Biotechnology Journal International

19(4): 1-14, 2017; Article no.BJI.36345
ISSN: 2456-7051
(Past name: British Biotechnology Journal, Past ISSN: 2231–2927, NLM ID: 101616695)

RNA-Seq Evaluating Several Custom Microarrays Background Correction and Gene Expression Data Normalization Systems

Noel Dougba Dago[1,2*], Martial Didier Yao Saraka[1], Nafan Diarrassouba[1], Antonio Mori[3], Hermann-Désiré Lallié[1], Edouard Kouamé N'Goran[1], Lamine Baba-Moussa[4], Massimo Delledonne[2] and Giovanni Malerba[3]

[1]Unité de Formation et de Recherche (UFR) des Sciences Biologiques, Département de Biochimie-Génétique, Université Peleforo Gon Coulibaly BP1328 Korhogo, Côte d'Ivoire.
[2]Laboratory of Functional Genomic, Department of Biotechnology, University of Verona, Laboratory of Functional Genomic, Strada Le Grazie 15 CàVignal 1, 37134, Verona, Italy.
[3]Department of Neurological, Biomedical and Movement Sciences, University of Verona, Strada Le Grazie 8, 37134, Verona, Italy.
[4]Laboratoire de Biologie et de Typage Moléculaire en Microbiologie, Faculté des Sciences et Techniques, Université d'Abomey-Calavi, Cotonou, Benin.

Authors' contributions

The present work was carried out in collaboration between all authors. Author NDD designed the study and wrote the first draft of the manuscript. Bioinformatics and statistical analysis have been performed by authors NDD and GM. This work and/or project has been supervised by authors MD and GM. Authors NDD and HDL wrote the protocol of the present study. Authors NDD, ND, MDYS and EKN managed the analyses of the study. Authors NDD, MDYS and ND managed the literature searches. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/BJI/2017/36345
Editor(s):
(1) Xing Li, Division of Biomedical Statistics and Informatics and Department of Health Sciences Research, Mayo Clinic College of Medicine, USA.
Reviewers:
(1) Yash Gupte, Ramnarain Ruia College, India.
(2) Samuel Ifedioranma Ogenyi, Nnamdi Azikiwe University, Nigeria.
Complete Peer review History: http://www.sciencedomain.org/review-history/20951

Original Research Article

Received 24th August 2017
Accepted 8th September 2017
Published 13th September 2017

ABSTRACT

Microarray gene expression technologies represents a widely used tool in transcriptomics and genomics studies worldwide. Even if this technology exhibits a low dynamic range as well as a feeble sensitivity and specificity (limited performances) with respect to RNA sequencing (RNA-seq)

**Corresponding author: E-mail: dgnoel7@gmail.com, noel.dago@upgc.edu.ci;*

methodology in whole transcriptomic and/or genomic studies; it is noteworthy to underline the stability of the former (microarrays) because of their well-established biostatistics and bioinformatics analysis schemes. Several studies shown that inadequate data pre-processing as regards microarray gene expression data analysis; i.e. inadequate gene expression data normalization (DN) and scarce noise background subtraction (BS), might compromise microarray aptitude in calling correctly significantly differentially expressed genes (DEGs). Here, we were interested in assessing the performance of 20 different microarrays background correction and gene expression data normalisation arrangements from R software “linear models for microarray and RNA-seq data analysis” package, by comparing the number of differentially expressed genes detected by our previous developed custom microarray designs and RNA-seq platform. The present study basing exclusively on several clustering and principal component analysis (PCA) as well as descriptive and inferential statistic surveys, developed in the R programing environment, suggested a predominance of microarray data normalisation systems with respect to noise background correction procedure. Although, all processed background subtraction and gene expression data normalization arrangement (BS+DN) claimed to improve the agreement (sensitivity) between microarrays and RNA-seq in calling DEGs; quantile normalisation procedure applied to our processed custom microarray designs has been recorded as exhibiting the best sensitivity (p-value<0.05), since discriminates the highest number of DEGs in agreement with RNA-seq as opposed to the others analysed microarray gene expression data normalisation systems. In conclusion our findings confirmed the pre-eminence of data pre-processing procedure in microarray gene expression profiling analysis according a priority to data normalisation procedure and suggested the stability of quantile normalisation system with respect to the others processed normalisation arrangements in the present executed gene expression comparative study.

Keywords: Microarrays; RNA-seq; data normalisation (DN); background subtraction (BS); differentially expressed genes (DEGs).

1. INTRODUCTION

DNA microarray is a technology that simultaneously evaluates quantitative measurements for the expression of thousands of genes. DNA microarrays have been used to assess gene expression between groups of cells of different organs or different populations. In order to understand the role and function of the genes, one needs the complete information about their mRNA transcripts and proteins [1]. Expression microarrays are designed to quantify the amount of mRNA in a specific sample. However, this can only be done indirectly through quantifying the color intensities returned by labeled mRNA molecules bound to the array surface. Translating pixel intensities into transcript expression requires a series of computations and/or operations, generically known as data pre-processing and normalization steps [2]. Typically, the first transformation applied to expression data, referred to as normalization, adjusts the individual hybridization intensities to balance them appropriately so that meaningful biological comparisons can be made. There are a number of reasons why data must be normalized, including unequal quantities of starting RNA, differences in labelling or detection efficiencies between the fluorescent dyes used, and systematic biases in the measured expression levels. Conceptually, normalization is similar to adjusting expression levels measured by northern analysis or quantitative reverse transcription PCR (RT–PCR) relative to the expression of one or more reference genes whose levels are assumed to be constant between samples. More than a decade, oligonucleotide microarrays have been the method of choice for transcriptional profiling studies, used to characterize biological systems. The power of microarray platforms depends on the number, identity and specificity of the oligonucleotide probes for their target gene models [3-4]. Gene expression microarrays are widely used as measurement tools in biologicalresearch by processing a wide range of methods for microarray data analysis,ranging from simple fold change (FC) approaches to testing for differentialexpression, to many complex and computationally demanding techniques [5]. Recognizing this allows investigator to choose procedure more judiciouslyand methodologist to direct their efforts more efficiently. In microarray the hybridization intensity is represented by the amount of fluorescence emission, which give an estimate of the relative amount of the different transcript that is represented. Several factors

should be considered when setting up a microarray experiment. The development of an experimental plan (experimental design) can contribute to maximize the quality and quantity of information. Experimental design affects the efficiency and internal validation of microarray experiments [6]. Also, processing of the microarray image and normalization of the data result to be a crucial steps to remove systematic variation measuring gene expression value in microarray gene expression differential analysis. Hence, several image-processing methods have been developed and are now available for expression microarray. These methods estimate the amount of RNA from fluorescent array images, while trying to minimize the extraneous variation that occurs owing to technical artefacts [7-8]. For example robust multi-array average (RMA), corrects arrays for background using a transformation, normalizes them using a formula that is based on a normal distribution, and uses a linear model to estimate expression values on a log scale. However, for accurate comparisons both within and among experimental sets it is critical to consider issues such as data quality and processing method prior to data analysis. After condensing the data, a box plot can be used to visualize the detection range of each array, and to compare it with the known dynamic range of the array type. Data outliers with high background, low intensity, or narrow detection range can be identified. Hierarchical clustering of experiments can also be used to assess data quality by determining if replicate or biologically related samples cluster together. Another important preprocessing step is normalization, a process by which non biological variation is minimized and standardized and which allows comparisons between microarray experiments. It also generally makes data more consistent with the assumptions that underlie many inferential procedures. Normalization can be applied multiple times at different levels of analysis for different purposes. There are many different methods for normalizing microarray data i.e. microarray and RNAseq normalization scheme from linear models for microarray data analysis (*limma*) [9] of R statistical package [10]. Users should be aware that certain condensing algorithms, such as MAS 5.0 or RMA, normalize the data during the condensing process. Depending on the hypothesis, experimental objectives, and experimental design, additional normalization may or may not be required. One way to account for experimental differences between arrays during normalization is to divide every value on the array by the arithmetic or logarithmic median of the entire array. This effectively establishes a common reference for array-to-array comparisons. This calculation is a linear transformation, specific to each array, so the relative expression level differences between genes on the same array do not change. So, the global normalization method is based on the assumption that the total amounts of labeled mRNA in all samples are similar. Also, the internal controls as regards microarray gene expression data normalization can be genes with housekeeping functions that are constitutively expressed, or spiked cRNA controls [11]. When using internal controls, it is important to validate the assumption that the control genes have a constant transcription level across samples. Then, considering the necessity of removing systematic variation in performing microarray gene expression profiling analysis, we were interested to compare all background subtraction (BS) + microarray gene expression data normalization (DN) arrangement from *limma* R package [9] as regards our previous developed custom microarrays design manufactures based on both ex-CombiMatrix (CMB.S and/or CMB.D microarray design based on single and/or multiple oligonucleotide short probe set per gene model transcript) and ex-Roche NimbleGen (NMG.S and/or NMG.D microarray design based on single and/or multiple oligonucleotide long probe set per gene model transcript) platforms by assessing the agreement between microarrays and RNA-seq approaches in calling significantly differentially expressed genes (DEGs), analyzing two *Vitis vinifera* berry developmental stages [4]. For this purpose several hierarchical clustering analysis based on principal component considerations and/or analysis [12-13] developed in R software programming environment [10] have been performed.

2. MATERIALS AND METHODS

2.1 Gene Expression Differential Analysis by Applying Several Microarray Background Subtraction (BS) + Data Normalisation (DN) Arrangements

Microarray gene expression differential analysis by processing two grape (*Vitisvinifera*) development stages (ripening and veraison) has been performed by processing 20 different BS + DN arrangements of the *limma*R package (version 3.10.3) [14]. In fact we combined Quantile, Cyclic Loess, Scale and None (Null) normalization methods with Saddle, Maximum

Likelihood Estimation, Robust Multiple-chip Average (RMA) and Robust Multiple-chip Average 75 (RMA 75) and None (Null) background subtraction (BS) methods for each considered microarray designs (in total four different microarray designs combined with 3 different probe set average method were processed). In addition, *Vitis vinifera* RNA samples were also analyzed by sequencing-based methods generally referred to as RNA-seq, whose results were used as reference values evaluating the impact and/or the influence of microarray BS+DN arrangement procedures on genes expression results. Concerning the RNA-seq experimentation, two technical replicates each for two grape berry development stages (ripening and veraison) were prepared and sequenced using an Illumina Genome analyzer II machine yielding more than 59 million reads of average length 36 bp. Reads were aligned onto the 12x grape genome assembly followed by genome reconstruction step by cufflinks package that measured gene expression levels. Also, read count was performed using the packages RSEM (v1.1.21) [15] and Cufflinks (1.2.0 release, http://cufflinks.cbcb.umd.edu/). Next DESeq (version 1.1.6) package has been used for the gene expression differential analysis. RNA-seq raw data are available at SRA009962 as well as at URL http: //ddlab.sci.univr.it/cgi-bin/gbrowse/grape [16].Indeed, differential gene expression (DGE) analysis between above mentioned grape development stages (*Vitis vinifera* repining and veraison development stages) was performed by comparing arrays processed with the same BS+DN combination and RNA-seq gene expression differential analysis results. In addition, DGE survey was conducted by applying linear models on the log-expression values followed by an empirical Bayes moderated t-statistics on each gene aiming to reduce data variability errors. The "*lmFit"* and "*eBayes"* functions of the *limma*R package (version 3.10.3) were used [14]. The False Discovery Rate (FDR) suggested by Benjamini and Hochberg [17] was adopted to control the FDR since gene expression differentially analysis usually englobes multiple comparisons statistical test. Significance of DGE analysis results of both custom (ex-CombiMatrix and ex-Roche NimbleGen) microarray designs based on multiple short and/or long probes per gene model transcript; CMB-D (CMB-D.fisher) and NMB-D (NMB-D.fisher) platforms respectively; when applying the mean and/or median values of the probe signals was also estimated by applying the Fisher's combined p-value method to combine evidence from multiple probes of the same gene [16-18]. A gene was considered as differentially expressed (DE) when showing a mean difference of the expression value greater than or equal to two folds between the 2 grape berry development stages at a False Discovery Ratio ≤ 0.05 (FDR≤0.05). Only genes shared among all the platforms were included in the present performance comparisons survey [16].

2.2 Hierarchical Clustering Survey, Principal Components Analysis (PCA) and Biplot Graphic Survey in R programing Environment

Principal component analysis (PCA) is a dimensionality reduction technique that is widely used in data analysis. Reducing the dimensionality of a dataset can be useful in different ways. Lower dimension can sometimes significantly reduce the computational time of some numerical algorithms. Besides, many statistical models suffer from high correlation between covariates, and PCA can be used to produce linear combinations of the covariates that are uncorrelated between each other [12]. There are many packages and functions that can apply principal component analysis (PCA) in R. In this study we used the function *prcomp* from the stats package. We also visualized PCA in R using Base R graphics. However, with purpose to improve PCA graphic visualization (biplot graphic) in R, "*ggbiplot*" script or function, which is implemented by Vince Q. Vu, and available on "*github*" library has been used.Since skewness and the magnitude of the variables influence the resulting PCs, it is good practice to apply skewness transformation, center and scale the variables prior to the application of PCA. Here, we applied a log transformation to the variables (number of DEGs) but we could have been more general and applied a Box and Cox transformation [13]. The *prcomp* function returns an object of class *prcomp*, which have some methods available. The print method returns the standard deviation of each PCs, and their rotation (or loadings), which are the coefficients of the linear combinations of the continuous variables. The summary method describe the importance of the PCs. The first row describe again the standard deviation associated with each PC. The second row shows the proportion of the variance in the data explained by each component while the third row describe the

cumulative proportion of explained variance. The plot method returns a plot of the variances (y-axis) associated with the PCs (x-axis). The Figure is useful to decide how many principal components to retain for further analysis. Also, R has an amazing variety of functions for cluster analysis. In this study, we used three of the many approaches: hierarchical agglomerative, partitioning, and model based. Indeed, model based approaches assume a variety of data models and apply maximum likelihood estimation and Bayes criteria to identify the most likely model and number of clusters. Specifically, the *Mclust()* function in the "*mclust*" package selects the optimal model by hierarchical clustering for parameterized Gaussian mixture models. One chooses the model and number of clusters with the largest BIC [19-20].

3. RESULTS

3.1 Principal Component Analysis Measuring the Variability of Detected DEGs Among Array Features and RNA-seq by Combining Several Custom Microarray Platforms BS and DN systems

Since skewness and the magnitude of the variables influence the resulting principal components (PCs) factors, it is good practice to apply skewness transformation, center and scale the variables prior to the application of principal component analysis (PCA) (see materiel and methods). Here we applied a log transformation to the variables (number of DEGs recognized as such by both array and RNA-seq, by combining several custom microarrays BS and DN systems). The plot method returned a plot of the variances (y-axis) associated with the PCs (x-axis) (Fig. 1). The Figure below (Fig. 1) is useful to decide how many PCs to retain for further analysis. Standard deviation associate to processed principal components by combining microarray features BS and DN systems ranged from 4.34 to 1.577e-15, while proportion of variance oscillated between 0.94 and 0.00. In addition, the cumulative proportion analysis explaining the variance (data variability) suggested (i) the first PC account for morethan 94% and (ii) the first two PCs accounts for more than 97% of the variance of the data (Fig. 1). Considering as a whole this funding opined the first two principal components as satisfactoriness factors explaining presently analysed data variability (variability as regards the number of commonly called DEGs between processed array features and RNA-seq), even if the first principal component claimed to explain more than 90% of analysed data variability (Fig. 1).

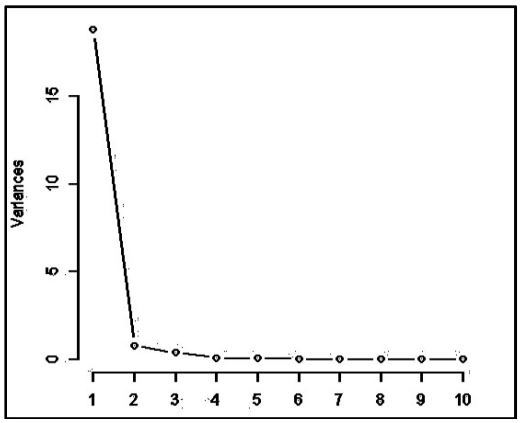

Fig. 1. PCA analysis explaining detected DEGs variability between several array features and RNA-seq approach

3.2 Biplot Analysis Assessing Analysed Microarrays Dispersion by Processing Background Correction/Subtraction (BS) and Gene Expression Signal Data Normalisation (DN)

There are many approaches to normalizing expression levels. Here, we processed 20 microarrays BS+DN arrangement of *limma* R package with the purpose to assess analysed microarray features aptitude in calling DEGs recognized as such by RNA sequencing (RNA-seq) approach. The present biplot survey graphic; providing an optimal visualization of the principal component analysis (PCA) reinforced the aptitude of the first principal component (CP) explaining more than 90% of the presently processed microarray features data variability (Fig. 2). Microarray design features based on long multiple probe set (NMG.D) exhibited a relative high stability as opposed to those based on long single probe per gene target (NMG.S) (Fig. 2). In the similar tendency, the same analysis evoked a major stability of microarray design based on short multiple probe set per gene model transcript (CMB.D) as opposed to array feature based on short single replicate oligonucleotide probe per gene model transcript (CMB.S) (Fig. 2). Then, this analysis suggested the stability of microarray tools in calling DEGs in agreement with RNA-seq approach, as

depending on the type of arrays features designs, rather than microarray probe size. In addition, array design features based on different and/or multiple probes in targeting a gene model transcript claimed to stabilize detected expressed gene signal when compared to RNA-seq (Fig. 2). However, the first principal component of the Fig. 2, indexed microarray features based on single long and/or short probe set targeting gene model transcript (NMG.S and CMB.S) as the substantial source of analysed data variability (number of DEGs between processed microarray designs). In the other word our findings supported a substantial disagreement between microarray features based on single (long or short) probe per gene model transcript and RNA-seq approach as opposed to microarray array design that include multiple (long and/or short) probe set per gene model transcript (Fig. 2). Interestingly, all analysed array features exhibited a consistent agreement in term of data pre-processing procedure by applying gene expression data normalization and background correction (background subtraction) (Fig. 2). Also, this analysis suspected a predominance of data normalisation (DN) procedure with respect to those of background subtraction (BS) in microarray data pre-processing survey (Fig. 2).

3.3 Clustering Analysis Assessing Euclidian Distance between Analysed Microarrays Features and Applied Background Correction and Gene Expression Data Normalisation Systems

Here we performed Euclidian distance clustering analysis with the purpose to assess the rearrangement of processed microarrays features and as well applied data normalization and background correction procedures. Cluster dendrogram graph referred to array features displayed two tendencies based on (i) the size of microarray oligonucleotide probe i.e. CMB: array design based on single and/or multiple short (35-40mer) probes per gene model transcript and NMG: array design based on single and/or multiple long probe (60 mer) per gene model transcript as well as on (ii) microarray design strategies i.e. CMB.S: array design based on single short probe per gene model transcript, CMB.D: array design based on multiple short probe per gene model transcript, NMG.S: array manufacture based on single long probe per transcript model and NMG.D: array design established on multiple long probe per gene transcript model (Fig. 3). In addition, performed cluster analysis regarding array features exhibited a contrasting behaviour among both array design manufactures centred on long (NMG.D) and short (CMB.S) probes per gene model transcript when compared to microarray feature based on probe set average procedure sort out by Fisher method (CMB.D Fisher and NMG.D Fisher) (Fig. 3). Considering as a whole the present survey evoked a heterogeneous reply and performance of processed microarray designs in gene expression differential analysis, when RNA-seq approach was assumed as reference. Next, cluster dendrogram analysis referred to both microarray background correction and/or subtraction (BS) and data normalisation (DN), evidenced three distinct situations, suggesting a consistence influence of both microarray DN and BS procedures on array gene expression profiling analysis. Also, a coherent clustering evidence has been observed between processed data normalisation practice in comparison to background correction process (Fig. 3). Furthermore, this analysis suspected a predominance of data normalisation (DN) with respect to those of background subtraction (BS) methodology in the presently microarray data pre-processing step as well as suggested a relative agreement between quantile, cycle Loess and scale normalisation systems (Fig. 3).

3.4 Impact of Background Correction and Gene Expression Data Normalization Arrangement Methods on Microarray Features Variability in calling accurately DEGs by Model Based Clustering Analysis

This clustering analysis based on "*mclust*" package of R software includes both univariate and multivariate mixture parameters [19]. Then, focusing on multivariate features parameter, our analysis (*mclust* function outcome graphic) recorded 14 multivariate mixtures in the present gene expression comparative study (Fig. 4). Indeed, this graphic attributed high BIC values (see material and method chapter) to Scale, Quantile and Cyclic Loess normalisation methods by processing ellipsoidal, equal volume and equal shape (EEV) multivariate parameter advising their high performance enhancing gene expression differential analysis quality as opposed to null normalisation factor. Next, focusing our attention on ellipsoidal, equal volume, shape, and orientation (EEE) multivariate mixture parameter, we were able to

demonstrate the high performance of both Quantile and Cyclic Loess normalization methods with respect to Scale normalization method (Fig. 4). Interestingly, the other analyzed multivariate mixture parameters (more than 75% of them) suggested a high performance as well as a reliable stability of Quantile normalization system in detecting accurately DEGs in agreement with RNA-seq approach in the present comparative gene expression differential analysis. In the other word, Quantile normalization system seems to be more tolerant as regards microarray background subtraction (BS) procedures as opposed to both Cyclic Loess and Scale normalization systems.

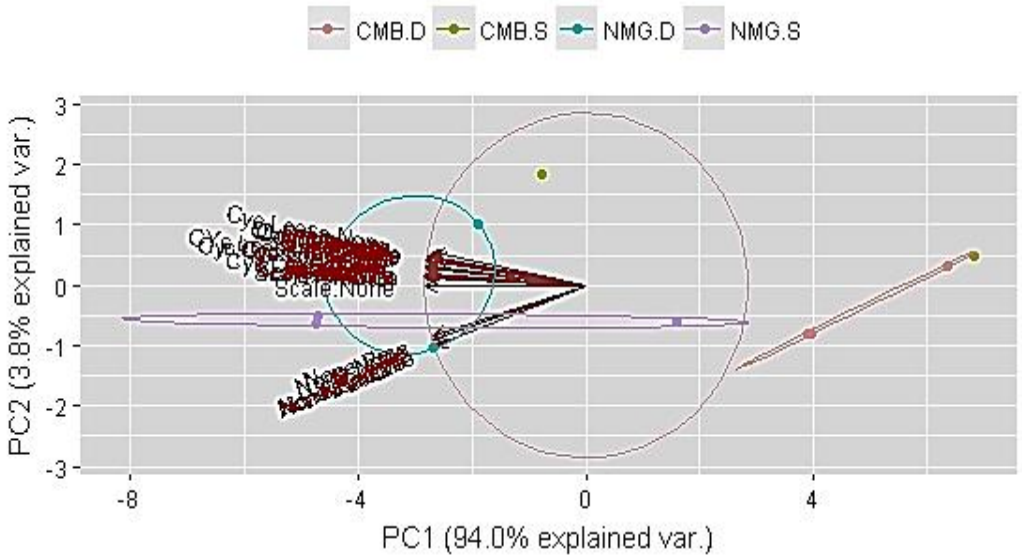

Fig. 2. Biplot graphic measuring microarray feature performance variability combining the former BS and DN procedures in discriminating DEGs in agreement with RNA-seq. *CMB.S/D NMG.S/D* are for ex CombiMatrix and ex Roche NimbleGen custom microarray platform designs based on short and/or long single and/or multiple probe set per gene model transcript respectively

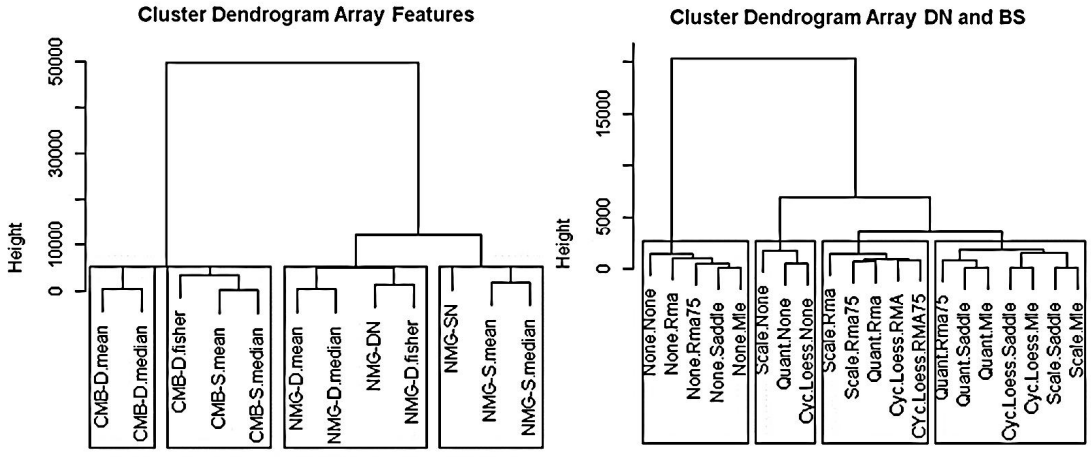

Fig. 3. Euclidean distance hierarchical clustering survey applied on both processed (i) microarray design features and (ii) BS+DN arrangement methods

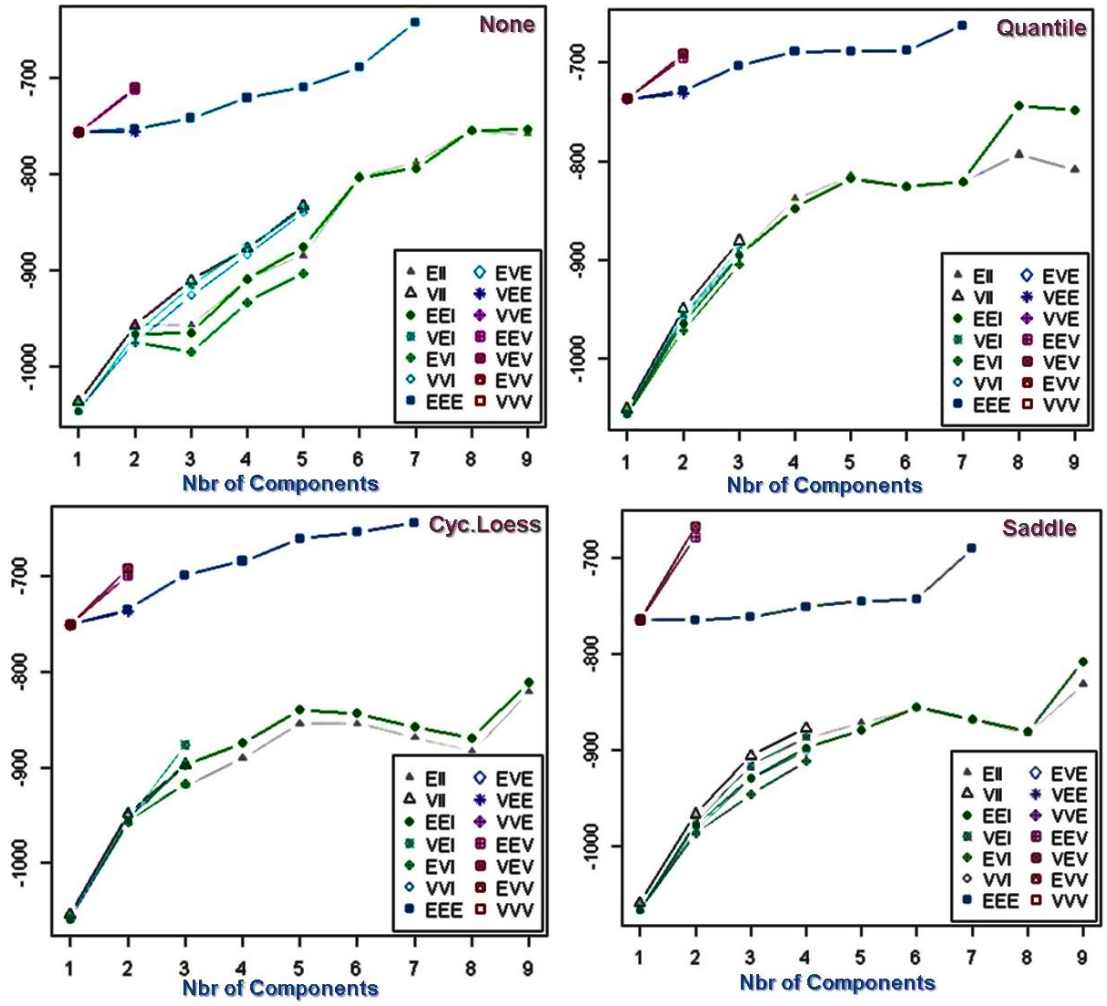

Fig. 4. BIC "*mclust*" clustering graphic by multivariate mixture survey assessing microarrays BS+DN arrangement assuming RNA-seq approach as reference

3.5 Detailed Evaluation of Inter-Microarray Features Data Variability by Processing Data Density Dispersion and Descriptive Statistic Survey

We performeddensity dispersion analysis as regards microarray performance in gene expression differential data by processing 20 BS+ DN arrangement (see material and methods chapter) assuming RNA-seq as reference. The present analysis exhibited an apparent stability of Quantile and Cyclic Loess normalisation system in calling DEGs in agreement with RNA-seq. The same analysis censured Null microarray gene expression data normalisation, since induces a relative high data variability in the present processed comparative gene expression profiling analysis (Fig. 5). In the other word, normalisation procedure in array pre-processing analysis is strongly required to enhance both quality and quantity information as regards microarray gene expression differential survey. The present analysis showed that normalisation procedure in microarray gene expression data analysis improved the number of detected and/or candidate differentially expressed genes (p-value <0.05). Indeed, detected DEGs in agreement between microarrays and RNA-seq, when microarray gene expression data normalisation methods were correctly applied, ranged from 3065 to 2861 against 2223 DEGs for non-normalized expression data (p-value<0.05) (Table 1). Interestingly, previous evoked

apparent stability displayed by Quantile and Cyclic Loess normalisation methods in calling DEGs in concordance with RNA-seq next generation sequencing approach was partially confirmed by descriptive statistic results reported in Table 1. In fact, Quantile normalization method exhibited a high stability with respect to the other analysed normalisation procedures and discriminated more DEGs in agreement with next generation sequencing approach (Table 1 and Fig. 5). Moreover Fig. 5 suggested a high data dispersion as regards Scale normalisation methodology with opposed to those of Cyclic Loess. This result was confirmed by our processed descriptive statistic evoking a relative difference between variance parameter refereed to the latter's (Table 1). Taking together the present findings suggested the stability of quantile normalisation system in array gene expression differential analysis and exhibited both Quantile and Cyclic Loess normalisation methodologies as discriminating a considerable number of differentially expressed gene in agreement with RNA-seq as opposed to array feature without any normalisation procedure (p-value <0.05).

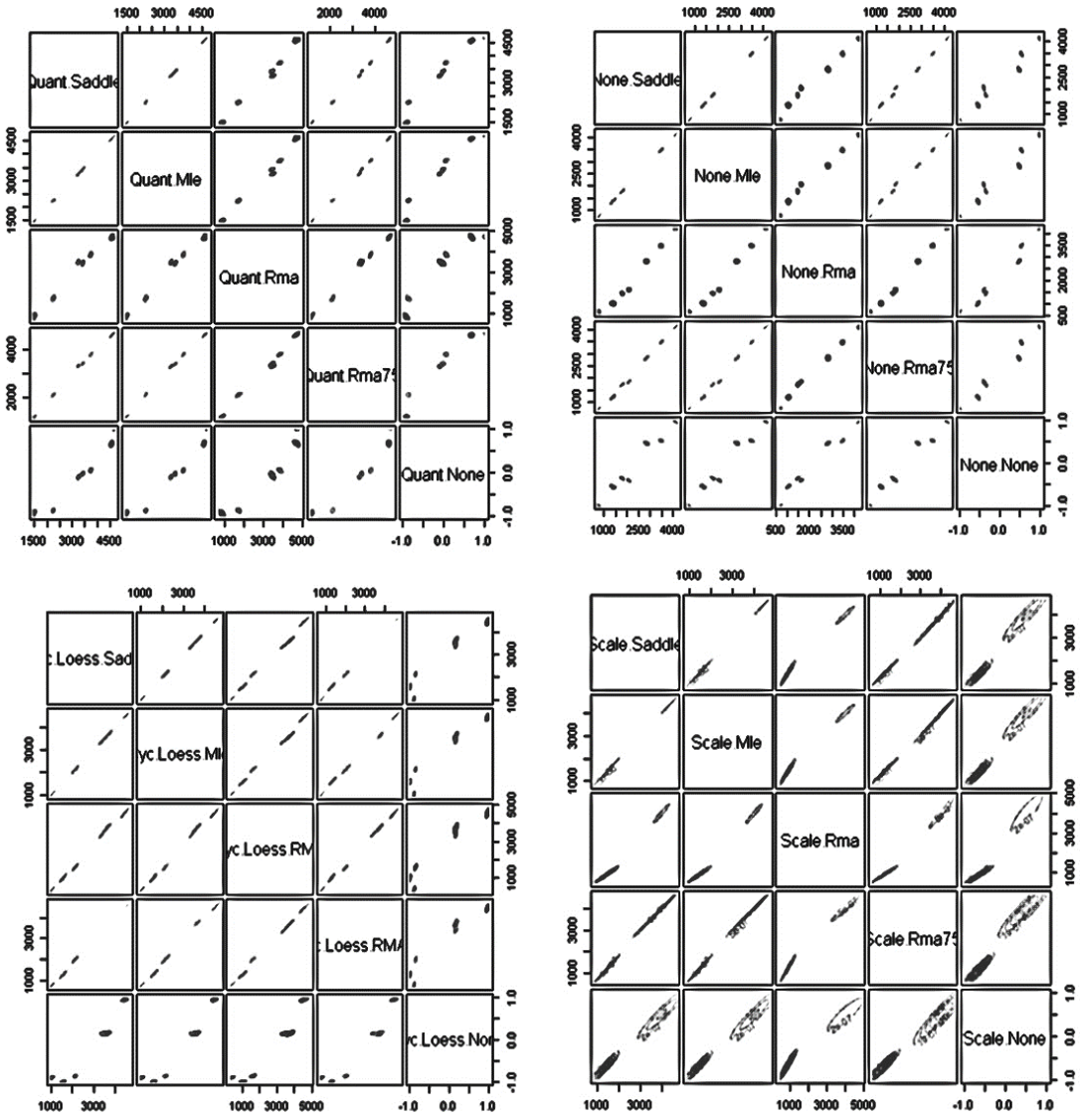

Fig. 5. Multivariate analysis assessing data variability as density dispersion as regards analysed microarray designs by processing array BS+DN arrangement

Table 1. Descriptive statistical analysis measuring the performance of both ex-CMB and ex-NMG microarray platforms BS+DN arrangement.

	Scale (Saddle, Mle, Ram, Ram75, None)	Quantile (Saddle, Mle, Ram, Ram75, None)	Cyc.Loess (Saddle, Mle, Ram, Ram75, None)	None (Saddle, Mle, Ram, Ram75, None)
Mean (DEGs)	2860.3	3065.28	2971.1	2222.97
Maximum (Number of DEGs)	4869	4875	4890	4305
Minimum (Number of DEGs)	453	673	283	617
Standard Deviation (Log DEGs)	0.68	0.56	0.67	0.60
Variance (Log DEGs)	0.45	0.29	0.41	0.38

Table 2. Descriptive statistic assessing intra-array designs data variability (detected DEGs in agreement with RNA-seq) by combining DN+BS methods

DN+BS methods	Statistical parameters	*CMB.S	*CMB.D	*CMB.S fisher	*NMG.S	*NMG.D	*NMG.D fisher
Scale (DN)+ Saddle, Mle, Rma, Rma75, Null (BS)	Mean of DEGs	1402.3	868.1	1875.4	3318.46	4469.66	4543
	Log. Variance	**0.08**	**0.17**	**0.05**	**0.00**	**0.00**	**0.00**
	Log. Standard Deviation (SD)	0.28	0.4	0.22	0.06	0.04	0.06
Quantile (DN)+ Saddle, Mle, Rma, Rma75, Null (BS)	Mean of DEGs	1469.7	1441	2145	3492.4	4551.87	4684.2
	Log. Variance	**0.09**	**0.21**	**0.02**	**0.00**	**0.00**	**0.00**
	Log. Standard Deviation (SD)	0.29	0.43	0.13	0.06	0.03	0.05
Cyclic-Loess (DN)+ Saddle, Mle, Rma, Rma75, Null (BS)	Mean of DEGs	1507.8	1073.4	2042	3556.53	4367.67	4676.2
	Log. Variance	**0.08**	**0.48**	**0.02**	**0.00**	**0.00**	**0.00**
	Log. Standard Deviation (SD)	0.27	0.66	0.14	0.04	0.03	0.05
Null (DN)+Saddle, Mle, Rma, Rma75, Null (BS)	Mean of DEGs	1302.6	735.5	1734.4	2228.45	3362.26	4092.8
	Log. Variance	**0.03**	**0.00**	**0.01**	**0.42**	**0.19**	**0.001**
	Log. Standard Deviation (SD)	0.16	0.07	0.1	0.004	0.002	0.04

**CMB.S and CMB.D: custom microarray designs based on single and/or multiple short oligo probe set (35-40 mer) per gene model transcript. NMG.S and NMG.D: custom microarray designs based on both single and multiple long oligo probe set (60 mer) per gene model transcript. Fisher indicates probe average by the Fisher probability method.*

3.6 Assessment of Intra-Microarray Variability *Vis-à-vis* of Detected Differential Expressed Genes (DEGs) in Agreement with RNA-seq Combining Microarray Gene Expression Data Normalisation (DN) and Background Subtraction (BS)

Assessment of intra-microarray data variability by processing previous evoked microarray DN+BS arrangement based on the R *limma* package, showed a relative stability of all analysed microarray features in discerning DEGs in agreement with RNA-seq. However, this study suggested the high susceptibility of microarray designs based on short oligonucleotide probe set per gene model transcript as regards applied DN+BS methods (Table 2). Indeed, array design based on multiple short probes set per gene model target exhibited a consistent versatility performances as regards considered and applied microarray DN+BS arrangement (Table 2) with respect to the other analysed array platforms. Nevertheless, microarray designs based on long oligonucleotide probe (60mer) per gene target claimed to be more stable with respect to those based on short oligo probe (35-40 mer) when we combined all microarray *limma* package background subtraction and expression data normalisation methods (Table 2). Apparent increasing of data variability has been observed between Null normalisation method and quantile, scale, and cyclic loess normalisation methods. This observed data variability could be explain by the fact that normalization in microarray experiment increases the number of differential gene expression candidates improving the quality and/or quantity of the experimentation results (Table 2). Then, the present results confirmed the relationship between microarray experimentation data pre-processing step and the quality of the results. Considering as a whole, our findings suspected a selective impact of microarray DN+BS methods; meaning depending on the microarray feature probes size.

4. DISCUSSION

In microarray experiments, removal of systematic variations resulting from array preparation or sample hybridization conditions is crucial to ensure sensible results from the ensuing data analysis. Then, normalizes expression intensities so that the intensities or log-ratios have similar distributions across a set of arrays. Linear models for microarray and RNA-seq data analysis (*limma*) implements a range of normalization methods for spotted microarrays. Smyth and Speed [21] describe some of the most commonly used methods. The methods may be broadly classified into methods which normalize the M-values for each array separately (within-array normalization) and methods which normalize intensities or log-ratios to be comparable across arrays (between-array normalization). Indeed, for single-channel arrays (our treated cases), within array normalization is not usually relevant and so normalize between microarray platforms is the sole normalization step.For single channel microarray data, the scale, quantile or cyclic loess normalization methods can be applied to the columns of data. So, scale normalization method scales the columns to have the same median [22-23], while quantile and cyclic loess normalization was originally proposed by Bolstad et al (2003) for Affymetrix-style single-channel arrays [24]. Quantile normalization forces the entire empirical distribution of each column to be identical. Cyclic loess normalization applies loess normalization to all possible pairs of arrays, usually cycling through all pairs several times [24-25]. Cyclic loess is slower than quantile, but allows probe-wise weights and is more robust to unbalanced differential expression. Also, array background correction result to be a fundamental step managing custom microarray platforms. The default background correction action is to subtract the background intensity from the foreground intensity for each spot on array. Usually *limma* package be default integrate array background correction process with normalize within arrays function and/or script. The present study processed the performance of our previous developed custom array designs based on the ex-Combimatrix and ex- Roche NimbleGen microarray platforms by combining and/or integrating R *limma* package normalization and background correction methods with the purpose to discern the agreement between the latter's (processed microarray platforms) and RNA-seq approach in gene expression differential analysis. This analysis was based exclusively on hierarchical clustering and principal component analysis (PCAs) and suggested the first component of above mentioned PCA survey, as able to fully explain the present analysed data variability (Fig. 1). However, standard deviation associate to our processed principal components (PC) by evaluating microarray BS+DN arrangements, in calling significantly differentially expressed genes (DEGs) in agreement with RNA-seq, ranged from 4.34 to 1.577e-15, while,

the cumulative proportion analysis explaining that variance (data variability) suggested the first PC account for more than 90% (Fig. 1). This result was confirmed by our developed biplot graphic by measuring performance variability as regards the analysed and/or processed microarray features (Fig. 2). This analysis suggested the highest stability of microarray designs based on multiple long and/or short oligonucleotide probe set per gene model target as opposed to those based on single probe per gene model transcript. Moreover, the present analysis enhanced and confirmed previous suspected instability regarding microarray design based on short single probe set per gene model transcript (CBM-S) [26]. We therefore showed that the use of different oligo nucleotides probe per transcript model by using adequately microarray DN+BS methods, provided a stable measure of transcript abundance and/or intensity in gene expression differential analysis. However, our findings showed that disregarding applied microarray gene expression data normalization (DN) and background subtraction (BS), the present analysed array features exhibited heterogeneity behaviours among themselves [16]. By contrast, the same analysis seems to favour microarray platforms clustering based on data normalisation (DN) procedure with respect to those based on background subtraction (BS) (Fig. 3). Then, Euclidean distance clustering analysis suggested the preponderance of microarray gene expression data normalisation as opposed to those of their background correction. This could may be explain the integration and/or combination by the linear models for microarray and RNA-seq data analysis (*limma*) package between array background subtraction (BS) function and those of gene expression data normalization (DN) methods under the "*normalizeWithinArrays*" application, since the latter performs array background correction by default [27]. Latter, we focused on a multivariate clustering analysis provided by "*mclust*" package, by selecting the optimal model, performing hierarchical clustering for parameterized Gaussian mixture models by favoring the model and number of clusters with the largest BIC parameter [19-20]. This clustering analysis recommended the needed of data normalization for our processed microarray platforms, since largest BIC value were calculated for Quantile, Cyclic Loess and Scale normalization methods respectively as oppose to Null normalization parameter (Fig. 4). The same analysis suspected the stability as well the tolerance of quantile normalization method as regard the*limma* package processed background correction methods. Quantile normalization is routinely used in the treatment of both oligonucleotide and cDNA microarray data, even though there might be some loss of information in the normalization process. We recognize that the ideal normalization, if it ever exists, would aim to keep the maximal amount of gene profile information with the lowest possible noise. Then, Hu J. et al. (2007) proposed a valuable enhancement to quantile normalization, and demonstrate through three Affymetrix experiments that the enhanced normalization can result in better performance in detecting and ranking differentially expressed genes across experimental conditions [28]. Next, our performed density clustering analysis (Fig. 5) and as well descriptive statistical survey (Table 1) proposed quantile normalization method applied to our developed microarray designs as exhibiting the highest agreement (sensitivity) with RNA-seq approach in calling accurately DEGs (p-value<0.05). Then, the lower variability observed between microarray pre-processed data by applying the background correction including quantile normalization system confirmed the advantage of merging all developed background correction systems of the R *limma* package with the quantile normalization method in gene expression differentially analysis. Taking together, the present analysis recognizing the needed of microarray gene expression data normalization in performing gene expression profiling survey, suggested this data pre-processing step as mandatory for improving array gene expression data quality and quantity (Fig. 5 and Table 2). Also this study suspect a relative selective effect of the combination of processed microarray gene expression data normalization (DN) and background correction (BS) methods on the performance of our developed microarray designs, since combination between analysed BS and DN methods applied to the same array exhibited contrasting results (Table 2). However, the same analysis supported an improvement of the quality of microarray gene expression analysis when we correctly applied data normalisation and background correction methods. In the other words, good concordance and/or agreement was observed between both microarray and RNA-seq platforms in calling DEGs candidates when analysed microarray platforms were rigorously submitted to normalization as well as to background correction procedures as opposed to Null normalization and background correction (Table 2 and Fig. 5). Moreover, our finding showed microarray design based on short oligo

probes as more versatile as opposed to microarray design based on long probe replying to the *limma* package DN+BS combination. However, quantile and scale normalisation methods seem to stabilize array design based these oligonucleotide probes (short oligo probe) as opposed to cyclic loess method. Experience with microarray data has repeatedly shown that normalization is a critical component of the processing pipeline, allowing accurate estimation and detection of differential expression (DE) [24]. The aim of normalization is to remove systematic technical effects that occur in the data to ensure that technical bias has minimal impact on the results.

5. CONCLUSION

The importance of microarray data normalization and background procedures in prelude to genomics and transcriptomics studies have been fully discussed and continued to captivate researcher community attention. The particularity of our study was to weight the performance of our previous developed microarray designs processed by all*limma*package background correction (BS) and data normalisation (DN) combination methods with the purpose to weigh the agreement between our previous developed microarray design strategies and RNA-seq approach in a comparative gene expression differential analysis. The present study confirmed the necessity of both gene expression data normalization and background correction procedures in microarray analysis data pre-processing step. Also, our findings preconized the preponderance of normalization methods with respect to background correction in the present performed microarray gene expression differential analysis. Finally, our results, exhibited quantile normalisation method as more tolerable as regards to applied *limma* package background correction systems with respect to Cyclic Loess and Scale normalisation methods.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Suárez E, Burguete A, Mclachlan GJ. Microarray data analysis for differential expression: A tutorial. P R Health Sci J. 2009;28(2):89-104. PubMed.
2. Calza S, Pawitan Y. Normalization of gene-expression microarray data. Methods Mol Biol. 2010;673:37-52. DOI: 10.1007/978-1-60761-842-3_3. PubMed.
3. Joseph DC, Ton Z. Microarray analysis of the transcriptomes as a stepping stone towards understanding biology system: Pratical consideration and perspectives. The Plant Biology Journal. 2006;45:630-650.
4. Dago DN, Alberto F, Diarassouba N, Fofana IJ, Silué S, Giovanni M, Massimo D. Probes specificity in array design influences the agreement between microarray and RNA-seq in gene expression analysis. African Journal of Sciences and Research. 2014;3(5):08-12.
5. Kerr MK. Design considerations for efficient and effective microarray studies. Biometrics. 2003;59(4):822-8.
6. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. 2002;30(4):e15.
7. Nielsen HB, Gautier L, Knudsen S. Implementation of a gene expression index calculation method based on the PDNN model. Bioinformatics. 2005;21:687–688.
8. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix Gene Chip probe level data. Nucleic Acids Res. 2003;31:e15.
9. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015;43(7):e47.
10. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013; ISBN 3-900051-07-0. Available:http://www.R-project.org
11. Evertsz EM, Au-Young J, Ruvolo MV, Lim AC, Reynolds MA. Hybridization cross-reactivity within homologous gene families on glass cDNA microarrays. Bio-techniques. 2001;31(5):1182, 1184, 1186 passim.
12. Venables WN, Brian DR. Modern applied statistics with S-PLUS Springer-varlg (Section 11.1)
13. Box G and Cox D. An analysis of transformations. Journal of the Royal

Statistical Society. Series B (Methodological). 1964;211-252.
14. Yoav B, Yosef H. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the royal Statistic Society. Series B (Methodological). 1995;57(1):289-300.
15. Li B, Dewey CN. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323. Available:https://doi.org/10.1186/1471-2105-12-323. PubMed
16. Noel DD, Alberto F, Luciano X, Antonio M, Massimo D, Giovanni M. Heterogeneity of global gene expression microarray designs in detecting differentially expressed genes. International Journal of Bioinformatics Research. 2016;7(2):349-357.
17. Yoav Benjamini, Yosef Hochberg. Journal of the royal Statistic Society. Series B (Methodological). 1995;57(1):289-300.
18. Ann H, Hari I. Fisher's combined p-value for detecting differentially expressed genes using Affymetrix expression arrays-1BMC Genomics. 2007;8(96):1-13 DOI: 10.1186/1471-2164. PubMed
19. Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association. 2002;97:611-631..
20. Fraley C, Raftery AE, Murphy TB and Scrucca L. Mclust version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation. Technical Report. 2012; No. 597, Department of Statistics, University of Washington.
21. Smyth G, Speed T. Normalization of cDNA microarray data. Methods. 2003;31:265–273.
22. Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. J. Biomed. Optics. 1997;2:364-374. PubMed
23. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Research. 2002;30(4):e15. PubMed
24. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. Bioinformatics. 2003;19:185-193. PubMed
25. Yang YH, Thorne NP. Normalization for two-color cDNA microarray data. In: D. R. Goldstein (ed.), Science and Statistics: A Festschrift for Terry Speed, IMS Lecture Notes - Monograph Series. 2003;40:403-418.
26. Cheng-Chung, CCH, Te-Tsui L, Konan P. Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression. Nucleic Acids Research. 2004;32(12):e99. DOI:10.1093/nar/gnh099.
27. Gordon KS, Matthew R, Natalie T, James W, Wei S, Yifang HT, Eliza H. Limma: Linear models for microarray and RNA-seq Data User's Guide First edition 2 December 2002 Last revised 16 October 2016.
28. Hu J, He X. Enhanced quantile normalization of microarray data to reduce loss of information in gene expression profiles. Biometrics. 2007;63(1):50-9.

Peer-review history:
The peer review history for this paper can be accessed here:
http://sciencedomain.org/review-history/20951