# Comparing Affymetrix Human Gene 1.0 ST preprocessing methods on tissue mixture data

Evgeniy Riabenko*†, Maria Kogadeeva*†, Kirill Gavrilyuk*, Evgeny Sokolov*,
Ivan Shanin*, Alexander Tonevitsky*‡
*Faculty of Computational Mathematics and Cybernetics, Moscow State University, Moscow, Russia
Email: microarray_msu@googlegroups.com
†LLC RND Bioclinicum, Bolshaya Polyanka 28/3, Moscow, Russia
‡The Institute of General Pathology and Pathophysiology, Moscow, Russia

*Abstract*—**Preprocessing is an essential step of gene expression level estimation in microarray data analysis. Although there are several benchmark studies comparing different preprocessing methods for several Affymetrix platforms, no such studies were performed for the Human Gene 1.0 ST microarray which are substantially different from the previous arrays. We aimed to compare several preprocessing methods on the tissue mixture data provided for the HuGene 1.0 ST platform to assess which methods perform best on this array. The key is the use of gene tissue specificity, tissue proportions and fold change estimates for creating assessment criteria for preprocessing methods. While on some stages of preprocessing it is possible to select uniformly best method, for the others methods could have different ranks according to different criteria.**

## I. Introduction

High density oligonucleotide array technology has been widely used for high-throughput quantitative measurements of gene expression for a long time. Over the past decade Affymetrix has developed several microarray platforms which differ in probe design, hybridization affinities, target properties etc. One of the most up-to-date platforms Affymetrix Human Gene 1.0 ST allows simultaneous quantification of expression levels of 28869 human genes.

Before expression levels from different arrays could be compared, data should be preprocessed to reduce technical noise caused by variation on different experimental steps. The preprocessing stage consists of three phases: background correction, normalization and summarization. Various preprocessing methods have been implemented which perform differently depending on the input data and experimental design. Thus it may be quite difficult for a researcher unfamiliar with peculiarities of these methods to identify one that suits his purposes best. Several benchmark studies have been carried out to assess the performance of different preprocessing methods based on the Affymetrix spike-in studies with Human U95 and U133 arrays (see, for example, [1], [2]). Although the HuGene 1.0 ST platform is widely used, no spike-in data on this platform was provided by Affymetrix. Some spike-in studies used for the HuGene 1.0 ST quality assessment were mentioned in the official Affymetrix whitepaper [3], but no further information or experimental data is available. In this paper we decided to compare existing preprocessing methods according to their performance on data obtained from HuGene 1.0 ST microarray platform. As the assessment dataset we used data from the Affymetrix tissue mixture experiment. Although there is no ground-truth information about gene expression level and only tissue proportions are known, we developed several assessment criteria for the preprocessing methods based on gene tissue specificity, replicate similarity and fold change estimates.

## II. Methods and Data

For the preprocessing methods comparison the Affymetrix Power Tools software (version 1.14.3.1) was used, which implements various basic algorithms for expression estimation. Background correction helps to measure non-specific binding and background noise, normalization is needed to minimize differences caused by scanning, hybridization and printing artifacts and summarization is a final step aimed to estimate gene expression according to specific probe intensities. We focused on the robust multi-array analysis (RMA) and GC-content methods for background correction, quantile and median normalization procedures, and PLIER, iterPLIER, median and median-polish methods for summarization of probes' intensities.

### A. Background correction methods

**RMA-BG.** Robust multi-array analysis (RMA) [4] provides one of the most popular methods for background adjustment. It assumes that the original intensity of a single probe may be represented as a sum of two values: useful signal which has exponential distribution and normally distributed background signal.

**PM-GCBG.** The main idea of the PM-GCBG method is estimating background signal as the median of the probes' intensity values that have the same GC-content as the given probe [5].

### B. Normalization methods

**Quantile normalization.** This method, as discussed by Bolstad and Irizarry [6], aims to reduce the intensity distributions of all the arrays from the sample to a common reference distribution.

**Median normalization.** The main principle of this method is adjusting intensities in such way that all the arrays have the same median value.

TABLE I
CELL-TYPE PROPORTIONS FOR EACH SAMPLE IN AFFYMETRIX DATA.

| Mix. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Brain | 0.00 | 0.05 | 0.10 | 0.25 | 0.50 | 0.75 | 0.90 | 0.95 | 1.00 |
| Heart | 1.00 | 0.95 | 0.90 | 0.75 | 0.50 | 0.25 | 0.10 | 0.05 | 0.00 |

### *C. Summarization methods*

**Median-polish.** This summarization method gives a robust estimate of gene expression level because median as opposed to mean allows to avoid taking outliers into account and the information from all the arrays in the sample is used. The robust multi-array summarization method also uses the median polish algorithm [4].

**Median summarization.** In this method the gene expression measure is estimated as the median of probes intensity values for probes from the given probeset.

**PLIER.** The probe logarithmic intensity error (PLIER) method produces an improved signal (a summary value for a probe set) by accounting for experimentally observed patterns for feature behavior and handling error appropriately at low and high abundance [7].

**IterPLIER.** The iterPLIER method uses feature sets from several exploratory annotations and iteratively discards those that appear to be performing poorly. This approach takes advantage of PLIERs ability to identify some of the signal at a particular locus and iteratively exclude features that are not correlated with that signal [8].

### *D. Data*

For the comparison analysis a publicly available dataset from Affymetrix [9] was used. The dataset contains mixtures of human brain and heart RNAs hybridized to Affymetrix HuGene 1.0 ST microarrays. There are 33 samples in the dataset in total, each sample contains specific proportions of the brain and heart cells (Table I). Each mixture is presented in triplicate, for mixture 5 nine replicates are available.

### *E. Tissue-specific genes detection*

For the comparison of preprocessed methods gene tissue specificity information was required. The data used for the analysis contained brain and heart tissues, thus two tissue-specific gene sets were created. Two tissue databases were used for tissue-specific genes extraction. The Gene Expression Barcode [10] database provides absolute measures of expression for most annotated genes for 131 human tissue types. An algorithm that leverages information from the GEO (Gene Expression Omnibus) and ArrayExpress public repositories to build statistical models that permit converting data from a single microarray into expressed/unexpressed calls for each gene was used. Database was created for Affymetrix HGU133-A and HGU-Plus 2.0 platforms. To get the list of brain- and heart-specific genes those transcript IDs were selected, which have zero expression level in heart tissue (atrial myocardium) and non-zero level in brain tissue (brain, caudate nucleus, cerebellum, lateral substantia nigra, hippocampus, cortex, neuroblastoma, gyrus, accumbens, amygdala, corpus callosum, lobe, thalamus, medulla, putamen) or vice versa. HGU133-A and HGU-Plus 2.0 Transcript IDs were matched to HuGene 1.0 ST Transcript IDs according to the official Affymetrix BestMatch table [9].

The other tissue database used in our analysis is TiGER (Tissue-specific Gene Expression and Regulation) [11]. The database contains gene expression patterns for UniGene and RefSeq genes in human tissues, which were calculated based on NCBI EST database. Tissue specific genes were identified based on the expression enrichment and statistical significance. As with the Barcode database, those UniGene and RefSeq genes which are expressed either in brain or in heart were selected. The RefSeq gene IDs were matched to HuGene 1.0 ST Transcript IDs according to Affymetrix HuGene Transcript Annotation file and the UniGene IDs were matched with the help of Ensembl BioMart database [12]. The final brain- and heart-specific gene lists were compiled as the intersection of corresponding lists from the Barcode database with those from the union of Tiger UniGene and Refseq databases.

## III. BENCHMARK

We estimated expression levels for all 33 microarrays with every combination of preprocessing algorithms described above, 16 variants in total. Expression densities are presented on Figures (1a), (1b). None of the considered algorithms was able to provide such expression estimates that technical replicates would cluster together (data not shown). To compare their performance the following criteria were developed.

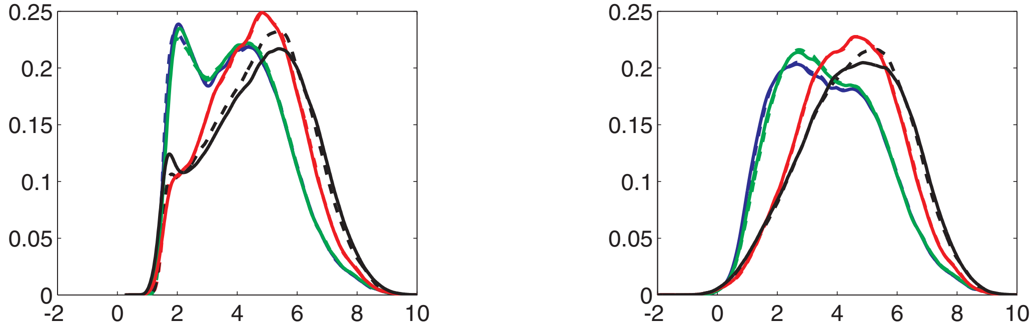
### *A. Variability between replicates*

One of the desired properties of the expression estimates is low variability on technical replicates. To measure this variability we calculated pure error mean square for each gene expression estimate and each algorithm:

$$s_p^2 = \frac{\sum_{j=1}^{9}\sum_{u=1}^{n_j}\left(C_{ju}-\bar{C}_j\right)^2}{27}.$$

Here $j$ is the number of mixture, $n_j$ is the number of replicates for the mixture $j$, $C_{ju}$ — expression level for the replicate $u$ of the mixture $j$, $\bar{C}_j$ — mean expression level between replicates corresponding to the mixture $j$. Their densities are presented on Figures (2a), (2b).

To further examine the effect of preprocessing on reproducibility of expression estimates we performed 3-way ANOVA for each gene. For about 26% of genes choice of background correction method affects reproducibility. For 22% of all the genes pure error mean square is lower for RMA-BG, while PM-GCBG performs better on the remaining 4%. Choice of normalization method affects 38% of probes, while for 35% of them quantile normalization outperforms median normalization. Summarization method seems to have the most crucial effect on reproducibility of expression estimates: only for less than 5% of genes its choice does not significantly influence pure error mean square. Median polish algorithm



(a) RMA background correction

(b) GC background correction

Fig. 1. Log(expression) estimation densities: (a) — algorithms that use RMA background correction, (b) — algorithms that use PM-GC background correction. Dotted lines represent algorithms which use quantile normalization, thick lines refer to median normalization. Colors refer to summarization methods: blue — median polish, green — median, red — PLIER, black — iterPLIER.

performance is not worse than any of its competitors on about 99.5% of genes. Median normalization is only slightly worse: its pure error is not significantly different from median polish on 91% of genes. PLIER gives more reproducible estimates than iterPLIER for about 99% of genes.

### B. Linearity

Suppose a gene is expressed in heart tissue at level $C_1$ and in brain tissue at level $C_9$, then its expression level in the mixture $i$ is defined by $C_i = \alpha_i C_1 + (1-\alpha_i)\, C_9$, where $\alpha_i\colon (1-\alpha_i)$ is heart to brain mixture ratio. Note that this expression is linear on alpha; good expression estimates $\hat{C}_i$ should be linear on alpha, too. To measure the linearity of expression estimates for each gene and each algorithm we fitted linear regression of $C$ on alpha and measured the lack of fit sums of squares which equals to:

$$s_l = \sum_{j=1}^{9} n_j \left(\hat{Y}_j - \bar{Y}_j\right)^2 ,$$

where $\hat{Y}_j$ is the expression level predicted by regression for the mixture $j$. Densities are presented on Figures (2c), (2d).

In the same manner as for the previous measure we performed 3-way ANOVA for each gene. For about 25% genes choice of background correction method affects linearity. For 16% of all the genes lack of fit is lower for RMA-BG, while PM-GCBG performs better on the remaining 9%. Choice of normalization method affects 24% of genes, while for 21% of them quantile normalization outperforms median normalization. Choice of summarization method makes bigger impact on linearity: it affects 73% of genes. Median polish, again, is the method of choice, since it is not worse than its competitors on 99.6–99.8% of genes. Median summarization is equal to median polish in terms of expression linearity for about 95% of genes and performs worse only on less than 5% of genes. PLIER outperforms iterPLIER on about 37% and performs equally on 61% of genes.

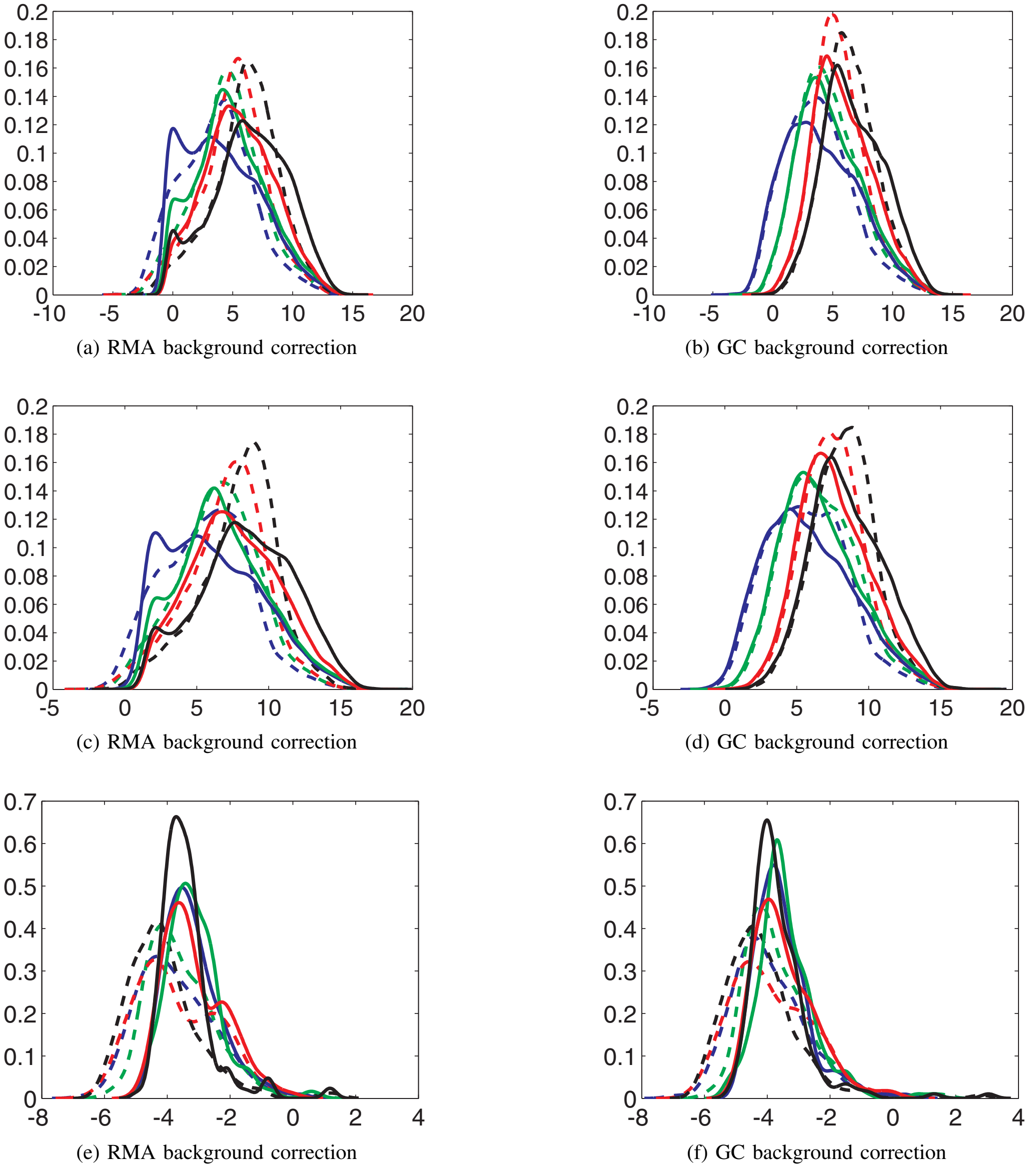
### C. Fold Change accuracy

It is easy to notice that for genes that are expressed in brain and not expressed in heart tissue $C_1 = 0, \;\; \frac{C_i}{C_9} = 1 - \alpha_i$, and for genes expressed in heart and not expressed in brain $C_9 = 0, \;\; \frac{C_i}{C_1} = \alpha_i$. Using tissue specificity data described in Methods and Data section we obtained two lists of genes — 33 heart-specific and 134 brain-specific. For each gene, each algorithm and each microarray we calculated the above-mentioned ratios and computed their mean squared deviation from the true $\alpha$. The deviations' densities are presented on Figures (2c), (2d).

Again, RMA-BG outperforms PM-GCBG for the most of the genes (71%, unsignificant for 28% of genes). Quantile normalization is significantly better than median on 74% of genes, unsignificant on 19%. Choice of summarization method does not affect accuracy of fold change estimation for 25% of genes. It is noticeable that median polish is not the best choice anymore, iterPLIER instead shows the best performance: the accuracy of fold change estimation is not worse than for the other methods for 93–95% of genes. Median polish is the second best, although its results are equal to the results of iterPLIER only for 66% of genes. Median summarization shows the worst result and is outperformed by median polish, PLIER and iterPLIER on 31%, 31% and 41% respectively.

## IV. Conclusion

Microarray data preprocessing is an important step in microarray data analysis which significantly influences gene expression level estimation. We provided a benchmark study for different preprocessing methods for Affymetrix Human Gene1.0 ST microarray platform based on the publicly available tissue mixture dataset from Affymetrix. The proposed quality criteria of expression estimations — variability between replicates, linearity of expression estimates, accuracy of fold change restoration on tissue-specific genes — might be used in further studies. We also compared the performance of different preprocessing methods according to these criteria. While on some stages of preprocessing it is possible to select



(a) RMA background correction

(b) GC background correction

(c) RMA background correction

(d) GC background correction

(e) RMA background correction

(f) GC background correction

Fig. 2. Different algorithms' logarithmic estimation errors: (a), (b) — pure expression values prediction errors; (c), (d) — lack of fit sum of squares; (e), (f) — fold change prediction errors. Colors and line styles refer to the same methods as in Fig.1.

uniformly best method, for the others methods could have different ranks according to different measures.

## Acknowledgment

This work is supported by Russian Ministry of Education and Science (contract 16.522.12.2015).

## References

[1] Cope L.M., Irizarry R.A., Jaffee H.A., Wu Z., Speed T.P. A benchmark for Affymetrix Genechip expression measures. *Bioinformatics.* 2004; 20: 323–331.

[2] Irizarry R.A., Wu Z., Jaffee H.A. (2006). Comparison of Affymetrix GeneChip expression measures. *Bioinformatics*, 22(7): 789–794.

[3] Affymetrix. (2007). Human Gene 1.0 ST Array Performance.

[4] Irizarry R.A., Hobbs B., Collin F., Beazer-Barclay Y.D., Antonellis K.J., Scherf U., Speed T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2): 249–264.

[5] Affymetrix. (2005). Exon Array Background Correction. Technical Note.

[6] Bolstad B.M., Irizarry R.A., Astrand M., Speed T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2): 185–193.

[7] Affymetrix. (2005). Guide to Probe Logarithmic Intensity Error.

[8] Affymetrix. (2005). Gene Signal Estimates from Exon Arrays.

[9] http://www.affymetrix.com/support/technical/sample_data/gene_1_0_array_data.affx.

[10] McCall M.N., Uppal K., Jaffee H.A., Zilliox M.J., Irizarry R.A. (2011). The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic acids research*, 39: D1011-D1015.

[11] Liu X., Yu X., Zack D.J., Zhu H., Qian J. (2008). TiGER: a database for tissue-specific gene expression and regulation. *BMC bioinformatics*, 9: D1011–D1015.

[12] http://www.ensembl.org/biomart/martview/