International Journal of Statistical Sciences *ISSN 1683-5603*
Vol. 16, 2017, pp 93 -102

Identification of Differentially Expressed Genes by Using the Robust Likelihood Ratio Test

Md. Mehedi Hasan[1*], Md. Manir Hossain Mollah[2] and Md. Nurul Haque Mollah[3*]

[1]Bioinformatics Centre, College of Biological Science, China Agricultural University, Beijing, China.

[2]Department of Biostatistics, Bangladesh University of Health Sciences (BUHS), Dhaka-1216, Bangladesh.

[3]Department of Statistics, University of Rajshahi, Rajshahi-6205, Bangladesh

[*]Correspondence should be addressed to Md. Nurul Haque Mollah

(Email: mollah.stat.bio@ru.ac.bd)

[Received May 8, 2017; Accepted October 15, 2017]

Abstract

Differentially expressed (DE) gene identification from microarray datasets is a challenging statistical problem due to the small sample sizes with a large number of transcripts surveyed. To decrease the dimensionality of gene transcripts, there are numerous statistical methods. Nevertheless, in presence of an irregular pattern of expressions or contaminated transcripts, most of them produce misrepresentative results. Few robust statistical algorithms existed for the identification of DE genes. Most approaches are not so appropriate for the detection of multi-class DE genes. In this study, we robustify the likelihood ratio test for revealing DE genes. Real dataset of gene expression and simulation analyses show that our method improves the performance over the Bayesian and as well as classical approaches.

Keywords: LRT-criterion, chi-square test, gene expression, normal distribution, minimum covariance determinant (MCD) estimators and differentially expressed genes.

AMS Classification: 62Fxx, 62P10.

1. Introduction

Differentially expressed (DE) genes analysis from microarray data is a challenging statistical problem due to the small sample of sizes with large number of transcripts surveyed. Reduce the dimensionality of transcripts from microarray

dataset is an important statistical problem. A straightforward approach is to identify of differentially expressed (DE) genes under the $k \geq 2$ groups. With the combine the genotypes of molecular markers, we may get useful information on the regulatory network for DE identification [7]. There are two types of statistical inferences in the literature for identification of DE genes (i) classical parametric (likelihood ratio, F-test, test t-test, and so on) and non-parametric [2] measures, and (2) empirical Bayes (EB) approaches [4, 5, 6, 7] and non-parametric [2, 4] procedures. Usually, classical techniques examine the DE genes using the levels of significance i.e. based on p-values by permutation distribution of their test statistic. In contrast to classical approach, EB measures the posterior probability of differential expressions. Though, the above mentioned approaches except BRIDGE [5] are not robust against outliers. Most of the existing microarray dataset, the assumption of normality does not follow [6]. Due to the gene expression dataset are contaminated by outliers. To solve this issues, in this study, a robust likelihood ratio test (LRT) approach based on the MCD estimator was proposed that an extension of classical LRT approach to detect DE genes [12].

2. Examination the equality of several means by Likelihood ratio test

Let $x_{j1}, x_{j2},...,x_{jnj}$ be a random samples of size n_j from the j^{th} normal population (j =1, 2, ..., k). Assume that the j^{th} population has mean μ_j and variance σ_j^2. Further assume that k random samples are independent. Assuming $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$(unknown), we want to test the following statistical hypothesis :

H_0: $\mu_1 = \mu_2 = ... = \mu_k = \mu$ (say) against H_1: $H_0$is not true.

Under H_1, the likelihood function is as follows

$$L_1 = L(\boldsymbol{\theta}_1 \mid X) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ji}-\mu_j\right)^2\right], \tag{1}$$

where $\theta_1=(\mu_1, \mu_2,\ldots \mu_k,\sigma^2)$ and $n = \sum_{j=1}^{k} n_j$.

Under $H_{0,}$ the likelihood function is as follows

$$L_0 = L(\boldsymbol{\theta}_0 \mid X) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ji}-\mu\right)^2\right], \tag{2}$$

where $\theta_0=(\mu,\sigma^2)$ and $n = \sum_{j=1}^{k} n_j$.

Then the likelihood ratio test (LRT) criterion for testing H_0 against H_1 can be written as,

$$\lambda = \frac{L(\hat{\theta}_0 \mid X)}{L(\tilde{\theta}_1 \mid X)} = \left[\frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ji} - \hat{\mu}\right)^2}{\sum_{j}^{k}\sum_{i=1}^{n_j}\left(x_{ji} - \tilde{\mu}_j\right)^2} \right]^{-n/2}, \quad (3)$$

where $\hat{\theta}_0 = (\hat{\mu}, \hat{\sigma}^2)$ and $\tilde{\theta}_1 = (\tilde{\mu}_1, \tilde{\mu}_{2,}, \ldots, \tilde{\mu}_k, \tilde{\sigma}^2)$ are the maximum likelihood estimates (MLE) of $\theta_0 = (\mu, \sigma)$ and $\theta_1 = (\mu_1, \mu_2, \ldots, \mu_k)$, respectively.

Then $\chi^2 = -2\log\lambda$ follows approximately chi-square distribution with [(k+1) -1] = k degrees of freedom and the approximation is usually good, even for small sample sizes. So, LRT procedure computes $\chi^2 = -2\log\lambda$ for testing H_0 against H_1 and rejects the H_0 if χ^2 is larger than a Chi-Square percentile with k degrees of freedom, where the percentile corresponds to the confidence level chosen by the analyst. However, the standard LRT criterion as defined in (3) is very much sensitive to outliers. So, in presence of outliers, it produces misleading results. Therefore, we would like to robustify the classical LRT criterion as discussed bellow.

2.1. Robustification of LRT

To robustify the classical LRT criterion (3), let us rewrite it as follows

$$= \left[\frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ji} - \hat{\mu}\right)^2}{\sum_{j}^{k}\sum_{i=1}^{n_j}\left(x_{ji} - \tilde{\mu}_j\right)^2} \right]^{-n/2} = \left[\frac{n\hat{\sigma}^2}{n_1\tilde{\sigma}_1^2 + n_2\tilde{\sigma}_2^2 + \quad \ldots + n_k\tilde{\sigma}_k^2} \right]^{-n/2} \quad (4)$$

where

$$\tilde{\mu}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ji} \quad , \text{the mean of } j\text{th group data } (j = 1,2, ..k)$$

$$\tilde{\sigma}_j^2 = \frac{1}{n_j}\sum_{i=1}^{n_j}\left(x_{ji} - \tilde{\mu}_j\right)^2, \text{ the variance of } j\text{th group data}$$

$$\hat{\mu} = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ji}, \text{ the grand mean}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ji} - \hat{\mu}\right)^2, \text{ the grand variance}$$

$$n = n_1 + n_2 + \quad \ldots \quad + n_k$$

It is obvious that $\tilde{\mu}_j$, $\tilde{\sigma}_j^2$, $\hat{\mu}$, $\hat{\sigma}^2$ $(j = 1,2,..,k)$ as defined in equation (4) are very much sensitive to outliers. So our proposal is to use any robust estimators like minimum β-divergence estimators [11] or MCD estimators [13]of μ_j, σ_j^2, μ and σ^2, $(j = 1, 2,..,k)$ for robust estimation of the LRT criterion (λ) and compute *p*-value for testing H_0 against H_1 assuming $-2\log\lambda$ as an approximate χ^2- distribution and also we can use permutation techniques for computing the p-values.

3. Simulation Results

To investigate the performance of the proposed method in a comparison of some popular methods t-test, SAM, GaGa, EBarrays-LNN and BRIDGE [5, 6, 14, 15] for detection of DE (important) genes, we generate gene expression profiles for 15000 genes such that 2000 genes are differentially expressed (DE) between two groups ***A*** and ***B,*** and the rest 13000 genes are equally expressed (EE) between this two groups. For each gene, we generate n_1=40 expressions from group ***A*** with density function $N(\mu_1,\sigma^2)$ and n_2=50 expressions from group ***B*** with density function $N(\mu_2,\sigma^2)$. For generating DE genes, we use μ_1= 2, μ_2=-2 and σ^2=1. For generating EE genes, we use $\mu_1=\mu_2=\mu$ =0 and σ^2=1.

Then we apply 6 methods namely t-test, SAM, GaGa, EBarrays-LNN and BRIDGE including the proposed one to detect DE and EE genes from the whole gene profiles. If a DE gene is detected as DE gene, it is called true positive (TP) and if a DE gene is detected as EE gene, then it is called false negative (FN). On the other hand, if an EE gene is detected as EE gene, then it is called True negative (TN) and if an EE gene is detected as DE gene, then it is called false positive (FP).

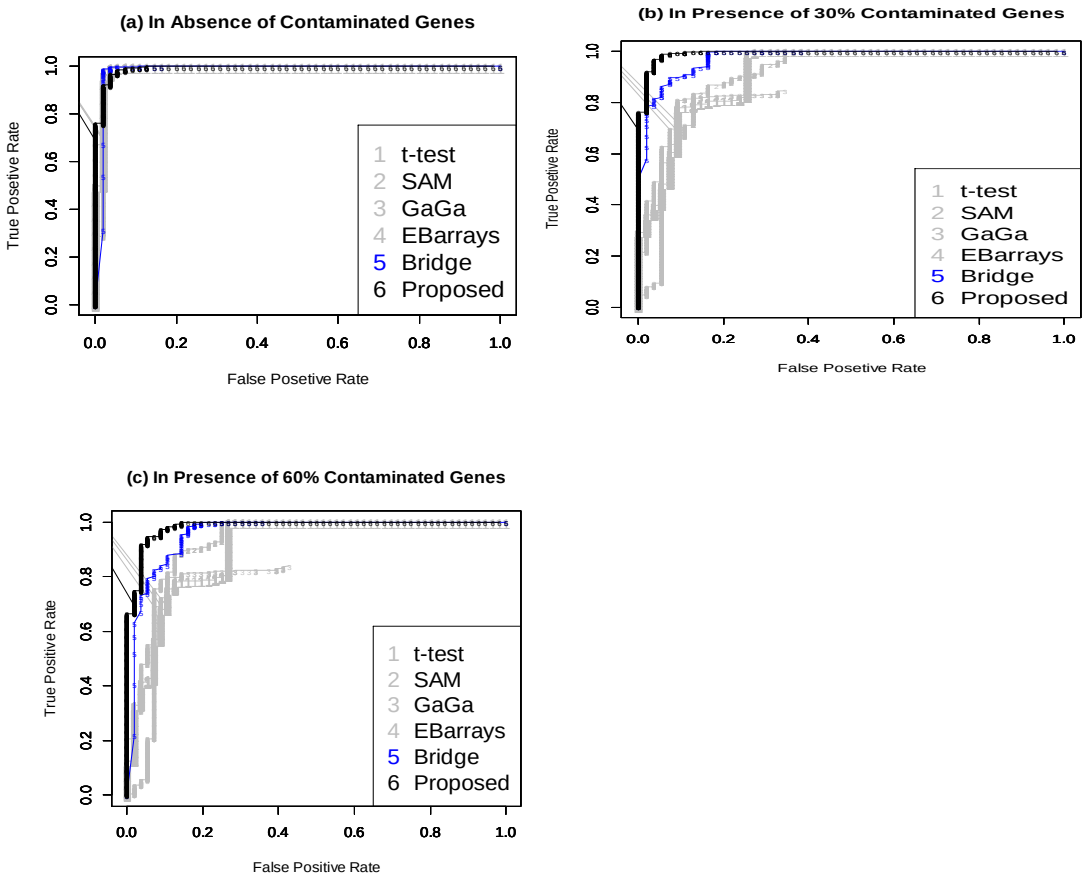

Figure 1: (a-c) Receiver operating characteristic (ROC) curves to investigate the performance of the proposed method in a comparison of t-test, SAM, GaGa, EBarrays and Bridge for identification of differentially expressed genes in presence of 0% , 30% and 60% contaminated genes, respectively.

An ROC curve represents the curve between TP rate (sensitivity) against the FP rate (1-specificity). ROC curves in figure-1 (a-c) represent the performance of t-test (blue line), SAM (gray line), GaGa (yellow line), the classical EBarrays-LNN (violet line), Bridge (black line) and the proposed method (red line) in presence of 0% , 30% and 60% contaminated genes, respectively. From these ROC curves, it is seen that all 6 methods show almost same performance in absence of outliers (0%), however, in presence of 30% and 60% contaminated genes, the proposed method shows better performance than other 5 methods though BRIDGE is

robust. Table 1 shows the area under the ROC curve (AUC) and partial area under the ROC curve (pAUC; at FPR≤ 0.2). It is seen that the values of both AUC and pAUC for t-test, SAM, GaGa, EBarrays-LNN and Bridge are not so stable in presence of 0% , 30% and 60% contaminated genes, respectively, while with regards to the proposed approach, both AUC and pAUC remain similar for each case. Thus simulation results shows that the proposed method improves the performance over the t-test, SAM, GaGa, EBarrays-LNN and Bridge approaches, otherwise, it keeps almost equal performance.

To investigate the performance of the proposed methods comparison with t-test, Sam, GaGa, EBarrays-LNN, BRIDGE, when FPR≤0.2 then calculated area under the ROC curve(AUC) and partial under the ROC curve area following results in case of moderate size(n_1=40, n_2=50) as shown in Table 1. In case our proposed method better performance than *Bridge* and other existing method see for values of AUC and pAUC.

Table 1: The calculated area under the ROC curve (AUC) and pAUC (with FPR≤0.2) calculated by t-test, SAM and empirical Bayes approaches (GaGa, EBarrays-LNN, BRIDGE) and the proposed approaches average over the 60 simulated datasets: for large sample cases

Method	T-test	SAM	GaGa	EBarrays-LNN	BRIDGE	Proposed
In Absence of Contaminated Genes						
AUC	0.9768 (0.0027)	0.9764 (0.0029)	0.9769 (0.0027)	0.9732 (0.0021)	0.9785 (0.0025)	0.9784 (0.0026)
pAUC	0.1861 (0.0015)	0.1856 (0.0015)	0.1842 (0.0015)	0.1851 (0.0019)	0.1857 (0.0014)	0.1860 (0.0014)
In Presence of 30% Contaminated Genes						
AUC	0.9219 (0.0034)	0.92332 (0.0034	0.9122 (0.0037)	0.9022 (0.0037)	0.9527 (0.0016)	0.9725 (0.0014)
pAUC	0.1536 (0.0021)	0.1573 (0.0022)	0.1476 (0.0025)	0.1565 (0.0026)	0.1852 (0.0012)	0.1862 (0.0011)
In Presence of 60% Contaminated Genes						
AUC	0.9100 (0.0049)	0.91425 (0.0041	0.9078 (0.0047)	0.8902 (0.0041)	0.9327 (0.0023)	0.9642 (0.0018)
pAUC	0.1531 (0.0019)	0.1542 (0.0017)	0.1466 (0.0015)	0.1534 (0.0023)	0.1849 (0.0009)	0.1858 (0.0007)

3.1 Analysis of the Head and Neck Cancer Data

We analyzed the publicly available microarray data in the study of head-and-neck cancer [8]. Most head-and-neck cancers are squamous cell carcinomas (HNSCC), originating from the mucosal lining (epithelium) of these regions. The data consist of tumor and normal tissues from 22 patients with histologically confirmed HNSCC. The expression levels of 12625 cellular RNA transcripts were assessed for this study. It also contains 42 head-and-neck cancer genes used as positive controls, i.e., genes known in advance to be DE.

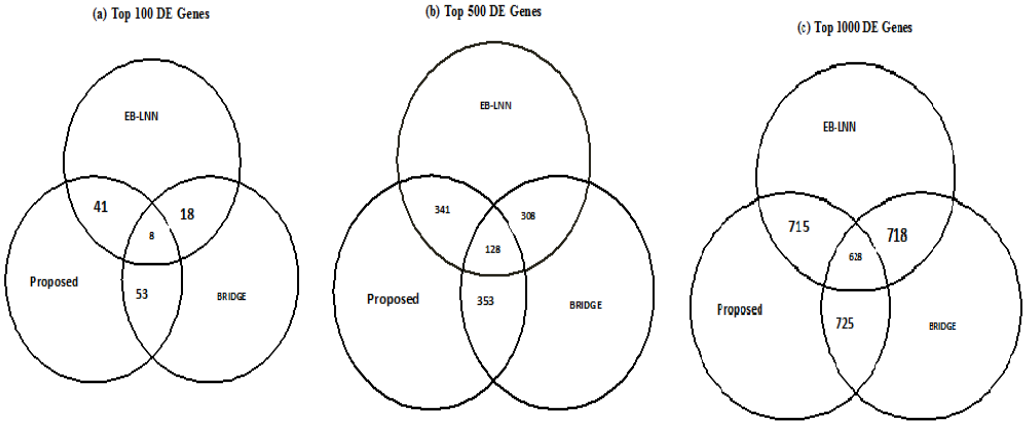

Figure 2: Ven-diagram of (a) Top 100 DE genes by EBarrays-LNN (EB-LNN), BRIDGE and Proposed approach, (b) Top 500 DE genes and (c) Top 1000 DE genes for head and neck cancer data.

We applied classical EBarrays-LNN, Bridge and Proposed robust approaches to analysis head and neck cancer data [12]. Figure 2 shows the ven-diagram of (a) top 100 DE genes by 3 methods, where number of common DE genes between EBarrays-LNN & BRIDGE is 18, number of common genes between proposed method & EBarrays-LNN is 41, number of common DE genes between Bridge & proposed method is 53, and number of common DE genes among 3 methods is 8. (b) top 500 DE genes by 3 methods, where number of common DE genes between EBarrays-LNN & BRIDGE is 308, number of common DE genes between EBarrays-LNN & proposed method is 341, the number of common DE genes between Bridge & proposed method is 353, and number of common DE genes among 3 method is 125. (c) Top 1000 DE genes by 3 methods, where

number of common DE genes between between EBarrays-LNN & BRIDGE is 718, where number of common DE genes EBarrays-LNN and proposed method common DE 715 but Bridge and proposed method common gene are 725 and both methods common DE genes 628.

3.2 Analysis of Lung Cancer Data

We also analyzed publicly available microarray data in a study of two types of lung cancer [9]. Non-small cell lung cancer (NSCLC) is the most common bronchial tumor, which can be classified into two major histological subtypes: adenocarcinoma (AC) and squamous cell carcinoma (SCC). After quality assessment of 60 microarray hybridizations, the data represent the gene expression profiles of 54675 cellular RNA transcripts in 40 AC and 18 SCC samples.

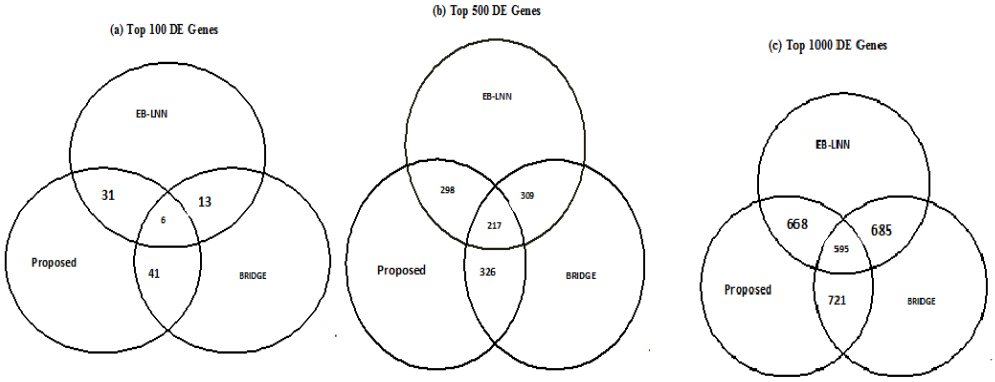

Figure 3: Ven-diagram of (a) Top 100 DE genes among EBarrays-LNN (EB-LNN), BRIDGE and Proposed approach, (b) Top 500 DE genes and (C) Top 1000 DE genes for lung cancer data

We compare the EBarrays-LNN and Bridge with proposed models to analysis lung cancer data [13]. All the three methods detected DE gene are follows in shown the Ven-diagram of (a) Top 100 genes among EBarrays-LNN, BRIDGE and Proposed approach we found that EBarrays-LNN and Bridge common 13 genes but our proposed method with EBarrays-LNN and Bridge common DE genes are 31 and 41 (b) Top 500 common DE genes among EBarrays-LNN, BRIDGE and Proposed approach in case found that EBarrays-LNN and Bridge common 309 genes but our proposed method with EBarrays-LNN and Bridge common DE genes are 298 and 326 and (c) Top 1000 DE genes among

EBarrays-LNN, BRIDGE and Proposed approach also found that EBarrays-LNN and Bridge common gene are 685genes but our proposed method with EBarrays-LNN and Bridge common DE genes are 668 and 721 and both methods common 595 genes

4. Conclusions

In this study, we propose a robust LRT criterion for the detection of DE genes. We used a simulation and a real dataset to examine the proposed model of LRT. Our resultant analyses show that the proposed method improves the performance over the classical existing and Bayesian approaches.

References

[1] Dean, N. and Raftery, A. E. (2005): Normal uniform mixture differential gene expression detection for cDNA microarrays. BMC Bioinformatics, 6:173.

[2] Do, K., Muller, P., and Tang1, F. (2005): A Bayesian Mixture Model for Differential Gene Expression. Journal of the Royal Statistical Society: Series-C (Applied Statistics), 54(3), pp. 627-644.

[3] Dudoit, S., Yang, Y. H. (2002), Callow, M. J., and Speed, T. P: Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica, 12, pp. 111-139.

[4] Efron B, Tibshirani R, Storey J, and Tusher V. (2001): Empirical Bayes analysis of a microarray experiment. Journal of the American Statistical Association, 96, pp. 1151-1160.

[5] Gottardo R, Raftery AE, Yeung KY and Bumgarner RE. (2006): Bayesian robust inference for differential gene expression (BRIDGE) in microarrays with multiple samples. Biometrics, 62, pp. 10-18.

[6] Kendziorski C, Newton M, Lan H and Gould MN. (2003): On parametric emparical Bayes methods for comparing multiple groups using replicated gene expression prfile. Statistical Medicine, 22, pp. 3899-3914.

[7] Kendziorski CM, Chen M, Yuan M, Lan H, Attie AD. (2006): Statistical methods for expression quantitative trait loci (eQTL) Mapping. Biometrics, 62, pp. 19-27.

[8] Kuriakose MA, Chen WT, He ZM, Sikora AG, Zhang P, Zhang ZY, Qiu WL, Hsu DF, McMunn-Coffran C, Brown SM, Elango EM, Delacure MD,

Chen FA. (2004): Selection and validation of differentially expressed genes in head and neck cancer. Cell Mol Life Sci, 61, pp. 1372-1383.

[9] Kuner R, Muley T, Meister M, Ruschhaupt M, Buness A, Xu EC, Schnabel P, Warth A, Poustka A, Sultmann H et al. (2008): Global gene expression analysis reveals specific patterns of cell junctions in non-small cell lung cancer subtypes. Lung Cancer, 63, pp. 32-38.

[10] Lo K and Gottardo R: Flexible empirical Bayes models for differential gene expression. Bioinformatics, 2007, 23, pp. 328-335.

[11] Mollah, M.N.H., Minami, M. and Eguchi, S. (2007): Robust prewhitening for ICA by minimizing β-divergence and its application to Fast ICA. Neural Processing Letters, 25(2), 91-110.

[12] M. H. Mollah, M. N. H. Mollah, H. Kishino (2012): β-empirical Bayes inference and model diagnosis of microarray data. BMC Bioinformatics, 2012, 13:135

[13] P. J. Rousseeuw and Katrien Van Driessen (1999): A Fast Algorithm for the Minimum Covariance Determinant (MCD) Estimator. Technometrics, 41, pp. 212-223.

[14] Rossell D (2009): GaGa: A parsimonious and flexible model for differential expression analysis. Ann. Appl. Statist., 3, pp. 1035-1051.

[15] Tusher, V., Tibshirani, R. and Chu, G. (2001): Significance analysis of microarrays (SAM) applied to the ionizing radiation response. Proc. Natl. Acad. Sci. (PNAS), U.S.A., 98, pp. 5116-5121.