Effect of normalisation on detection of differentially expressed genes in cDNA microarray data analysis

C. Dimauro, N. Bacciu, N. P. P. Macciotta

Dipartimento di scienze zootecniche. Università di Sassari

Corresponding author: Corrado Dimauro, Dipartimento di Scienze Zootecniche, Università degli Studi di Sassari, Via De Nicola 9, 07100 Sassari, Italia - tel. 079.229298 - Fax: 079.229302 - Email: dimauro@uniss.it

ABSTRACT: Four different normalisation techniques were applied for the corrections of fluorescence data generated by a cDNA microarray experiment. Correction for inaccurate signals and possible bias induced by fluorescence intensity, background intensity and dye effect were used in different combinations. Results of the present study highlight a pronounced role for the normalisation techniques in the absolute number of genes different expressed and a low concordance between different methods. Moreover, a significant effect of the dependent variable used, mean or median fluorescence intensity, was observed.

Key words: cDNA, Normalisation techniques, Statistical test.

INTRODUCTION - cDNA microarrays allow for the monitoring of the expression of many genes in parallel, providing substantially more information in comparison with standard molecular techniques where one or few genes at time are investigated. Use of cDNA microarrays in animal science is continuously increasing, due to their great potential to study the level of expression of thousand of genes in relation to environmental and physiological challenges. Obviously, the usefulness of these tools is based on the reliability of results. The generation of gene expression raw data is straightforward and quite easy to do. On the contrary there is not a consensus on methodology for the statistical analysis of fluorescence-intensity data (Do and Choi, 2006; Barbacioru et al., 2006). In particular, a large number of procedures for preliminary normalization and transformation of data have been proposed. However, the relative impact of these methods on the detection of gene expression level remains to be assessed. Aim of this work is to compare the effects of four different data correction methods and of their combinations in the detection of differentially expressed genes on a prototype data set.

MATERIAL AND METHODS - *Data and experimental design*. A reduced data set was extracted from the Stanford Microarray Database (http://microarray-pubs.stanford.edu/androgencluster). Raw data were fluorescence intensities of mRNA from human prostate carcinoma cell line (LNCaP) treted with two types and seven doses of human androgen. In the original experiment (Bebermeier et al., 2006), mRMA from androgen treated LNCaP cells was reverse-transcribed and fluorescently labeled with the Cy5 dye and compared directly on the same microarray to that from control cells (treated with ethanol alone) labeled with the Cy3 dye. Furthermore, the Cy5-labeled cDNAs were co-hybridized with a Cy3-labeled common reference mRNA pooled from several immortalized cell lines and normal genital skin fibroblasts (Holterhus et al., 2003).
Normalization methods. A first correction aimed to eliminate inaccurate signals. The technique suggested by Tran et al. (2002), and based on the correlation between mean and median signal intensities was used: spots that had a mean/median ratio <0.85 were removed from the dataset. The second correction dealt with the background subtraction that is advisable when the washing leaves a portion of the array covered by a high background signal. However, the simple subtraction of background intensity can lead to missing log intensity, especially when the expression levels are low. In this paper, the smoothing function proposed by Edwards (2002) was used. The background corrected signals may still have a systematic dependence on fluorescent intensity. To remove this bias, a robust local non parametric regression (LOWESS) was used (Cleveland, 1979). Finally, the competitive hybridization with the two dyes could result in a systematic bias due to differences in the efficiency of labeling. The most

direct approach to remove this bias is to center the data such that the mean value of each intensity measure is zero. Constructing residuals from a simple linear mixed model that included dye as fixed effect and array, array*dye as random effects is a simple way to achieve this goal. All the above described procedures were used to correct both mean and median foreground intensities of each spot according to the combinations reported in table 1.

Table 1. Different plans of correction techniques used for data normalization.

Plan	Techniques
1	All
2	All except the correction for inaccurate signal
3	All except the correction for the dye effect
4	All except the correction for fluorescence intensity

Statistical analysis. The detection of differentially expressed genes was carried out by analyzing normalized data with a gene-specific mixed model (Wolfinger *et al.,* 2001) that included the fixed effects of treatment and dye and the random effect of the array. A gene was assumed to be differentially expressed if it had a fold change greater than 1.5 and a *p*-value lower than 0.05. Moreover, the statistical significance of the treatment effect was also tested by using the Bonferroni test, in order to account for multiple comparison.

RESULTS AND CONCLUSIONS - The effect of the correction plan adopted was clearly evident (Table 2), with an obvious increase of the number of detected genes as the number of corrections decreases. The correction for the dye effect seemed to have a lower impact in comparison with the other two normalisation techniques.

Table 2. Number of differentially expressed genes detected in the different plans using unadjusted and Bonferroni adjusted pairwise comparisons.

Plan	Mean		Median	
	Unadjusted P	Bonferroni adjusted P	Unadjusted P	Bonferroni adjusted P
1	1338	60	1390	77
2	2324	140	2326	125
3	1467	97	1597	121
4	2474	412	2394	428

Together with the absolute number of genes detected, also a remarkable change in the identity of genes can be observed between different correction plans (Table 3). Actually, the frequency of genes commonly detected in different plans ranges from 0.01 to 0.85. Moreover, although the same plan generally resulted in a similar number of differentially expressed genes regardless of whether the mean or median fluorescence intensity was used (table 2), large differences were observed in the identity of loci detected. For example, 71% of genes were in common with plan 4 when using both the variable, whereas the percentage decreased to 25% when all the corrections were applied, i.e. plan 1. Finally, a relevant effect of the type of statistical test can be observed, with an abrupt reduction of the number of detected genes (>80%) with the more conservative correction.

Research on statistical techniques used to analyse data generated by cDNA microarrays experiments has usually focused on the optimisation of models for the detection of differentially expressed genes. Results of the present study highlight a pronounced effect of the normalisation technique used. Moreover, a significant effect of the dependent variable used, mean or median fluorescence intensity, was observed. However, the issue about the most appropriate set of normalisation techniques cannot be addressed, being strictly dependent on the specific data structure. The most intuitive choice to use all the possible corrections does not seem to be an advisable option because a not specifically required correction may add serious bias to the results (Qin *et al*., 2004).

Table 3. Absolute (and relative) frequencies of differentially expressed genes commonly detected by different correction plans.

Plan	Mean		Median	
	Unadjusted P	Bonferroni adjusted P	Unadjusted P	Bonferroni adjusted P
1*2	371 (0.28)	4 (0.07)	404 (0.29)	0 (0)
1*3	1138 (0.85)	46 (0.77)	1184 (0.85)	62 (0.80)
1*4	575 (0.43)	8 (0.13)	558 (0.40)	12 (0.16)
2*3	396 (0.27)	3 (0.03)	463 (0.29)	1 (0.01)
2*4	686 (0.29)	15 (0.11)	640 (0.27)	9 (0.07)
3*4	599 (0.41)	10 (0.10)	640 (0.40)	9 (0.07)

REFERENCES - Barbacioru, C.C., Wang, Y., Canales, R.D., Sun, Y.A., Keys, D.N., Chan, F., Poulter, K.A., Samaha, R.R., 2006. Effect of various normalization methods on applied biosystems expression array system data. BMC Bioinformatics 7, 533. **Bebermeier**, J.H., Brooks, J.D., DePrimo, S.E., Werner, R., Deppe, U., Demeter, J., Hiort, O., Holterus P.M., 2006. Cell-line and tissue-specific signatures of androgen receptor-coregulator transcription. J. Mol. Med. 84: 919-931. **Holterhus**, P.M., Hiort, O., Demeter, J., Brown, P.O., Brooks, J.D., 2003. Differential gene-expression patterns in genital fibroblasts of normal males and 46,XY females with androgen intensitivity syndrome: evidence for early programming involving the androgen receptor. Genome Biol. 4:R37. **Tran**, P.H., Peiffer, D.A., Shin, Y., Meek, L.M., Brody, J.P., Cho, K.W.Y., 2002. Microarray optimizations: increasing spot accuracy and automated identification of true microarray signals. Nucleic. Acids Res. 30, 12 *e54*. **Edwards**, D., 2002. Non-linear normalization and background correction in one-channel cDNA microarray studies. Bioinformatics 19 (7) 825-833. **Cleveland**, E.S., 1979. Robust locally weighted regression and smoothing scatterplots. J. Amer. Stat. Assoc. 74, 829-836. **Wolfinger**, R.D., Gibson, G., Wolfinger, E., Bennett, L., Hamadeh, H., Bushel, P., Asfhari, C., Paules, R.S., 2001. Assessing gene significance from cDNA microarray expression data via mixed models. J. Compu. Biol. 8. **Qin**, L.X., Kerr, K.F., et al., 2004. Empirical evaluation of data transformations and ranking statistics for microarray analysis. Nucleic. Acids Res. 32, 18 5471-5479. **Do**, J.H. and Choi D.K., 2006. Normalizzation of microarray data: single-labeled and dual-labeled arrays. Mol. Cells 22, 3, 254-261