Applied Medical Informatics
Original Research Vol. 37, No. 4 /2015, pp: 29-37

The Impact of the Preprocessing Methods in Downstream Analysis of Agilent Microarray Data

Loredana BĂLĂCESCU[1,2,*], Ovidiu BĂLĂCESCU[1], and Ioana BERINDAN-NEAGOE[1,2]

[1]The Oncology Institute "Prof. Dr. Ion Chiricuţă", Department of Functional Genomics and Experimental Pathology, 34-36 Republicii Street, 400015 Cluj-Napoca, Romania
[2]Iuliu Haţieganu University of Medicine and Pharmacy, Research Center for Functional Genomics, Biomedicine and Translational Medicine, 23 Marinescu Street, 400337 Cluj-Napoca, Romania
E-mails: loredana_balacescu@yahoo.com; obalacescu@yahoo.com; ioananeagoe29@gmail.com

* Author to whom correspondence should be addressed; Tel.: +4-0264-590638

Received: 30 July 2015 /Accepted: 15 September 2015 / Published online: December 15, 2015

Abstract
Over the past decades, gene expression microarrays have been used extensively in biomedical research. However, these high-throughput experiments are affected by technical variation and biases introduced at different levels, such as mRNA processing, labeling, hybridization, scanning and/or imaging. Therefore, data preprocessing is important to minimize these systematic errors in order to identify actual biological changes. The aim of this study was to compare all possible combinations of two normalization, four summarization, and two background correction options, using two different foreground estimates. The results shows that the background correction of the raw median signal and summarization methods used here have no impact in downstream analysis. In contrast, the choice of the normalization method influences the results; the quantile normalization leading to a better biological sensitivity of the data. When Agilent processed signal was considered, regardless of the summarization and normalization options, there were consistently identified more differentially expressed genes (DEG) than when raw median signal was used. Nevertheless, the greater number of DEG didn't result in an improvement of the biological relevance.

Keywords: Microarray; Gene expression; Preprocessing; Agilent

Introduction

Over the past decades, microarray technologies have become routine tools in biomedical research with a broad range of applications in molecular classification of cancers [1-3], discovery of biomarkers [4-6], identification of novel drug targets [7, 8] and prediction of disease outcome [9-11]. The DNA microarray technology is based on the ability of the nucleic acids to form hydrogen bonds with complementary sequences (DNA, PCR products, oligonucleotides) chemically attached to a solid surface, usually a microscope slide or silicon chip. These technologies provide means of simultaneous measurement of mRNA levels of tens of thousands of genes, allowing the detection of changes in gene expression between two or more conditions (disease states, treatment with different drugs, etc). However, due to different handling procedures and experimental artifacts, noise and bias are often introduced in microarray experiments and these systematic errors are

reflected as differences in gene expression profiles. Therefore, data preprocessing is crucial for minimizing such systematic errors to identify actual biological changes.

Currently, there are many available gene-expression microarray platforms that differ in array production, type of targets onto the slides or dye selection. However, just a few technologies have arisen as leaders in the field due to their reliability and widespread use; cDNA microarrays, high-density oligonucleotide microarrays (Affymetrix), and long oligonucleotide microarrays (Agilent) [12]. The preprocessing methods depend on the type of microarray platform used but, in general, consist of three steps: background correction, normalization and summarization. The selection of the methods for each step can affect microarray results. While preprocessing approach for cDNA microarrays [13-15] and Affymetrix high-density oligonucleotide microarrays [16-18] were extensively studied, not so much attention is paid to Agilent microarrays.

In this study, different preprocessing strategies for Agilent microarray data were compared in terms of their ability to identify differentially expressed genes and to improve the biological sensitivity of the results.

Material and Method

Samples Collection and Processing

Peripheral blood collected from 29 breast cancer patients (BC) and 7 healthy women (CTR) was used for this study. The study was approved by the ethical committee of the University of Medicine and Pharmacy "Iuliu Hatieganu" Cluj-Napoca and all subjects signed an informed consent. The total RNA was isolated from nucleated cells according to the classical protocol with TriReagent and purified with RNeasy Mini kit (Qiagen, Germany). The integrity of RNA samples was evaluated with the Bioanalyzer 2100 (Agilent Technologies, USA). All samples had the RNA Integrity Number (RIN) greater than 8 and were further considered for the microarray experiment.

Microarray Assay

Cy3-labeled microarray probes were synthesized from 100 ng of total RNA according to manufacturer's protocol (Agilent Low Input Quick Amp Labeling Kit, Agilent Technologies, USA). The probes were hybridized on human gene expression 4x44k v2 microarray slides (G4845A, Agilent Technologies, USA). Each microarray slide comprises four individual arrays with over 44k features containing probes sourced from RefSeq, Ensemble, UniGene and GenBank databases. Slides were scanned with an Agilent G2565CA scanner at 5 microns resolution.

Microarray Data Analysis

Agilent Feature Extraction (AFE) software v.11.0 was used for gridding, quantification of foreground and background intensities and quality assessment. Preprocessing and differential analysis of raw data generated by AFE were done in R/Bioconductor (https://www.bioconductor.org/) using standard routines included in the limma package as well as custom written routines.

Two foreground estimates were used as inputs into microarray analysis; the raw median signal of feature from inlier pixels (gMedianSignal) and the signal left after AFE processing steps (gProcessedSignal). The gProcessedSignal was generated using a multiplicative detrending algorithm for background subtraction, which is described in detail in the AFE reference guide.

The data preprocessing pipeline included the following steps: background correction (optional), removal of the control probes, normalization between arrays, probeset summarization and filtering probes by flags. Two different normalization and four different summarization methods were used, with and without a background correction step when gMedianSignal was considered. The raw median signal was background corrected using *normexp+offset* method, implemented in *backgroundCorrect* function, which has been shown to be superior to a simple subtraction of the background [19]. This method uses a convolution model which is fitted to the background

subtracted signals and results in positive adjusted intensities. Data were normalised between arrays with either quantile or cyclic loess method using the *normalizeBetweenArrays* function. The role of normalization is to reduce the effect of systematic errors caused by experimental factors and make the arrays comparable to reveal actual biological differences. Quantile normalization forces the distribution of intensities on all analyzed arrays to be identical while cyclic loess is a method based on curve fitting, which iteratively applies multiple loess normalizations over all possible pairs of arrays. *BackgroundCorrect* and *normalizeBetweenArrays* functions are available through the package limma. Summarization was done at the probe level with a custom written function who computes mean, geometric mean, median or the highest normalized intensity of duplicated probesets.

Differentially expressed genes (DEG) between BC and CTR were assessed using moderated t-test from limma package. The Benjamini and Hochberg method was used to adjust p-values for multiple testing [20]. The genes were considered differentially expressed when gene expression changes exceeded 1.5-fold in BC compared to CTR and the adjusted p-values < 0.05.

Functional analysis of DEG was performed in Ingenuity Pathway Analysis (IPA) software. Fisher's exact test was used to evaluate the significance of the associations between the DEG and the canonical pathways or the biological functions. A p-value < 0.05 was considered statistically significant.

Results and Discussion

The data were analyzed with 24 different preprocessing strategies by varying one parameter (foreground estimate, background correction method, normalization method, summarization method) at each run in order to evaluate the impact of these preprocessing methods for selecting differentially expressed genes.

First, the effect of array normalization was assessed. In Figure 1a, b, the distribution of the raw and normalized intensities across all arrays for both foreground estimates was plotted. The boxplots of the non-normalized data revealed a higher dispersion of the signal intensities after AFE processing for all arrays, a pattern that is also preserved after normalization. Our data showed that quantile method performed well for both estimates; the boxplots were centered and had identical distributions. Cyclic loess tends to center the medians of arrays, but slight differences can still be observed mainly for the gProcessedSignal. However, an asymmetry was observed for both methods when dealing with gMedianSignal.

To further explore the effect of background adjustment, a background correction step before normalization of the gMedianSignal was introduced. The basic rationale of background adjustment is the assumption that the signal of a spot is not only due to the fluorescence of the labeled probes hybridized in that spot but also to a background noise introduced by a variety of experimental factors. The decision to remove or not the influence of background is controversial, with no clear criteria to favor one or another approach. Background correction could reduce bias but, on the other side, this additional estimate increases the variability in the data. Scharpf and coworkers pointed out that a high correlation of foreground to background (0.3 or greater), in two color arrays, could indicate the need for background adjustment [21]. When the correlations are borderline (0.1-0.3), it is difficult to decide whether the background adjustment is the right choice. However, in this case the authors prefer a strategy with no background correction [21]. In line with this view, the correlation of foreground to background intensities was assessed. Our data were borderline, with values in the range of 0.08 and 0.13. Therefore we considered a strategy with a background correction step for gMedianSignal. Background adjustment led to a compression of probe intensities for array 28 but, as expected, both normalization procedures uniformed the distributions (Figure 1c). However, the cyclic loess method increased spreading of the data and number of outliers for array 28. Looking back at the data, a small correlation coefficient for this array (0.08) was observed. According to above considerations, it seems that background correction was not a proper option for this array. This situation is also reflected in the MA plots (Figure 2a).

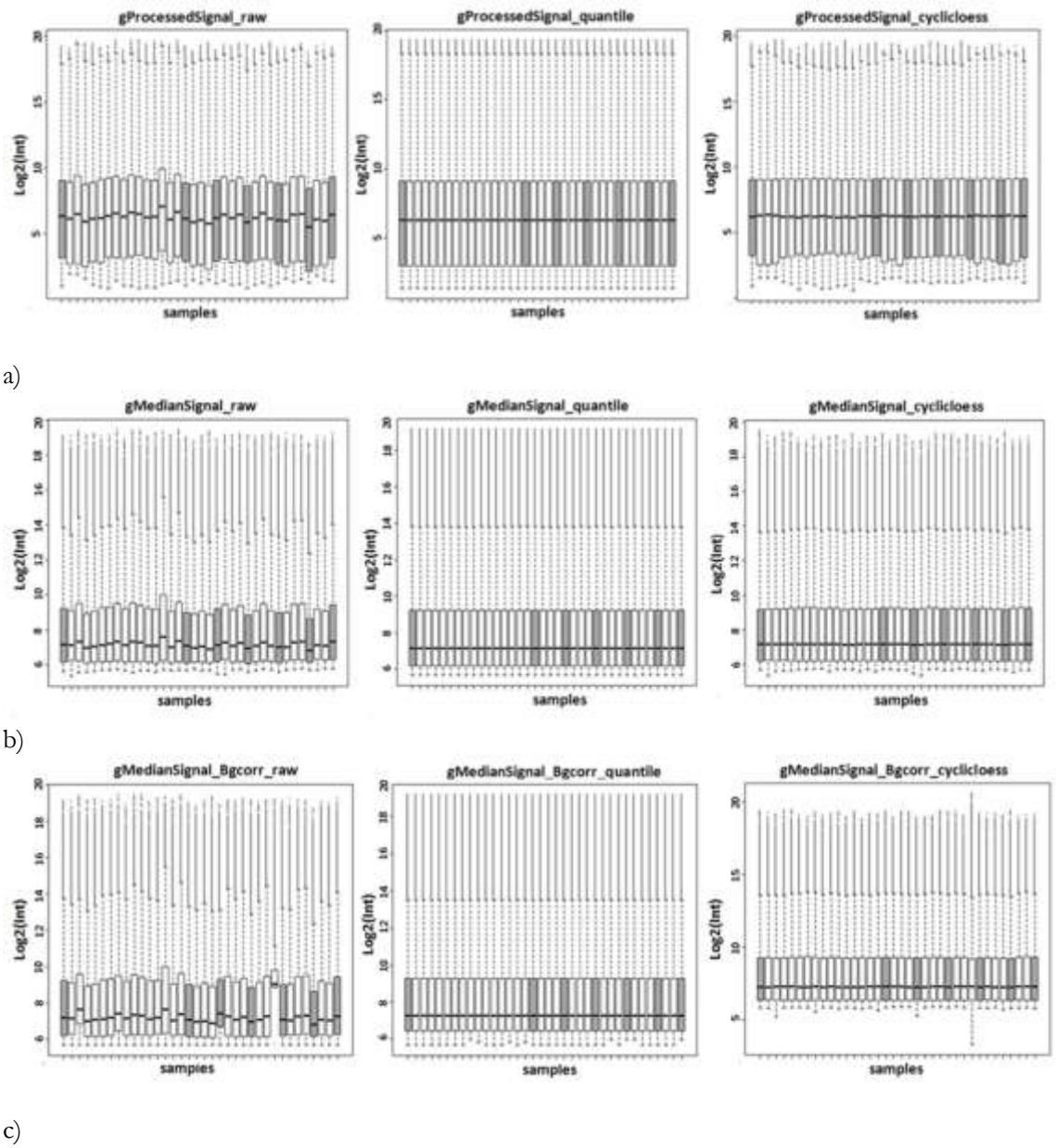

Figure 1. The distributions of the raw and normalized log2 intensities for **(a)** AFE processed signal **(b)** median signal without background correction **(c)** median signal with background correction. Each box plot represents a sample (array) and the line at the center of each box represents the median value of the distribution.

The MA plots produced with the two normalization methods were similar and overall background correction seems to not affect the variability of the data, excepting array 28 (Figure 2, Figure 1). However, the AFE processing steps seem to increase signal variability to a greater extent. At low intensities, normalized gProcessedSignal exhibited more variability than gMedianSignal, regardless of the normalization method (Figure 2).

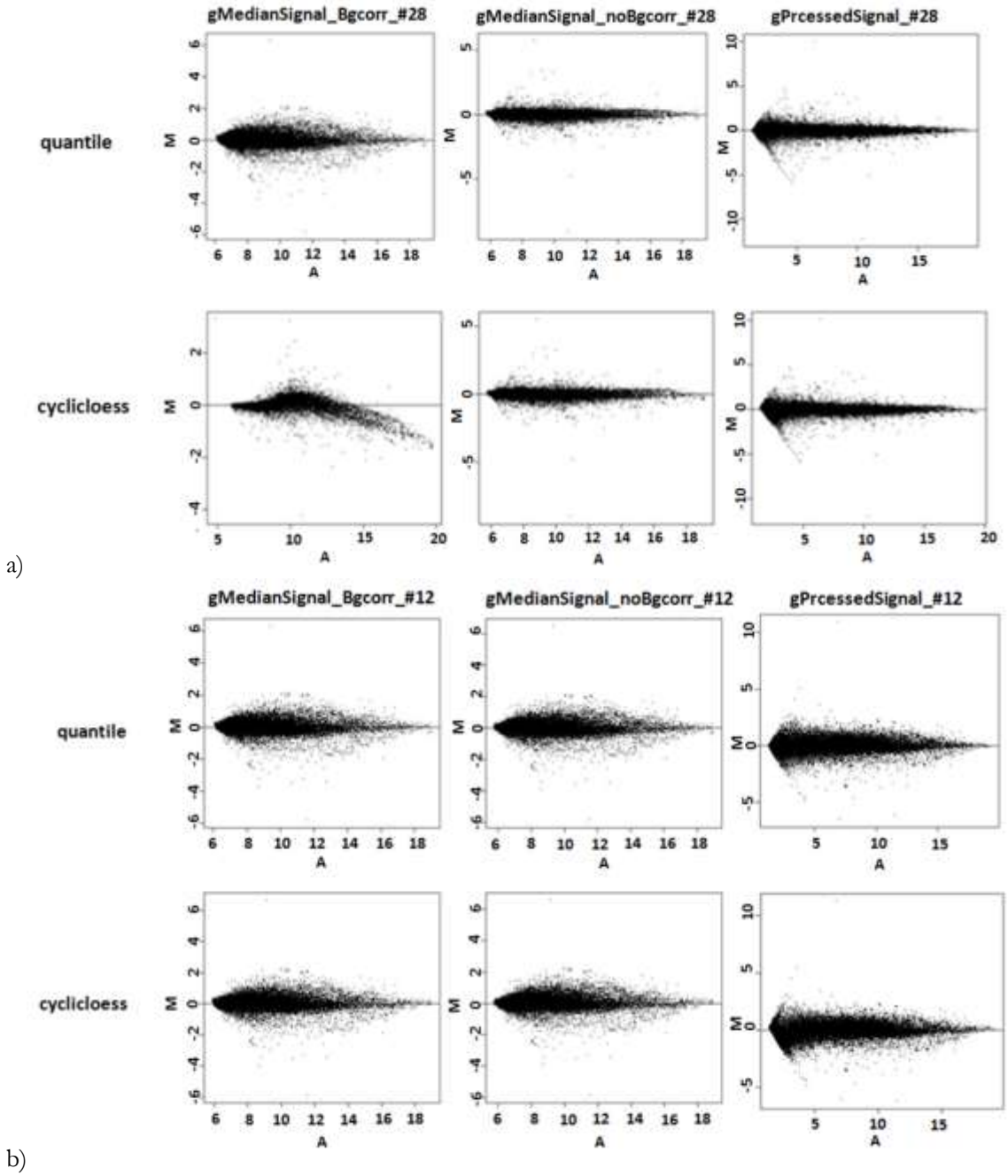

Figure 2. MA plots produced with the two normalization methods, with and without background correction for gMedianSignal **(a)** array 28 **(b)** array 12. The MA plots were generated relative to a virtual reference array where intensity of each probe is represented as its median in all analyzed samples (arrays). On y coordinate are represented the M-values as differences between log2 of each probe intensity on a certain array and corresponding intensity on the virtual array while on x coordinate are represented the A-values as the average value of these intensities.

Further, probe-level data were summarized into a single signal using four different approaches, as described in Materials and methods section. The sequences flagged as saturated and non-uniform by AFE, in 85% of the samples, were removed. After summarizing and filtering step 34127 sequences were subjected to differential analysis. The results are summarized in Table 1.

- Quantile normalization of gProcessedSignal resulted in the identification of 312 to 317 DEG, depending on the summarization method, whereas 273-274 DEG were identified

with cyclic loess. A large overlapping between the number of genes identified with cyclic loess and quantile (258-259) regardless of the summarization method was found.

- Very similar results were obtained when gMedianSignal with and without background correction was quantile normalized (100-101 DEG and 107-108 DEG). The same pattern was observed when cyclic loess was applied (82-84 DEG and 86-87 DEG). The quantile method identified almost all genes detected with cyclic loess across all summarization methods, except one gene identified just with cyclic loess.
- The majority of the DEG identified when gMedian was normalized with cyclic loess was also identified when gProcessedSignal was normalized with the same method (78-79 DEG). A large overlap was also observed for quantile method (91-92 DEG). Thus, using quantile normalized gMedianSignal we uniquely identified 15-16 DEG and when gMedianSignal was normalized with cyclic loess, we uniquely identified 8-9 DEG, whatever summarization method was used.

Table 1. The results from differential analysis obtained with different processing strategies (↑- up-regulated genes, ↓ - down-regulated genes)

		median		geometric mean		mean		max	
		#DEG	*regulation*	*#DEG*	*regulation*	*#DEG*	*regulation*	*#DEG*	*regulation*
gProcessedSignal	*quantile*	317	162↓ 155 ↑	315	161↓ 154 ↑	314	161↓ 153 ↑	312	161↓ 151 ↑
	cyclic loess	273	134 ↓ 139 ↑	274	134 ↓ 140 ↑	273	134 ↓ 139 ↑	273	135↓ 140 ↑
gMedianSignal	*quantile*	107	37 ↓ 70 ↑	108	38↓ 70 ↑	108	38↓ 70 ↑	108	39↓ 69 ↑
	cyclic loess	87	21 ↓ 66 ↑	87	21 ↓ 66 ↑	86	21 ↓ 65 ↑	87	22 ↓ 65 ↑
gMedianSignal_ BgCorrected	*quantile*	100	34 ↓ 66 ↑	101	35 ↓ 66 ↑	101	35 ↓ 66 ↑	101	36↓ 65 ↑
	cyclic loess	84	22 ↓ 62 ↑	84	22 ↓ 62 ↑	83	22 ↓ 61 ↑	82	22 ↓ 60 ↑

Although the majority of the preprocessing methods presented here were assessed in terms of performance of the microarray data, to our knowledge there are no reports that describe their impact for selecting differentially expressed genes. Our results showed that regardless of the preprocessing methods, background correction of the raw median signal seems to not affect the results, but the algorithm of background correction implemented in AFE has a great impact in downstream analysis. Our data showed that AFE preprocessing led to a higher dispersion of the signal intensities and resulted in a large increase in the number of DEG. The summarization methods used in this paper have no impact regarding the number of DEG, but as expected, they have a small effect in terms of magnitude of the fold changes for some genes. On the other hand, the normalization methods appear to influence to a greater extent the results. If in terms of reducing variability, both methods performed comparably, in terms of impact in the selection of DEG, quantile method performed better than cyclic loess. Using quantile, more genes than with cyclic loess normalization were consistently identified. Our results are in accordance with the literature, which indicate quantile as a robust method for normalization of microarray data [22].

Although some of the above approaches led to a higher number of DEG, it is interesting to verify if these additional genes are biological relevant. To check whether some of these strategies result in better biological sensitivity, functional analysis in IPA was performed for three gene datasets:

(i) gene dataset 1: DEG obtained using gMedianSignal, quantile normalized, median summarized

(ii) gene dataset 2: DEG obtained using gProcessedSignal, quantile normalized, median summarized

(iii) gene dataset 3: DEG obtained using gProcessedSignal, cyclic loess normalized, median summarized

Using IPA we identified: (i) 36 canonical pathways (CP) and 21 molecular and cellular functions (MCF) for IPA dataset 1, (ii) 25 CP and 20 MCF for IPA dataset 2 and (iii) 24 CP and 18 MCF for IPA dataset 3. For each gene dataset, five of the most significant CP and MCF are listed in Table 2.

Table 2. The top five canonical pathways and molecular and cellular functions obtained in IPA for three different strategies (three gene datasets)

	gene dataset 1 Input: gMedianSignal Normalization: quantile Summarization: median		**gene dataset 2 Input: gProcessedSignal Normalization: quantile Summarization: median**		**gene dataset 3 Input: gProcessedSignal Normalization: cyclic loess Summarization: median**	
Top Canonical Pathways (CP)						
Name	*p-value*	*overlap (%)*	*p-value*	*overlap (%)*	*p-value*	*overlap (%)*
Granulocyte Adhesion and Diapedesis	7.36e-08	5.5	2.68e-07	8	1.68e-05	6.1
Agranulocyte Adhesion and Diapedesis	1.56e-06	4.6	3.4e-06	6.9	1.58e-04	5.2
Differential Regulation of Cytokine Production in Macrophages and T Helper Cells by IL-17A and IL-17F	7.84e-05	16.7	1.61e-03	16.7	1.03e-03	16.7
Communication between Innate and Adaptive Immune Cells	2.26e-05	6.9	2.65e-03	6.9	1.33e-03	6.9
Differential Regullation of Cytokine Production in Intestinal Epithelial Cells by IL-17A and IL-17F	1.67e-04	13	3.34e-03	13	2.13e-03	13
Total number of CP	***36***		***25***		***24***	
Top Molecular and Cellular Functions (MCF)						
Name	*p-values range*	*#molecules*	*p-values range*	*#molecules*	*p-values range*	*#molecules*
Cellular Movement	9.48e-04 – 1.30e-11	33	4.64e-03 – 5.13e-10	61	4.35e-03 – 1.46e-10	55
Cellular Development	9.22e-04 – 1.35e-16	52	4.64e-03 – 9.52e-09	99	3.57e-03 – 4.19e-08	84
Cell_to-Cell Signaling and Interaction	8.52e-04 – 7.72e-09	33	4.64e-03 – 1.02e-07	64	4.35e-03 – 9.18e-07	54
Cellular Growth and Proliferation	9.31e-04 – 5.15e-09	48	3.94e-03 – 2.37e-07	95	3.57e-03 – 1.84e-06	54
Cell Death and Survival	8.64e-04 – 1.79e-10	46	4.65e-03 – 2.24e-06	82	4.35e-03 – 7.88e-07	77
Total number of MCF	***21***		***20***		***18***	

One might think that the more DEG in a dataset are subjected to functional analysis, the more significant CP and MCF are over-represented in that dataset. Conversely, we identified more statistically significant CP and MCF for gene dataset 1 (gMedianSignal) than for gene dataset 2 (gProcessedSignal), although the second dataset contains with 200 DEG more than the first one (Table 1). On the other hand, the p-values for CP and MCF in gene dataset 1 were consistently more significant (smaller values) than p-values for CP and MCF in gene dataset 2 (Table 2). These results show that using the signal preprocessed by AFE instead of the raw median signal doesn't lead to a better biological sensitivity, despite increased number of DEG.

When the normalization method was taken into account, we noticed that a small increase in the number of DEG (n=44) led to an increase in the number of MCF, CP and statistical significance (Table 2). These results confirm that the choice of normalization method impacts the results and show that quantile method improves the biological relevance of the data.

Conclusions

Our study showed that microarray results are affected to a greater or a lesser extent by the choice of the preprocessing methods. We found that background correction of the raw median signal and summarization methods have no impact in downstream analysis while the choice of the normalization method influences the results. Using quantile method we consistently identified more DEG than with cyclic loess normalization and this increase in the number of DEG result in an improvement of the biological relevance. Our data highlighted that preprocessing algorithm implemented in AFE led to identification of a larger number of DEG but that is not reflected in a better biological sensitivity.

List of abbreviations

AFE – Agilent Feature Extraction software
BC – breast cancer patients
CP - canonical pathways
CTR – control group (healthy female)
DEG - differentially expressed genes
IPA - Ingenuity Pathway Analysis
MCF - molecular and cellular functions

Conflict of Interest

The authors declare that they have no conflict of interest.

Acknowledgements

Dr. Loredana Bălăcescu is a fellow of POSDRU grant no. 159/1.5/S/138776 "Model colaborativ institutional pentru translarea cercetarii stiintifice biomedicale in practica clinica – TRANSCENT".

This study was supported by a POSCEE grant 709/2010.

References

1. Geyer FC, Rodrigues DN, Weigelt B, Reis-Filho JS. Molecular classification of estrogen receptor-positive/luminal breast cancers. Adv Anat Pathol. 2012;19(1):39-53.
2. Comprehensive molecular portraits of human breast tumours. Nature 2012;490(7418):61-70.
3. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature 2000;406(6797):747-52.
4. Milioli HH, Vimieiro R, Riveros C, Tishchenko I, Berretta R, Moscato P. The Discovery of Novel Biomarkers Improves Breast Cancer Intrinsic Subtype Prediction and Reconciles the Labels in the METABRIC Data Set. PLoS One 2015;10(7):e0129711.
5. Sorensen KD, Orntoft TF. Discovery of prostate cancer biomarkers by microarray gene expression profiling. Expert Rev Mol Diagn. 2010;10(1):49-64.
6. Cooper CS, Campbell C, Jhavar S. Mechanisms of Disease: biomarkers and molecular targets from microarray gene expression studies in prostate cancer. Nat Clin Pract Urol. 2007;4(12):677-687.
7. Kim YW, Liu TJ, Koul D, Tiao N, Feroze AH, Wang J, et al. Identification of novel synergistic targets for rational drug combinations with PI3 kinase inhibitors using siRNA synthetic lethality screening against GBM. Neuro Oncol. 2011;13(4):367-375.

8. Jayapal M, Melendez AJ. DNA microarray technology for target identification and validation. Clin Exp Pharmacol Physiol. 2006;33(5-6):496-503.
9. Reis-Filho JS, Pusztai L. Gene expression profiling in breast cancer: classification, prognostication, and prediction. Lancet 2011;378(9805):1812-1823.
10. Kao KJ, Chang KM, Hsu HC, Huang AT. Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization. BMC Cancer 2011;11:143.
11. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415(6871):530-536.
12. Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet. 2006;7(1):55-65.
13. Fan J, Tam P, Vande Woude G, Ren Y. Normalization and analysis of cDNA microarrays using within-array replications applied to neuroblastoma cell response to a cytokine. Proc Natl Acad Sci USA 2004;101(5):1135-1140.
14. Smyth GK, Speed T. Normalization of cDNA microarray data. Methods 2003;31(4):265-273.
15. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. 2002;30(4):e15.
16. Cope LM, Irizarry RA, Jaffee HA, Wu Z, Speed TP. A benchmark for Affymetrix GeneChip expression measures. Bioinformatics 2004;20(3):323-331.
17. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. 2003;31(4):e15.
18. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003;4(2):249-264.
19. Ritchie ME, Silver J, Oshlack A, Holmes M, Diyagama D, Holloway A, et al. A comparison of background correction methods for two-colour microarrays. Bioinformatics 2007;23(20):2700-2707.
20. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Statist Soc Ser B (Methodological) 1995;57:289-300.
21. Scharpf RB, Iacobuzio-Donahue CA, Sneddon JB, Parmigiani G. When should one subtract background fluorescence in 2-color microarrays? Biostatistics 2007;8(4):695-707.
22. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003;19(2):185-193.