Comparative Analysis of the Annotation Systems of Mus Musculus 3' High Density Expression Microarray

Anna Cichońska[1], Roman Jaksik[1], Wiesława Widłak[2], Joanna Polańska[1],

[1] Institute of Automatic Control, Silesian University of Technology,
44-100 Gliwice, Poland
{Anna.Cichonska, Roman.Jaksik, Joanna.Polanska}@polsl.pl
[2] Institute of Oncology, 44-100 Gliwice, Poland
wwidlak@io.gliwice.pl

Abstract. Annotations created by microarrays producer Affymetrix are based on the genomic and transcriptomic knowledge that was available at a time of construction of a particular chip. However scientific databases are regularly updated so there is a need of performing re-annotation processes. The goal of this work was to assess the relevance of the probes located on the given Affymetrix 3' expression or promoter microarray and to perform a re-annotation procedure. The use of updated definition files proved that they can improve the results of the microarray data analysis comparing to the standard Affymetrix approach.

1 Introduction

A microarray has become an invaluable research tool as it allows for an effective observation of the presence and the amount of many particular nucleic acids molecules or proteins within the analyzed biological sample in a single experiment. Among all the types of chips the popular and frequently used ones include the expression and the promoter microarrays. Annotation in this context is a term for the probe sets' definitions as it describes the probes' layout on a given chip, linking probes to the corresponding genes, transcripts or promoters. Annotation system for each expression microarrays is included in a Chip Definition File (CDF), while for the promoter chip it is stored in a Binary Probe MAP file (BPMAP). CDF file consists of sections, each describing the layout of the probes belonging to one probe set characterized by a unique name. BPMAP file is not divided up into sections, each row contains information about one probe's layout, nucleotide sequence and localization in the genome.

Annotation systems created by microarrays producer Affymetrix are based on the genomic and transcriptomic knowledge that was available at a time of a particular chip construction. In case of expression arrays producer provided updates rely only on the reassignments of the probe sets to other genes but there are no changes in the sets' composition. The studies performed until now suggests that the probes located on many of the popular microarrays can in fact hybridize with the other fragments of genomes than with the sequences that they were designed for. Due to the above facts the need of performing re-annotation processes, that involve verifying on the basis of which probes it is possible to accurately determine the presence of certain nucleic

acids molecules within a biological sample of an interest and then creating custom CDF or BPMAP files based on the current genomic and transcriptomic knowledge, seems to be reasonable. [1-4]

2 Aim

The aim of this study was to assess the quality and relevance of all the probes located on the Affymetrix promoter *GeneChip Mouse Promoter 1.0R* and 3' high density expression chip *Mouse430_2* and to carry out a re-annotation process based on the preferred build of the genome. The second goal of the work relied on performing a comparative analysis of three different annotation systems of *Mouse430_*2 microarray: the original one provided by Affymetrix, the one created in this work and the one downloaded from a popular repository of custom CDF files – Brainarray [2,5].

3 Material and Methods

Microarrays examined in this work were the 3' expression chip *Mouse430_2* and the promoter array *Mouse Promoter 1.0R*, both designed by Affymetrix company. *Mouse430_2* contains 495 374 probes organized in 45 101 sets, while the analyzed promoter chip includes 4 104 464 probes. On both microarrays there are located also control probes that have not been taken into account in this study.

3.1 Data

Complete mouse genome sequence that was published in December 2011 (*mm10* build) and information about mouse transcripts coming from RefSeq database updated on 23rd July 2012 were downloaded from the UCSC (University of California, Santa Cruz) server. The original CDF file designed for *Mouse430_2* microarray, sequences of all the probes located on that expression chip, stored in FASTA file, and the original BPMAP file for *Mouse Promoter 1.0R* were obtained from the Affymetrix website.

The microarray data used in this survey are coming from the experiment carried out in the Maria Sklodowska-Curie Memorial Cancer Center and Institute of Oncology in Gliwice using *Mouse430_2* chips. A global gene expression profiling analysis was performed in order to determine the response of mouse spermatocytes and hepatocytes to a heat shock. Spermatocytes were treated with the high temperature of 37°C (HS37) and 42°C (HS42), hepatocytes were subjected to 42°C. In both cases control samples (CG) were also used. Data coming from the experiment with mouse 3T3 cells using *Mouse Promoter 1.0R* chip and an anti-trimethyl histone H3 antibody were downloaded from the Affymetrix website. Each experiment was conducted in three technical replicates.

3.2 Software

In order to assess the quality and relevance of the probes located on chosen expression or promoter microarray and then to construct a custom CDF or BPMAP file, a software in Python under the Linux platform was created.

Probes Quality Assessment. Towards assessing molecular probes quality a set of criteria was proposed, separately for 3' expression and promoter microarrays. The criteria are shown in the Tables 1-2 together with the results of the probes quality assessment. The quality assessment of the probes is conducted based on the results of alignment performed using blastn version of BLAST tool. It searches a chosen nucleotide database using a nucleotide sequence query and finds areas of the local similarity between two sequences. Program call parameters were selected towards searching the genome database for the short probes' nucleotide sequences with the maximum sensitivity: *reward* = 2, *penalty* = -4, *gapopen* = 6, *gapextend* = 10. An option allowing for masking the low complexity regions in the query was turned off. The alignment is done separately for a sense and an antisense DNA strand. Obtained matches are being linked to the information about transcripts and exons stored in the RefSeq database. In case of promoter microarrays promoter's regions, according to the information providing by Affymetrix, cover 6 kb upstream through 2.5 kb downstream of 5' transcription start sites. In case of both examined types of microarrays, files storing a list of all the probes, except control ones, placed on a given chip, detailed information about their alignments and the identifiers of the quality assessment group to which each probe is classified are created based on the chosen by the user build of the genome and preferred version of refGene.txt file coming from the RefSeq database.

Custom CDF and BPMAP files Construction. A construction of a custom CDF file is based on the probes that belong to the quality assessment groups preferred by the user except *4-7*. The user can choose if the probes included in the sets should be specific for genes or transcripts. A single molecule is always represented by one probe set. In order to build a probe set, a gene or a transcript must be characterized by at least three probes. In case of the gene-specific custom CDF file, each probe is included into only one set, in opposite to the transcript-specific file where a particular probe may be incorporated into any number of sets. Custom BPMAP file is designed based on the probes that are classified to the quality assessment group labeled *1P*.

3.3 Microarray Data Analysis

The custom annotation file for *Mouse430_2* 3' expression chip, proposed in this work, contains information about the probe sets representing genes and it is based on the probes that were found as the most reliable ones – assigned to the class labeled *1a*. The comparative analysis of the three microarray annotation systems of *Mouse430_2* chip was done by performing the microarray gene expression data analysis using different CDF files: the original file provided by Affymetrix, the one

proposed in this work and the one downloaded from the Brainarray repository. The dataset has been preprocessed using RMA algorithm.

Mouse Promoter 1.0R microarray data were normalized using rMAT open source R package implementing MAT algorithm that is based on the probe sequence information [6,7]. Normalization was done using the original BPMAP file and using thr BPMAP file created in this work.

4 Results

4.1 Results of Probes Quality Assessment

The results of the quality assessment of the probes located on the *Mouse430_2* and *Mouse Promoter 1.0R* microarrays are summarized in the Tables 1-2.

Table 1. Criteria for quality assessment of the probes located on 3' expression chips and number of probes (*Mouse430_2*) in each quality assessment group

Alignment to the region of interest	Matches of the probe to any other regions in the genome		Region of interest		
			3'-most exon of only one gene	the 2nd from the 3' end exon of only one gene	exons of only one gene, other than 1st and 2nd from the 3' end
			(*a*)	(*b*)	(*c*)
None mismatch	No match	(*1*)	203 876	18 071	33 880
	At least one match	(*01*)	24 229	2 032	3 553
One mismatch	No match	(*2*)	418	164	356
	At least one match	(*02*)	77	22	55
More than one mismatch	No match	(*3*)	824	368	688
	At least one match	(*03*)	152	39	101
Other groups					
(*4*)	probe is not aligned to any exon but has at least one match to the other region in the genome		168 549		
(*5*)	probe has matches to the exons of more than one gene with no more than one mismatch, additionally probe can be aligned to other regions in the genome than exons		4 986		
(*6*)	probe has matches to the exons of more than one gene, at least to one of the exons with more than one MM, in addi-		404		

	tion probe can be aligned to other regions in the genome than exons	
(*7*)	probe is not aligned to the genome	32 530

Table 2. Criteria for quality assessment of the probes located on promoter chips and number of probes (*Mouse Promoter 1.0R*) in each quality assessment group

Criteria		Code	No of probes
probe is aligned to only one sequence in the genome	the match is within a promoter's region and no mismatch in the alignment	(*1P*)	1 432 557
	the match is not within a promoter's region or the match is within a promoter's region but the alignment is with at least one mismatch in the alignment	(*1N*)	2 365 680
probe has more than one match to the genome		(*0*)	304 066
probe is not aligned to the genome		(*00*)	2 161

The results showed that the most reliable probes, classified to the group labeled *1a*, that were used to create a custom annotation file for *Mouse430_2* chip constitute 41.16% of all the probes located on that microarray, while in case of *GeneChip Mouse Promoter 1.0R* array this percentage is equal to 34.90% (group 1P).

Proposed custom CDF file allows to calculate expression levels of 50.76% of the protein-coding mouse genes.

4.2 Results of *Mouse Promoter 1.0R* Microarray Data Analysis

The distribution of the differences between probes' signals coming from the microarrays where the antibody was used during the experiment (AB1) and where it was not used (AB0) is presented on the Figure 1. The modified BPMAP file (named newBPMAP) allows for identification of higher number of activated promoter regions and gives for them by average higher signal value. The signal distribution is right shifted while comparing newBPMAP versus AffymetrixBPMAP, medians were equal to 0.0997 (95% CI: 0.0982÷ 0.1012) and 0.0753 (95% CI: 0.0745÷0.0761) respectively.

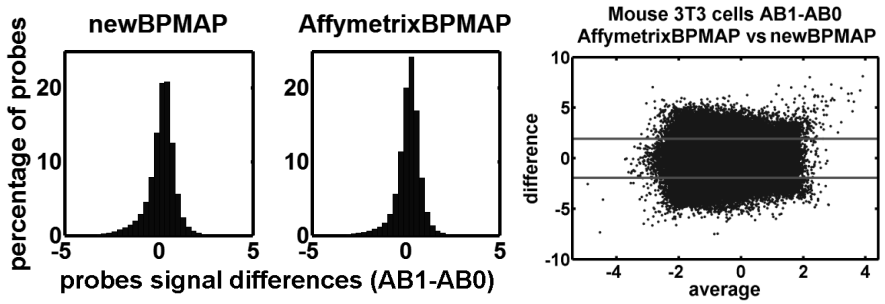

Fig.1. The distribution of signal difference (AB1-AB0) depending on applied BPMAP file and the appropriate Bland-Altman plot

4.3 Results of Comparative Analysis of *Mouse430_2* Annotation Systems

Table 3 summarizes the information about analyzed in this work CDF files.

Table 3. Comparison of three different CDF files for *Mouse430_2* chip

	newCDF	**BrainarrayCDF**	**AffymetrixCDF**
number of probe sets	15 107	17 306	45 037
total number of probes	202 051	266 041	495 374

Only in the Affymetrix annotation file a single gene can be represented by more than one probe sets.

T-test was performed in order to demonstrate the impact of CDF file used during microarray data analysis on the proportion of genes expressed differentially between control and treated with the high temperature samples. The significance level was set to 0.05. According to the z-test for two proportions in case of spermatocytes' samples using our or Brainarray custom CDF file leads to increase in the percentage of probe sets with p-values lower than 0.05 and significantly decreases FDR (False Discovery Rate) values comparing to the standard Affymetrix approach (Table 4).

Table 4. Results of discriminative gene discovery: in each cell first there is a percentage of discriminative probe sets over all sets and next - an average FDR value.

	Spermatocytes CG vs HS37		**Spermatocytes CG vs HS42**		**Hepatocytes CG vs HS42**	
AffymetrixCDF	17.65%	0.2863	9.89%	0.3910	5.34%	0.9404
newCDF	24.38%	0.1859	11.76%	0.3428	5.07%	0.9875
BrainarrayCDF	24.67%	0.1842	11.71%	0.3479	4.99%	0.9910

In the next part of the study only genes common to three examined CDF files were taken into the account in order to analyze the differences between gene expression levels obtained on the base of various annotation systems. In case of the Affymetrix CDF file the values coming from the probe sets representing the same gene were averaged. On the Bland-Altman plots shown in the Figure 2 values on the x-axis cor-

responds to the mean, while values on the y-axis to the difference of two gene expression measurements. Each point corresponds to one gene and the horizontal lines represents limits of agreement that are equal to an average difference ± 1.96 standard deviation of the difference. As an example in the figures there are presented results for spermatocytes' control and subjected to 37°C samples.

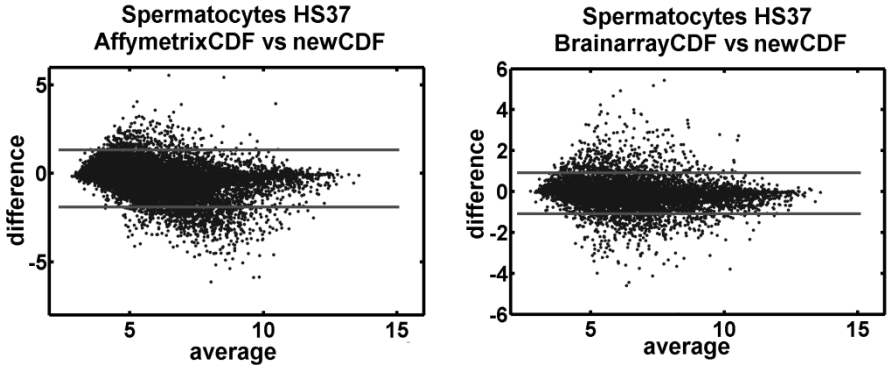

Fig 2. Bland-Altman plots for spermatocytes' sample subjected to 37°C

Additionally the average values of the gene expression levels' variances between all the groups being compared are summarized in the Table 5.

Table 5. The average values of the gene expression levels' variances between the groups being compared

compared groups	control sample vs sample treated with the high temperature	newCDF vs AffymetrixCDF	newCDF vs BrainarrayCDF
average variance	0.0049	0.0359	0.0150

Expression levels calculated for the same biological sample based on the information included in the varied CDF files differ more than the values obtained between the control sample and the treated with the high temperature sample but in the same time the concordance among the custom annotations is equal in average to 70.26% and is higher than between the one created in this work and provided by Affymetrix, equal in average only to 45.52%. It was observed that gene expression values gained using the original CDF file are more frequently underestimated than overestimated in comparison to those obtained by using the probe set definitions created in this study. Values of the signals coming from the bad quality probes, especially from these that do not have the match to any exon (*4*) or any sequence present in the genome (*7*), should be mostly located at the level of noise. In case of the Affymetrix annotation system these probes are taken into the account, often together with the reliable probes, while calculating the gene expression levels, thus lowering the final values comparing to obtained by using the custom CDF file proposed in this work, particularly in the domain of medium and relatively high values of the signals, what is well visible in the Bland-Altman plots presented in the

Figure 2. The chance for reducing the gene expression level by the bad quality probe, that is included in the set consisting of the probes from which coming signals are also situated at the low level, is lower. Furthermore at the Figure 2 it is possible to observe that in the domain of high values of the signals the majority of the data is located within the limits of agreement. It is related to the fact that a single wrong signal plays a larger role in the area of low signals' values than in the case of high values.

5 Conclusions

The performed analysis proved that the probe set definitions provided by the chip manufacturer Affymetrix are outdated as they contain irrelevant probes. Using them may lead to incorrect conclusions while analyzing data coming from the microarray experiments. The survey demonstrated that corrected for probe irrelevance custom CDF files can increase the validity of the results of the microarray experiments comparing to the Affymetrix annotation system. Presented in this work approach has an advantage over the Brainarray custom CDF files repository as it allows the user to choose the groups of probes that should be included in the alternative CDF file, based on the proposed quality assessment criteria. Additionally in case of the promoter microarrays the software allows to create custom BPMAP files. It was observed that arranging the information about probes can increase the quality of the results coming from the microarray experiments.

Funding. This work was financially supported by grant no. NN519579938

References

1. Affymetrix official website: http://www.affymetrix.com
2. 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 (2005)
3. Marczyk, M., Jaksik, R., Polańska, J., Polański, A.: Affymetrix Chip Definition Files Construction Based on Custom Probe Set Annotation Database. In: Katarzyniak, R., Chiu, T.F., Hong, C..F, Nguyen, N., editors. Semantic Methods for Knowledge Management and Communication: Springer Berlin/Heidelberg (2011) 135-144
4. Nurtdinov, R.N., Vasiliev, M.O., Ershova, A.S., Lossev, I.S., Karyagina, A.S.: PLANdbAffy: Probe-Level Annotation Database for Affymetrix Expression Microarrays. Nucleic Acids Research (2010) 726-730
5. Brainarray official website: http://brainarray.mbni.med.umich.edu/brainarray/Database/ CustomCDF/genomic_cur ated_CDF.asp
6. Cheung, Ch., Droit, A., Gottardo, R.: R implementation from MAT program to normalize and analyze tiling arrays and ChIP-chip data. R package version 2.4.2. http://www.rglab.org
7. Johnson, W.E., Li, W., Meyer, C.A., Gottardo., R, Carroll, J.S., Brown, M., Liu, X.S.: Model-based analysis of tiling-arrays for ChIP-chip. Proc. Natl. Acad. Sci. USA (2006) 12457-12462.