Automatic detection of outlying microarrays using multi-array quality metrics

Michal Marczyk, Lukasz Krol, and Joanna Polanska

Data Mining Group, Institute of Automatic Control, Silesian University of Technology,
Akademicka 16, 44-100 Gliwice, Poland
{Michal.Marczyk, Lukasz.Krol, Joanna.Polanska}@polsl.pl

Abstract. Gene-chip microarrays are used to measure changes on transcript level induced by various experimental conditions. Due to many sources of potential measurement errors quality control of data is critical in these experiments. We propose an automatic method for finding poor quality microarrays which is based on two known multi-array quality metrics and linear regression model. Removing of low quality samples has a big influence on increasing the amount of association between measurements and it increases the number of obtained differentially expressed genes decreasing false discovery rate. It also leads to obtaining more significant ontology terms found by functional analysis of differentially expressed genes and helps discovering genes which are biologically consistent with analyzed problem.

Keywords: Gene-chip, microarrays, NUSE, QC, RLE.

1 Introduction

Microarray technology allows measuring multiple genes in a single experiment using hundreds of thousands 25-mer oligonucleotide probes grouped into gene specific sets. Microarrays are one of the most widely used measurement techniques for the characterization of transcript level changes induced by various factors. A typical approaches of microarray experiment are searching for genes which behave differently in the various populations of cells (called differentially expressed genes, DEGs) and classification of different data groups. These studies also support discovering of interactions between genes by its functional analysis.

Major advantage of oligonucleotide microarrays is its high-throughput nature; obtaining multiple measurements of gene expressions at a relatively short time. But highly automated acquisition of a large number of data increases the risk of receiving poor quality signals. Microarray experiment involve multiple steps that can lead to introducing unwanted random or systematic variation in the data by technical, instrumental or computational factors [1]. Quality control (QC) step is necessary before data analysis in order to disregard changes which originate only from technique specificity and meas-

urement inaccuracy. Complicated and multi-step process of obtaining final gene expressions requires creating sophisticated statistical methods for detecting outlying microarrays.

Quality control of oligonucleotide microarray data is a widely discussed problem in the literature and different solutions were proposed [2-6]. Despite that there remains a lack of consensus in defining poor quality measurements. In [2] there are introduced two measures of the precision of a probe set expression estimate. Authors of [3] propose an outlier detection method based on principal component analysis (PCA) and robust estimation of Mahalanobis distances. In [4] authors used the Pearson correlation coefficient and the percentage of outlier data points on an array to find suspected fail arrays. There exist a few R packages designed especially for quality control of oligonucleotide microarrays [7-8]. ArrayQualityMetrics can assess reproducibility, identify apparent outlier arrays and compute measures of signal-to-noise ratio [7]. Simpleaffy provides access to a variety of QC metrics for assessing the quality of RNA samples and of the intermediate stages of sample preparation and hybridization [8]. But most of the introduced techniques require an expert knowledge from the user to set a threshold for removing microarrays or choose proper criterion used in QC of data.

In this study we propose a fully automatic tool for finding poor quality microarrays based on the linear regression model (LR) and one of the best multi-array quality metrics proposed in [2]: the normalized unscaled standard error (NUSE), which provides a measure of the precision of expression of a given gene on a given array relative to other arrays in the experiment and the relative log expression (RLE), which is calculated by subtracting the median gene expression estimate across arrays from each gene expression estimate. We prove that using our tools we can properly exclude outlying signals from the data to guarantee the reliability of discovered patterns and reproducibility of microarray experiments.

2 Methods

2.1 Data

In this work we used Affymetrix GeneChip microarray data from three studies with different number of samples measured on Human Gene 1.0 ST platform. Genepi dataset consists 60 breast cancer samples of lymphocytes from radiosensitivity study, which was exposed to low dose radiation (200 mGy). Myeloma dataset (E-MTAB-1038, Human Gene 1.0 ST array) with 73 samples was used for determining the molecular signature characteristic for centrosome abnormalities in patients with multiple myeloma. Atheroma dataset (E-MTAB-1470, Human Gene 1.0 ST array) consists 64 samples from study of human carotid atheroma; a plaque of degenerated thickened arterial intima, occurring in atherosclerosis.

2.2 Data pre-processing

Probes are annotated using CDF from Brainarray repository ver.18 [9] based on the latest genome and transcriptome information from EntrezG database. Using updated

annotations we get 19657 probe sets for Human Gene 1.0 ST platform. Datasets are normalized separately by using the robust multichip average algorithm (RMA) that includes background correction, quantile normalization and summarization by the median polish approach [10] which gives expression data in log2 scale.

2.3 QC algorithms description

To check the quality of microarrays we used multi-array measures which are able to determine if an array's quality is better or worse than the typical array being analyzed in the same experiment or batch. We choose two quality metrics: the normalized unscaled standard error (NUSE) and the relative log expression (RLE). Both metrics are calculated for each gene on a given microarray separately. The median value of NUSE characterizes the average quality of probe sets within the microarray. To find if there exist a shift in the median value of NUSE for a given array from other samples in the same experiment we introduced a robust version of *z* statistic incorporating the median value and the median absolute deviation instead of parametric statistics. The variance of RLE is a measure of the variability associated with measurement errors. We introduced chi^2 statistic to calculate differences in variance estimate of RLE for a given array.

In order to distinguish outlier microarrays in the dataset we fit the linear regression model to values of obtained statistics (*z* and chi^2 separately) sorted in ascending order (Fig. 1). Proposed algorithm is as follows:

- We start fitting the LR model with the half of the values of statistic sorted in ascending order.
- By increasing the number of samples used for modeling by one and calculating R^2 each time we find the model with the best fit.
- If the obtained number of samples used for modeling is less than N-3, where N is the number of samples in the experiment, we introduce second LR model.
- We start the fitting procedure of the second model with use of first three highest statistic values.
- We increase the number of points used for modeling (we take the samples from the highest to the median value of the statistic) by one and calculate R^2 for each LR model.
- Threshold for finding outlying measurements is found by maximizing sum of R^2 for both models.

2.4 Statistical and functional analysis

To quantify influence of quality control step on data integrity we calculated the Spearman correlation coefficient between samples within the same experimental group. We summarized obtained values by calculating mean value and performing one-sided T-test for significance of average coefficient. For discovery of DEGs we use the modified version of Welch test designed for microarray data [11]. Additionally we calculated false discovery rates (FDR) of estimated DEGs. The significance level was set to 0.05.

Functional analysis was done by using the gene ontologies (GO). The Gene Ontology project is a collaborative effort to address the need for consistent descriptions of gene products in different databases. It provides a structured controlled vocabulary of gene and protein biological roles by comprising three hierarchies that define functional attributes of gene products: Molecular Function (MF), Biological Process (BP), and Cellular Component (CC) [12]. 13927 GO terms was found for all genes from Human Gene 1.0 ST platform and they were used as a universe. To identify significantly overrepresented GO terms Fisher's exact test was performed for all terms separately at 0.05 significance level.

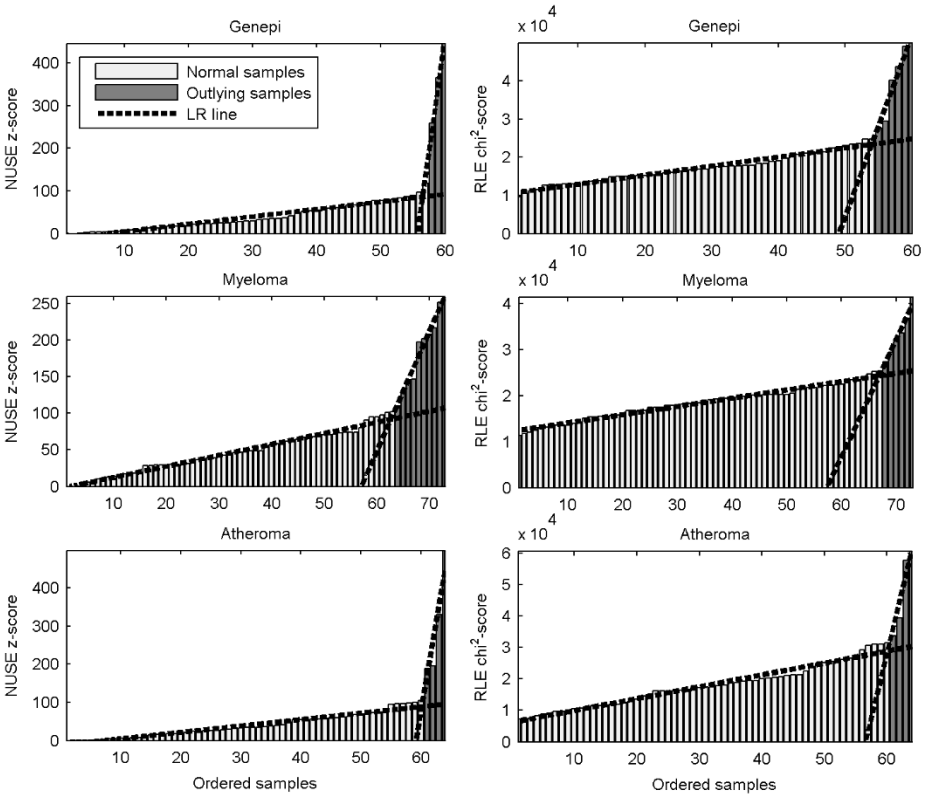

Fig. 1. Bar plots of ordered z-score of NUSE measure and chi^2 statistics of RLE measure with fitted linear regression lines. In the succeeding rows there are models for different datasets: Genepi, Myeloma and Atheroma

3 Results

In many cases applying of LR modeling to *z* statistic of NUSE and *chi*^{2} statistic of RLE leads to finding different number of outlying microarrays (Table 1). Considering different potentials of two QC measures it happens that a sample, which was marked as

outlier with use of NUSE *z* statistic, was not selected by an algorithm based on RLE *chi*^{2} statistic. In that case we propose four scenarios for removing samples which was marked as outliers by: (I) NUSE - only NUSE *z* statistic, (II) RLE - only RLE *chi*^{2} statistic, (III) NUSE+RLE - NUSE *z* statistic or RLE *chi*^{2} statistic, (IV) NUSE&RLE - NUSE *z* statistic and RLE *chi*^{2} statistic. Using QC based on single measures leads to removing from 3% to 10% of the data, while after NUSE+RLE method we remove 11-19% of samples and after NUSE&RLE method only one or two microarrays.

Table 1. Number and percentage of microarrays marked as outliers using different QC methods. After name of the dataset (in the bracket) there is a number of all microarrays in a set.

Dataset	NUSE	RLE	NUSE +RLE	NUSE &RLE
Genepi (60)	3 (5%)	6 (10%)	7 (12%)	2 (3%)
Myeloma (73)	10 (14%)	6 (8%)	14 (19%)	2 (3%)
Atheroma (64)	4 (6%)	4 (6%)	7 (11%)	1 (2%)

Intra-group correlation between microarray expression signals was measured by averaging correlation coefficients from pairwise comparisons of samples (Table 2). We additionally checked influence of introducing QC methods on correlation significance measured by T test statistic, since the number of samples in the set is decreased. For all datasets removing of outlying microarrays marked by all four methods increases average correlation coefficient. There is also an increase of T statistic, which denotes increase of significance of the correlation coefficient.

Table 2. Average Pearson correlation coefficient (Av. r) between microarrays within the same experimental group and value of T statistic from a test of correlation coefficient significance.

Dataset	Group	Meas.	None	NUSE	RLE	NUSE +RLE	NUSE &RLE
Genepi	RR	Av. R	0.987	0.988	0.988	0.988	0.988
		T stat	245.6	327.6	298.5	297.8	325.7
	RS	Av. R	0.986	0.988	0.986	0.986	0.988
		T stat	249.1	331.5	302.7	301.9	329.6
Myeloma	MM	Av. R	0.941	0.942	0.941	0.942	0.941
		T stat	77.5	80.0	77.5	80.0	77.5
	N	Av. R	0.931	0.938	0.933	0.934	0.937
		T stat	59.6	99.7	68.0	66.5	95.5
Atheroma	Intact	Av. R	0.974	0.975	0.974	0.974	0.976
		T stat	80.0	96.5	80.0	77.4	99.3
	Ather.	Av. R	0.976	0.978	0.977	0.977	0.977
		T stat	78.2	98.3	95.1	95.2	98.3

In Genepi dataset we found about 4%, in Myeloma dataset about 15% and in Atheroma dataset about 40% of DEGs from all of the genes on microarray (Table 3). For first two datasets introducing all scenarios for removing outlying microarrays resulted in an increased number of DEGs, which leads to decreased FDR. The biggest increase was observed for NUSE+RLE method. Opposite results were obtained for Artheroma dataset, where introducing QC step decreased the number of DEGs.

Table 3. Number of differentially expressed genes (DEGs, $p<0.05$) and false discovery rate (FDR) before and after QC using different methods

Dataset	Quan.	None	NUSE	RLE	NUSE +RLE	NUSE &RLE
Genepi	DEGs	686	823	685	821	685
	FDR [%]	100	100	100	100	100
Myeloma	DEGs	2620	4817	3280	4702	3500
	FDR [%]	37,51	20,40	29,96	20,90	28,08
Atheroma	DEGs	7848	7718	6950	6888	7741
	FDR [%]	12,52	12,73	14,14	14,27	12,70

For each GO term describing different processes in human organism we found number of related genes which was marked as DEG and checked its significance (Table 4). For Genepi dataset after QC we obtain about 30 more significant GO terms, while for Myeloma dataset this number ranges from 40 GO terms less to 60 more depends on QC method used. For Atheroma dataset we also observe a divergent result of removing outlying signals; for two methods there is an increase in the number of significant GO terms and for two methods decrease. NUSE+RLE method in all dataset gives increase of the number of significant GO terms.

Table 4. Number of significant GO terms ($p<0.05$) before and after QC using different methods. There are 13927 GO terms for all genes on HuGene platform

Dataset	None	NUSE	RLE	NUSE +RLE	NUSE &RLE
Genepi	319	340	348	354	304
Myeloma	260	318	224	286	295
Atheroma	1128	1092	1179	1149	1096

We also calculated number of significant GO terms found only in the case with no data cleaning or after QC and the one that remain significant (Table 5). In Genepi dataset we observed from 82 to 233 new significant GO terms after removing outliers. Similar results was obtained for Myeloma dataset. In Atheroma dataset most of the significant GO terms are common for the cases before and after QC.

Table 5. Number of significant GO terms after QC using different methods in comparison to the case without removing outlying microarrays (none)

Dataset	No. of sig. GO	NUSE	RLE	NUSE +RLE	NUSE &RLE
Genepi	only in none	125	179	198	97
	Common	194	140	121	222
	only after QC	146	208	233	82
Myeloma	only in none	169	143	181	130
	Common	91	117	79	130
	only after QC	227	107	207	165
Atheroma	only in none	169	248	278	128
	Common	959	880	850	1000
	only after QC	133	299	299	96

4 Discussion

For all datasets removing of outlying microarrays increased the amount of association between the samples obtained in the same experimental conditions and increased significance of founded DEGs, which proves validity of proposed algorithm. Providing better similarity of data within the same group may lead to better classification results and finding more significant panels of genes describing analyzed problem.

In Atheroma dataset we found from 1 outlying microarray using NUSE&RLE method to 7 using NUSE+RLE. QC increased the average correlation between samples in this set , but the number of estimated DEGs was smaller by 107 to 960 genes comparing to the case with no cleaning. It results from the fact that when we remove a microarray from the dataset we decrease a power of the test for finding DEGs, because of smaller number of measurements used for testing.

Multi-array QC metrics used in the proposed algorithm can be calculated despite of microarray platform used which makes the method a very universal solution. This creates an opportunity to integrate data from different experiments related to the same disease or the same biological process to achieve best possible quality of prediction and perform QC on connected researches.

Removing of outlying microarrays in most cases leads to increasing number of DEGs. But discovering of DEGs in oligonucleotide microarray data brings a risk of obtaining false positive results. Concerning that we believe that 40% of DEGs found in microarray experiment (like in Atheroma dataset) is too many; it contain many false positive genes and removing of outlying samples may reduce this number. In Genepi and Myeloma datasets estimated number of DEGs was similar to the one that we expect in microarray analysis. To prove this statement we performed functional analysis of obtained DEGs.

Due to the QC of Genepi dataset using proposed methods we can find significant GO terms characteristic to iron metabolism (“2 iron 2 sulfur cluster binding”, “iron-

sulfur cluster assembly", "iron-sulfur cluster binding"). These are rare GO terms connected with less than 20 genes on the microarray. It proves that there exist a relationship between patient radiosensitivity and iron metabolism, which was also found in the literature [13-14]. Additionally we found significant GO term associated with repair of DNA damages induced by UV radiation ("pyrymidine dimer repair"). Removing of outlying microarrays indicates marking 2 from 4 genes on the microarray connected with this term as DEGs, what was not possible for raw dataset.

Functional analysis of Myeloma dataset after QC indicates significant GO terms which could be related to enhanced ability of cancer cells to growth ("positive regulation of transcription from RNA polymerase II promoter", "cell-cell signaling", "glucose homeostasis", "growth factor activity"). Also we found more genes related to GO terms connected with intensive transcription, like "Sequence-specific DNA binding transcription factor activity", "transcription factor complex", "RNA polymerase II distal enhancer sequence-specific DNA binding transcription factor activity" increasing their significance.

Cleaning of Atheroma dataset brings smaller number of DEGs, so it is important to remain biologically relevant genes marked as significant. After QC we found significant GO terms connected with the fact of an accumulation and swelling in artery walls lipids (cholesterol and fatty acids) and calcium due to atheroma ("calcium ion transmembrane transport", "cytosolic calcium ion homeostasis", "cholesterol transport"). Also we found significant GO terms connected with inflammatory processes ("cytokine binding", "negative regulation of interleukin-6 production", "negative regulation of interleukin-2 production"). Finding smaller number of DEGs after QC leads to increasing significance of more general, but important to atheroma problem GO terms ("blood coagulation", "regulation of immune response").

5 Conclusion

Obtained results show that QC of measured data is a necessary step in microarray experiments. Incorporating information from NUSE and RLE metrics to find outlying microarrays has positive effect on data integrity, increases number of estimated DEGs and decreases FDR. GO terms analysis proved that using our algorithm leads to discovering more biologically meaningful genes. From four proposed methods of QC we distinguish NUSE+RLE, which gave the best results concerning above mentioned criterions. Suggested method is fully automatic and we proved that it is a suitable tool for quality control of oligonucleotide microarray data.

Acknowledgments. The authors kindly thank prof. Wieslawa Widlak for biological interpretation of GO analysis and helpful discussions. This work was financially supported by Silesian University of Technology internal grant BK/214/RAU-1/2013 t.10 (MM) and National Science Centre grant no. DEC-2013/08/M/ST6/00924, HARMONIA 4 (LK, JP).

References

1. Jaksik, R., Marczyk, M., Polanska, J., Rzeszowska-Wolny, J.: Sources of High Variance between Probe Signals in Affymetrix Short Oligonucleotide Microarrays. Sensors (Basel, Switzerland) 14, 532-548 (2013)
2. Bolstad, B.M., Collin, F., Simpson, K.M., Irizarry, R.A., Speed, T.P.: Experimental design and low-level analysis of microarray data. Int Rev Neurobiol 60, 25-58 (2004)
3. Shieh, A.D., Hung, Y.S.: Detecting outlier samples in microarray data. Statistical applications in genetics and molecular biology 8, Article 13 (2009)
4. Yang, S., Guo, X., Yang, Y.C., Papcunik, D., Heckman, C., Hooke, J., Shriver, C.D., Liebman, M.N., Hu, H.: Detecting outlier microarray arrays by correlation and percentage of outliers spots. Cancer informatics 2, 351-360 (2006)
5. Kauffmann, A., Huber, W.: Microarray data quality control improves the detection of differentially expressed genes. Genomics 95, 138-142 (2010)
6. Brettschneider, J., Collin, F., Bolstad, B.M., Speed, T.P.: Quality Assessment for Short Oligonucleotide Microarray Data. Technometrics 50, 241-264 (2008)
7. Kauffmann, A., Gentleman, R., Huber, W.: arrayQualityMetrics--a bioconductor package for quality assessment of microarray data. Bioinformatics 25, 415-416 (2009)
8. Wilson, C.L., Miller, C.J.: Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis. Bioinformatics 21, 3683-3685 (2005)
9. Dai, M., Wang, P., Boyd, A.D., Kostov, G., Athey, B., Jones, E.G., Bunney, W.E., Myers, R.M., Speed, T.P., Akil, H., Watson, S.J., Meng, F.: Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Research 33, e175 (2005)
10. Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., Speed, T.P.: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4, 249-264 (2003)
11. Demissie, M., Mascialino, B., Calza, S., Pawitan, Y.: Unequal group variances in microarray data analyses. Bioinformatics 24, 1168-1174 (2008)
12. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics 25, 25-29 (2000)
13. Haro, K.J., Sheth, A., Scheinberg, D.A.: Dysregulation of IRP1-mediated iron metabolism causes gamma ray-specific radioresistance in leukemia cells. PloS one 7, e48841 (2012)
14. Haro, K.J., Scott, A.C., Scheinberg, D.A.: Mechanisms of resistance to high and low linear energy transfer radiation in myeloid leukemia cells. Blood 120, 2087-2097 (2012)