *et al.* (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**(Suppl. 1), S96–S104.

Huber,W. *et al.* (2003) Parameter estimation for the calibration and variance stabilization of microarray data. *Stat. Appl. Genet. Mol. Biol.*, **2**, Article3.

Kan,T. *et al.* (2001) Gene expression profiling in human esophageal cancers using cDNA microarray. *Biochem. Biophys. Res. Commun.*, **286**, 792–801.

Ling,Z.Q. *et al.* (2007) Optimization of comparative expressed sequence hybridization for genome-wide expression profiling at chromosome level. *Cancer Genet. Cytogenet.*, **175**, 144–153.

Luo,L. *et al.* (1999) Gene expression profiles of laser-captured adjacent neuronal subtypes. *Nat. Med.*, **5**, 117–122.

Motakis,E.S. *et al.* (2006) Variance stabilization and normalization for one-color microarray data using a data-driven multiscale approach. *Bioinformatics*, **22**, 2547–2553.

Quackenbush,J. (2002) Microarray data normalization and transformation. *Nat. Genet.*, **32**(Suppl.), 496–501.

Shi,L. *et al.* (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.*, **24**, 1151–1161.

Smyth,G. (2005). Limma: linear models for microarray data. In Gentleman,R. *et al.* (eds) *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Springer, New York, pp. 397–420.

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

Yang,Y.H. *et al.* (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.