International Journal of Bioinformatics Research
ISSN: 0975-3087&E-ISSN: 0975-9115, Volume 7, Issue 2, 2016, pp.-349-357.
Available online at http://bioinfopublication.org/jouarchive.php?opt=&jouid=BIJ0000001

Research Article

HETEROGENEITY OF GLOBAL GENE EXPRESSION MICROARRAY DESIGNS IN DETECTING DIFFERENTIALLY EXPRESSED GENES

NOEL DOUGBA DAGO*[1,3], ALBERTO FERRARINI[1], LUCIANO XUMERLE[1], ANTONIO MORI[2], MASSIMO DELLEDONNE[1] AND GIOVANNI MALERBA[2]

[1]Department of Biotechnology, University of Verona, Italy Strada le Grazie 15, Cà vignal 1, 37134, Verona Italy
[2]Department of Neurological, Biomedical and Movement Sciences University of Verona, Strada Le Grazie 8, 37134, Verona, Italy
[3]Unité de Formation et de Recherche (UFR) Sciences Biologiques Université Péléforo Gon Coulibaly, BP 1328 Korhogo, Cote d'Ivoire
*Corresponding Author: Email- dgnoel7@gmail.com

Received: May 25, 2016; Revised: July 31, 2016; Accepted: August 01, 2016; Published: August 21, 2016

Abstract- Microarray is widely used for gene expression studies by many laboratories worldwide. Microarrays vary for the type and number of oligonucleotide probes implemented and for the procedure to subtract background noise (BS) and normalize data (DN) among samples consenting to make these reliable tools somewhat heterogeneous, as heterogeneity may play an important role identifying differentially expressed (DE) genes in global gene expression studies. We essayed four different microarray design strategies based on either single replicate or multiple probes per gene model transcript and on different probes size (long and/or short) to analyze two *Vitis vinifera* berry developmental stages. Microarray data were processed basing on 20 different BS-DN arrangements. 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. Microarray performances in detecting DE genes were evaluated by several measures comprising correlation between estimated fold-change values, classification functions and the Area Under Curve (AUC) of receiver operating characteristic (ROC) curves. The number of DE genes changed from one microarray design to another, suggesting their heterogeneous performances in gene expression differential analysis. However our findings suggested a good agreement between microarrays and RNA-Seq technologies for gene expression level higher than 10 fpkm discriminating differentially (DE) expressed genes. The present results warn researchers that even if different microarray designs can lead different results, both RNA-Seq and array approaches can exhibit comparable performance in gene expression analysis for higher expressed gene. Then, the present survey provided a powerful methodology helping researchers choosing microarrays and/or RNA-Seq approaches in their transcriptomic studies.

Keywords- Microarray designs, RNA-Seq, Differentially Expressed (DE) gene, Oligonucleotide Probes, *Vitis vinifera*.

Citation: Noel Dougba Dago, et al., (2016) Heterogeneity of Global Gene Expression Microarray Designs in Detecting Differentially Expressed Genes. International Journal of Bioinformatics Research, ISSN: 0975-3087 & E-ISSN: 0975-9115, Volume 7, Issue 2, pp.-349-357.

Introduction

Nowadays the approaches to discover genes differentially expressed between various biological conditions at genome level are based on microarray (MA) or next generation sequencing (NGS) technologies known as RNA-Seq. MAs are used over a long time and represent a widely used and reliable technology in transcriptomic studies [1].Generally, all MA platforms offer highly inter-platform reproducible results indicating that this technology is mature having developed stable analytical setup. RNA-Seq has considerable advantages for examining transcriptome fine structure but data processing requires skilled bioinformatics analysis. Moreover, the latter attribute facilitates comparison measuring gene expression levels between different experiments favoring the expression of different genes within the same sample to be compared, which is useful, for instance in modeling regulatory networks [2]. However, research studies are proposing novel approach to microarray analysis that attains many of the advantages of RNA-Seq [3].Sensitivity of RNA-Seq in detecting expressed genes depends on the sequencing depth and it is still more expensive than MAs [4-7]. Although the reliability of the arrays is considered high and satisfying, it is not still clear if their design heterogeneity affect the final results when searching for differential expressed (DE) genes. This issue has been marginally studied by earlier studies. The Micro-Array Quality Control (MAQC) consortium measured the correlation of results obtained from different MA platforms in terms of gene expression levels and concluded that the constancy of selected gene lists correlates with endpoint expectedness, subtly suggesting that MA platforms are all similar in defining gene expression profiles [8]. However, despite their high degree of inter-platform data reproducibility, MAs are extremely heterogeneous tools due to their makeup, and both bioinformatics and statistical approaches. Moreover customizable MA designs are now available to the scientist community making these tools extremely suitable for several purposes but still more heterogeneous. We inquired whether results of differential gene expression analyses on the same RNA from different custom MA designs can change significantly from one another or differences are rather minimal. In this study we investigated the performance of four microarray design strategies based on two different custom microarray platforms (ex- Roche NimbleGen; *NMG* and ex-Combi-Matrix; *CMB*) paying attention on DE genes. Analyzed microarray design platforms were based on either duplicate or different long (60 bp) and/or small (35-40 bp) oligonucleotide probes for the same gene. The present analyzed MA platforms shared a large number of common genes across them. Experiments were conducted profiling 2 development stages of grape *Vitis vinifera* plant (veraison and ripening). Microarray data were analyzed using combinations of the most common procedures to subtract background noise (BS) and normalize data (DN) among

samples. We then compared the concordance of the results between each of the four microarrays with the results from the analysis of the same samples with sequencing-based methods (RNA-Seq).

Materials and Methods

The same samples of Zenoni *et al*. (2010) [9] were used for microarray experiments, corresponding of grapevine (*Vitis vinifera*) berry tissue at veraison and ripening stages. Microarrays data were processed and analyzed in combination of several background subtraction and normalization procedures. Results of DE gene analysis from each microarray were then compared with the results obtained from RNA-Seq. The four microarray designs were tested using several performance measures.

RNA Preparation

Sample of *Vitis vinifera* at veraison and ripening growth phase were collected as reported in Zenoni *et al*. (2010) [9] and total RNA has been extracted as described in Zamboni *et al*. (2008) [10]. RNA amount and integrity were essayed by Nanodrop 2000 instrument (Thermo Scientific) and an Agilent Bio-analyzer Chip RNA 6000, respectively.

Microarrays (MAs)

Grape Custom-Array designs based on short probe (35-40 bp) per gene model transcript (CMB)

We used the 29971 transcript sequence annotations of the *Vitis vinifera* grape 12x assembly [11] to design custom probes using the software Oligoarray v2.1 [12] under the following parameters: oligonucleotides length range between 35 and 40 nucleotides; melting temperature varies between 80 and 86°C; GC content fluctuates between 40 and 60%; threshold to reject oligonucleotides folding into stable secondary structures or forming putative cross-hybridizations set to 65°C; rejections of oligonucleotides containing homo-polymers of at least 5 base; maximum distance between 5' end of an oligonucleotide and 3' end of input sequence set to 1500 bp.

Best oligonucleotide per input sequence were chosen for Grape Custom-Array based on single small probes (35-40 oligonucleotide) per gene model transcript with 3 replicates per probe (*CMB-S*).

Best 3 oligonucleotides per input sequence (with 100 bp minimum distance between 5' ends of two contiguous oligonucleotide probes) were chosen for Grape Custom-Array based on small multiple probes per gene model transcript with a single replicate per probe (*CMB-D*). Then, for the present analysis, a total of 29464 and 82326 oligonucleotide were processed for *CMB-S* (MA design with single triplicate probe per gene model transcript) and *CMB-D* (MA design with 3 different probes per gene model transcript) MA designs, respectively.

Grape Custom-Array based on long probes (60 bp) per gene model transcript (NMG)

Two Grape Custom-Array designs based on single and/or multiple long oligonucleotide probes per gene model transcript from transcript sequence file of grapevine 12x assembly [11] have been performed by Roche Nimble-Gen applying the following parameters: oligonucleotide length of 60 nucleotides; melting temperature range between 76 and 79°C; GC content range between 40 and 47%; threshold to reject oligonucleotides folding into stable secondary structures or forming putative cross-hybridizations set to 65°C; maximum distance between 5' end of an oligonucleotide and 3' end of input sequence set to 1500 bp; checking of probes uniqueness against grapevine 12x assembly. For *NMG-S* design, best oligonucleotide per input sequence were chosen with 4 replicates per probe, while for *NMG-D* design, best 4 oligonucleotides per input sequence were chosen with a single replicate per probe. Then, a total of 29877 and 118328 oligonucleotide probes were processed for *NMG-S* (single quadruplet probe per gene transcript model) and *NMG-D* (4 different or multiple probes per gene transcript model) MA designs respectively.

Hybridization experiment of Grape Custom-Array based on short probes (CMB)

The same quantity of total RNA sample (2 µg) from the three analyzed technical replicates of veraison and ripening development stage was handled for Grape Custom-Array based on short probe (CMB) hybridization by using the Universal Labeling System (ULS) based on cy5 one color fluorescent system. Hybridization and chips washing step were accomplished following ex-CombiMatrix Custom-Array 90k Microarray manufacturer's instructions. The following image scanning step has been performed by the Axon Scanner Instruments GenePix 4200A at 632 wave length.

Hybridization experiment of Grape Custom-Array based on long probes (NMG)

The same quantity of total RNA sample (10 µg) from the three analyzed technical replicates of veraison and ripening development stage was processed for Grape Custom-Array based on long probes hybridization experiment by using One Color-DNA labeling system with Cy3 fluorescent. RNA processing, labeling, hybridization and chip washing phase were performed basing on Nimble-Gen Arrays User's Guide Gene Expression Analysis version 3.1 protocol manufacturer's instructions. Hybridization images scanning have been achieved by the axon scanner Instruments GenePix 4200A at 535 wave length.

Microarray Data Preprocessing

Data preprocessing comprises computer methods adjusting ambient intensity (background subtraction, BS) across arrays as well as removing variation sources between arrays due to external biological factors (data normalization, DN). Therefore, the present microarray gene expression data were preprocessed using all the combinations of background subtraction (BS) and data normalization (DN) procedures available in the library package *limma* (version 3.10.3) [13]. BS methods include none (i.e. no background subtraction) and normexp methods. Normexp method depends on saddle, mle and robust multichip average (rma or rma75) parameter estimation strategies. DN procedure was applied using none (i.e. no data normalization), scale, quantile or cyclic loess normalization method (4 methods). Designs are reported across the paper referring to their (i) background correction (BS) + (ii) data normalization procedure (DN). Array designs were therefore preprocessed with 20 different combinations of BS+DN methods. However, Grape-Array designs based on long oligonucleotide probes per gene model transcript (NMG) were also processed for BS using the proprietary method included in the software Nimblegenscan. Indeed, Nimblegenscan BS treated designs, (*NMG-SN, NMG-DN*), underwent the DN preprocessing only. However to facilitate the comparisons across this study we shall indicate a BS preprocessing for the NMG-SN, NMG-DN designs even if BS preprocessing was not applied. Expression (i.e. intensity) values of each gene were expressed applying either mean or median values of the probe signals of the same gene across each array.

Differential Gene Expression Analysis

Differential gene expression (DGE) analysis between 2 grape development stages was performed by comparing arrays processed with the same BS+DN combination. DGE analysis 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 [13]. The False Discovery Rate (FDR) suggested by Benjamini and Hochberg [14] was adopted to control the FDR since gene expression differentially analysis usually englobes multiple comparisons statistical test. Significance of DGE analysis results of CMB-D (CMB-D.fisher) and NMB-D (NMB-D.fisher) platforms when applying the mean 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 [15]. 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 berry development stages at a False Discovery Ratio ≤ 0.05 (FDR≤0.05). Only genes shared among all the platforms were included in the performance comparisons.

RNA-Seq Experiment

The RNA-Seq data used in this study was generated during our previous study [10]. Briefly, 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. Here, read count was performed using the packages RSEM (v1.1.21) [16] and Cufflinks (1.2.0 release, http://cufflinks.cbcb.umd.edu/). Next DESeq (version 1.1.6) package has been used for DGE analysis. RNA-Seq raw data is available at SRA009962 as well as at URL http: //ddlab.sci.univr.it/cgi-bin/gbrowse/grape.

Comparisons of Microarray Designs Performance

RNA-Seq data results were set as the reference values to compare the microarray BS+DN design performances. Performance of each microarray design was tested by:

1) Comparing the Pearson's correlation of the fold-change (FC) values of the all expressed genes and the number of detected DE genes.
2) Estimating the association of DE genes between RNA-Seq and microarray platforms. Association was estimated by contingency tables and CHIsq test. CHIsq value was used to score the performance of the array in detecting DE genes.
3) Estimating specificity, sensitivity, accuracy, positive predictive (PPV) and negative predictive (NPV) values to identify DE genes.
4) Estimating the area under the curve (AUC) of receiver operating characteristic (ROC) curves [17].

PCR-Real Time Validation

We designed RT-PCR primers (forward and reverse) for 10 randomly selected genes on within their 1 kb upstream of the 3'end region validating above mentioned RNA-Seq and microarray expression data. As template for primer design we used the 12x grape genome assembly [12]. RNA samples were treated with DNase using the Turbo DNA-free kit (Applied Biosystem).

Superscript II reverse Transcriptase of Invitrogen kit for cDNA synthesis was used for cDNA synthesis (3 different reactions were performed for each considered grape development stage). Quantitative RT-PCR was performed in 25 µl reaction containing SYBER green master mix (Invitrogen), 1 µl of each primer and 2 µl of above prepared cDNA template.

PCR survey was achieved in a MX 3000 recognized as a Fast Real Time PCR system (ABI Instrument) in three technical replicates for each sample. PCR run cycle was as following: 50°C hold for 2 min and a 95°C hold for 10 min followed by 40 cycles at 95 °C for 30 s, 55 °C for 30 s and 72 °C for 20 seconds. Detection of threshold cycle for each reaction was determined using a standard curve, after normalization procedure of the results by using quantitative RT-PCR result of actine primes TC81781 (TIGR, Release 6.0), which exhibits constant expression level between ripening and veraison berry development stage. We estimated RT-PCR amplification efficiency basing on raw data by using Ling Reg PCR software [18]. The relative expression ratio value and Standard Error (SE) were calculated according to the Pfaffl equation [19].

Results

Grape Microarray Oligonucleotide Probe Sequence Designs

Gene expression studies were conducted using 2 custom microarray platforms including 2 probe customizations each. Customizations were based on either duplicate and/or different probes for the same gene. Probe sequence designs were based on the grape genome assembly [11] and differ in size length (*NMG* probe length 60 bp and *CMB* probe length 35-40 bp). In this study four different microarray designs based on either different (*NMG-D* or *CMB-D*) and single replicate (*NMG-S* and *CMB-S*)probes per gene model transcript were developed. 29464, 82326, 29877 and 118328 oligonucleotide probes were selected for *CMB-S*, *CMB-D*, *NMG-S* and *NMG-D* microarray designs respectively [Table-1]. Compressively, more than 85% selected oligonucleotide probes resulted specific (with no mismatch) to their respective gene model transcripts when aligned against *Vitis vinifera* whole genome. Both mRNA and cDNA microarray hybridization process based on one color cy5 and cy3 fluorescence were performed for *CMB* (design with short probes) and *NMG* (design with long probes) microarray designs respectively. Details of microarray platform technologies (microarray design) used are reported in Materials and Methods chapter and summarized here in [Table-1].

Table-1 *Overview of the four microarray platforms and RNA-Seq in gene expression assay*

	CMB-S	CMB-D	NMG-S	NMG-D	RNASeq *
Platform Technology	RNA microarray hybridization	RNA microarray hybridization	cDNA microarray hybridization	cDNA microarray hybridization	mRNA sequencing
Probes or reads length	35-40 mer	35-40 mer	60 mer	60 mer	36- 44 bp
Substrate	Ceramic slide	Ceramic slide	Glass slide	Glass slide	-
Deposition	*In situ* synthesis	*In situ* synthesis	*In situ* synthesis	*In situ* synthesis	-
Detection	One color cy5 Fluorescence	One color cy5 Fluorescence	One color cy3 Fluorescence	One color cy3 Fluorescence	Reads count
Software for DE statistical analysis	*Limma* package	*Limma* package	*Limma* package	*limma* package RMA Nimblescan 2.5	*RSEM, Cufflinks* *DESeq* packages
Number of probe per transcript	1	3	1	4	-
Replicated probe per transcript	3	1	4	1	-
Number of targets	29464	29464	29582	29582	29971
Total number of probe/ total number of reads	29464 probes	82326 probes	29877 probes	118328 probes	41899518 read count

(*) RNA-Seq mRNA sequencing platform results used as reference assessing microarrays performance calling DEGs.

Differential Expression Analysis

Before accomplishment of differential expression analysis, we assessed microarrays intra-platform data reliability by Pearson correlation analysis between processed technical replicate samples of each considered veraison and ripening viticulture development stages. This survey revealed a good intra-platform data reliability for each analyzed microarray designs ($R \geq 0.9$; p-value < 0.05). Next, we performed differential expression analysis of developed microarray designs. In total, 17,446 genes were common across all microarrays and were detected for the subsequent survey. Microarray data were preprocessed using all the BS-DN combinations available in the library package *limma* (v. 3.10.3) and differentially gene expression analysis was conducted for each combination. [Table-2] shows the number of DE genes detected by each analyzed microarray according to different BS-DN combinations, ranging from 296 to 15,146 genes. RNA-Seq performed with Illumina Genome Analyzer II, yielding 41,899,518 reads (36-44 bp) for both veraison and ripening viticulture development stage [Table-1] allowed to identify 5,650 DE genes. The number of DE genes common to all microarray platforms and within the same BS-DN combination ranged from 114 to 317 and from 129 to 372 when gene expression levels were summarized by the mean and

Table-2 *Numbers of Differentially expressed genes by the microarray designs*

Normalization method:	CYCLIC LOESS					QUANTILE					SCALE					NONE				
BS method:	*A*	*B*	*C*	*D*	*E*	*A*	*B*	*C*	*D*	*E*	*A*	*B*	*C*	*D*	*E*	*A*	*B*	*C*	*D*	*E*
Platform																				
CMB-S.mean	3869	6570	2977	2327	3861	3467	6316	2828	2287	3461	3296	5190	2674	2130	3307	3105	3350	2733	2137	3112
CMB-S.median	3773	6425	2961	2327	3766	3373	6136	2847	2305	3368	3268	5045	2634	2105	3266	3071	3355	2690	2162	3075
CMB-D.mean	1179	4093	863	396	1202	1975	4837	1445	763	1996	1080	1689	643	490	1094	830	828	840	814	826
CMB-D.median	1039	3764	722	296	1052	1897	4634	1378	735	1914	905	1551	600	580	921	765	824	758	676	762
NMG-S.mean	7214	8257	7186	7379	7232	5746	7130	5806	6359	5744	5482	6216	5367	4569	5466	1944	2650	1596	1434	1929
NMG-S.median	7788	8676	7659	7801	7804	7515	8324	7325	7508	7523	7040	7432	6728	5118	7026	3573	2919	2835	2151	3611
NMG-D.mean	9954	11372	9804	9959	9965	9771	11234	9671	9972	9764	10657	11495	10327	8484	10650	3509	4696	3559	3618	3511
NMG-D.median	9351	10554	9105	8981	9370	9319	10541	9107	9176	9338	10339	11035	9871	7936	10361	3784	5022	3758	3771	3805
NMG-DN	8225	8225	8225	8225	8225	9818	9818	9818	9818	9818	10957	10957	10957	10957	10957	12951	12951	12951	12951	12951
NMG-SN	6216	6216	6216	6216	6216	6092	6092	6092	6092	6092	6248	6248	6248	6248	6248	6274	6274	6274	6274	6274
CMB-D.fisher	3558	6078	3164	2375	3587	4310	7049	3821	2819	4347	2866	3533	2560	2086	2885	2440	2638	2388	2148	2434
NMG-D.fisher	13897	15146	13645	13489	13905	13390	14959	13140	13379	13410	13917	14955	13545	12005	13925	6609	10719	6585	7334	6636

Microarray were processed using all background subtraction methods (see columns A, B, C, D, and E) and data normalization (cyclic loess, quantile, scale and none groups of columns) combinations. The name of the platform in the Platform column is followed by a ".mean", ".media" indicating that the probe signals of each gene have been summarized by the mean or median value respectively. The ending ".fisher" refers to array having different probes for the same gene whose signal has been summarized by the mean value and with DE analysis conducted by applying the Fisher method (see Materials and Methods).

median values of probe signals, respectively (data not shown). For clarity we show throughout the paper the results of the arrays preprocessed with normexp (rma) + quantile (randomly chosen) or with BS-DN approaches suggested by the company. Overall results from all BS-DN combinations are reported in the supplementary material and they are indicated when needed.

Correlation of the Fold-Change Values and Association of DE Genes

[Fig-1] shows plots of fold change values estimated from microarrays and RNA-Seq platforms. Plot representation reports the correlation of fold change values between microarray and RNA-Seq. In all cases, microarray designs including different probes per gene, showed a stronger fitting (i.e., correlation of fold change values and association of differential expressed genes) with the results from RNA-Seq than the microarray designs made up of the same probes per gene. The number of genes labeled as differentially expressed by microarray platforms but not by RNA-Seq ranges from 70 to 3932. The number of genes that resulted to be differentially expressed for both microarray and RNA-Seq ranges from 693 to 4683 depending on the microarray design. Supplementary [Fig-1] (S1) shows correlation plots of fold change values between microarray and RNA-Seq for each of the BS-DN combinations.

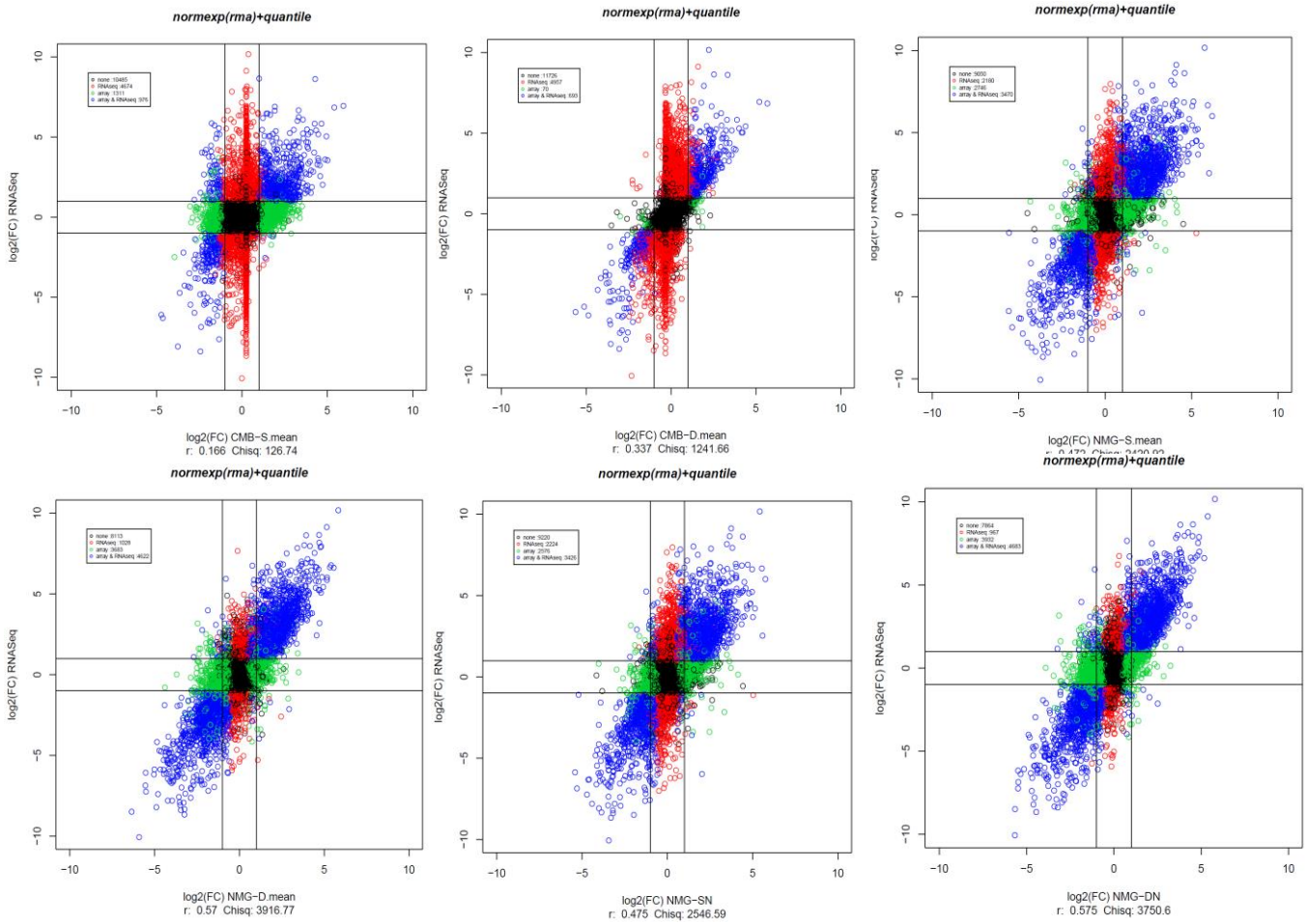

Fig-1 Correlation of FC between microarrays (MAs) and RNA-Seq. Circles are plotted according to the log2FC values of MA and RNA-Seq analyses for each gene. FC values of MA were estimated using the mean values of the log intensity of the probes targeting the same gene transcript and processed using *normexp(rma) + quantile* procedure (results for the other BS-DN combinations are reported in Figure S2). The strength of the association between the results of MA and RNA-Seq platforms are expressed by the CHIsq values estimated from contingency tables of DE genes. The r values of the Pearson correlation between log2FC values of MA and RNA-Seq are reported in each plot. Blue circles, DE genes in both MA and RNA-Seq; green circles, DE genes in MA only; red circles, DE genes on RNA-Seq only; black circles, no DE genes.

Classification Functions (Specificity, Sensitivity, Accuracy, Positive Predictive and Negative Predictive Values) for DE Gene Analysis

[Table-3] reports values from classification functions of comparisons between each microarray design and RNA-Seq results. Data processed by Fisher test showed a higher numbers of DE genes (from 757 to 1472 for CMB-D and from 4530 to 4696 to for NMG-D design) and lower true positive rate (from 90.75 to 82.20 for CMB-D and from 74.75 to 73.53 for NMG-D). Performances from all BS-DN combinations are reported in the supplementary table S1. Table S1 also report values from classification functions when studying subgroup of genes grouped in 4 samples (quartiles) according to their expression levels. [Fig-2] shows radar plots of true positive rate (TPR) values from each analyzed microarray design. The BS methods normexp (saddle) and normexp (mle) show overlapping performances (same TPR values). In all the plots, we can observe concentric shapes showing that BS and DN procedures are all important factors that influence the final TPR values of the present analyzed microarray design platforms.

Estimating the Area Under the Curve (AUC) of Building Receiver Operating Characteristic (ROC) Curve

Microarray design performance were also evaluated in terms of AUC of ROC curves. [Table-4] shows the AUC values of ROC curves estimated considering all expressed genes and genes grouped by their expression (FPKM) level. AUC values across platforms range from 0.526 (CMB-S) to 0.837 (NMG-D) when all expressed genes are considered. AUC values are higher (from 0.565 to 0.906) when ROC curve are estimated for highly expressed genes (FPKM > 10) suggesting that microarrays perform better with highly expressed genes, even if variability among platforms is high. Table S2 shows the AUC values of ROC curves for all the microarray designs.

Table-3 *Comparisons of microarray platforms performance*

Platform	DE genes (N)	Sensibility (%)	Specificity (%)	Accuracy (%)	TPR (%)	TNR (%)
CMB-S	2269	17.22	89.01	65.76	42.88	69.18
CMB-D	757	12.16	99.41	71.15	90.75	70.26
CMB-D.Fisher	1472	21.45	97.78	73.08	82.20	72.24
NMG-S	3820	46.27	89.78	75.68	68.43	77.72
NMG-D	4530	59.93	90.30	80.47	74.75	82.47
NMG-D.Fisher	4696	61.12	89.46	80.28	73.53	82.77
NMG-SN	3512	43.66	91.14	75.77	70.24	77.16
NMG-DN	4451	58.35	90.22	79.90	74.07	81.89

The number of DE genes detected by each array (column "DE genes") was compared with the 5650 DE genes detected by RNA-Seq. The performance of the arrays were estimated according their sensibility, specificity, accuracy, TPR and TNR. Sensitivity: % of the DE genes of RNA-Seq detected by MA; Specificity: % of the non-DE genes of RNA-Seq detected by MA; Accuracy: % of genes that were labelled in the same way (DE or non-DE) by both MA and RNA-Seq; TPR: % of DE genes of MA that results to be DE by RNA-Seq analysis.

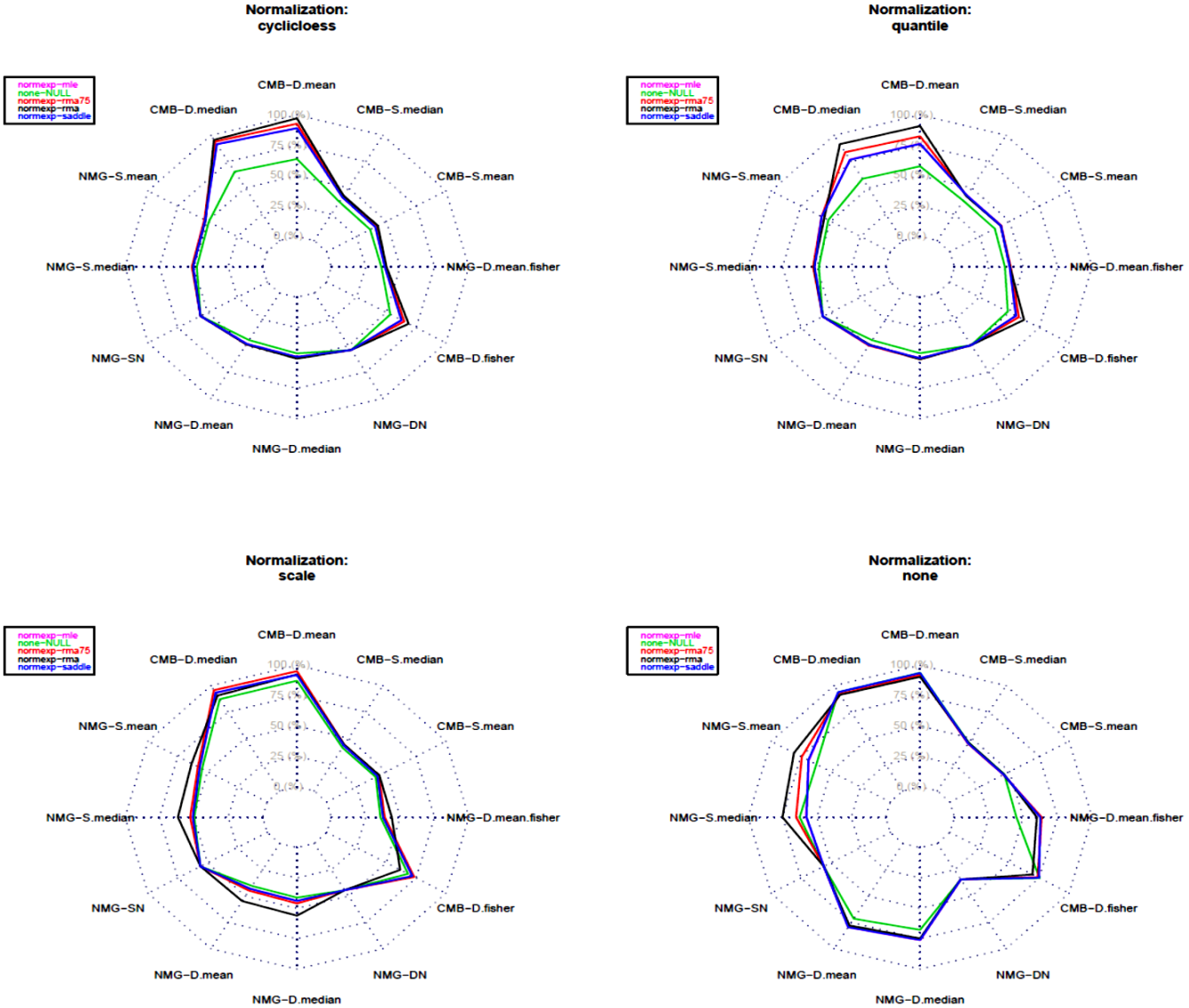

Fig-2 Radar plot of TPR values of MA designs. The figure shows the values of TPR (ranging from 0 to 1) achieved by different platforms analyzed with the 5 procedures BS (plus proprietary arrays Nimblegen) and DN (plot a, b, c, and d). The value 0 (low performance) is located at the center of the plot and the value 1 (high performance) is located on the outer margin. The plot a, b, c, and d show the TPR values when the arrays are normalized (DN) with "cyclic loess", "quantile", "scale" and "none" respectively (see materials and methods).

Table-4 *Area under the ROC curve (AUC) of Microarrays*

MA PLATFORM	AUC							
	All genes	*FPKM value*						
		<5	5-10	10-20	20-30	30-40	40-50	>50
CMB-S	0.526	0.513	0.548	0.565	0.582	0.568	0.675	0.675
CMB-D	0.654	0.545	0.612	0.756	0.857	0.843	0.900	0.906
NMG-S	0.751	0.651	0.750	0.795	0.843	0.841	0.850	0.836
NMG-D	0.842	0.781	0.861	0.892	0.882	0.889	0.897	0.845
NMG-SN	0.756	0.658	0.756	0.798	0.842	0.843	0.854	0.838
NMG-DN	0.837	0.787	0.849	0.879	0.870	0.872	0.880	0.836

AUC of ROC curves was estimated considering all expressed genes (column "All genes") and groups of genes according their FPKM expression levels (Columns "FPKM value"). Microarray data were processed using *normexp (rma) + quantile* procedure (results for the other BS-DN combinations are reported in Table S2).

Microarray and RNA-Seq PCR-Real Time Validation

Ten randomly chosen genes have been tested by real time RT-PCR to assess microarray and RNA-Seq accuracy discriminating DE genes. [Fig-3] shows a bar-plot reporting fold changes (log2FC) estimated by the technologies (i.e. RNA-Seq, NMG-D Microarray platform, and RT-PCR) for each of the 10 genes chosen to be studied by real time RT-PCR. The results were concordant (sign of log2 FCs and significant p-values) with the results of RNA-Seq in 8 of the 10 genes. For the other 2 genes (JGVV151.6, JGVV61.51) detected log2 FC values by real time RT-PCR results were not statistically significant [Fig-3]. Moreover, discordance have been observed between microarray and real time RT-PCR for 4 genes (JGVV151.6, JGVV129.66, JGVV4.362, and JGVV61.51). RT-PCR appears to exhibit a relative high agreement with RNA-Seq as opposed to microarray designs suggesting the latter as an acceptable reference assessing microarray designs performance calling DE genes

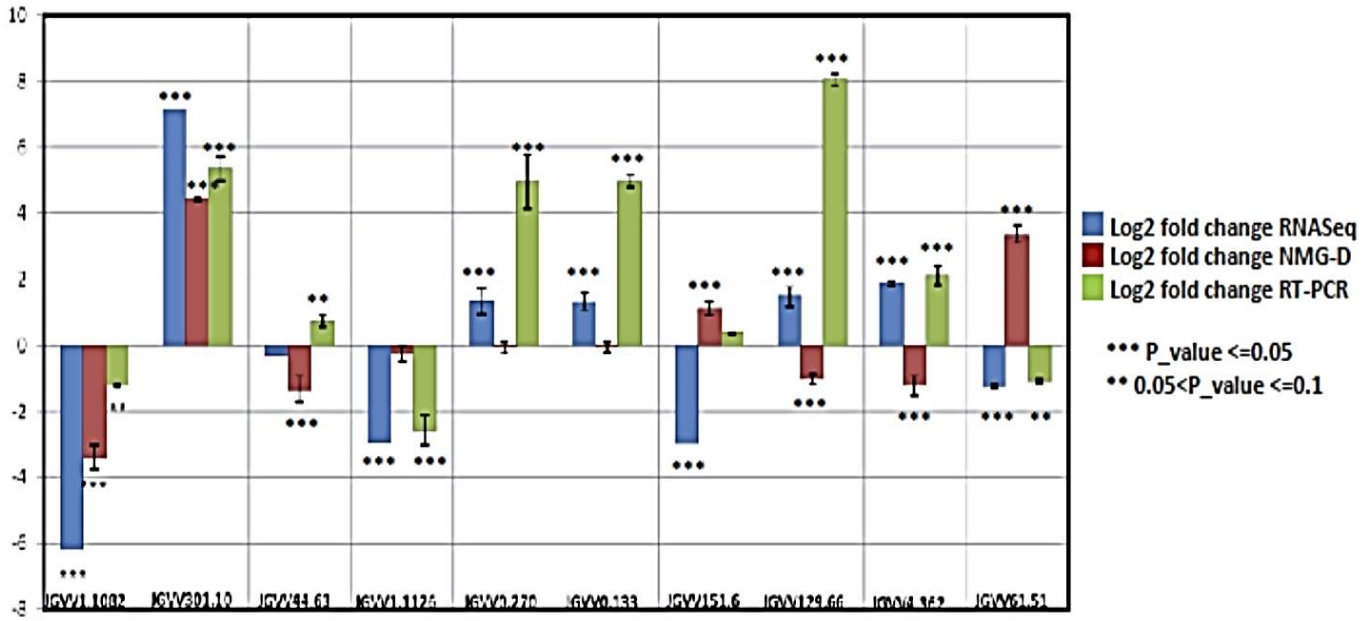

Fig-3 Quantitative RT PCR. Real time RT-PCR of 10 genes randomly selected which fold changes were in agreement or in disagreement between Microarrays and RNA-Seq. Histograms represent the fold change between of two development stages (veraison and ripening) of viticulture assessed by real time RT-PCR, Microarray (NMG-D) and RNA-Seq.

Discussion and Conclusion

Several authors showed the high reproducibility of microarray data and further indicate that the criteria used to define statistically significantly modulated genes can have a dramatic impact on the overlap of the resulting gene lists [8, 20].

The MAQC consortium, aiming to address concerns on studies reporting dissimilar or contradictory results obtained using different MA platforms, showed a high level of inter-platform concordance in terms of genes identified as differentially expressed when analyzing the same mRNA under well-controlled conditions [8, 21].

Although the reliability of microarray platforms is considered high and satisfying we would test whether the heterogeneity existing among different custom microarray platforms plays a role when searching for DE genes. Assuming that all the microarray platform designs are all the same in term of capability to detect DE genes may be inappropriate as it might not be true and hence this may lead to unsatisfying results. In this study we investigate if the efficiency in identifying DE genes is actually the same or not according to the microarray designs employed.

In this work, four custom microarrays based either on single replicate and different probes (multiple probes) per gene model transcript were processed with several bioinformatics procedures to identify DE genes between 2 viticulture stages (veraison, ripening). The 2 viticulture stages were previously investigated profiling the transcriptome (RNA-Seq) by Illumina sequencing as described in Zenoni *et al.* (2010) [9]. RNA-Seq results were used as reference to compare the performance of the developed microarray platforms detecting DE genes. In total, 17,446 genes common across all microarrays were selected to test the microarray performances. Since it is commonly accepted that microarrays are not recommended for discriminating small fold changes [21], and following the requirements imposed for a transcript or gene to be called differentially expressed [8] we arbitrarily set the two-fold change requirement [8]to claim that a gene was

differentially expressed. RNA-Seq gene expression differential analysis performed by DESeq package [22] detected 5650 DE genes. Results from the various microarray designs were compared with results obtained by RNA-Seq.

Microarray performances were assayed evaluating several parameters including correlation of gene expression fold-changes, association of DE genes, classification functions, and AUC of ROC curves.

We observed that different designs showed different results from one another in term of number of DE genes. The present study showed that fold change values estimated from the same experiments (comparison of the same RNA) conducted with different microarray platforms present a good correlation with one another. However, the different microarray platforms identify a different number of DE genes with a variable rate of true positives when standard adjustments for multiple testing are applied. We believe that these differences reflect a different accuracy of the microarray platforms in measuring the gene expression level (high or low coefficient of variability) and a different specific sensitivity in detecting the absolute intensity of each probe signal. Bioinformatics and statistical analysis (i.e combining evidence from multiple different probes of the same gene, usage of the mean or median values of the probe signals) represent other minor factors affecting on the final results.

However, this study presents some limits since we tested the performance of microarray designs using only a single comparison between samples of the same type (same tissue) in which many genes changed their expression. We cannot know if the relative performance of the arrays would be the same when varying the number of DE genes and the experiment setting (for instance, comparison of the cells from different tissues). Moreover, we used the results from RNA-Seq as reference for the comparisons assuming that they were the best choice because deriving from a technology believed superior to microarrays [2, 3 and 22]. Nevertheless, it is becoming clear that RNA-Seq also has limits. For example, the normalization between samples was believed to be not necessary but now it is coming out that standardization across samples is an important step [23].

Anyway, our study shows that the heterogeneity of the array designs affects the efficiency of finding DE genes, warning about the importance of knowing that different array platforms can give different results. In the absence of standard references it becomes crucial to know that different arrays behave differently and that the choice of the array to be used will impact heavily on the final results. Researchers should know in advance if they would benefit of a tool that provides a rich list of candidates even if including a certain percentage of false positives, or they would be rather interested in getting a list with only true positives though this can lead to have a short list of candidates. This may help in choosing the most appropriate microarray design.

In conclusion, we observed that performances in detecting differentially expressed (DE) genes of the different analyzed microarray design strategies are extremely variable despite the high correlation of the FC values of the microarray platforms with RNA-Seq. The most important factor affecting the performance in detecting DE genes resulted to be the microarray platform followed by the design (BS-DN combination, usage of mean or media probe value to synthesize the gene expression value, statistical analysis) adopted to conduct the DE analysis. Researchers should choose the microarray platform and then the bioinformatics and statistical methods by which to conduct their experiments with care and knowing in advance the performance the different microarray design available.

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 RNA-Seq gene expression row data as well as for their technical support for the present work.

References

[1] Bammler T, Beyer RP, Bhattacharya S, Boorman GA, Boyles A, Bradford BU, Bumgarner RE, Bushel PR, Chaturvedi K, Choi D, Cunningham ML, Deng S, Dressman HK, Fannin RD, Farin FM, Freedman JH, Fry RC, Harper A, Humble MC, Hurban P, Kavanagh TJ, Kaufmann WK, Kerr KF, Jing L, Lapidus JA, Lasarev MR, Li J, Li YJ,Lobenhofer EK, Lu X, Malek RL, Milton S, Nagalla SR, O'malley JP, Palmer VS, Pattee P, Paules RS, Perou CM, Phillips K, Qin LX, Qiu Y, Quigley SD, Rodland M, Rusyn I, Samson LD, Schwartz DA, Shi Y, Shin JL, Sieber SO, Slifer S, Speer MC, Spencer PS, Sproles DI, Swenberg JA, Suk WA, Sullivan RC, Tian R, Tennant RW, Todd SA, Tucker CJ, Van Houten B, Weis BK, Xuan S and Zarbl H (2005) *Nature Methods,* 2 (5), 351-356.

[2] Fu X, Fu N, Guo S, Yan Z, Xu Y, Hu H, Menzel C, Chen W, Li Y, Zeng R, Philipp K (2009) *BMC Genomics,* 10 (1),161.

[3] Paul K Korir, Paul Geeleher and Cathal Seoighe (2015) *BMC Bioinformatics,* 16,286.DOI: 10.1186/s12859-015-0712-z.

[4] Malone JH, Oliver B (2011) *BMC Biol.,* 31(9), 1-9. doi: 10.1186/1741-7007-9-34.

[5] Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A (2011) *Genome Res.,* 21(12), 2213-2223.

[6] Nalpas NC, Park SD, Magee DA, Taraktsoglou M, Browne JA, Conlon KM, Rue-Albrecht K, Killick KE, Hokamp K, Lohan AJ, Loftus BJ, Gormley E, Gordon SV, Machugh DE (2013) *BMC Genomics,* 14(1), 1-19.

[7] Daniel Hebenstreit, Miaoqing Fang , Muxin Gu , Varodom Charoensawan , Alexander van Oudenaarden and Sarah A Teichmann (2011) *Molecular Systems Biology,* 7(497), 1-9 doi: 10.1038/msb.

[8] MAQC Consortium (2006) *Nat Biotechnol.,* 24(9), 1151–1161. Doi: 10.1038/nbt1239.

[9] Sara Zenoni, Alberto Ferrarini, Enrico Giacomelli, Luciano Xumerle, Marianna Fasoli, Giovanni Malerba, Diana Bellin, Mario Pezzotti and Massimo Delledonne (2010) *Plant Physiology,* 152 (4), 1787-1795.

[10] Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyère C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, Felice N, Paillard S, Juman I, Moroldo M, Scalabrin S, Canaguier A, Le Clainche I, Malacrida G, Durand E, Pesole G, Laucou V, Chatelet P, Merdinoglu D, Delledonne M, Pezzotti M, Lecharny A, Scarpelli C, Artiguenave F, Pè ME, Valle G, Morgante M, Caboche M, Adam-Blondon AF, Weissenbach J, Quétier F and Wincker P (2007) *Nature,* 449 (7161), 463-467.

[11] Rouillard JM, Zuker M and Gulari E (2003) *Nucleic Acids Res.,* 31(12), 3057-62.

[12] Smyth G. (2004) *Statistical Applications in Genetics and Molecular Biology,* 3 (3), doi 10.2202/1544-6115.1027.

[13] Yoav Benjamini and Yosef Hochberg (1995) *Journal of the royale Statistic Society. Series B (Methodological),* 57(1), 289-300.

[14] Ann Hess and Hari Iyer. (2007) *BMC Genomics*, 8(96), 1-13 doi: 10.1186/1471-2164.

[15] Li B, Dewey CN. RSEM. (2011) *BMC Bioinformatics,* doi:10.1186/1471-2105-12-323: 1-16.

[16] Tom Fawcett (2006) *Pattern Recognition Letters,* 27, 861–874

[17] Ramakers C, Ruijter JM, Deprez RH and Moorman AF (2003) *Neurosci Lett.,* 339(1), 62-6. PMID: 12618301

[18] Pfaffl MW. (2001) A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Res 29: 45th edition.

[19] Bosotti R, Locatelli G, Healy S, Scacheri E, Sartori L, Mercurio C., Calogero R and Isacchi A (2007) Cross platform microarray analysis for robust identification of differentially expressed genes. BMC Bioinformatics. PubMed PMID: 17430572; PubMed Central PMCID: PMC1885857: 1-12.

[20] Leming Shi, Weida Tong, Hong Fang, Uwe Scherf, Jing Han, Raj K Puri, Felix W Frueh, Federico M Goodsaid, Lei Guo1, Zhenqiang Su, Tao Han, James C Fuscoe, Z Alex Xu, Tucker A Patterson, Huixiao Hong, Qian Xie, Roger G Perkins, James J Chen and Daniel A Casciano. (2005) *BMC Bioinformatics,* 6 (Suppl 2), 1-14.

[21] Dago Dougba Noel, Giovanni Malerba, Alberto Ferrarini and Massimo Delledonne (2014) *Journal of Multidisciplinary Scientific Research,* (2)6, 5-9. Available online http://jmsr.rstpublisher.com/
[22] Simon Anders (2010) Analysing RNA-Seq data with the "DESeq" package. EMBL Heidelberg.
[23] Robinson MD and Oshlack A. **(**2010) *Genome Biol.,* 11(3), 1-9 DOI: 10.1186/gb.