ACTA ABP BIOCHIMICA POLONICA

Vol. 58, No 4/2011
573–580
on-line at: www.actabp.pl

Regular paper

Impact of DNA microarray data transformation on gene expression analysis — comparison of two normalization methods

Marcin T. Schmidt[1#✉], Luiza Handschuh[2,3#], Joanna Zyprych[4], Alicja Szabelska[4], Agnieszka K. Olejnik-Schmidt[1], Idzi Siatkowski[4] and Marek Figlerowicz[3,5]

[1]Department of Biotechnology and Food Microbiology, Poznan University of Life Sciences, Poznań, Poland; [2]Institute of Bioorganic Chemistry PAS, Poznań, Poland; [3]Poznan University of Medical Sciences, Department of Hematology, Poznań, Poland; [4]Department of Mathematical and Statistical Methods, Poznan University of Life Sciences, Poznań, Poland; [5]Institute of Computing Science, Poznan University of Technology, Poznań, Poland

Two-color DNA microarrays are commonly used for the analysis of global gene expression. They provide information on relative abundance of thousands of mRNAs. However, the generated data need to be normalized to minimize systematic variations so that biologically significant differences can be more easily identified. A large number of normalization procedures have been proposed and many softwares for microarray data analysis are available. Here, we have applied two normalization methods (median and loess) from two packages of microarray data analysis softwares. They were examined using a sample data set. We found that the number of genes identified as differentially expressed varied significantly depending on the method applied. The obtained results, i.e. lists of differentially expressed genes, were consistent only when we used median normalization methods. Loess normalization implemented in the two software packages provided less coherent and for some probes even contradictory results. In general, our results provide an additional piece of evidence that the normalization method can profoundly influence final results of DNA microarray-based analysis. The impact of the normalization method depends greatly on the algorithm employed. Consequently, the normalization procedure must be carefully considered and optimized for each individual data set.

Keywords: microarray, gene expression profiling, transcriptome analysis, data normalization, adhesion, probiotic, enterocyte

Received: 16 May, 2011; revised: 25 November, 2011; accepted: 12 December, 2011; available on-line: 20 December, 2011

INTRODUCTION

DNA microarrays are well-established tools for biological, medical and pharmaceutical research (Trevino *et al.*, 2007; Cowell & Hawthorn, 2007). DNA microarray technology enables a simultaneous analysis of thousands of genes/transcripts/genome sequences. It can be applied for many purposes from genotyping to gene expression profiling (Howbrook *et al.*, 2003; Venkatasubbarao, 2004). The last one is still the most popular application of DNA microarrays. In the simplest approach, this technique involves RNA isolation from cells, reverse transcription, fluorescent labeling of the resultant cDNA followed by its hybridization with probes immobilized on a solid surface, usually a glass slide (Simon *et al.*, 2007). Particular genes are represented by one or more specific probes on the array. After washing out of unbound cDNAs the microarray is scanned to determine the level of fluorescence emitted by each probe. Then the obtained results are digitalized. In a dual-label approach the application of two fluorescent dyes enables a direct comparison of two RNA samples (tested and control/reference samples). The ratio between the intensity of signals detected for the examined and reference samples hybridized with the same probe reflects the difference in the level of a single gene's expression. Such ratios need to be determined for each individual probe on the array. Usually, a microarray experiment involves a number of hybridizations and produces a large amount of data. Moreover, the experiments are often affected by numerous factors that can lead to unwanted, random or systematic (non-biological), variation (Chua, 2006). To obtain reliable results the collected data have to be normalized and analyzed using proper statistical methods (Yang *et al.*, 2001; 2002; Quackenbush, 2002; Simon *et al.*, 2007; Ness, 2007; Baker, 2008). Due to the wider and wider application of microarray-based techniques, a broad spectrum of programs devoted to microarray data analysis is available. Among them, Bioconductor, operating in R environment (R Development Core Team, 2009), is one of the most commonly used and recommended (Gentleman *et al.*, 2004 and 2005; Hahne *et al.*, 2008). In fact, Bioconductor offers much more than all other programs, beeing an open source and open development software project, designed for genomic data handling and analysis (http://Bioconductor.org). Based on various statistical approaches, several methodologies and many software packages have been developed in Bioconductor for the analysis of data generated with different types of microarrays (http://Bioconductor.org/packages/release/Software.html; Gentleman *et al.*, 2005; Hahne *et al.*, 2008). There are, however, several observations suggesting that some theoretically equipotent methods of data transformation offered by various software packages can differently influence the results of microarray experiments. To verify the above presumption we have compared two apparently similar normalization methods (median and loess) implemented in two types of software: GenePix Pro/Acuity (Molecular Devices)

✉e-mail: mschmidt@up.poznan.pl
#both authors contributed equally to the study
Abbreviations: FC, fold change.

and Bioconductor. As a probe data set we used representative results of Caco-2 cell transcriptome analysis. The tested intestinal epithelial cells were subjected to interactions with selected probiotic microorganisms. The changes in their transcriptome were analyzed with Operon DNA microarrays (Human V4.0 OpArrays).

Probiotic microorganisms (mainly lactic acid bacteria) have been shown to have beneficial effects on human and animal health. The majority of the effects are attributed to their interaction with intestinal epithelium, mainly through adhesion to enterocytes (Heczko *et al.*, 2006). Among the most widely used and best characterized are *Lactobacillus rhamnosus* GG (ATCC 53103) and *Bifidobacterium animalis* subsp. *lactis* Bb12 (Nestle). Both strains are commercial probiotic bacteria, which are known to be able to adhere to enterocytes (Gopal *et al.*, 2001). Since it is very difficult to study bacterial adhesion *in vivo*, intestinal cell lines are used as *in vitro* models. One of them is Caco-2 cell line derived from human colon carcinoma. The cell line differentiates spontaneously under standard cell culture conditions and expresses several markers that are distinctive of normal small intestinal villi. The Caco-2 cell line has become a standard tool in the pharmaceutical industry, applied e.g. for investigation of drug transport through intestinal epithelium (Sambuy *et al.*, 2005; Delgado *et al.*, 2008).

MATERIALS AND METHODS

Strains and growth conditions. Probiotic microorganisms: *Bifidobacterium animalis* subsp. *lactis* Bb12, *Lactobacillus rhamnosus* GG (ATCC 53103), *L. acidophilus* LA-5, *L. plantarum* PL02, *L. rhamnosus* KL53A, *L. delbruecki* subsp. *bulgaricus* LBY-27, and *Lactococcus lactis* PB411 were cultured in glucose-free Brain and Heart Infusion broth (BHI; BTL) supplemented with 2% fructooligosaccharides (FOS; Sigma-Aldrich) as a carbon source. Probiotic yeast *Saccharomyces cerevisiae* subsp. *boulardi* was grown in YAPD broth (Merck). All microorganisms were cultivated at 37 °C for 20 hours in anaerobic conditions (Anaerocult A, Merck).

Caco-2 cell culture and *in vitro* adhesion assay. The Caco-2 cell line (ATCC HTB37) was cultured in Dulbecco's Modified Eagle's Minimal Essential Medium (DMEM; Sigma-Aldrich) supplemented with 1x Non-Essential Amino Acids (Sigma-Aldrich), 10% heat-inactivated fetal calf serum (Gibco-Invitrogen), 50 μg/ml gentamycin (Gibco-Invitrogen) at 37 °C in an atmosphere of 10% CO_2/90% air. The cells were cultured on PTFE filter (1 μm pore size, Millipore) at a concentration of 4×10^5 cells/cm^2 to obtain confluence. The culture medium was changed every second day and cells were maintained for 21 days, until differentiation.

The apical side of differentiated Caco-2 cell monolayer was washed twice and overlayed with 2 ml of 10 mM HEPES-buffered Hank's Balanced Salts Solution preconditioned at 37 °C in an atmosphere of 10% CO_2/90% air. Approximately 10^8 cfu of bacteria was added to the Caco-2 apical side and incubated for 4 h at 37 °C in an atmosphere of 10% CO_2/90% air. A mock-control was performed as well. Four separate adhesion experiments were carried out with: *L. rhamnosus* GG (L), *B. animalis* Bb12 (B), *S. cerevisiae* subsp. *boulardi* (Y) and probiotic mixture (M) consisting of equal amounts of: *L. acidophilus* LA-5, *L. plantarum* PL02, *L. rhamnosus* KL53A, *L. delbruecki* LBY-27, *L. lactis* PB411, and *B. animalis* Bb12.

Microarray experiment. After the adhesion assay the Caco-2 cells were suspended in Trizole Reagent (Invitrogen) and total RNA was extracted according to the standard Trizole procedure (Invitrogen). The DNA contamination was removed from the samples by digestion with DNase (TURBO DNA-free kit, Applied Biosystems). The quantity of the total RNA was evaluated using a NanoDrop ND-1000 spectrophotometer (Nanodrop Technologies) and its integrity was verified on a BioAnalyzer 2100 (Agilent Technologies). Reverse transcription involving anchored-oligo$(dT)_{20}$ and amino-allyl-modified nucleotides was performed with SuperScript Plus Indirect cDNA Labeling System (Invitrogen). Amino-allyl-cDNA was labeled with AlexaFluor 555 or AlexaFluor 647. Human V4.0 OpArray (Operon Biotechnologies GmbH) microarrays containing over 35k 70-nt-long oligonucleotide probes were hybridized with 1–2 μg of labeled cDNA. Hybridization and washing were performed in automatic hybridization station HybArray12 (Perkin Elmer), in a buffer containing 5×SSC, 0.1% SDS, and 0.1 mg BSA/ml. Step-down hybridization protocol (5.5 h incubation at 60 °C, 55 °C and 50 °C each, 16.5 h in total) was followed by 3 washes: (I) 2× SSC, 0.1% SDS at 35 °C, (II) 2x SSC at RT, (III) 0.2× SSC at RT. The slides were dried through centrifugation and scanned with a ScanArrayExpress (Perkin Elmer) laser scanner at 5-μm resolution.

Data analysis. Microarray images were processed using GenePix Pro v. 6.0 software (Molecular Devices) in order to obtain numerical data (raw data). Spots of poor quality ("bad" and "not found") were removed from further analysis by automatic flagging and filtration. Raw data (gpr files) were then submitted to a normalization procedure. Within array normalization was performed using two methods (median and loess), both implemented in different software tools as depicted in Fig.1. Median normalization was executed in GenePix Pro v. 6.0 software (Molecular Devices) and limma software (a package of R Bioconductor; Smyth, 2005). Loess method was performed in Acuity v. 4.0 software (Molecular Devices) and limma software (R Bioconductor) as well. Bioconductor R platform v.2.10.0 (R Development Core Team, 2009) was used.

Median normalization. The median (as well as mean) method is the most simple and the earliest normalization approach. This kind of global normalization approach treats all spots on a microarray equally, subtracting a constant from all intensity log-ratios, usually their mean or median value (Churchill, 2002; Dudoit *et al.*, 2002). It is implemented in the majority of programs designed for microarray data analysis. In the GenePix Pro software tested here, the median local background values were simply subtracted from the corresponding median spot intensities. For the data normalization the "Ratio of Medians" method was selected. An inter-channel normalization factor was set as equal to 1.

In the limma package the median method subtracts the weighted median from the M-values for each array. It was performed using "normalizeWithinArrays" function with the "median" method declared.

Loess normalization. This method is a combination of linear least squares regression and nonlinear regression. It builds a function that describes the deterministic part of the variation in the data by fitting simple models to localized subsets of the data point by point. The loess function depends on the set of parameters. Parameter 'smoothing' ('bandwidth') is set between 0 and 1 and determines the proportion of the data that is

used to fit each local polynomial. Parameter 'iterations' is the number of iterations of loess fit. Parameter 'delta' is connected with the speed of computations, representing the proportion of the data which are grouped in a single bin during local regression fit. For the points within the bin the fitted values are filled in based on linear interpolation. The aforementioned parameters for Acuity and R softwares were specified as follows: 'smoothing' = 0.4, 'iterations' = 3, 'delta'=0.01 of the range of data. To learn more about the method used in the Acuity software see Dudoit *et al.*, (2002) and Yang *et al.*, (2002). In the case of limma package in R the normalization was performed with the "normalizeWithinArrays" function with the "loess" method declared. A detailed description of this implementation can be found in Smyth and Speed (2003).

Gene selection. The four variously-normalized data sets of M and A values ("median" from GenePix Pro, "median" from limma, "loess" from "Acuity" and "loess" from limma) were then separately submitted to the same sequence of further analysis in R software. Normalization between arrays was done using the aquantile method implemented in the function "normalizeBetweenArrays" in limma package. To select differentially expressed genes, we calculated the fold change (FC) of every gene. The obtained M and A values from each normalization method had to be first transformed again to the R and G values. Afterwards, the fold change was derived as a ratio of the R and G values. The genes with an FC value greater than 1.7 or smaller than 1/1.7 were considered as differentially expressed.

RESULTS

The study was conducted to check how the microarray data normalization methods influence the results of microarray experiment. A long-term aim of our research is to discover gene expression changes in human intestinal epithelial cells induced by probiotic microorganisms colonizing the human gut. Prior to the main experiment we needed to establish the optimal data normalization procedure.

Using Caco-2 cell culture we performed four separate cell adhesion assays with various probiotic microorganisms: *B. animalis* Bb12 (experiment B), *L. rhamnosus* GG (experiment L), mixture (experiment M) consisting of equal amounts of six probiotic bacteria strains (*L. acidophilus* LA-5, *L. plantarum* PL02, *L. rhamnosus* KL53A, *L. delbruecki* LBY-27, *L. lactis* PB411, and *B. animalis* Bb12) and *S. cerevisiae* subsp. *boulardi* (experiment Y). RNA samples (isolated from the variously treated or control Caco-2 cells) were reverse transcribed, fluorescently labeled and hybridized to the Human V4.0 OpArray oligonucleotide arrays. Standard two-color approach was applied to compare transcriptomes of the treated cells versus non-treated control, used as a common reference. Four microarray hybridizations (array B, L, M and Y), corresponding to four cell adhesion assays, were then used as a trial set for testing of the normalization procedures. In fact, only one normalization step — within-array normalization — was tested, as the crucial step in two-color microarray data pr. We compared two methods of within-array-normalization: median and loess, both implemented in two different types of software: commercial (GenePix Pro and Acuity, supported by Molecular Devices) and free (limma, a package of R Bioconductor) (Fig. 1). As a result we

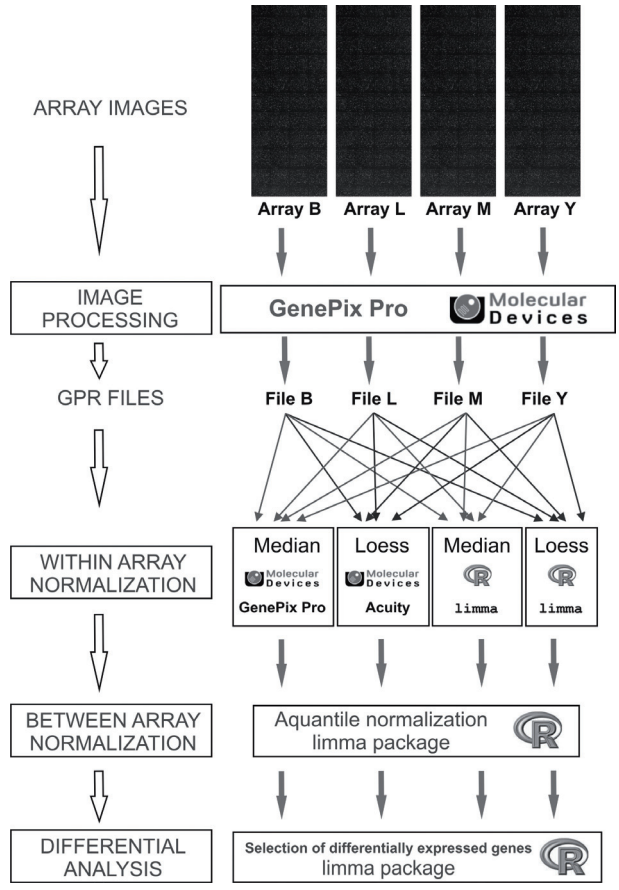

Figure 1. General scheme of experiment.
Four microarrays were used for gene expression profiling in Caco-2 cells treated with probiotic microorganisms: *B. animalis* Bb12 (Array B), *L. rhamnosus* GG (Array L), mixture of six probiotic bacterial strains (Array M), probiotic yeasts (Array Y). All images were processed using GenePix Pro software (Molecular Devices). Raw data files (gpr files) were then separately submitted to four different normalization approaches: two methods (median and loess), each performed with two types of software: commercial programs from Molecular Devices (GenePix Pro/Acuity) or open-source R Bioconductor (limma package). The next steps of analysis, performed only in R Bioconductor (limma package), were identical for all datasets. The end results were sixteen lists of genes selected as differentially expressed in probiotic-treated cells when compared to an untreated control.

obtained 16 within-array-normalized files (4 arrays multiplied by 4 methods). All were subjected to the same further steps of analysis — between-array normalization and differential analysis — using limma software only. Between-array normalization was done with the "Aquantile" method. Then, the fold change approach was applied for identification of differentially expressed genes in each set of normalized data. In Table 1, besides the total numbers of differentially expressed genes for each experiment, we indicate the number of down- and up- regulated genes.

The numbers of genes identified as differentially expressed between treated and untreated Caco-2 cells depend on the software and method used for normalization within microarrays. The most substantial differences appear between the two loess methods implemented in the two softwares compared, and the median and loess methods implemented in the Molecular Devices softwares. Similar discrepancies can be observed for all four microarrays.

Table 1. Numbers of differentially expressed genes obtained from each microarray experiment.
Microarray analysis of Caco-2 cell transcriptome was performed after treatment with probiotic bacteria: B. animalis Bb12 (B), *L. rhamnosus* GG (L), mixture of six selected strains (M), and probiotic yeast (Y), as compared with an untreated control. Total numbers of differentially expressed (all), and up (↑) and down (↓) regulated genes (absolute fold-change greater than or equal to 1.7) were identified in the datasets normalized with two different methods (median and loess) in two types of software (Molecular Devices GenePix Pro for median and Molecular Devices Acuity for loess (MD) and R Bioconductor limma for both median and loess (R)).

Software type and normalization method			Microarray analysis			
			B	L	M	Y
MD	median	all	6958	4353	5796	5525
		↑	2484	179	1619	2431
		↓	4474	4174	4177	3094
	loess	all	14667	12199	17049	15155
		↑	6290	2226	7914	6908
		↓	8377	9973	9135	8247
R	median	all	6986	3734	5685	5531
		↑	2576	403	1816	2479
		↓	4410	3331	3869	3052
	loess	all	3435	1984	3479	4115
		↑	1455	497	1688	2332
		↓	1980	1487	1791	1783

Table 2 presents the numbers of differentially expressed genes included into a common set of genes that was determined for particular microarrays, programs and normalization methods. The data presented in Table 2 clearly shows a high compatibility only between the two median methods (81–99%).

Venn diagrams, presented in Fig. 2, also demonstrate a higher conformity between the two median methods for all four microarrays (Fig. 2.1–2.4). The majority of genes determined as differentially expressed are shared between the data sets normalized with the median methods in the two softwares. The common sets of genes are definitely less numerous in the cases of the two loess methods (Fig. 2.5–2.6).

The number of genes indicated as differentially expressed after loess normalization with the R Bioconductor limma software is much smaller than after the loess normalization performed with the Molecular Devices Acuity software. The loess normalization executed in limma seems to be more restrictive. However, not all genes selected after limma loess normalization are found among those assigned as differentially expressed after loess normalization with the Molecular Devices Acuity software. This suggests more general differences in the algorithms implemented in the two loess methods.

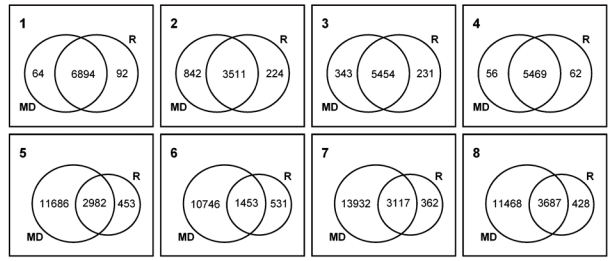

Figure 2. Overlap of lists of genes found to be differentially expressed.
Probiotic microorganisms used in Caco-2 treatments were as follows: *B. animalis* Bb12 (1, 5), *L. rhamnosus* GG (2, 6), mixture of six selected strains (3, 7), and yeast (4, 8). The normalization approaches were: median (1-4) and loess (5-8) performed with Molecular Devices GenePix Pro/Acuity software (MD), or R Bioconductor (R).

Taking into account the biological background of the experiment, we expected that microarrays B, L, and M should give similar results as they all concern treatment of human epithelial intestinal cells with probiotic bacterial strains. The Venn's diagrams presented in Fig. 3 show common sets of differentially expressed genes for these three microarrays for each normalization method and software. Surprisingly, only 3–8% of the differentially expressed genes obtained after the two median normalization approaches are shared by the B, L and M data sets (Fig. 3.1–3.2). In the case of the loess normalization method from Molecular Devices Acuity, the common set of genes comprised about 21% of differentially expressed genes selected for each microarray separately (Fig. 3.3). At first sight, the Acuity loess method seems to be the best one, since it results in the most abundant set of shared genes. However, it must be remembered that this normalization method led to the most numerous lists of differentially expressed genes, comprising between one third and half of the genes present on the microarrays (see Table 1).

Table 2. Common sets of differentially expressed genes determined using two types of software and two methods of normalization.
Sample and software notation as in Table 1. Two percentages are counted per microarray — in each case they are derived as a quotient of common differentially distributed gene sets and the total number of differentially expressed genes in the case considered.

Z ∩ W	Microarray analysis							
	B		L		M		Y	
MDmedian ∩ MDloess	4755		2775		4327		4505	
	68% of Z	32% of W	64% of Z	23% of W	75% of Z	25% of W	82% of Z	30% of W
Rmedian ∩ Rloess	2299		1524		2500		3189	
	33% of Z	67% of W	41% of Z	77% of W	44% of Z	72% of W	58% of Z	77% of W
MDmedian ∩ Rmedian	6894		3511		5454		5469	
	99% of Z	99% of W	81% of Z	94% of W	94% of Z	96% of W	99% of Z	99% of W
MDloess ∩ Rloess	2982		1453		3117		3687	
	20% of Z	87% of W	12% of Z	73% of W	18% of Z	90% of W	24% of Z	90% of W

Z ∩ W, a pair of compared methods, different in each line of the table; MDmedian, Molecular Devices GenePix Pro median normalization; Rmedian, R limma median normalization; Rloess, R limma loess normalization; MDloess, Molecular Devices Acuity loess normalization; B, L, M, Y, symbols of the arrays.

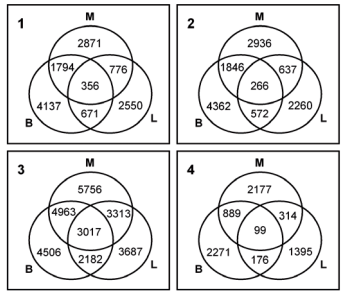

Figure 3. Overlap of differentially expressed gene lists generated using two normalization methods.
Probiotic microorganisms used in Caco-2 treatments were as follows: *B. animalis* Bb12 (B), *L. rhamnosus* GG (L), and mixture of selected six strains (M). The normalization approaches were: median (1-2) and loess (3-4) method performed by Molecular Devices GenePix Pro/Acuity (1, 3) and R Bioconductor (2, 4) software.

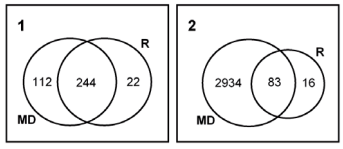

Figure 4. Intersection of differentially expressed gene lists generated using two normalization methods.
Lists of differentially expressed genes common to experiments B, L, and M (B ∩ L ∩ M) were obtained from comparison of lists of differentially expressed genes from microarrays: B, L, and M (see Fig. 3) obtained after median (1) and loess (2) normalization performed by Molecular Devices GenePix Pro/Acuity (MD) and R Bioconductor (R) software.

Even though the numbers of differentially expressed genes selected for each microarray normalized with a median method are similar (Table 1), the number of overlapping genes for arrays B, L and M normalized in limma R Bioconductor (Fig. 3.2) is by 25% smaller than the one obtained after median normalization performed with Molecular Devices GenePix Pro software (Fig. 3.1). However, a comparison of the lists of overlapping differentially expressed genes (indicated in the two median normalized datasets) showed that the set of genes common for the three microarrays (B, L and M) comprised 92% of the genes from limma (R) and 69% of the genes from GenePix Pro (Fig. 4.1). A similar comparison of the results for loess normalization revealed that the common set of differentially expressed genes included 84% of genes obtained from limma (R) and only 3% of genes from Molecular Devices Acuity software (Fig. 4.2). This divergence results from the difference in the overall number of genes assigned as differentially expressed after loess normalization in the two software types.Lists of differentially expressed genes were extracted from the two softwares and the fold-change values were analyzed in pairs: MD median *vs.* R median, MD loess *vs.* R loess. The distribution of fold-change was very coherent for the genes selected as differentially expressed after the median normalization methods (Fig. 5a–d). On the contrary, in the case of the loess normalization methods the fold-change values of shared differentially expressed genes were dissimilar or sometimes even opposite between the two methods (Fig. 5e–h). A small subset of genes showed upregulation after normalization by one software and downregulation after normalization with the other, and vice versa. This tendency was observed mainly for genes with low fold-change values. Table 3 presents median and standard deviation values of differences between the fold-changes obtained for the data obtained after the same normalization method performed in the two programs. The high values of standard deviation indicate how divergent were the results obtained using the loess methods implemented in the two programs analyzed.

DISCUSSION

The design of the experiment and data normalization are crucial steps of microarray-based gene-expression profiling. Two-color hybridization is commonly used and has many advantages, allowing for direct comparison between control (reference) and tested sample (Knapen *et al.*, 2009). Ratiometric data analysis minimizes different sources of variation related to the construction and hybridization of a microarray, thus providing the highest level of precision in the comparison of gene expression profiles (Quackenbush, 2002). Several principal methods of microarray data normalization are in use but which one is the optimal remains an open question (Chua *et al.*, 2006). It depends on the type of the array and the biological background of the assay. In the case of two-color hybridization, the most important step is within-slide (inter-channel) normalization that aims to correct discrepancies resulting from variable dye incorporation, fluorescence intensity and sensitivity to degradation. It can similarly affect all the genes or only genes with similar intensity (Berger *et al.*, 2004).

Here, we used a simple model of four biological replicates (samples treated with probiotic microorganisms) and a common reference (untreated control) to test the selected methods of within-array normalization. Application of a whole-genome microarray (35 thousand spots) and a special biological model entitled us to use global normalization methods. Although the probiotic micro-

Table 3. Statistical comparison of normalized fold-change data.
Mean and standard deviation of differences between fold-changes obtained using the same normalization method implemented in the two programs, calculated for each microarray separately. *B. animalis* Bb12 (B), *L. rhamnosus* GG (L), mixture of six selected probiotic bacterial strains (M) and probiotic yeasts (Y) as compared with control.

	Characteristic	Microarray			
		B	L	M	Y
MDmedian ∩ Rmedian	Mean	0.0137	0.1441	0.0449	0.0091
	Standard deviation	0.0232	0.0451	0.0466	0.0470
MDloess ∩ Rloess	Mean	1.3549	2.2030	1.9116	1.5667
	Standard deviation	1.3539	2.0425	1.5655	1.0644

MDmedian, Molecular Devices GenePix Pro median normalization; Rmedian, R limma median normalization; Rloess, R limma loess normalization; MDloess, Molecular Devices Acuity loess normalization.

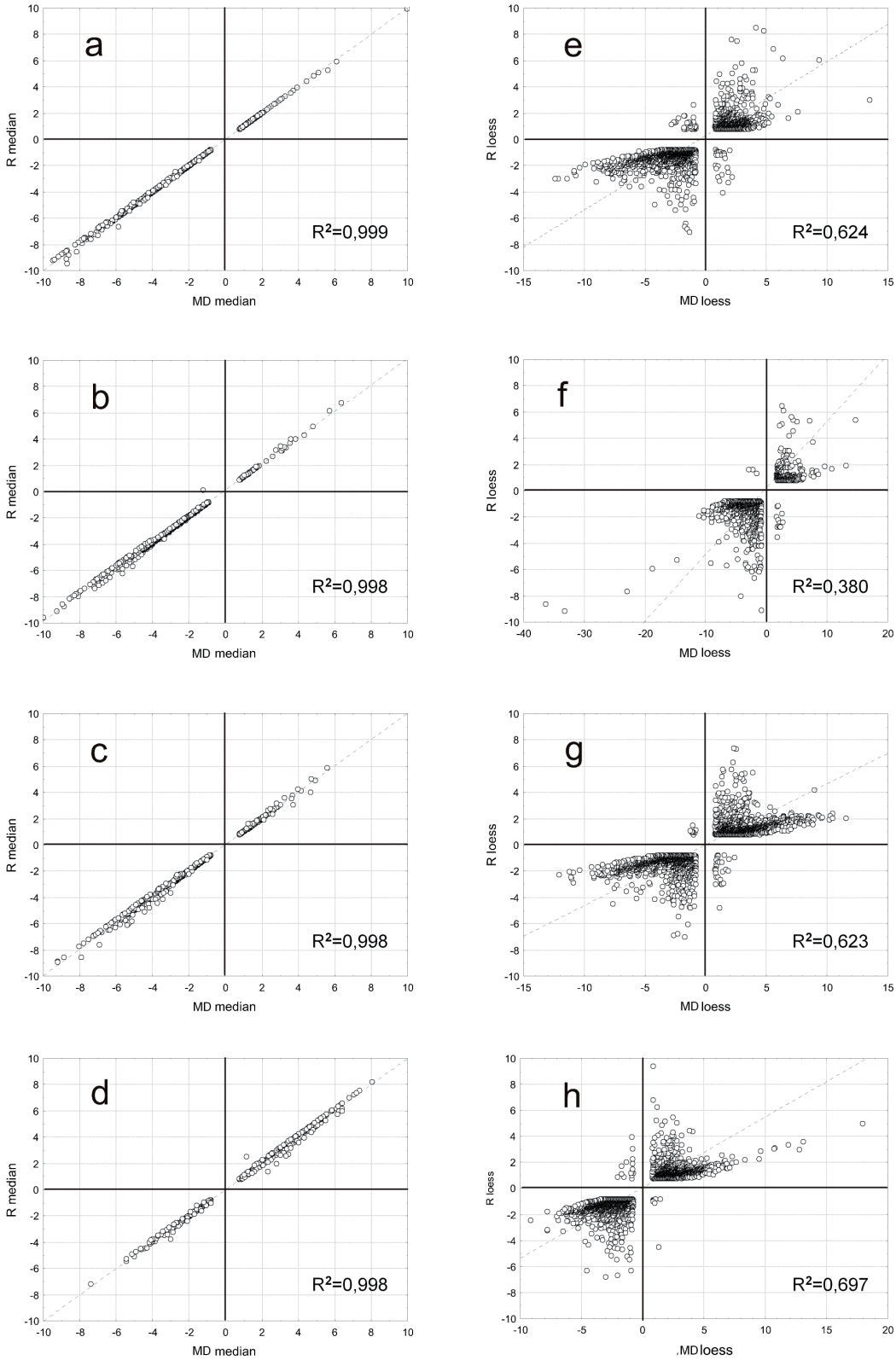

Figure 5. Comparison of log-transformed fold-change distribution of differentially expressed genes for the four analyzed microarrays.
Graphs: **a** (array B), **b** (array L), **c** (array L), and **d** (array Y) illustrate the consistency of median normalization performed by the Molecular Devices GenePix Pro (MD median) and R Bioconductor (R median). The median-normalized data points for each gene calculated by the two software types are positioned at the diagonal. Graphs: **e** (array B), **f** (array L), **g** (array L), and **h** (array Y) illustrate the disparity of results obtained with loess normalization performed by Molecular Devices Acuity (MD loess) and R Bioconductor (R loess). Lower left and upper right quadrants represent genes in statistical agreement between both programs. Upper left and lower right quadrants show genes that were attributed opposite expression change direction by the two programs and the same normalization method. Coefficients of determination (R^2) are given for each dataset on graphs.

organisms are generally beneficial for the host and are recommended as a component of the diet, they presumably do not drastically change the physiology of human epithelial intestinal cells. Similar studies showed that interactions with bacteria (either probiotic, commensal or pathogenic) modified the expression of a relatively small fraction of epithelial cell genes (0.35 to 13%) (Eckmann *et al.*, 2000; Belcher *et al.*, 2000; Rosenberger *et al.*, 2000; Pedron *et al.*, 2003; Fukushima *et al.*, 2003; Panigrahi *et al.*, 2007). Consequently, we expected only a slight effect of the probiotics on the gene expression in Caco-2 cells.

There are many advanced approaches to solve the gene selection problem, such as classical or moderated t-statistics, significance analysis of microarrays (SAM), analysis of variance (ANOVA), between group analysis (BGA) or area under the ROC curve (Jeffery *et al.*, 2006; Hsu *et al.* 2008). It is known that the feature selection process, as well as the data preprocessing, the number of genes, the number of samples and the noise in the dataset, all have a profound impact on the results of microarray experiments. This problem has been discussed in details in several papers (Jeffery *et al.*, 2006; Hsu *et al.*, 2008; Jirapech-Umpai & Aitken, 2005; Jung *et al.*, 2011). Here, we focused on one of the earliest stages of microarray data preprocessing to show how much one element of the microarray experiment puzzle could affect the end result. The main aim of our work was to identify differences in the results coming from application of distinct within-array normalization methods. The small size of our dataset and the lack of technical replicates prompted us to use fold change, the simplest approach to select differentially expressed genes. Shi *et al.* (2005) and Guo *et al.* (2006) indicated that differential analysis based on fold-change results in more reproducible gene lists than the ordinary and modified t-statistics. Our test has revealed a dramatic impact that the within-array normalization methods have on the results of a microarray experiment. Table 1 shows that the 'median' method implemented in both analyzed programs results in the identification of similar numbers of differentially expressed genes. For this method the ratios of up- to down-regulated genes and their positions in the ranking are also similar for each microarray. A comparison of the two variants of this type of normalization indicated not only very similar numbers of differentially expressed genes (Table 1) but also a very high proportion of genes shared between the two median methods, comprising approximately 95% of the differential genes (Table 2, Fig. 2.1–4). Furthermore, the sets of common genes are coherent between the biological experiments. On average, 81% of the genes determined as differentially expressed were common for all three microarrays that examined treatment with probiotic bacteria, regardless of the software used for median normalization (Figs. 3.1–2 and 4.1). However, median normalization is one of the simplest methods and as such is not always recommended as it treats all the genes equally, regardless their fluorescence intensities.

More sophisticated algorithms, such as "loess", capable of removing intensity-dependent bias, produce much more divergent results when implemented in different softwares. The Molecular Devices Acuity software finds up to six times more differential genes than limma from R Bioconductor and their fold-change values are also higher. Following Acuity loess normalization even half of the genes presented on the array were indicated as differentially expressed. This result stands in contradiction with the idea of global normalization, based on the assumption that only a small subset of genes reveal up- or down- regulation. Such tendency is usually true for microarrays covering the whole genome/transcriptome of a studied organism. From the biological point of view, such a huge number of genes responding to treatment with probiotic microorganisms would not be expected either. Furthermore, in the case of the Molecular Devices software the "median" method compared to "loess" resulted in a smaller number of differentially expressed genes whereas in the R Bioconductor limma software the tendency was opposite — the "loess" method was more restrictive than 'median'. Irrespective of the normalization method and software the ratios of the numbers of up- and down-regulated genes are similar (Table 1). However, after loess normalization in the two types of software tested some of the genes reveal opposite modes of regulation (Fig. 5). Taking into account the fact that it concerned mainly genes with a low level of expression, it would be useful to apply an additional step of filtration in order to avoid distorted results. In our opinion, it is more desirable to obtain a shorter list of genes selected as differentially expressed than to get a long list with a high number of false-positive results.

Summarizing, data normalization performed using each of the software types tested (two Molecular Devices programs and R Bioconductor limma package) and methods (median and loess) gave divergent results of the analysis of the same microarrays. Furthermore, the two loess normalization methods produced opposite changes for some of the genes. The high values of the mean standard deviation of fold-change of shared differentially expressed genes (Table 3) indicate significant differences between the algorithms applied for loess normalization in Molecular Devices Acuity and R Bioconductor limma. This information has to be taken into account before the analysis of microarray data. Researchers who start their work with microarray-based gene expression profiling must be aware that each data transformation step can remove technical bias but, on the other hand, it can also introduce major changes influencing later biological interpretation of the results. The choice of the normalization method should be carefully considered based of the demands stremming from both (biological and technical) points of view as well as the aims of the experiment. A more restrictive method seems to be more reliable, as it probably produces less false-positive results. On the otherhand, some biologically relevant information can be lost. This is especially important in experiments where substantial modulation of gene expression is not expected. Nevertheless, once the method is chosen it should be applied consequently for normalization of all the microarrays from the studied dataset.

Acknowledgements

This work was supported by the Ministry of Science and Higher Education grant no. N312 047 32/2668 in years 2007–2010.

REFERENCES

Baker SG (2008) Using microarrays to study the microenvironment in tumor biology: the crucial role of statistics. *Semin Cancer Biol* **18:** 305–310.

Belcher CE, Drenkow J, Kehoe B, Gingeras TR, McNamara N, Lemjabbar H, Basbaum C, Relman DA (2000) The transcriptional responses of respiratory epithelial cells to *Bordetella pertussis* reveal host defensive and pathogen counter-defensive strategies. *Proc Natl Acad Sci USA* **97:** 13847–13852.

Berger JA, Hautaniemi S, Järvinen AK, Edgren H, Mitra SK, Astola J (2004) Optimized LOWESS normalization parameter selection for DNA microarray data. *BMC Bioinformatics* **5:** 194.

Chua S, Vijayakumar, P, Nissom P, Yang H (2006) A novel normalization method for effective removal of systematic variation in microarray data. *Nucleic Acids Res* **34:** e38.

Churchill G A (2002) Fundamentals of experimental design for cDNA microarrays. *Nat Genet* **32:** 490–495.

Cowell JK, Hawthorn L (2007) The application of microarray technology to the analysis of the cancer genome. *Curr Mol Med* **7:** 103–120.

Delgado S, O'Sullivan E, Fitzgerald G, Mayo B (2008) *In vitro* evaluation of the probiotic properties of human intestinal Bifidobacterium species and selection of new probiotic candidates. *J Appl Microbiol* **104:** 1119–1127.

Dudoit S, Yang YH, Callow MJ, Speed TP (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiment. *Statistica Sinica* **12:** 111–139.

Eckmann L, Smith JR, Housley MP, Dwinell MB, Kagnoff MF (2000) Analysis by high density cDNA arrays of altered gene expression in human intestinal epithelial cells in response to infection with the invasive enteric bacteria Salmonella. *J Biol Chem* **275:** 14084–14094.

Fukushima K, Ogawa H, Takahashi K, Naito H, Funayama Y, Kitayama T, Yonezawa H, Sasaki I (2003) Non-pathogenic bacteria modulate colonic epithelial gene expression in germ-free mice. *Scand J Gastroenterol* **38:** 626–634.

Gentleman R, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5:** R80.

Gentleman R, Irizarry RA, Carey VJ, Dudoit S, Hubrer W (2005) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York.

Gopal PK, Prasad J, Smart J, Gill HS (2001) *In vitro* adherence properties of Lactobacillus rhamnosus DR20 and Bifidobacterium lactis DR10 strains and their antagonistic activity against an enterotoxigenic *Escherichia coli*. *Int J Food Microbiol* **67:** 207–216.

Guo L, Lobenhofer EK, Wang C, Shippy R, Harris SC, Zhang L, Mei N, Chen T, Herman D, Goodsaid FM, Hurban P, Phillips KL, Xu J, Deng XT, Sun YMA, Tong WD, Dragan YP, Shi LM (2006) Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nature Biotechnology* **24:** 1162–1169.

Hahne F, Huber W, Gentleman R, Falcon S (2008) *Bioconductor Case Studies*. Springer, New York.

Heczko PB, Strus M, Kochan P (2006) Critical evaluation of probiotic activity of lactic acid bacteria and their effects. *J Physiol Pharmacol* **57** (Suppl 9)**:** 5–12.

Howbrook DN, van der Valk AM, O'Shaugnessy MC, Sarker DK, Baker SC, Lloyd AW (2003) Developments in microarray technologies. *Drug Discov Today* **8:** 642–651.

Hsu HH, Lu MD (2008) Feature Selection for Cancer Classification on Microarray Expression Data. *Eighth International Conference on Intelligent Systems Design and Applications,* **IEEE 2008**: 153–158.

Jeffery IB, Higgins DG, Culhane AC (2006) Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics* **7:** 359.

Jirapech-Umpai T, Aitken S (2005) Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. *BMC Bioinformatics* **6:** 148–158.

Jung K, Becker B, Brunner E, Beissbarth T (2011) Comparison of global tests for functional gene sets in two-group designs and selection of potentially effect–causing genes. *Bioinformatics* **27:** 1377–1383.

Knapen D, Vergauwen L, Laukens K, Blust R (2009) Best practices for hybridization design in two-colour microarray analysis. *Trends Biotechnol* **27:** 406–414.

Ness SA (2007) Microarray analysis: basic strategies for successful experiments. *Mol Biotechnol* **36:** 205–219.

Panigrahi P, Braileanu GT, Chen H, Stine OC (2007) Probiotic bacteria change *Escherichia coli*-induced gene expression in cultured colonocytes: Implications in intestinal pathophysiology. *World J Gastroenterol* **13:** 6370–6378.

Pedron T, Thibault C, Sansonetti PJ (2003) The invasive phenotype of *Shigella flexneri* directs a distinct gene expression pattern in the human intestinal epithelial cell line Caco-2. *J Biol Chem* **278:** 33878–33886.

Quackenbush J (2002) Microarray data normalization and transformation. *Nat Genet* **32** (Suppl)**:** 496–501.

R Development Core Team (2009) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. URL http://www.r-project.org.

Rosenberger CM, Scott MG, Gold MR, Hancock RE, Finlay BB (2000) *Salmonella typhimurium* infection and lipopolysaccharide stimulation induce similar changes in macrophage gene expression. *J Immunol* **164:** 5894–5904.

Sambuy Y, De Angelis I, Ranaldi G, Scarino ML, Stammati A, Zucco F (2005) The Caco-2 cell line as a model of the intestinal barrier: influence of cell and culture–related factors on Caco–2 cell functional characteristics. *Cell Biol Toxicol* **21:** 1–26.

Shi LM, Tong WD, Fang H, Scherf U, Han J, Puri RK, Frueh FW, Goodsaid FM, Guo L, Su ZQ, Han T, Fuscoe JC, Xu ZA, Patterson TA, Hong HX, Xie Q, Perkins RG, Chen JJ, Casciano DA (2005) Cross-platform comparability of microarray technology: Intraplatform consistency and appropriate data analysis procedures are essential. *BMC Bioinformatics* **6:** S12.

Simon RM, Korn EL, McShane LM, Radmacher MD, Wright GW, Zhao Y (2007) *Design and Analysis of DNA Microarray Investigations*. Springer, New York.

Smyth GK (2005) Limma: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Gentleman R, Carey V, Huber W, Irizarry R, Dudoit S, eds. pp 397–420. Springer, New York.

Smyth GK, Speed TP (2003) Normalization of cDNA microarray data. *Methods* **31:** 265–273.

Trevino V, Falciani F, Barrera–Saldana HA (2007) DNA microarrays: a powerful genomic tool for biomedical and clinical research. *Mol Med* **13:** 527–541.

Venkatasubbarao S (2004) Microarrays — status and prospects. *Trends Biotechnol* **22:** 630–637.

Yang YH, Dudoit S, Luu P, Speed TP (2001) Normalization for cDNA microarray data. In *Microarrays: Optical Technologies and Informatics*. Bittner ML, Chen Y, Dorsel AN, Dougherty ER eds. *Proceedings of SPIE* **4266:** 141–152.

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucl Acids Res* **30:** e15.