DOI: 10.21276/ijabpt.2016.7.3.16

INTERNATIONAL JOURNAL OF APPLIED BIOLOGY AND PHARMACEUTICAL TECHNOLOGY

www.ijabpt.com

ISSN : 0976-4550

Volume-7, Issue-3, July-Sept-2016 **Coden IJABFP-CAS-USA** **Copyrights@2016**

Received: 25th June 2016 ***Revised: 10th July 2016*** ***Accepted: 14th July 2016***

Research article

ROC CURVE ASSESSING MICROARRAY OLIGONUCLEOTIDES SIZE CALLING DIFFENTIALLY EXPRESSED GENES BY HIGH-THROUGHPUT SEQUENCING APPROACH

Dago Dougba Noel*[1, 2], Lallié Hermane Désiré M.N[1], N'Goran Kouamé Edouard[1], Mori Antonio [3], Diarrassouba Nafan[1], Massimo Delledonne[2] and Giovanni Malerba[3].

[1]UFR Sciences Biologiques Université Péléforo Gon Coulibaly BP 1328 Korhogo, Ivory Coast.
[2]Department of Biotechnology University of Verona Strada le Grazie 15, Cà-vignal 1, Italy.
[3]Department of Neurological, Biomedical and Movement Sciences University of Verona, Strada Le Grazie 8, 37134, Verona, Italy.

ABSTRACT: Even if RNA-seq high-throughput sequencing technology has been nowadays adopted by many researchers in their genomic and transcriptomic studies, microarrays technologies remained capable tool assessing gene expression differential analysis because of their well established bioinformatics and biostatistics methods as well as their wide usage in transcriptomic and molecular diagnosis process. The purpose of this article in contrast to other's is to assess the relationship between custom microarrays gene expression manufactures and their oligonucleotide probe set design strategies discriminating accurately significantly differentially expressed genes (DEGs) by high-throughput sequencing RNA-seq approach performing Receiver Operating Characteristics (ROC) curve analysis exclusively. For the present study, our previous developed microarray design strategies based on long and/or short single (unique replicate) and/or multiple oligonucleotide probes per gene model transcripts have been recycled. Indeed, the aptitude of analysed microarrays calling DEGs has been measured through ROC curve analysis exclusively assuming RNA-seq as reference because of their high performance and high dynamic range in gene expression profiling survey. Then, ROC curve analyse evidenced microarray manufactures based on short probes size as sturdily influenced by the sort of array probe set design strategy with respect to those based on long oligonucleotide probes per gene model transcript (p-value ≤ 0.05). Moreover, our findings showed that optimizing and monitoring type I statistical error (monitoring false discovery rate parameter; FDR$\leq$0.01) of microarray platforms detecting significantly modulated genes, allowed to improve the specificity of the latter's (microarray probe set design) calling truly DEGs candidates especially for microarray design strategies based on multiple probes per gene unit. In conclusion, the present study suggested (i) high heterogeneity of microarray manufactures based on short oligonucleotide probes in gene expression differential survey as opposed to those based on long oligonucleotides, and (ii) showed that disregarding microarray platforms as well as oligonucleotide probes size, microarray manufactures based on multiple probes per gene model transcript exhibited better genes signal detection as well as high agreement with RNA-seq discriminating significantly DEGs reducing false discovery rate (FDR) ratio, suggesting the latter as satisfactory parameter integrating both microarrays and RNA-seq tools transcriptomic data in gene expression profiling survey.

Key words: Microarrays, RNA-seq, ROC curve, Oligonucleotides, False Discovery Rate (FDR), Differentially Expressed Genes (DEGs).

*Corresponding author: Dago Dougba Noel, [1]UFR Sciences Biologiques Université Péléforo Gon Coulibaly BP 1328 Korhogo, Ivory Coast Email: dgnoel7@gmail.com

Copyright: ©2016 Dago Dougba Noel. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

INTRODUCTION

Microarrays are a powerful tools for monitoring gene expression change under various conditions and have been widely used for genome wide transcriptional analysis (De Risi J.L et al, 1997; Gao H.Y et al, 2004; Wan X. et al, 2004; Wodicka L et al, 1997), discovery of gene function (Hughes T.R et al, 2000), cancer study (Ochs M.F et al, 2003; Petricoin E. F et al, 2002), neuroscience (Luo Z. and Geschwind D.H, 2001), discovery of drug target and environmental studies (Rhee S.K et al.2004; Taroncher-Oldedburg et al, 2003; Tiquia S.M et al, 2004). While the multiplicity of microarray platforms offered an opportunity to expand the use of the methodology and make it more easily available to different laboratories, the comparison and integration of data sets obtained with different microarray platforms as well as with other genes expression measurement tools like high throughput sequencing (RNA-seq) technology has been partially solved (MAQC, 2006; Nalpas N.C et al, 2013; SEQC/MAQC-III consortium, 2014). Sources of diversity arise from the technology features intrinsic to chip and/or platform manufacturing, from the protocols used for sample processing, from detection systems, as well as from approaches applied to data analysis. On one hand, the combined use of multiple platforms can overcome the inherent biases of each approach, and may represent an alternative that is complementary to RT-PCR for identification of the more robust changes in the gene expression profiles on the other hand, the comparison of data generated using different platforms may represent a significant challenge, particularly when considering very different systems (i.e. microarray and RNA-seq). Indeed, publication of studies with dissimilar or altogether contradictory results, obtained using different microarray platforms as well as different gene expression approaches to analyze identical RNA sample, has raised concerns about of the reliability of these technology. The Microarray Quality Control (MAQC) project was initiated to address these concerns, as well as other performance and data analysis issues generating a rich data set that, when appropriately analyzed, reveals promising result regarding the consistency of microarray data between laboratories and cross platforms especially measuring gene expression level between the latter and RNA-seq (SEQC/MAQC-III consortium, 2014). Moreover, searching for determinants of a phenotype using gene expression levels requires suitable exposure of the genome coupled with reasonable reproducibility, accuracy and sensitivity in the technology employed. These limitations matter less if microarrays are used for screening because changes in gene expression can be verified independently. However, the stakes were raised when microarrays were suggested as a diagnostic tool in molecular disease classification (Wang Y. et al, 2005) because regulatory agencies, such as the Food and Drug Administration (FDA), require solid, empirically supported data about the accuracy, sensitivity, specificity, reproducibility and reliability of diagnostic techniques. An ideally precise technique would have all measurements exactly equal (zero variance). Accuracy and precision are completely independent. A technique can be accurate but not precise (the mean of several measurements is close to the actual value but the individual measurements vary considerably), precise but not accurate (the individual measurements are close to each other but their mean is far from the actual value) neither or both. If a result is both accurate and precise, it is valid. Then, specificity in the context of DNA microarrays, refers to the ability of a probe to bind to a unique target sequence. A specific probe will provide a signal that is proportional to the amount of the target sequence only. A non-specific probe or probe set will provide a signal that is influenced by the presence of other molecules. The specificity of a probe (probe set) can be diminished by cross-hybridization, a phenomenon in which sequences that are not strictly complementary according to the Watson Crick rules bind to each other. Microarray experiment generally depend on the hybridization intensity measurement for an individual probe to infer a transcript abundance level for a specific gene. This relationship raises several difficult issues, including which gene correspond to which probe or probe set, and how sensitive and specific is the probe or the probe set. The Receiver Operating Characteristics (ROC) graphs are a useful technique for organizing classifiers and visualizing microarray probe performance (Dago N., 2012; Dago D.N et al, 2014^{a}). ROC graphs were commonly used in medical decision making and in recent years have been increasingly adopted in the machine learning and data mining research communities. Moreover, they have long been used in signal detection theory to depict the tradeoff between hit rates and false alarm rates of classifiers (Egan, 1975, Swets et al, 2000). It is noteworthy to underline that ROC analysis have been extended for use in visualizing and analyzing the behavior of diagnostic and molecular biology systems (Swets, 1988) allowing medical decision making community having an extensive literature on the use of ROC graphs for diagnostic testing (Zou, 2002). Swets, Dawes and Monahan (2000) brought ROC curves to the attention of the wider public with their Scientific American article. So, a common method applying ROC analysis is to calculate the area under the ROC curve, abbreviated AUC (Bradley, 1997; Hanley and McNeil, 1982). Since the AUC is a portion of the area of the unit square, its value will always be between 0 and 1. However, because random guessing produces the diagonal line between (0; 0) and (1; 1), which has an area of 0.5, no realistic classifier should have an AUC less than 0.5. The AUC has an important statistical property since is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. Indeed, this is equivalent to the Wilcoxon test of ranks (Hanley and McNeil, 1982).

Taking together and considering our previous investigative results comparing both RNA-seq and microarrays approaches measuring transcript expression level in gene expression profiling analysis (Dago D.N et al, 2014[a]), we assessed microarray oligonucleotide probes size and probe set design strategies influencing accurate calling of statistically significantly DEGs candidates by using ROC curve survey analysis exclusively, paying attention setting false discovery rate (FDR) statistical parameter threshold setting since traditional transcriptomic approaches studies testing many hypotheses use this estimation for inference, and often for classification technique that have thousands of transcript.

MATERIAL and METHODS

The same samples of Zenoni S. et al (2010) were used for microarray and RNA-seq experiments, corresponding of grapevine (*Vitis vinifera*) berry tissue at veraison and ripening stages. Results of differential analysis from each microarray manufactures were then compared with the results obtained from high throughput RNA-seq by ROC curve analysis exclusively assessing microarray probe set behaviors calling accurately DEGs since ROC graphs are a very useful tool for visualizing and evaluating classifiers. They are able to provide a richer measure of classification performance than scalar measures such as accuracy, error rate or error cost. Because they decouple classifier performance from class skew and error costs, they have advantages over other evaluation measures such as precision-recall graphs and lift curves (Tom F. 2005).

RNA Preparation.

Sample of *Vitis vinifera* at the development stages of veraison and ripening were collected as reported in Zenoni et al (2010) and total RNA has been extracted as described in Zamboni A. et al (2008). RNA amount and integrity were essayed by Nanodrop 2000 instrument (Thermo Scientific) and a Bio-analyzer Chip RNA 6000 (Agilent), respectively.

Microarray Design Strategies and Hybridization Experiment.

Grape microarray (Custom-Array) design strategies based on short (35-40bp) and long (60bp) probe per gene model transcript and hybridization experiment process have been reported by Dago Noel (2012).

Microarray Data Preprocessing.

Data preprocessing comprises computer methods to adjust for the ambient intensity (background subtraction) across the array and to remove sources of variation between arrays of non-biological origin (data normalization). Microarray data were preprocessed using (i) Normexp background subtraction method and (ii) quantile data normalizing method available in the library package *limma* version 3.10.3 (Smyth, G 2004). Expression (i.e. intensity) values of each gene were expressed applying mean values of the probe set signals of the same gene across each microarray platform.

RNA-seq Experiment.

RNA-Seq data used in this study was generated during Sara Zenoni *et al* (2010) study. To summarize, two technical replicates each for both ripening and veraison grape development stages 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 (8.4 fold draft sequence of the Pinot Noir 40024 genome) and then analyzed to measure gene expression levels. Here, read count was performed using the packages RSEM V1.1.21 (Li B, 2011). Differential expression analysis was conducted using the computer package DESeq (Version 1.6.1, http://bioconductor.org/packages/release/bioc/html/DESeq.Html). RNA-seq raw data are available at SRA009962 (or data can also be accessed on Genome Browser at URL http: //ddlab.sci.univr.it/cgi-bin/gbrowse/grape).

Differential Gene Expression Analysis.

Differential expression analysis between 2 grape development stages was performed by comparing arrays. This analysis were conducted by applying linear models on the log-expression values and then an empirical Bayes moderated t-statistics on each gene by using both *lmFit* and *eBayes* functions of *limma* package (Smyth, G 2004) from R software (version 3.10.3). The False Discovery Rate (FDR) suggested by Benjamini and Hochberg (1995) was adopted to control the FDR since multiple comparisons were computed. A gene was considered as differentially expressed when showing a mean difference of the expression value greater than or equal to two folds between the 2 considered berry development stages at a FDR$\leq$0.05. Only genes shared among all platforms (microarrays and RNA-seq) were processed by the present ROC curve analysis evaluating microarray probe set designs performance discriminating accurately DEGs by RNA-seq approach.

Microarray Oligonucleotide Probes Specificity and Sensitivity Analysis (ROC Analysis)

We estimated the accuracy, the sensibility and the specificity of each analysed microarray manufactures probe detecting DEGs in gene expression differential analysis assuming RNA seq gene expression data as reference (Dago N, 2012). Then, we summarized the scheme of this examination and/or analysis in Table 1. In statistics, a receiver operating characteristic (ROC), or ROC curve, is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied.

The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, or recall in machine learning. The false-positive rate is also known as the fall-out and can be calculated as (1 - specificity). Hence, we reported both sensitivity and specificity mathematical explanation as following: (i) Sensitivity: measures the proportion of actual positives which are correctly identified as such. Sensitivity= Number of true positive ÷ (Number of true positives + Number of false negatives); (ii) Specificity: measures the proportion of negatives which are correctly identified. Specificity=Number of true negatives ÷ (Number of true negatives + Number of false positives).

RESULTS

Comparison between microarrays designs based on short oligonucleotide probes and RNA-Seq in gene expression differential survey.

In total 17674 genes were analysed between microarrays and high throughput sequencing RNA-seq platforms. Statistically significantly DEGs have been discriminated at a FDR≤0.05 assessing genes fold change between both repining and veraison grape development stages (see material and methods chapter). Microarray design based on single short oligonucleotides per gene model transcript, detected 1.46 fold more DEGs as opposed to the same manufacture based on multiple short probes per gene model transcripts (Figure 1). Furthermore, disregarding microarray design strategies used, Venn diagram suggested a high agreement between microarrays and RNA-seq platforms instead of between above mentioned microarray platforms. Although, microarray manufacture with probe set design strategy based on single replicate oligonucleotide per gene unit discriminated more DEGs in comparison to microarray design strategy based on multiple short probes per gene model transcript. However, the same analysis showed that microarray manufacture with multiple short probes per gene model transcript exhibited better agreement with RNA-seq calling DEGs candidates (Figure 1). Moreover, the number of DEGs discriminated exclusively by microarray based on single short replicate oligonucleotide probe per gene unit, resulted 8.86 fold more than those detected by the same manufacture with multiple oligonucleotide probes (Figure.1). Taking together, these results suggested heterogeneous replies of above mentioned and analysed microarray probe set design strategies in gene expression differential profiling analysis.

Comparison between microarrays manufactures based on long oligonucleotide probes and RNA-seq in gene expression differential profiling analysis.

The present analysis showed that microarray design strategy with multiple long oligonucleotide probes per gene model transcript detected 1.77 fold more DEGs than the same manufacture with single probe per gene unit (Figure 2). It is interesting to underline that the present scheme and/or situation contrast with the previous one (see Figure 1). Generally, microarray design with long probe set per gene unit called more DEGs candidates in comparison with microarray manufacture based on short oligonucleotide (Figures 1 and 2).These results suggested and confirmed high sensitivity of long probes with respect to short oligonucleotide probes measuring gene expression intensity. Even if microarray manufacture based on multiple long oligonucleotide probes per gene model transcript discriminated more DEGs as previously mentioned, it is noteworthy to discern that around 25% of these detected DEGs were not recognized as such by RNA-seq approach (reference), against 16% for the same array manufacture with single long replicate oligonucleotide probe per gene unit (Figure 2). This result supposed high sensitivity as well as low specificity of the latter in gene expression profiling analysis.

ROC Curve analysis assessing the aptitude of microarray designs with short oligonucleotide probes calling DEGs by RNA-seq.

We were interested to evaluate the capacity of microarray design strategies based on short oligo, calling accurately DEGs through RNA-seq approach. So, combining both specificity and sensitivity parameters of these microarray manufactures, we showed that probe set exhibiting multiple oligonucleotide probes detected better DEGs signal in agreement with high throughput sequencing approach as opposed to microarray design strategy with probe set that based on single replicate oligonucleotide per gene model transcript. Indeed, area under curve (AUC) assessing the ability of microarray manufactures based on short probes per gene unit detecting right gene expression signal, was higher for design strategy with multiple oligonucleotide probes per gene model transcript (Figure 3). In other word microarray manufacture with short multiple oligonucleotide probes, exhibited high specificity with respect to microarray design strategy based on short single replicate probe per gene model transcript. Then, merging the present results with the previous one (Figure 1), we showed that (i) microarray design strategy that exhibited single replicate probe set per gene unit displayed a good sensibility (low specificity) while (ii) microarray manufacture with multiple short oligonucleotide probes per gene unit tend to display high specificity (low sensitivity).These results confirmed the strong inconstant performance of these two microarray design strategies and/or manufactures in gene expression profiling analysis (Figure 1 and 3).

ROC Curve analysis weighing the propensity of array manufactures based on long probes calling DEGs by RNA-Seq.

ROC curve analysis based on area under curve (AUC) parameter evidenced weak difference between the two analysed microarrays manufactures based on long oligonucleotide probes per gene model transcript. In fact microarray design strategy based on multiple long oligonucleotide probe set resulted 0.81/0.75 ~1 fold more specific than those displaying single long replicate probe set per gene model transcript (Figure 4), while the same ratio estimation comparing the two microarray manufactures with short oligonucleotide probes was 0.70/0.57 ~1.32 (Figure 3). Considering as a whole, the present results showed that microarrays with long oligonucleotide probe set were weakly influenced by the nature of microarray design strategy. However, microarray manufacture based on multiple long probes per gene model transcript exhibited a discrete high specificity with respect to those based on single long replicate probe per gene model transcript (p-value≥0.05). Integrating these results with previous one (Figure 2), it is emerged that microarray design and/or manufacture with multiple long oligonucleotide probes per gene unit exhibited, high sensitivity (p-value≤0.05) as well as relative high specificity (p-value≥0.05) than the same array with single replicate probe set per transcript. Furthermore, these results highlighted the substantial difference between microarray platforms based on both (i) long and (ii) short oligonucleotide probes assessing DEGs proportion in gene expression profiling analysis.

Relationship between detected DEGs False Discovery Rate parameters and gene expression levels and ROC area under curve (AUC) ratio.

Our previous analysis highlighted the relationship between gene expression level and microarray aptitude discriminating DEGs recognized as such by RNA-seq. Here we investigated the relationship between (i) gene expression levels, (ii) statistical inference survey calling DEGs by monitoring FDR parameter and (iii) area under curve (AUC) from ROC curve analysis, aiming to evaluate the ability of microarray probe set design manufactures in gene expression profiling analysis. Our results showed that all analysed microarrays probe set design strategies displayed a descent agreement with RNA-seq technology at lower false discovery rate ratio (Table 2). This result suggested that adjustment of statistical parameter processing gene expression differential analysis can help obtaining high true positive ratio as well as reducing false positive signal calling DEGs candidates. Moreover, we evidenced that stringent adjustment of type I statistical error probability value resulting in a FDR value threshold ≤ 0.00001 (arbitrary value) allowed all analysed microarray probe design strategies to exhibit a quite similar performance with RNA-seq calling accurately DEGs candidates displaying an AUC mean value ≥0.90 (Table 2).Taking together, these results suggested microarrays as valuable tool performing gene expression profiling analysis suggesting that adequate false discovery rate interval threshold setting can help and improve both microarray and RNA-seq data integration in gene expression survey.

Performance assessment of microarray oligonucleotides size calling DEGs handling statistical False Discovery Rate parameter.

Next we were interested to assess analysed microarray manufactures performance in gene expression differential analysis monitoring DEGs false discovery rate parameter. Then, for this purpose and basing on previous results (Table 2), we arbitrary choose FDR ≤0.001 as parameter threshold assessing microarray ability calling accurately DEGs. As previously mentioned, disregarding considered microarrays manufacture as well as oligonucleotide size, all analysed microarrays displayed a good agreement with RNA-seq high throughput sequencing approach rejecting DEGs targeted as false positive (Table 2 and Fig. 5). Both panel B and D in figure 5 suggested that microarray designs based on multiple long and/or short oligonucleotide probes per gene model transcript, at FDR ≤ 0.001 (probability committing type I error ≤0.001) exhibited the same performance than RNA-seq platform discriminating DEGs. These results suggested highest true positive rate and/or sensitivity (~1) as well as lowest false positive rate and/or specificity (~ 0) of microarray manufactures with probe set including multiple oligonucleotide probes per gene unit. Moreover, it is interesting to note that AUC ratio values between microarray manufactures based on (i) long (~1) and (ii) short (~ 1.32) oligonucleotide probe set remained unchanged committing type I error at 1‰ (FDR≤0.001), when compared with previous results where FDR thresholds were set at 0.05 (see Figure 3 and 4; FDR ≤0.05).This result supported that stringency of statistical parameters in microarray gene expression analysis cannot improve as well as influence the former's intra platform gene expression analysis results. In other words, the present analysis highlighted and confirmed the heterogeneous behaviours of microarray manufacture based on short oligonucleotide in gene expression profiling analysis as opposed to those established on long oligonucleotide probe.

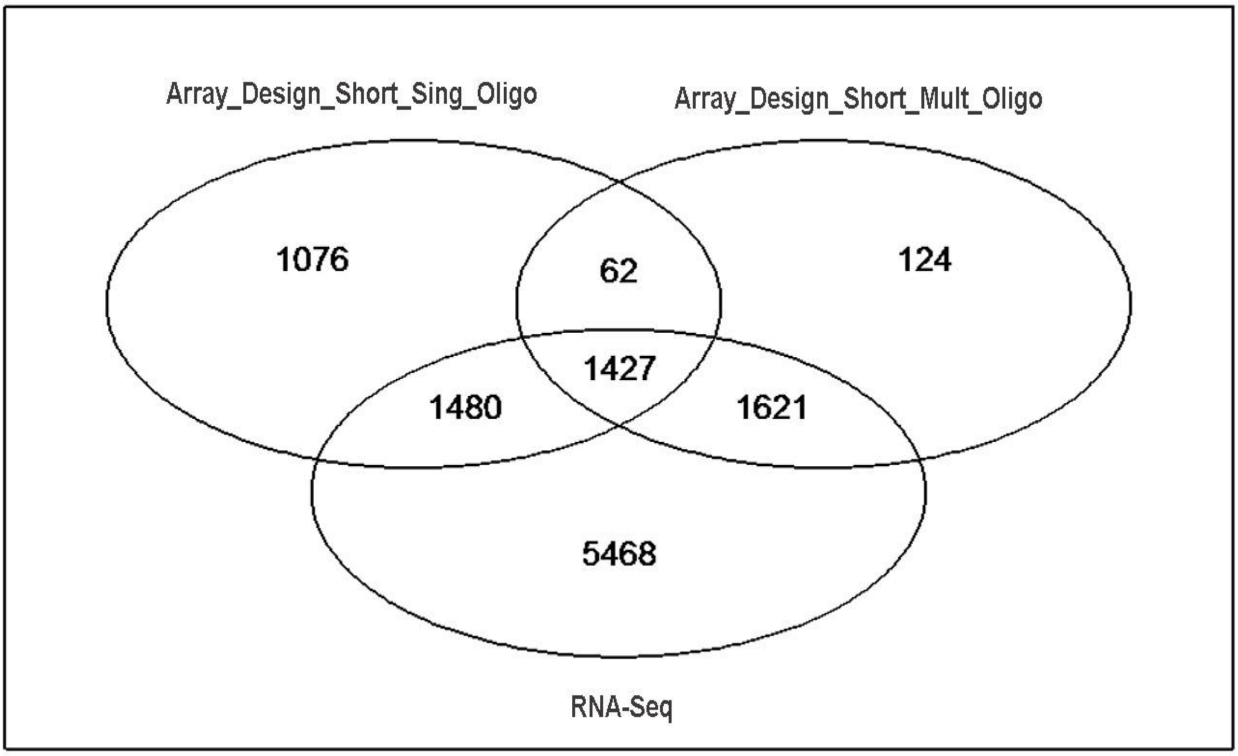

Figure 1. Venn diagram comparing the number of DEGs between microarray manufacture with short probe size per gene model transcript and RNA-seq approach at a FDR ≤ 0.05 in gene expression profiling survey.

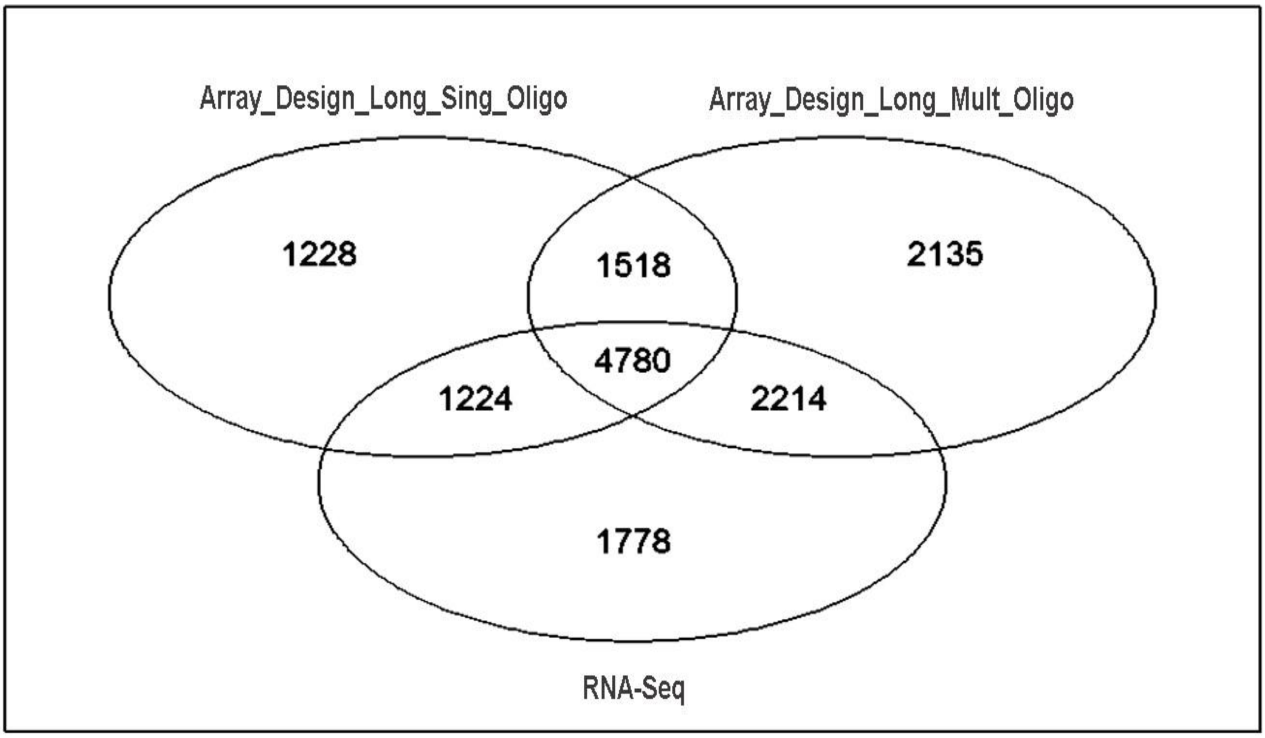

Figure 2. Venn diagram comparing the number of DEGs between microarray manufacture based on long oligonucleotide probes per gene model transcript and RNA-seq methodology at a FDR≤ 0.05in gene expression differentially profiling analysis.

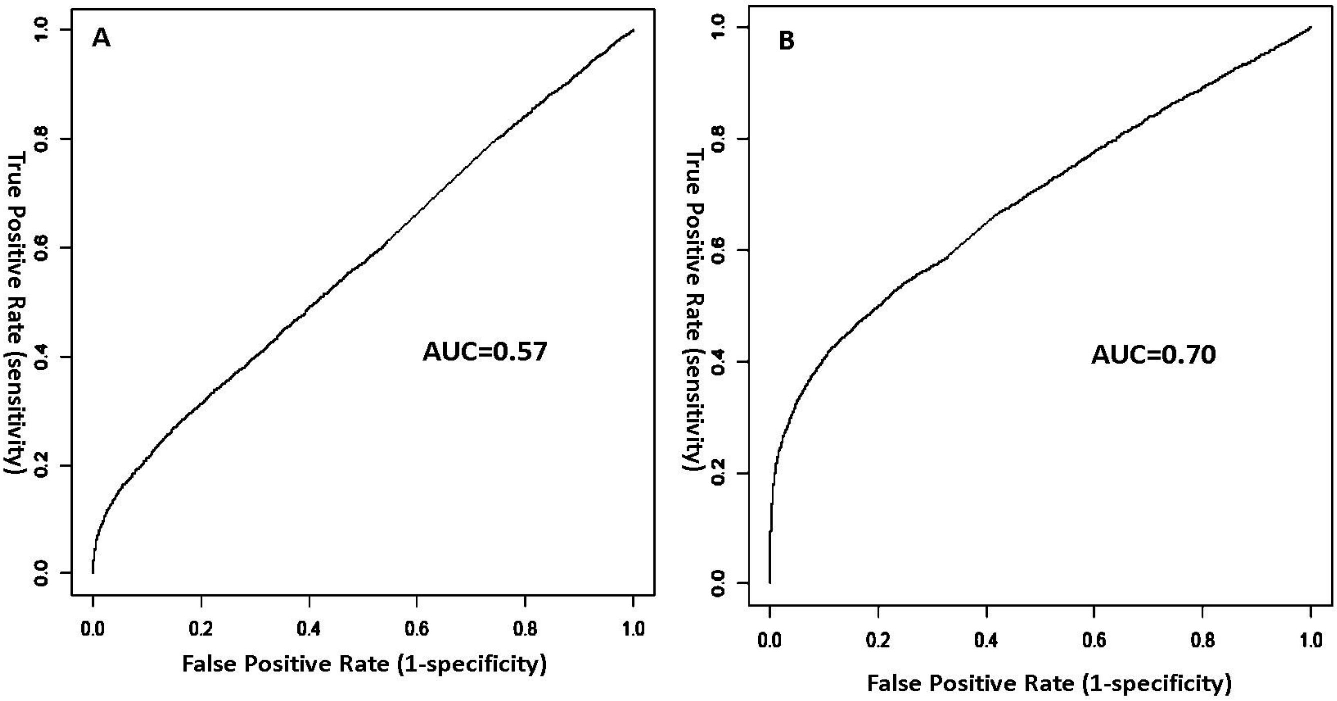

Figure 3. Panel A and B: ROC Curve graph of microarray manufactures based on single and multiple short probe(s) per gene unit discriminating statistically significantly differentially expressed genes (FDR ≤0.05 and log2 Fold Change value≥1) respectively.

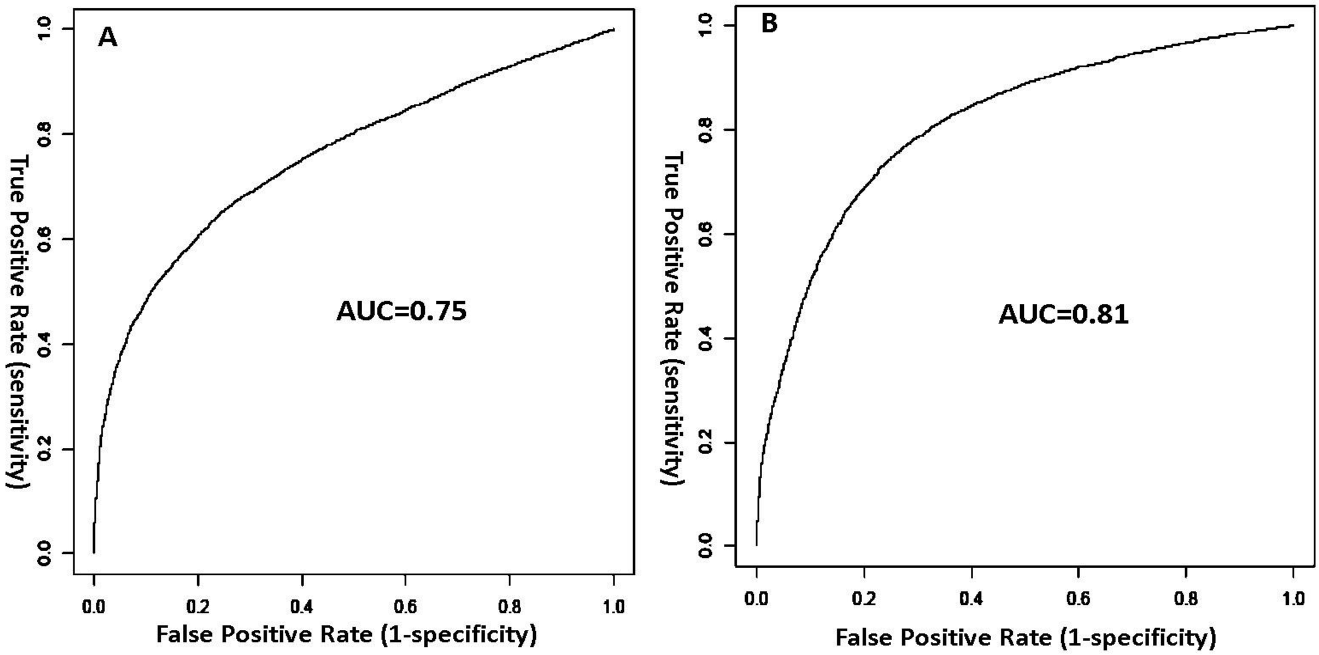

Figure 4. Panel A and B represent ROC Curve graphs of both microarray manufactures based on single and multiple long probe(s) per gene model transcript discriminating statistically significantly differentially expressed genes (FDR ≤0.05 and log2 Fold Change value≥1) respectively.

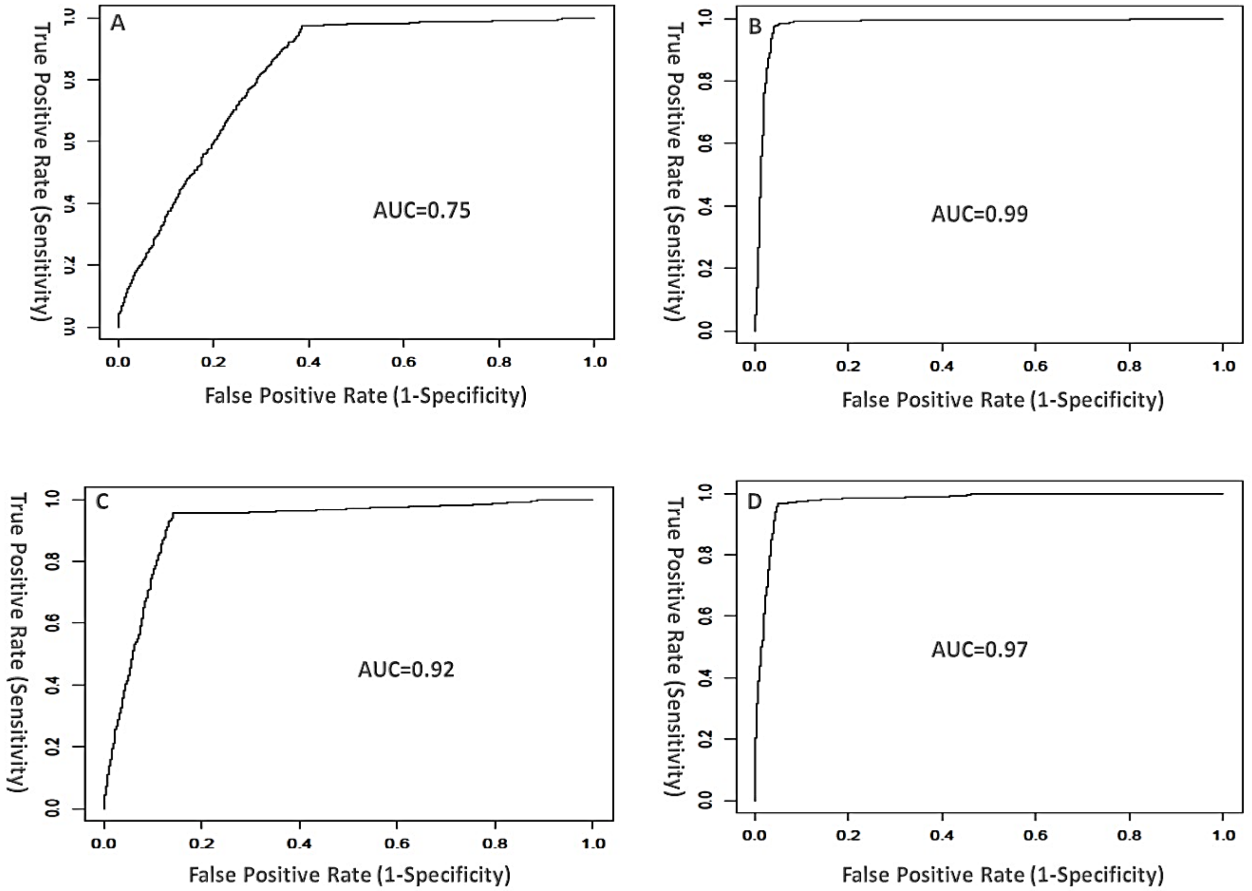

Figure 5. Area under curve (AUC) identified by Microarray designs with short unique replicate (A), short multiple (B), long unique replicate (C) and long multiple (D) probes detecting DEGs at False Discovery Rate (FDR) ≤0.01 with |log$_2$-Fold Change| ≥1.

Table 1. 2x2 contingency table and/or confusion matrix assessing microarray probe set designs performance selecting DEGs by next generation sequencing RNA-seq approach.

		Number of differentially expressed genes (DEGs) called by RNA-seq approach	
		True	False
Number of differentially expressed genes (DEGs) detected by analysed microarray probe set design strategies (test outcome).	True	True positive	False positive
	False	False negative	True negative

Table 2. Relationship between measured gene expression levels, false discovery rate probability (probability committing type I statistical error calling DEGs) and ROC curve analysis area under curve (AUC) parameters.

DEGs at False Discovery Rate (FDR)	[0.01;0.05]	]0.01;0.001]	]0.001;0.0001]	]0.0001;0.00001]	<0.00001
Gene Expression Level assessed by reads count sequence mean from RNA-seq Deseq analysis	9.26	20.26	26.21	33.31	>33.31
All Microarray ROC curve AUC mean*	0.73	0.81	0.86	0.87	>0.90
Number of Expressed Gene by RNA-seq	3422	2464	2673	1741	7374

* Log 2 fold change parameter has not been considered calculating AUC mean value as opposed to in Figure 5.

DISCUSSION

In this paper we have illustrated a feature selection method using a combination of standard area under curve (AUC) parameter from Receiver Operating Characteristics (ROC) curve analysis with several microarray's gene expression analysis manufacture assessing the impact of the latter's oligonucleotide probe set design strategies discriminating accurately statistically significantly differentially expressed genes (DEGs) by RNA-seq, since high throughput sequencing approach provides a powerful tool for transcriptome-based applications beyond the limitations of microarrays paying attention committing statistical type I error (Dago DN et al, 2014^{b}). Moreover, several studies demonstrated that RNA-seq outperforms microarrays in determining the transcriptomic characteristics (i.e. in cancer study), while RNA-seq and microarray based models perform similarly in clinical endpoint prediction (Zhao S. et al, 2014; Zhang W et al, 2015). Though, considering the vast amount of additional information provided by RNA-seq in comparison to microarrays, it is tempting to speculate that RNA-seq may outperform microarray; a comprehensive comparison of RNA-seq and microarray based predictive models is requested considering the fact that both technologies are widely used in genomics and/or transcriptomics studies as well as molecular diagnostic process nowadays (Song L et al, 2011; Zhang W. et al, 2015). In this predisposition, we proposed here to assess the agreement threshold between these two technologies by using receiver operating characteristics (ROC) curve analysis highlighting their similarity and/or dissimilarity evaluating DEGs calling event ratio monitoring false discovery rate (type I statistical error). For this purpose, we considered our previous developed microarray design manufactures based either on long and/or short single replicate and/or multiple oligonucleotide probe set per gene model transcript (Dago N. 2012). Then, we compared these microarray manufactures with RNA-seq by processing two *Vitis vinifera* grape development stage samples in gene expression profiling analysis basing exclusively on ROC curve analysis assessing FDR factor aiming to provide accurate statistical parameter threshold integrating both array and RNA-seq next generation sequencing data. Our analysis revealed strong heterogeneous behaviours amongst microarray manufactures based on short single and multiple oligonucleotide probes per gene unit discriminating DEGs as opposed to microarray design strategies with long single oligonucleotide probe set (Figures 1 and 2). Moreover, the good performance of microarray manufactures with long oligonucleotide probes per gene model transcript detecting right gene expression level in expression survey has been suggested by performed ROC curve analysis (Noel DD et al., 2016). These results are in agreement with Chung Chou (2004) investigation supporting that accurate gene expression measurement can be achieved with multiple probes per gene unit and fewer probes are needed if longer probes rather than shorter probes are used. Moreover the present results confirmed Roman J. et al (2013) results, suggesting that differences in the nucleotide composition of probes and high distances between their target sites in a transcript sequence are the main reasons for very high intra-probe set signal variance.Then, while microarrays design strategies based on long (single and multiple) oligonucleotide probes per gene model transcript exhibited a moderate accurate performance in detecting DEGs, microarray manufactures based on short (single and multiple) probes per gene model transcript resulted strongly influenced (p-value≤0.05) by the sort of microarray design strategies (Figures 3 and 4). Moreover, it is interesting to underline that our investigation revealed that analysed microarrays gene expression differential analysis results can be improved by rigorous statistical analysis reducing DEGs false discovery as well as avoiding committing type I statistical error without compromise differential analysis final results among analysed gene expression platforms. Also, we were able to show that disregarding applied microarray manufactures as well as microarray design strategies, all analysed microarray platforms based on short and/or long single replicate and/or multiple probes per gene unit exhibited a good agreement with RNA-seq approach detecting true positive DEGs at lowest false discovery rate values (Table 2). In other word, improvement of microarray's specificity calling true positive rate of DEGs candidates by RNA-seq needed stringent false discovery rate statistical parameters. In other word microarray probe set heterogeneity measuring gene expression level can be adjusted by a good monitoring of false discovery rate statistical parameter. Then, we showed that good predictive model based on microarray's manufacture required rigorous statistic parameter reducing and preventing false discovery rate events in gene expression differential profiling analysis. In the same tendency, our findings through area under curve by ROC curve survey evidenced strongest agreement between both microarrays and RNA-seq, assessing microarrays probe set performance calling accurately DEGs at a false discovery rate 10^{-3} (FDR ≤0.001) exhibiting these values as reasonable threshold avoiding committing type I error, comparing and integrating both microarray and RNA-seq differential gene expression analysis data and results (Table 2). Taking together, though all analysed microarray's displayed a good performances calling DEGs applying stringent FDR value, the present analysis suggested (i) best agreement between microarray manufactures based on multiple oligonucleotide probes per gene unit and RNA-seq measuring gene expression signal intensity as opposed to microarrays with probe set design strategy based on single replicate probe per gene model transcript (Figure 5); and (ii) microarray design based on long oligonucleotides were weakly influenced by the sort of array design with respect to microarray manufacture with short oligonucleotide probes per gene model transcript (Chen-Chung Chou et al., 2004).

This study, as our knowledge exhibited a particularity, since emphasized microarray oligonucleotide probe design strategies sensitivity and/or specificity executing differential analysis evaluating the risk committing statistical type I error when the former's were compared with RNA-seq gene expression approach in gene expression differential survey. However, despite the superior benefits of RNA-seq, microarrays are still the more common choice of researchers when conducting transcriptional profiling experiments. This is likely because RNA-seq sequencing technology is new to most researchers, relatively more expensive than microarray, data storage is more challenging and analysis is more complex. We expect that once these barriers are overcome, the RNA-seq platform will become the predominant tool for transcriptome analysis (Song L. et al, 2011).

CONCLUSION

Numerous studies evidenced microarray limits in gene expression profiling analysis as opposed to RNA-seq, neglecting their complementary aspects. We provided through the present study a methodology monitoring microarrays probes design strategies calling accurately DEGs in gene expression differential survey by using ROC curve enquiry exclusively, showing that microarray design strategies based on short oligonucleotide probe set (either single replicate or multiple probes) per gene model transcript exhibited strong dissimilar replies selecting correctly DEGs as opposed to microarray manufactures with long probe set per gene unit. The present study also suggested that despite their high heterogeneous responses in expression profiling analysis, microarrays could exhibit equal performance with respect to RNA-seq by an adequate monitoring of false discovery rate statistical parameter, suggesting a potential high complementary between above mentioned gene expression profiling technologies calling true positive event in transcriptomic survey. Finally these results as well as observations, combined with previous studies (Zhao S.et al, 2014), suggested that microarray and RNA-seq gene expression data integration were possible and needed an adequate statistical parameters setting aiming to reduce type I statistical error since the complementarity among these technologies could reinforce their widely used in transcriptomic and genomic analysis helping to accurate molecular diagnostic decision. Finally the present study highlighting the variableness of microarray probe set response in differential analysis as opposed to others studies provided a statistical guideline to overcome this limit exploiting next generation sequencing data performance.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

ACKNOWLEDGMENTS

The authors thank Genomic Center of the University of Verona (Italy) for providing microarrays and high throughput sequencing (RNA-Seq) gene expression row data as well as for the technical support for the present work.

REFERENCES

Bradley, A.P (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. 30 (7), 1145–1159.

Chou C.C, Chun-Houh C., Te-Tsui L. and Konan P. (2004). Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression. Nucleic Acids Research Vol.32 N°12 e99 doi10.1093/nar/gnh099.

Dago D.N, Alberto F., Diarassouba N., Fofana I.J, Silué S., Giovanni M. and Massimo D. (2014[a]). Probes specificity in array design influences the agreement between microarray and RNA-seq in gene expression analysis. Africa Journal of Science and Research 3 (5): 8-12.

Dago D.N, Malerba G., Ferarrini A. and Delledonne M. (2014b). Evaluation of Microarray Sensitivity and Specificity in gene Expression differential Analysis by RNA-seq and Quantitative RT-PCR. Journal of Multidisciplinary Scientific Research, 2 (6):05-09. http://jmsr.rstpublishers.com/.

Dago N. (2012). Performance assessment of different microarray designs using RNA-Seq as reference. Id prodotto: 67051; Id Ugov: 404537.

De Risi J.L, Iyer V.R, and Brown P.O (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278:680–686.

Egan, J.P (1975). Signal detection theory and ROC analysis, Series in Cognition and Perception. Academic Press, New York.

Gao H., Wang Y., Liu X., Yan T., Wu L., Alm E., Arkin A., Thompson D.K, and Zhou J. (2004). Global transcriptome analysis of the heat shock response of *Shewanella oneidensis*. J. Bacteriol. 186:7796–7803.

Hanley, J.A, McNeil, B.J (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29–36.

Hughes T.R, Marton M.J, Jones A.R, Roberts C.J, Stoughton R., Armour C.D, Bennett H.A, Coffey E., Dai H., He Y.D, Kidd M.J, King A.M, Meyer M.R, Slade D., Lum P.Y, Stepaniants S.B, Shoemaker D.D, Gachotte D., Chakraburtty K., Simon J., Bard M., and Friend S.H (2000). Functional discovery via a compendium of expression profiles. Cell 102:109– 126.

Li B., Dewey C.N RSEM. (2011). Accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics.doi:10.1186/1471-2105-12-323: 1-16.

Luo Z. and Geschwind D.H (2001). Microarray application in neuroscience. Neurobiol. Dis. 8:183–193.

MAQC Consortium (2006). The MicroArray Quality Control (MAQC) project shows inter- and intra-platform reproducibility of gene expression measurements. Nat Biotechnol. 24(9): 1151–1161. Doi: 10.1038/nbt1239.

Nalpas N.C, Park S.D, Magee D.A, Taraktsoglou M, Browne J.A, Conlon K.M, Rue-Albrecht K, Killick K.E, Hokamp K, Lohan A.J, Loftus B.J, Gormley E, Gordon S.V, Machugh D.E (2013). Whole transcriptome, high-throughput RNA sequence analysis of the bovine macrophage response to Mycobacterium bovis infection in vitro. BMC Genomics. 14(1): 1-19.

Noel D.D, Ferrarini A., Xumerle L., Mori A., Delledonne M. and Malerba G. (2016). Heterogeneity of Global Gene Expression Microarray Designs in Detecting Differentially Expressed Genes. In press in International Journal of Bioinformatics Research.

Ochs M.F and Godwin A.K (2003). Microarray in cancer: research and application. BioTechniques 34:S4–S15.

Petricoin E., Hackett J.L, Lesko L.J, Puri R.K, Gutman S.I, Chumakov K., Woodcock J., Feigal D.W, Zoon K.C, and Sistare F.D (2002). Medical application of microarray technologies: a regulatory science perspective. Nat. Genet. 32:474–479.

Rhee S.K, Liu X., Wu L., Chong S.C, Wan X., and Zhou J (2004). Detection of biodegradation and biotransformation genes in microbial communities using 50-mer oligonucleotide microarrays. Appl. Environ. Microbiol. 70:4303–4317.

SEQC/MAQC-III Consortium. (2014). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat Biotechnol.

Smyth, G. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology.3 (3): doi 10.2202/1544-6115.1027.

Song L., Lan L., Peng J., Dan W. and Yi X. (2011). A comparison of RNA-Seq and high-density exon array for detecting differential gene expression between closely related species Nucleic Acids Res; 39(2): 578–588. doi: 10.1093/nar/gkq817 PMCID: PMC3025565.

Swets, J. (1988). Measuring the accuracy of diagnostic systems. Science 240,1285–1293.

Swets, J.A, Dawes, R.M, Monahan, J. (2000). Better decisions through science. Scientific American 283, 82–87.

Taroncher-Oldedburg G., Griner E.M, Francis C.A, and Ward B.B (2003). Oligonucleotide microarray for the study of functional gene diversity in the nitrogen cycle in the environment. Appl. Environ. Microbiol. 69:1159–1171.

Tiquia S.M, Wu L.,Chong S.C, Passovets S., Xu D., Xu Y., and Zhou J. (2004). Evaluation of 50-mer oligonucleotide arrays for detecting microbial populations in environmental samples. BioTechniques 36:664–675.

Tom Fawcett (2006). An introduction to ROC analysis. Pattern Recognition Letters 27 861–874

Wan X., Ver B.N.C, McCue L.A, Stanek D., Connelly H., Wu L., Liu X., Yan T., Leaphart A., Hettich R.L, Zhou J., and Thompson D.K (2004). Defining the *Shewanella oneidensis* FUR regulon: integration of genome-wide expression analysis, proteome characterization, and regulatory motif discovery. J. Bacteriol. 186:8385–8400.

Wenqian Z., Ying Y., Falk H., JeanThierry M., Wenwei Z., Danielle T.M, Jian W., Viswanath D., Jie C., Youping D., Barbara H., Huixiao H., Meiwen J., Li L., Simon M.L, Yuri N., André O., Tao Q., Zhenqiang S., Ruth V., Charles W., May D.W, Junmei A., Davide A., Shahab A., Smadar A., Wenjun B.M.B, Murray H.B, Benedikt B., Marco C., Tzu-Ming C., Jibin Z., Richard G.G, MinMax H., Scott H., Howard L.K, Samir L., Lee J.L, Yan L., Xin X.L, Heng L., XiwenMa B.N, Rosa N., Martin P., John H.P, Frederik R., Carolina R., Susan S., Jie S., Jessica T., Gian P.T, Jo V., Po-Yen W., Wenzhong X., Xu J., Xu W., Xuan J., Yang Y., Ye Z., Dong Z., Zhang K.K, Yin Y., Zhao C., Zheng Y., Wolfinger R.D, Shi T., Malkas L.H, Berthold F., Wang J, Tong W., Shi L., Peng Z. and Fischer M. (2015). Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. Genome Biology 16:133 DOI 10.1186/s13059-015-0694-1.

Wodicka L., Dong H., Mittmann M., Ho M.H, and Lockhart D.J (1997). Genome-wide expression monitoring in *Saccharomyces cerevisiae*. Nat. Biotechnol.15:1359–1367.

Yoav B; Yosef H. (1995). Controlling the False Discovery Rate: A Powerful Approach to multiple testing. Journal of the royal Statistic Society. Series B (Methodological), 57(1): 289-300.

Zamboni A., Di Carli M., Guzzo F., Stocchero M., Zenoni S., Ferrarini A., Tononi P., Tofalli K., Desiderio A., Kathryn S.L, Pè M.E, Benvenuto E., Delledonne M. and Pezzotti M. (2010). Identification of putative stage-specific grapevine berry biomarkers and omics data integration into networks. Plant Physiology. vol. 154 (3): 1439-1459.

Zenoni S., Ferrarini A., Giacomelli E., Xumerle L., Fasoli M., Malerba G., Bellin D., Pezzotti M. and Delledonne M. (2010). Characterization of Transcriptional Complexity during Berry Development in *Vitis vinifera* Using RNA-Seq. Plant Physiology 152 (4): 1787-1795.

Zhao S, Fung-Leung W-P, Bittner A, Ngo K, Liu X (2014) Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. PLoS ONE 9(1): e78644. doi:10.1371/journal.pone. 0078644.

Zou, K.H (2002). Receiver operating characteristic (ROC) literature research. On-line bibliography available from: <http://splweb.bwh. harvard.edu:8000/pages/ppl/zou/roc.html>.

ISSN : 0976-4550

INTERNATIONAL JOURNAL OF APPLIED BIOLOGY AND PHARMACEUTICAL TECHNOLOGY

Email : ijabpt@gmail.com

Website: www.ijabpt.com