2007

Biostatistical Considerations of the Use of Genomic DNA Reference in Microarrays

Yunfeng Yang
Oak Ridge National Laboratory

Mengxia (Michelle) Zhu
Southern Illinois University Carbondale, mengxia@cs.siu.edu

Liyou Wu
University of Oklahoma Norman Campus

Jizhong Zhou
University of Oklahoma Norman Campus

Published in Yang, Y., Zhu, M.M., Wu, L., & Zhou, J. (2007). Biostatistical considerations of the use of genomic DNA reference in microarrays. Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, 2007. BIBE 2007, 593-600. doi: 10.1109/BIBE.2007.4375621

Recommended Citation

Yang, Yunfeng, Zhu, Mengxia (., Wu, Liyou and Zhou, Jizhong. "Biostatistical Considerations of the Use of Genomic DNA Reference in Microarrays." (Jan 2007).

Biostatistical Considerations of the Use of Genomic DNA Reference in Microarrays

Yunfeng Yang*
Biosciences Division
Oak Ridge National Laboratory
Oak Ridge, TN, USA, 37831
Email: yangy@ornl.gov
*Corresponding author

Michelle M. Zhu
Computer Science Department
Southern Illinois University
Carbondale, IL, USA, 62901
Email: mengxia@cs.siu.edu

Liyou Wu and Jizhong Zhou
Institute for Environmental
Genomics and Department of
Botany and Microbiology
University of Oklahoma
Norman, OK, USA, 73019
Email: lwu@rccc.ou.edu;
jzhou@rccc.ou.edu

Abstract

Using genomic DNA as common reference in microarray experiments has recently been tested by different laboratories (2, 3, 5, 7, 9, 20, 24-26). While some reported that experimental results of microarrays using genomic DNA reference conformed nicely to those obtained by cDNA: cDNA co-hybridization method, others acquired poor results. We hypothesized that these conflicting reports could be resolved by biostatistical analyses. To test it, microarray experiments were performed in a γ-proteobacterium Shewanella oneidensis. Pair-wise comparison of three experimental conditions was obtained either by direct cDNA: cDNA co-hybridization, or by indirect calculation through a Shewanella genomic DNA reference. Several major biostatistical techniques were exploited to reduce the amount of inconsistency between both methods and the results were assessed. We discovered that imposing the constraint of minimal number of replicates, logarithmic transformation and random error analyses significantly improved the data quality. These findings could potentially serve as guidelines for microarray data analysis using genomic DNA as reference.

1. Introduction

DNA microarray technology has been quickly adapted by mainstream laboratories to explore gene expression profiling of part or whole-genome for an organism (18, 19). A number of microarray studies use an experimental design in which experimental and reference RNA samples are transcribed into cDNA molecules, labeled with different fluorescent dyes (typically Cy5 and Cy3) and simultaneously hybridized to an microarray slide (8). This approach is very costly and tedious for samples of large numbers, for which comparison across all samples are often desired. Pairing all of the possible pairs for *n* samples results in a total of $n*(n-1)/2$ combinations. As *n* escalates, the polynomially increasing number could become unmanageable for individual laboratory. In addition, it is nearly impossible to compare data across experiments since the cDNA reference sample composition is subjected to differences of experimental design and hence not universal. It has been desired for a long time to develop novel strategies to integrate data across multiple, initially unrelated studies between laboratories or over a long period of time to promote data sharing and integration.

A conceptually sound solution to this problem is to use "reference design", which requires cohybridization of a common reference with all samples of the microarrays. Typically, the ratio ($\gamma 1$) from cDNA: common reference is compared to another ratio ($\gamma 2$) from cDNA: common reference. The computed "ratio of ratios" ($\gamma 1/\gamma 2$) is considered to be equivalent to direct cDNA: cDNA comparisons. Only *n* microarrays are needed to calculate the ratios of any possible pairs of *n* samples, if biological and technical replicates are not considered. Apparently, this strategy greatly reduces the costs and time incurred by traditional microarray experiments.

An ideal reference should fulfill the criteria of universality, reproducibility and uniformity, meaning that it should be universal across diverse microarrays, reproducible over a long time frame and in different laboratories, and represents each gene at a uniform level. One kind of such references is common RNA pools assembled from a number of different cell lines, tissues and conditions. Commercial universal RNA references are now available for mouse and human samples (Stratagene). However, the RNA references fall well short of the aforementioned criteria. Although RNA pools are more comprehensive than a single source of RNA sample, it still partially represents the whole genome; there is inherent biological variability among different RNA samples; and RNA could be degraded over time. Therefore, data quality across multiple studies is inevitably compromised. To address these issues, genomic DNA has been proposed to replace universal RNA reference (4). It is easy and economic to prepare genomic DNA in large amount with low variations between different laboratories. Furthermore, genomic DNA is stable and could be stored over a long period of time. It is independent of variations from one

1-4244-1509-8/07/\$25.00 ©2007 IEEE

preparation to another, which is a desirable feature of universal reference. In addition, genomic DNA represents entire genome completely and uniformly, due to the fact that the majority of genes are presented once in the prokaryotic genomes, or twice in most eukaryotic genomes. It is especially useful for microbial functional genomics because of low representation of repetitive sequences and intergenic regions in the genome. Several recent studies have proven that genomic DNA reference is indeed very effective and faithful to report gene expression profiles (2, 3, 7, 20, 25, 26). Furthermore, a comparative study between genomic DNA reference and universal RNA reference has reached the conclusion that genomic DNA is superior for routine use (25).

Nevertheless, adopting genomic DNA as reference also creates new challenges. It is conceivable that though this strategy enables the integration of disparate studies, it brings in new variations. For example, spots with low signal intensity from labeled genomic DNA are prone to high standard errors for measurements, and spots with high intensity considerably interfere with the hybridization of cDNA samples to the probes, leading to low fidelity in the ratio of cDNA to genomic DNA. For quality control purpose, it is critical to identify these variances and remove ambiguous values by biostastistical analyses. However, to our best knowledge, up to now this problem has not been unequivocally tackled and there is no consensus among the scientific community for the data analysis methods of microarray using genomic DNA reference. For instance, some researchers conducted array-to-array comparison with no data treatment except for background subtraction and removal of poor or negative spots (9, 24), while the others employed extensive techniques including setting minimum number of replicates and complicated statistical models (3, 7, 20, 26). It is thus necessary to appraise the performance of different biostatistical techniques.

In this study, we aim to fulfill this need by conducting a comparative study of genomic DNA reference and standard cDNA: cDNA co-hybridization. Microarray experiments were carried out for a γ-proteobacterium *Shewanella oneidensis*, which was capable of respiring with oxygen, fumarate, trimethylamine-N-oxide (TMAO), manganese (IV) oxides and ferric oxides as terminal electron acceptors (13-15). Gene expression profiles of *S. oneidensis* were generated under three growth conditions – aerobic growth or anaerobic growth with fumarate or ferric citrate as electron acceptor. Variations among gene expression profiles were compared and we concluded that biostatistical techniques, including setting minimal number of replicates, logarithmic (log) transformation and random error analyses, appeared to be valuable to improve data quality.

2. Materials and Methods

2.1. Sample preparation and microarray scanning

Shewanella oneidensis whole-genome microarray was constructed as described previously (6). Strain DSP10, a rifampin-resistant derivative of strain MR-1, was used in this study because this strain has been widely used for generating *Shewanella* mutants. It is thus of interest to catalog DSP10's gene expression behaviors in order to interpret the phenotypes of mutants derived from this strain.

DSP10 was grown aerobically in 100ml Luria-Bertani medium (LB, Difco) to mid-logarithmic phase at 30°C. Alternatively, DSP10 was grown anaerobically to mid-logarithmic phase in 200ml LB liquid supplemented with 20mM lactate, and with either 10mM fumarate or 10mM ferric citrate as electron acceptor. Mid-logarithmic phase was determined by measuring the turbidity at 600nm in a spectrophotometer for aerobic or anaerobic 10mM fumarate cultures, or by epifluorescence microscopy using acridine orange staining (11) for anaerobic 10mM Fe(III) citrate cultures. Cells were then collected by centrifugation at 4krpm for 10minutes. After discarding the supernatant, the pellets were immediately lysed by Trizol (Invitrogen), or chilled in liquid nitrogen and then kept at -80°C for later use. Total RNA was extracted as described previously (23). RNA samples were treated with RNase-free DNase I (Ambion) to digest residual chromosomal DNA and then purified with RNeasy Kit (Qiagen) prior to spectrophotometric quantification at 260 and 280nm. For cDNA: cDNA co-hybridation, cDNA was produced in a first-strand reverse transcription (RT) reaction and labeled with Cy5 or Cy3 dUTP (Amersham Biosciences) by direct labeling in the presence of random hexamer primers (Invitrogen). Fluorescein labeled probes were purified using a PCR purification kit (Qiagen). Slides were pre-hybridized at 50°C for about one hour to remove unbound DNA probes in a solution containing 50% (V/V) formamide, 9% H_2O, 3.33% SSC (Ambion), 0.33% sodium dodecyl sulfate (Ambion), and 0.8μg/μL bovine serum albuminin (New England Biolabs). Slides were hybridized at 50°C over night with Cy5- and Cy3- labeled probes in the above solution, minus BSA and with the addition of 0.8 μg/μL herring sperm DNA (Invitrogen) to prevent random binding. Pre-hybridization and hybridization were carried out in hybridization chambers (Corning). Slides were then washed on a shaker at room temperature in the following order: 7 minute in 1x SSC, 0.2% SDS; 7 minute in 0.1x SSC, 0.2% SDS; and 40 second in 0.1x SSC. For genomic DNA reference, 100ng *S. oneidensis* MR-1 genomic DNA (gDNA) was amplified by incubated at 37°C for 3 hours using Klenow fragment of DNA polymerase (Invitrogen) and random primers followed by transferring on ice to stop the labeling. Cy3 dUTP was incorporated in the product (Amersham Biosciences) and then Cy3-labeled genomic DNA was co-hybridized with Cy5-labeled cDNA on pre-hybridized microarray slides as described above.

A total of 12 replicates were prepared for both cDNA and genomic DNA reference methods. A program ImaGene version 5.5 (Biodiscovery) was used to grid and quantify microarray images. Background signals around each spot were calculated and subtracted from the signal intensity of each spots. Spots of Signal/background ratios < 3 were

1-4244-1509-8/07/$25.00 ©2007 IEEE

regarded as negative spots. All negative, poor and empty spots were flagged and discarded.

2.2. Data analysis of cDNA reference method

Data analysis of cDNA reference method has been previously established (23). Quantified microarray was loaded onto GeneSight-Lite, a plug-in program of ImaGene 5.5 for background subtraction, flagged spots removal, floor of 20 and normalization by mean. The results after processing were subsequently transferred onto software ArrayStatTM (Imaging Research), in which extensive statistical tools were available. In general, minimal number of replicates was set as 4, proportional model and small sample model were selected before outlier removal at $p < 0.05$. The significance of differential expression was determined by two-way *t*-test.

2.3. Data analysis of genomic DNA reference method

Local background subtraction and flagged spots removal were implemented in the same way as cDNA reference method. If no other biostatistical technique was used, inferred ratio was calculated by $T_2/T_1 = (T_2/R_2) / (T_1/R_1)$, where T and R represented the mean value of cDNA and genomic DNA reference signals from all of 12 replicates, respectively. To evaluate various biostatistical techniques, data were processed in the same way as cDNA reference method. Certain parameters were specified as: floor of 20 and normalization by mean, data with less than the minimal number of replicates of 4 were removed, and then followed by execution of proportion model and small sample model. Then outlier was removed by p <0.05 and finally, expression ratios were obtained by calculating the division of two ratios. At each time, only one parameter was allowed to change in order to test the corresponding technique; all of the rest remained unchanged.

2.4. Pearson correlation coefficient, the number of genes in opposite categories and One-way ANOVA

Pearson correlation coefficient (*r*) was computed between two sets of ratios acquired from cDNA and genomic DNA reference methods. To obtain the number of genes in opposite categories, 2-fold change was used as criterion for change of gene expression. We consequently categorized the differential expression values from both cDNA and genomic DNA reference methods into three classes: "up" for expression ratios of more than 2, "down" for expression ratios of less than 0.5, and "no change" for all other ratios. A gene was considered to be in opposite categories if its expression ratio was classified as "up" in cDNA reference method and "down" in genomic DNA reference method, or *vice versus.*

For one-way ANOVA, the logarithmic transformation was applied to the ratio values to normalize the expression variation among genes and equalize the data scale intervals for ANOVA test. The significance level of p value < 0.05 was used as criterium to reject or accept the null hypothesis "The two reference methods are not significantly different from each other".

3. Results

As indicated in the introduction, using genomic DNA reference could add an additional layer of variance, resulting in less reliable results than direct cDNA: cDNA comparison. To evaluate the impact, RNA was extracted from mid-logarithmically grown *Shewanella oneidensis* strain DSP10 under aerobic condition (O_2), or under anaerobic conditions with fumarate (Fum) or ferric citrate (Fe) as electron acceptors. cDNA were subsequently transcribed and labeled with Cy5 or Cy3, and any pair of two conditions was co-hybridized on microarray slides, yielding three direct ratios, namely Fum/O_2, Fe/O_2 and Fe/Fum. Meanwhile, RNA from each condition was reversely transcribed and labeled by Cy5 and co-hybridized with Cy3-labeled *Shewanella* genomic DNA. To obtain expression ratios of Fum/O_2, Fe/O_2 and Fe/Fum, the ratios of cDNA: gDNA were calculated, and then the inferred (indirect) ratios were obtained by calculating the "ratio of ratios" as ($cDNA_1$/gDNA) over ($cDNA_2$/gDNA).

Results from direct cDNA: cDNA co-hybridization ($cDNA_1/cDNA_2$) were analyzed according to a standard procedure in our laboratory (See Methods and Materials section for details), and compared to those obtained by genomic DNA reference. Two previous studies employed no biostatistical techniques except for basal ones such as background subtraction and removal of poor or negative spots (9, 24). Therefore, the same procedures were applied to generate the inferred ratios. Two criteria were used to judge the similarity between both methods. First of all, the overall similarity was determined by correlation coefficient derived from both sets of expression ratios over the entire genome, which provides a comprehensive view of the impact when a biostatistical method is evaluated. Secondly, to identify the most inconsistent data, the ratios were categorized as "induction (ratio>2)", "repression (ratio<0.5)" and "no change (0.5<=ratio<=2). From biological viewpoint, if the ratio is 3 for one method and 30 for the other, the data can still be considered as consistent despite ten fold differences. However, if the ratio is 3 for one method and 0.3 for the other, they should be considered as inconsistent because they represent two opposite categories as induction and repression, respectively. In this study, we focus on this type of inconsistency because they have the greatest impact on the biological interpretation.

The Pearson correlation coefficients of these two results fell in the range of 0.72-0.77 (Fig 1), which indicated that the both methods were not very similar. A careful inspection of the plots in Fig. 1 showed that many ratios from two methods (11 values for Fum/O_2, 17 values for Fe/O_2 and 8 values for Fe/Fum) fell into two opposite categories (induction vs. repression), as illustrated by dots located in the 2nd and 4th quadrants and away from the origin. Therefore, there were clear inconsistencies between these two methods.

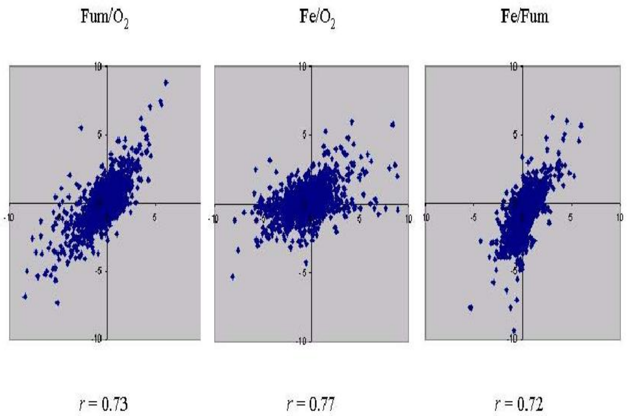

Figure 1. Comparison of direct ratios from cDNA reference method and inferred ratios from genomic DNA reference method. Each dot represents a gene whose complimentary probe is available on the microarray slides. X axis is the direct ratio and Y axis is the inferred ratio.

To provide quantitative criteria on the consistency of the both methods from statistical viewpoint, one-way ANalysis Of VAriance (ANOVA) was applied to a few selected genes. It is a powerful statistical approach and can be used to determine differences between the ratio means from cDNA and genomic DNA reference methods. Table 1 shows two representative genes with multiple replicates for the expression ratios of Fum/O_2. Analysis of OmcA leads to the *p* value of 0.0198, which is smaller than significance level of 0.05, inferring that two reference methods are significantly different. In contrast, the *p* value is 0.5054 for gene NapG, which fails to reject the null hypothesis.

A previous study has identified a number of genes previously regulated under Fum and Fe-reducing conditions in *S. oneidensis* (1). While the cDNA reference method was generally consistent with existing knowledge, genomic DNA reference method was not (Examples are shown in Table 2). *c*-type cytochromes OmcA and OmcB exist as a complex on outer membrane and function to reduce extracellular Fe(III) and U(VI) as terminal electron acceptors. Their expression is induced for several folds under anaerobic conditions ((12) and unpublished results in our laboratory). Expression of fumarate reductase FccA and its paralog IfcA is induced under Fe-reducing condition (12, 17). All of these expression patterns have been correctly confirmed by cDNA reference method but not the genomic DNA reference method. Therefore, the basal biostatistical analyses are not sufficient to remove potentially noisy values from genomic DNA reference method.

Minimal number of replicates. Minimal number of replicates serves as a threshold to remove genes without sufficient number of observations. If the number is lower than the threshold, the data for that gene will be disregarded. We first tested whether setting minimal number of replicates could improve the quality of the data. Figure 2A demonstrated that setting the minimal number of slides indeed had a significant impact on the correlation coefficients. As expected, when the minimal number of slides was set higher, Pearson correlation coefficient (***r***) of both methods increased under all conditions (Fig. 2A). Therefore, this technique is critical to improve data quality at global level.

Table 1. ANOVA for Genes napG and omcA. n: number of replicates; SD: standard deviation; SE: standard error; ssq: sum of square; DF: degree of freedom; MSq: mean square; F: F test value; and *p*: probability value.

napG	n	Means	SD	SE	
cDNA	12	-3.234	1.627	0.4698	
gDNA	10	-2.674	2.241	0.7087	
Source Variance	ssq	DF	MSq	F	*p*
napG	1.709	1	1.709	0.46	**0.5054**
Within-cells	74.338	20	3.717		

omcA	n	Means	SD	SE	
cDNA	12	1.667	0.527	0.1521	
gDNA	2	0.461	1.047	0.7400	
Source Variance	ssq	DF	MSq	F	*p*
omcA	2.494	1	2.494	7.21	**0.0198**
Within-cells	4.149	12	0.346		

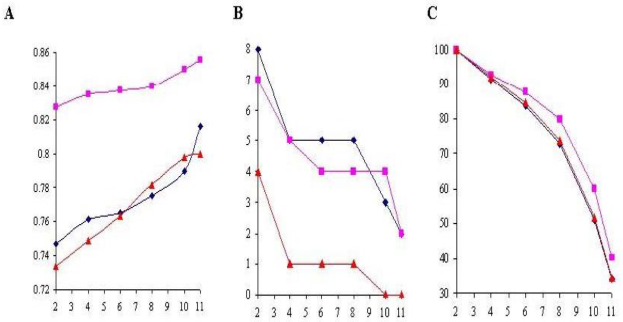

Figure 2. Effect of minimal number of replicates. Blue line: Fum/O_2; pink line: Fe/O_2; and red line: Fe/Fum. X-axis: minimal number of replicates. (A) Plot of *r* values with different minimal number of replicates. Y-axis: *r* values comparing cDNA and genomic DNA reference methods. (B) Number of genes in opposite categories (induction vs. repression) with different minimal number of replicates. Y-axis: numbers of genes. (C) Total number of genes with different minimal number of replicates. Total number of genes was set to 100% when minimal number of replicates was 2, and the total number of genes at other minimal number of replicates was normalized accordingly. Y-axis: numbers of genes.

We further calculated the number of genes in opposite categories in both methods. Fig. 2B shows that the number

1-4244-1509-8/07/$25.00 ©2007 IEEE

of genes in opposite categories was reduced at higher minimal number of replicates. However, significant amount of original data was lost at the same time (Fig. 2C). In this study, over 60% of values were lost when minimal number of replicates was set to be 11.

Logarithmic transformation. If there is a positive relationship between the standard deviation (SD) of the replicates and their mean, a log transformation is often conducted to remove a large portion of the relationship between the SD and mean. This approach is called proportional model. If there is no relationship between SD and mean, no log transformation should be applied and the data are analyzed in the raw form. This is called additive model. It is interesting to test whether applying log transformation makes differences or not in microarray analyses.

Table 2. Comparison of ratios using cDNA and genomic DNA reference methods. Values in boldface are consistent with previous reports, while values underlined are not. N/A: data not available.

	Fum/O_2		**Fe/Fum**		**Fe/O_2**	
Gene	Direct ratio	Inferred ratio	Direct ratio	Inferred ratio	Direct ratio	Inferred ratio
ifcA-1	N/A	**1.79**	**4.70**	**2.06**	**4.36**	**1.15**
ifcA-2	**4.16**	**8.52**	**24.16**	**19.88**	**7.87**	**2.33**
SO1427	**2.93**	**5.4**	**18.35**	**12.81**	4.19	2.37
mtrB	**2.81**	**6.15**	3.56	1.98	1.56	0.32
mtrA	**3.83**	**2.53**	2.51	2.64	1.36	1.04
omcB	N/A	**0.75**	1.40	0.89	1.30	1.19
omcA	**2.61**	**0.96**	1.70	1.11	1.41	1.15
mtrF	1.24	2.06	1.97	1.15	1.66	0.56
SO1781	0.92	1.84	1.15	0.48	1.16	0.26
mtrD	N/A	1.96	N/A	0.30	N/A	0.15
FccA	1.50	1.25	**2.34**	**0.79**	1.68	0.63
napB	**0.26**	**0.36**	0.66	0.61	3.85	1.68
napH	**0.1**	**0.28**	1.07	0.21	3.23	0.77
napG	**0.17**	**0.24**	0.54	0.38	5.28	1.61

Table 3. Summary of statistical results from application of outlier removal and flooring.

Techniques	***r***			**Number of genes in opposite categories**		
	Fum/O_2	Fe/O_2	Fe/Fum	Fum/O_2	Fe/O_2	Fe/Fum
No_outlier	0.76	0.84	0.76	5	1	1
Outlier_p<0.01	0.76	0.84	0.75	4	3	1
Outlier_p<0.05	0.76	0.84	0.75	5	5	1
No floor	0.76	0.85	0.77	2	6	2
Floor of 20	0.76	0.84	0.75	5	5	1
Floor of 50	0.77	0.84	0.78	5	4	1
Floor of 1%	0.75	0.83	0.72	7	7	2

Figure 3A indicated that proportional model clearly had a better performance than additive model in our data sets. Applying proportional model resulted in *r* values of 0.73, 0.80 and 0.72 for Fum/O_2, Fe/O_2 and Fe/fum conditions, respectively. In contrast, applying additive model resulted in much lower *r* values in the range of 0.39-0.53. Furthermore, proportional model resulted in fewer genes in opposite categories than additive model (Fig. 3B). To explain it, correlation coefficient of SD and mean value was calculated for each microarray dataset. There was clear positive relationship between SD and mean, as indicated by correlations of 0.81-0.88. These results suggested that log transformation should be applied.

Random error analyses. No measurement is entirely accurate. It is hence important to estimate the amount of measured value that could randomly deviate from the true value. This technique is called random error analyses or uncertainty analyses. A series of repeated measurements are usually used to make a reasonable estimate. The repeated measurements include biological and technical replicates. It is necessary to take random error into account to determine the significance of the results. One method, called small sample method, estimates random error on the replicates of individual genes, regardless of all of the other genes in the array. In contrast, random error could also be estimated on the entire array because this might be more accurate than estimation of small number of replicates for individual genes. A common error approach makes the assumption that the SD of replicates is unrelated to mean signal intensity. Alternatively, a curve fit approach could recognize the relationship between SD and mean by a regression line (curve fit).

Figure 4 demonstrates that small sample method has the best performance, as judged by the highest r values (Fig. 4A) and the fewest genes in opposite categories (Fig. 4B). Moreover, it is better than cases without random error analyses, suggesting that small sample method could improve data quality. In contrast, common error method yields the lowest r values and the most genes in opposite categories. This observation is expected because we have shown a positive relationship between SD and mean in the previous section.

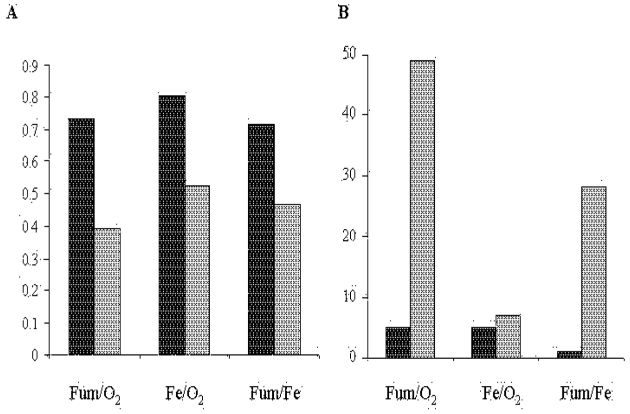

Figure 3. Effect of log transformation. Black and gray columns represent proportional and additive models, respectively. (A) The histogram representing the r values. (B) The histogram showing the number of genes in opposite categories.

Other biostatistical techniques. Outliers are data points that are not faithfully reproducible among replicates, yet have a disproportionately large effect on the average values. Thus removal of outlier from source file might improve data quality. Hence we compared results with and without outlier removal. Two criteria, namley $p < 0.05$ and $p < 0.01$, were applied to outlier removal. P refers to the possibility of making a Type I error in identifying outliers. Accordingly, $p < 0.05$ will detect more outliers than $p < 0.01$. Table 3 shows that the consistency between cDNA and genomic DNA reference methods is not improved by removing outliers, as demonstrated by the little change of r values and number of genes in opposite categories. Notably, when outlier removal was tested, minimal number of replicates was set as 4. It is thus possible that signal fluctuation is already fairly limited. Indeed, at outlier filter of $p < 0.05$, only ~15 values (i.e. 0.3% of total values) were removed from each of Fum/O$_2$, Fe/O$_2$ and Fe/Fum comparisons. If minimal number of replicates is not used, outlier removal has a slight impact on data quality (data not shown).

We also tested the effect of flooring. Low signals in direct comparison of RNA samples often produce spurious expression ratios, thus signals below a certain threshold level was often set to the threshold level (10, 21). To test if it is necessary to do so for genomic DNA, different flooring strategies were employed: no floor, floor an absolute value of 20 or 50, or floor 1% lowest signals. As shown in Table 3, data quality is not improved, indicating that flooring does not appear to be an effective technique.

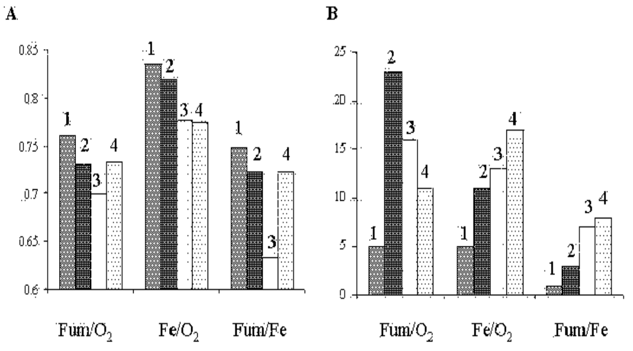

Figure 4. Effect of random error analyses. Column 1: small sample methods; 2: Curve fit method; 3: common error method; and 4: no method applied. (A) The histogram representing the r values. (B) The histogram showing the number of genes in opposite categories.

4. Discussions

It is often desirable to compare results from any two experimental conditions in microarray studies. Using a common reference such as genomic DNA allows for interconditional comparisons. However, the reliability of the comparison is often questionable since microarray is notorious for its considerable fluctuation of signals. In this report, we test different biostatistical techniques for improving data quality. Two criteria are used to evaluate these techniques: by (1) correlation coefficient to the cDNA reference method data; and (2) classifying differential expression values of genes into "up", "down" and "constant" categories, and then focus on genes in opposite categories. The first criterion evaluates the impact of techniques at the whole-genome level, while the latter addresses the most inconsistent data. Two-fold was used as threshold to classify the categories, which was reported to be a solid benchmark for induction or repression of gene expression (22). However, it is still likely that ratio changes of less than two fold are both statistically significant (judged by z-test or t-test) and biologically meaningful.

Thus two-fold is used here as a general guideline to simplify our study.

One underlying assumption in our study is that direct ratios from cDNA reference method are more reliable than the inferred ratios from genomic DNA reference method. One way to comprehend it is to analogize with triangle inequality relation for metric spaces: errors of an indirect path should be no less than errors of a direct path. The reliability of data from cDNA reference method has been extensively studied. It has been estimated that over 90% of the results could be verified by other techniques such as quantitative reverse transcription PCR or northern blot (16). Therefore, it is reasonable to believe that inferred ratios will be more reliable when their inconsistency to direct ratios is reduced. The prediction that direct ratios are more reliable than inferred ratios has been confirmed by the existing information of gene expression ratios (Table 2).

Among various biostatistical techniques, minimal number of replicates seems to be a critical one, as studies including minimal number of replicates for data analyses showed remarkable consistency between cDNA and genomic DNA reference methods (3, 20), while studies without it gave unfavorable opinion of genomic DNA reference method (9, 24). This is confirmed by our finding that using minimal number of replicates filters out a lot of inconsistency. When the minimum number of replicates is increased from 2 to 11, the average *r* values of all of pairs increased from 0.77 to 0.82. Meanwhile, there is a persistent drop in the number of genes in opposite categories. However, such improvement comes along with the expense of losing potentially biologically meaningful information.

A large number of biostatical methods are available for microarray data analyses. In this short report, we could only examine several of them. Moreover, caution should be taken to extend conclusions from our study to other microarray experiments. It is likely that some of our conclusions would not hold when biostatistical parameters are altered. Nevertheless, since we show that certain biostatical analyses affect data quality using genomic DNA reference; it is thus advisable for other researchers to evaluate their biostatistical methods when genomic DNA or another common reference is used in microarray experiments.

5. References

1. **Beliaev, A. S., D. M. Klingeman, J. A. Klappenbach, L. Wu, M. F. Romine, J. M. Tiedje, K. H. Nealson, J. K. Fredrickson, and J. Zhou.** 2005. Global Transcriptome Analysis of Shewanella oneidensis MR-1 Exposed to Different Terminal Electron Acceptors. J Bacteriol **187:**7138-45.
2. **Belland, R. J., G. Zhong, D. D. Crane, D. Hogan, D. Sturdevant, J. Sharma, W. L. Beatty, and H. D. Caldwell.** 2003. Genomic transcriptional profiling of the developmental cycle of Chlamydia trachomatis. Proc Natl Acad Sci U S A **100:**8478-83.
3. **Bina, J., J. Zhu, M. Dziejman, S. Faruque, S. Calderwood, and J. Mekalanos.** 2003. ToxR regulon of Vibrio cholerae and its expression in vibrios shed by cholera patients. Proc Natl Acad Sci U S A **100:**2801-6.
4. **Eisen, M. B., and P. O. Brown.** 1999. DNA arrays for analysis of gene expression, p. 179-205, Mthods Enzymol., vol. 303.
5. **Gadgil, M., W. Lian, C. Gadgil, V. Kapur, and W. S. Hu.** 2005. An analysis of the use of genomic DNA as a universal reference in two channel DNA microarrays. BMC Genomics **6:**66.
6. **Gao, H., Y. Wang, X. Liu, T. Yan, L. Wu, E. Alm, A. Arkin, D. K. Thompson, and J. Zhou.** 2004. Global transcriptome analysis of the heat shock response of Shewanella oneidensis. J Bacteriol **186:**7796-803.
7. **He, Q., K. H. Huang, Z. He, E. J. Alm, M. W. Fields, T. C. Hazen, A. P. Arkin, J. D. Wall, and J. Zhou.** 2006. Energetic consequences of nitrite stress in Desulfovibrio vulgaris Hildenborough, inferred from global transcriptional analysis. Appl Environ Microbiol **72:**4370-81.
8. **Hegde, P., R. Qi, K. Abernathy, C. Gay, S. Dharap, R. Gaspard, J. E. Hughes, E. Snesrud, N. Lee, and J. Quackenbush.** 2000. A concise guide to cDNA microarray analysis. Biotechniques **29:**548-50, 552-4, 556 passim.
9. **Kim, H., B. Zhao, E. C. Snesrud, B. J. Haas, C. D. Town, and J. Quackenbush.** 2002. Use of RNA and genomic DNA references for inferred comparisons in DNA microarray analyses. Biotechniques **33:**924-30.
10. **Lashkari, D. A., J. L. DeRisi, J. H. McCusker, A. F. Namath, C. Gentile, S. Y. Hwang, P. O. Brown, and R. W. Davis.** 1997. Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc Natl Acad Sci U S A **94:**13057-62.
11. **Lovley, D. R., and E. J. Phillips.** 1988. Novel mode of microbial energy metabolism: organic carbon oxidation coupled to dissimilatory reductio of iron or manganese. Appl Environ Microbiol **54:**1472-80.
12. **Meyer, T. E., A. I. Tsapin, I. Vandenberghe, L. de Smet, D. Frishman, K. H. Nealson, M. A. Cusanovich, and J. J. van Beeumen.** 2004. Identification of 42 possible cytochrome C genes in the Shewanella oneidensis genome and characterization of six soluble cytochromes. Omics **8:**57-77.
13. **Myers, C. R., and J. M. Myers.** 1994. Ferric iron reduction-linked growth yields of Shewanella putrefaciens MR-1. J Appl Bacteriol **76:**253-8.
14. **Myers, C. R., and J. M. Myers.** 1992. Localization of cytochromes to the outer membrane of anaerobically grown Shewanella putrefaciens MR-1. J Bacteriol **174:**3429-38.
15. **Myers, C. R., and K. H. Nealson.** 1990. Respiration-linked proton translocation coupled to anaerobic reduction of manganese(IV) and iron(III) in Shewanella putrefaciens MR-1. J Bacteriol **172:**6232-8.
16. **Quackenbush, J.** 2003. Genomics. Microarrays--guilt by association. Science **302:**240-1.
17. **Reyes-Ramirez, F., P. Dobbin, G. Sawers, and D. J. Richardson.** 2003. Characterization of transcriptional regulation of Shewanella frigidimarina Fe(III)-induced flavocytochrome c reveals a novel iron-responsive gene regulation system. J Bacteriol **185:**4564-71.
18. **Schena, M., D. Shalon, R. W. Davis, and P. O. Brown.** 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science **270:**467-70.
19. **Shoemaker, D. D., and P. S. Linsley.** 2002. Recent developments in DNA microarrays. Curr Opin Microbiol **5:**334-7.

20. **Talaat, A. M., S. T. Howard, W. t. Hale, R. Lyons, H. Garner, and S. A. Johnston.** 2002. Genomic DNA standards for gene expression profiling in Mycobacterium tuberculosis. Nucleic Acids Res **30:**e104.
21. **Tao, H., C. Bausch, C. Richmond, F. R. Blattner, and T. Conway.** 1999. Functional genomics: expression analysis of Escherichia coli growing on minimal and rich media. J Bacteriol **181:**6425-40.
22. **VanBogelen, R. A., K. D. Greis, R. M. Blumenthal, T. H. Tani, and R. G. Matthews.** 1999. Mapping regulatory networks in microbial cells. Trends Microbiol **7:**320-8.
23. **Wan, X. F., N. C. Verberkmoes, L. A. McCue, D. Stanek, H. Connelly, L. J. Hauser, L. Wu, X. Liu, T. Yan, A. Leaphart, R. L. Hettich, J. Zhou, and D. K. Thompson.** 2004. Transcriptomic and proteomic characterization of the Fur modulon in the metal-reducing bacterium Shewanella oneidensis. J Bacteriol **186:**8385-400.
24. **Weil, M. R., T. Macatee, and H. R. Garner.** 2002. Toward a universal standard: comparing two methods for standardizing spotted microarray data. Biotechniques **32:**1310-4.
25. **Williams, B. A., R. M. Gwirtz, and B. J. Wold.** 2004. Genomic DNA as a cohybridization standard for mammalian microarray measurements. Nucleic Acids Res **32:**e81.
26. **Yang, D. H., M. Barari, B. M. Arif, and P. J. Krell.** 2007. Development of an oligonucleotide-based DNA microarray for transcriptional analysis of Choristoneura fumiferana nucleopolyhedrovirus (CfMNPV) genes. J Virol Methods **143:**175-85.