Academy of Scientific Research and Technology
27th National Radio Science Conference
Faculty of Electronic Engineering, Menoufia Univ., Menouf, Egypt
16-18 March 2010

# The Influence of Pre-processing and Gene Rank Aggregation on Microarray Data Analysis

Vidan Fathi Ghoniem[1], Nahed H. Souluma[2], Yasser M. Kadah[3]
[1]*Biomedical Engineering Department, Misr University for Science and Technology, Egypt*
[2]*Engineering Applications Department, NILES, Cairo University, Giza, Egypt*
[3]*Systems and Biomedical Engineering Department, Cairo University, Giza, Egypt*

## Abstract

Preprocessing of microarray data might influence the precision of any further calculations based on these raw data. Important procedures should be considered, concerning preprocessing microarray data. This involves background correction, normalization, and detection of faulty spots from microarray data prior to analysis. In this study various methods for normalization and background correction were applied to a set of twenty samples presenting thyroid papillary cancer tissues versus patient-matched adjacent non-tumor thyroid tissues, in a trial to stand over the impact of preprocessing on the analysis results. To assess our results, we employed classification techniques to analyze the data after being preprocessed. This emphasizes the impact of preprocessing microarray data on gene expression analysis. Another challenge is to combine lists coming from different sources and platforms, for example different microarray chips, which may or may not be directly comparable otherwise. This is one of the major strengths of rank-based aggregation. Work is under way in this study to use genetic algorithm in the context of meta-analysis of several microarray cancer studies on liver where the goal is to determine the combined set of genes indicative of the cancer status. Genetic algorithm is applied to a set of sixty one samples presenting hepatitis C, another set of ten normal liver tissue samples, and a set of twenty seven liver cancer (primary tumor), in an attempt to predict biomarkers for progression towards hepatocellular carcinoma (liver cancer) in hepatitis C virus patients.

## 1. Introduction

Microarray technology offers a powerful tool for modern biomedical research. Using microarrays, expression levels of thousands of genes can be measured simultaneously. Two technologies are widely employed: high-density oligonucleotide- based chips produced by Affymetrix (Lockhardt et al., 1996) and cDNA microarrays, which are microscope slides spotted with thousands of cDNA fragments. In both cases the arrays are hybridized to fluorescent-labeled cDNAs generated by reverse transcription of RNA isolated from the cell sample or tissue under investigation. In standard terminology, the cDNAs spotted onto the arrays are called probes, and those in the samples are called target genes. cDNA microarrays generate one- or two-channel data. In two-channel use, the arrays are hybridized to a mixture of two samples, each labeled with a different dye (Cy3 and Cy5). In one-channel use, each array is hybridized to a single sample, labeled with a single dye. The arrays are laser-scanned at the wavelength(s) appropriate to the dye(s) used, and the images are processed to extract data for analysis. In one channel studies, these usually consist of a measure for the spot intensity and its local background, for each spot on the array. In two-channel studies, this is available for both dyes.

Measurements of microarrays may be biased by diverse effects such as efficiency of RNA extraction, reverse transcription, label incorporation, exposure, scanning, spot detection, etc. Furthermore, there are systematic effects due to characteristics of the array, such as effects of different probes (i.e. cDNAs or oligos), spotting effects, region effects, pin effects, etc... Gene activity estimation has an impact on subsequent data analysis and interpretation. If the gene's measured activity is not due to the activity itself, subsequent analysis using this

erroneous estimate will, of course, be misleading. Before analysis, the spot intensities are corrected for the background intensities. The background-corrected spot intensities reflect the abundance of the corresponding target genes in the samples. Although, there are many methods for the background correction, the results of their application are nor very similar. However, often the relation is not that of simple proportionality: the signals may be distorted in various ways. The process of correcting for bias prior to analysis is called normalization. The purpose is to promote uniformity within arrays and reproducibility between arrays. Normalization has profound effects on subsequent analysis, irrespective of the methodology used. Failure to normalize appropriately will generally lead to misleading conclusions.

A multitude of cancer mRNA profiling studies has stratified certain types of cancer and defined gene sets that correlate with outcome. These studies have resulted in plans for prospective application of nucleic acid–based tests to select patients that do not need further therapy after their primary resection. However, the number of genes used to predict patient outcome or define tumor subtypes by RNA expression studies is variable, non-overlapping, and generally requires specialized technologies. Immunohistochemical studies can be done with many fewer markers, but suffer from the inherent flaw of subjective analysis and variable reproducibility. Thus, it would be ideal if the familiarity and streamlined nature of immunohistochemistry could be combined with the rigorously quantitative and highly specific properties of nucleic acid–based analysis to predict patient outcome. This can be achieved using genetic algorithm-based objective quantitative analysis of tissue microarrays toward the goal of discovery of a minimal number of markers with maximal prognostic or predictive value that can be applied to the conventional formalin-fixed, paraffin-embedded tissue section [1].

## 2. Background Correction

The most common method of background correction is the subtraction method. The usual assumption of two-color microarray data; with red and green dyes; is that the background signals, Rb and Gb, are additive to the true signals, R and G on the raw intensity scale. Given the observed foreground intensities, Rf and Gf, this allows the true signal to be estimated by subtraction, such that R = Rf - Rb and G = Gf - Gb. Because of the very wide range of intensities, the corrected intensities are then used to form the log-ratio, M, and average log intensity, A for each spot.

Most image analysis programs return ‘local’ background intensities, obtained from the mean or median of the pixel intensity values surrounding each spot. Local background is arguably an unbiased estimate of the local non-specific signal, so subtracting it from the foreground intensity gives in principle an unbiased estimator of the true signal due to hybridization. Although well motivated, this traditional approach produces corrected intensities with undesirable statistical properties. It produces negative intensities whenever the background intensity is larger than the foreground intensity, leading to missing log-ratios, sometimes for a substantial proportion of probes on an array. Even when not missing, the log-ratios are highly variable for low intensity spots [2].

In the Kooperberg’s method, a Bayes model is used to solve the problem with negative background corrected spots. Relying on the prior knowledge that the true spot intensity should be non-negative, the posterior distribution of the true spot intensity is calculated. Thereby not only negative spots, but also spots with intensities slightly above background, are estimated more accurately .This model involving a convolution of normal distributions to background, adjusts the signal from each spot. Observed foreground and background mean intensities and their standard deviations, along with the number of foreground and background pixels for each spot in a given channel are used in this model. Numerical integration is applied to obtain the expected value of the true signal in each channel for each spot [3].

A simpler strategy to avoid negative intensities is suggested by Edwards (2003), who adjusts the foreground intensities by subtracting the background when the difference between the foreground and background is larger than a small threshold value. When the difference is less than the threshold, subtraction is replaced by a smooth monotonic function. This method is applied with local median background estimates from Genepix.

The normexp method is based on the same normal plus exponential convolution model. This model has been used to background correct Affymetrix data as part of the popular RMA (Robust Multi-array Analysis) algorithm [4]. Where, RMA summarizes probe intensities for each probe set, using an effective expression measure motivated by a log scale linear additive model. Two changes have been made to the method for use with two-color arrays. First, the convolution model is fitted to the background subtracted signals for each channel separately. Second, the kernel density parameter estimation method used in RMA has been replaced by maximum-likelihood estimation, which is more sensitive to the true parameter values. In order to make maximum likelihood numerically feasible, a saddle-point approximation is used to simplify the mathematical form of the likelihood function. GenePix data was also corrected using this model [2].

A slight variation on the normexp method is to add a small positive offset, k, to move the corrected intensities away from zero. This is a simple variance-stabilizing technique. It should reduce the variation of the low intensity M-values, since log2 [(R + k)/ (G + k)] will be close to 0 for R and G both small relative to k. The use of an offset is effective here because normexp produces corrected intensities which are positive but may be close to zero [2].

The variance stabilization method (VSN) of Huber et al. (2002) [5] calibrates the data from each channel between arrays and uses a generalized arcsinh transformation of the data instead of the logarithm. The arcsinh function is defined for negative values, which ensures negative corrected signals can be handled. At high intensities, the arcsinh transform is equivalent to the regular log-ratio, whereas at low intensities it is close to the difference between the transformed intensities. As it transforms the data to have equal variance for all intensities, this transformation tries to adjust for effects that are often observed after background subtraction (i.e. high variance for lowly expressed genes). VSN can be used with cDNA or affymetrix data, and is advisable if unstable results with lowly expressed genes are observed.

In contrast to the other presented methods, VSN operates on all the arrays together rather than for each array separately, and background corrects and normalizes the intensity data in one operation. For all other methods, a separate normalization step is required.

## 3. Normalization

Normalization is essential to compare the varying conditions of microarray experiments. While there are several methods for normalization, the choice of which method gives the best results really depends on local settings. During array spotting, different types of controls are used. This involves: negative controls, spiked in controls, and housekeeping genes. They enable to observe possible problems with the hybridizations. Whenever possible normalization on the majority of genes or using a sufficiently large number of non-differentially expressed controls are recommended to normalize the data.

Transforming the microarray intensities using the logarithmic (log2) scale should roughly adjust the variance to be the same for all intensities. Differences of log2 intensities reflect the log2 ratios for comparison.

Loess smoothing, also known as local regression is a technique for fitting smooth non-linear functions of a set of predictor variables to a continuous response variable. In the one-dimensional case, given a set of (x, y) values, the method fits a smooth function of x to y. In the two dimensional case, given a set of (x, y, z) values, it fits a smooth function of x, y to z. The procedure is quite complex [6]. A window of neighboring points about each x (or (x, y) in the two dimensional case) is defined. Within the window, robust regression techniques are applied, in which observations are down-weighted the further they are from x. Observations outside the window are excluded. A smoothing parameter $\alpha$ is used to size the window, by specifying the proportion of data to be included. The wider the window, the smoother the curve obtained. This type of normalization is most commonly used for two-color arrays.

The concept of quantile is also used for normalization. It is based on the rationale behind the QQ-plots where the quantiles (i. e. the sorted measurements or values) of a data set X is plotted against the quantiles of another data set Y. If X and Y both come from the same distribution then their QQ-plot approximately shows a line

along the diagonal. However, in the case of probe level intensities the quantiles of two arrays usually do not lie on the diagonal even though their true underlying expression values are (or at least should be) identically distributed in replicate samples. One could argue that the respective distribution functions were transformed during the microarray experiment due to technical reasons. In order to regain a common distribution, one simply could project the quantiles onto the diagonal of the QQ-plot. This method is recommended to correct for relative intensity bias, on grounds of computational efficiency [6], [7].

## 4. Rank Aggregation

Microarray cancer studies often attempt to identify genes related to a specific cancer. Their most common output is a list of genes ordered by corresponding p-values. Different studies, even the ones analyzing the same cancer type (i. e liver cancer), almost never produce identical gene lists. Meta-analysis of multiple microarray studies is difficult, especially if different experimental platforms have been used. Rank aggregation, however, avoids the issue of multiple experimental conditions by dealing with the final product: the ordered list of genes.

The goal of rank aggregation is to combine these lists into the overall top-k gene list which hopefully would be more accurate than any individual list by itself. The genetic algorithm performs reasonably well for the aggregation problem but one has to be careful with the selection of important tuning parameters which control the rate of the learning process [1].

To cast the rank aggregation in the framework of an optimization problem, we first need to define the objective function. In this context, we would like to find a "super"-list which would be as "close" as possible to all individual ordered lists simultaneously. This is a natural requirement and the objective function, at least in its most abstract form, is very simple and intuitive. The idea is to find $\delta^*$ which would minimize the total distance between $\delta$ and $L_i$'s

$$\delta^* = \arg\min \sum_{i=1}^{m} w_i d(\delta, L_i) \tag{1}$$

Where $\delta$ is a proposed ordered list of length $k = |L_i|$, $w_i$ is the importance weight associated with list $L_i$, d is a distance function and $L_i$ is the $i^{th}$ ordered list. Selecting the appropriate distance function d to measure the "distance" between ordered lists is very important. One of the most popular ones is Spearman's footrule distance.

The Spearman's footrule distance between Li and any ordered list $\delta$ can be defined as

$$S(\delta, L_i) = \sum_{t \in L_i \cup \delta} \left| r^{\delta}(t) - r^{L_i}(t) \right| \tag{2}$$

Where $r^{Li}(A)$ is the rank of A in the list $L_i$ (1 means "best") if A is within the top k, and be equal to k + 1, otherwise; $r^{\delta}(A)$ is defined likewise.

## 5. Data Analysis

The data set used in this study with preprocessing methods was downloaded from: http://www.ncbi.nlm.nih.gov/geo. The dataset includes twenty samples presenting thyroid papillary cancer tissues versus patient matched adjacent non-tumor thyroid tissues (GEO accession: GSE3950). The datasets for applying rank aggregation were downloaded from Stanford microarray database. It include a hundred and one samples representing hepatitis C, another ten normal liver tissue samples, and twenty seven liver cancer samples (primary

tumor) .The images have been analyzed using GenePix. GenePix results data are saved as GPR files in Axon Text File (ATF) format. A Results' file contains general information about image acquisition and analysis, as well as the data extracted from each individual feature.

The above mentioned methods of background correction and normalization were implemented using R limma package and were applied on the first dataset representing thyroid papillary cancer tissues acquired by genepix from cDNA microarrays spotted by a total of 16200 genes. To assess the impact of preprocessing, we tried to extract the genes having higher differential expression. Gene selection was done using Fisher discriminate analysis (FDA).FDA is a simple algorithm applied mainly to reduce the dimensionality of the data thus outputting the most discriminate features (genes expression levels), according to the value of the Fisher factor j given by:

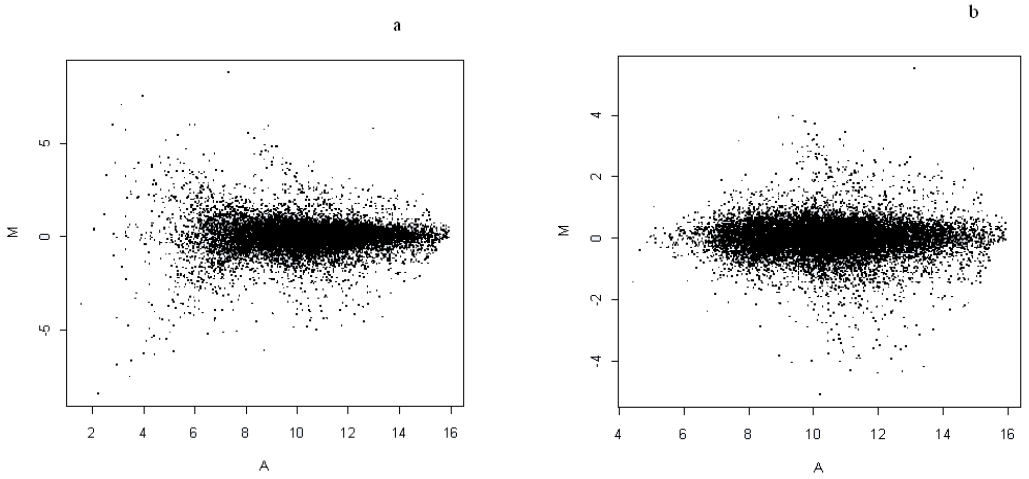
$$J(gene) = (\mu_1 - \mu_2)^2 \Big/ \sigma_1 + \sigma_2 \quad (3)$$

Where, $\mu_1$, $\sigma_1$ and $\mu_2$, $\sigma_2$ are the means and variances of the two classes; normal and tumor sets respectively. It is clear from (3) that, J has a higher value when the feature value differs greatly in the two classes and vice versa. Using the cross validation approach, we trained and tested the system using the available dataset. The simple Euclidian distance classifier, as a supervised classifier was used to classify the data as belonging to normal or tumor tissue.

The work in progress now, is to apply rank aggregation techniques to develop biomarkers for progression towards hepatocellular carcinoma in hepatitis C virus patients. The predicted biomarkers will be validated using classification techniques. In this perspective, we are using an R RankAggreg package available through CRAN (http://cran.r-project.org/web/packages/RankAggreg/) which provides two different algorithms for rank aggregation: the Cross-Entropy Monte Carlo algorithm (CE) and the Genetic algorithm (GA). Genetic algorithm is applied on the datasets mentioned earlier representing liver cancer, normal liver, and hepatitis C virus liver tissues.

## 6. Results and Discussion

When applying the six methods for background correction discussed earlier, we have obtained the MA-plots which represent M versus A, where M are the differences between corresponding gene expression values and A are their averages. Figure 1 shows the MA-plots of one patient notated by N1. It is obvious that some background correction methods produce M-values which are much more variable than others, and this fanning is most apparent at low A-values. The striking feature of Figure 1 is that the background methods with less variable M-values also give compressed ranges of A-values. The most extreme is VSN, for which the A-values start at nearly 8 rather than at 0.



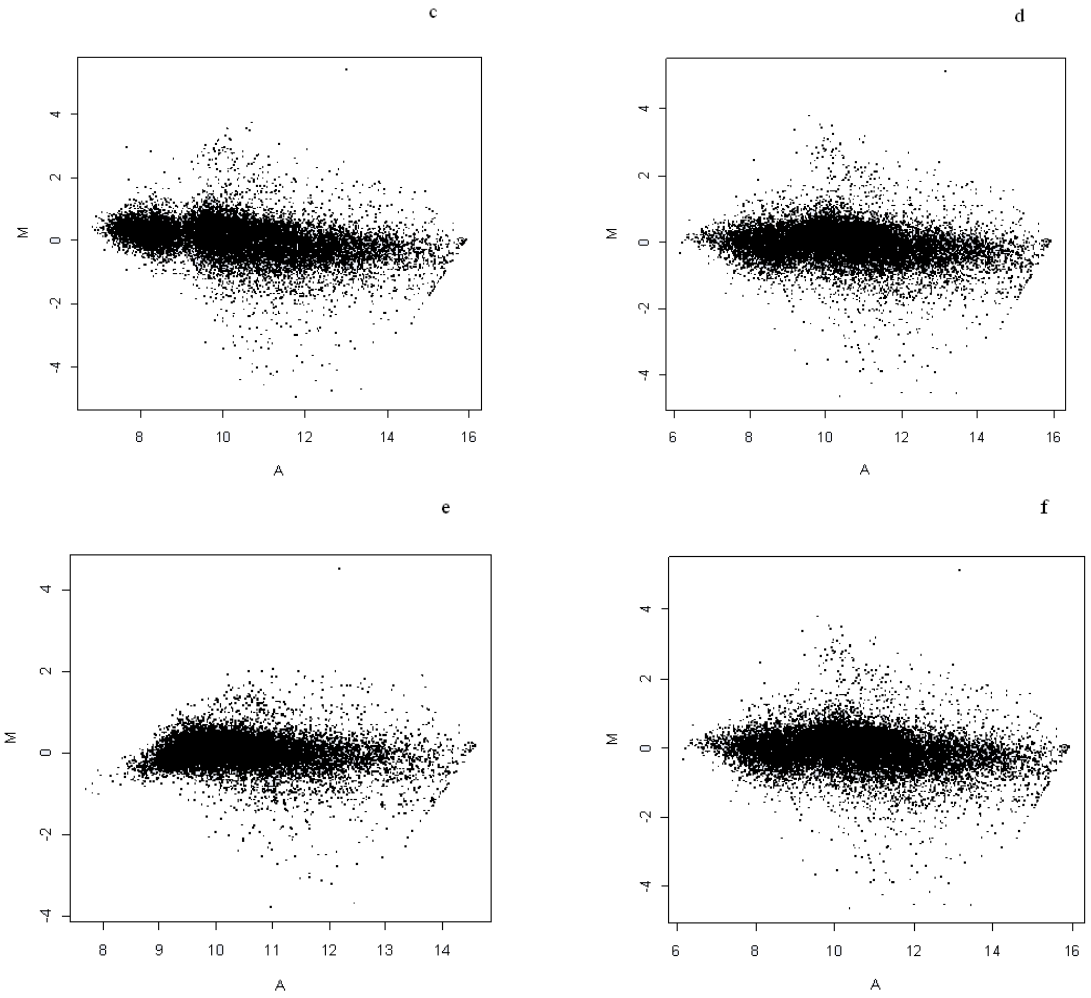



**Figure1. MA-plots obtained using different background correction methods on the same patient sample N1: (a) subtraction, (b) normexp+offset, (c) kooperberg, (d) Edwards, (e) VSN and (f) normexp.**

Normalization methods may be broadly classified into methods which normalize the M-values for each array separately (within-array normalization) and methods which normalize intensities or log-ratios to be comparable across arrays (between-array normalization). An important issue to consider before normalizing between arrays is how background correction has been handled. For between-array normalization to be effective it is important to avoid missing values in log-ratios which might arise from negative or zero corrected intensities. Figure 2 is a boxplot of the whole dataset including normal samples denoted by N1, N2,…..and tumor samples denoted by T1, T2,……It shows the result of using 'quantile' normalization after normexp+offset (offset=50) background correction. A feature which distinguishes most of the between-array normalization methods from within-array normalization is the focus on the individual red and green intensity values rather than merely on their log-ratios. These methods might therefore be called individual channel or separate channel normalization methods. Density plot displays smoothed empirical densities for the individual green and red channels on all the arrays. Without any normalization there is considerable variation between both channels and between arrays as illustrated in Figure 3. Applying quantile normalization to the A-values makes the distributions essentially the same across arrays as well as channels.

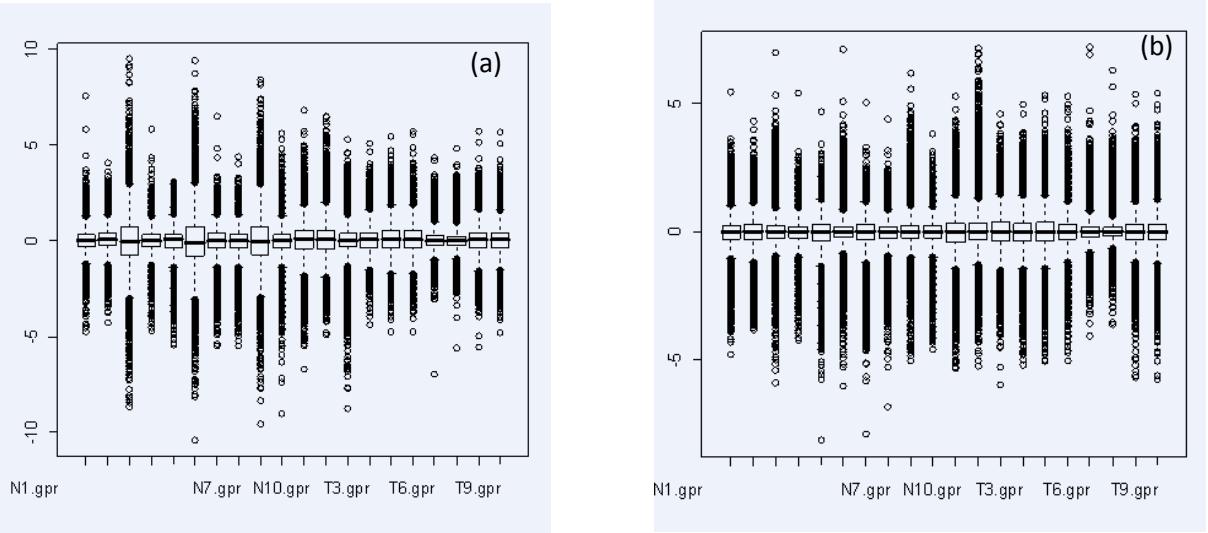



**Figure2. (a) box-plot of M-values of background-corrected dataset before quantile normalization, (b) box-plot of M-values of background-corrected dataset after quantile normalization.**

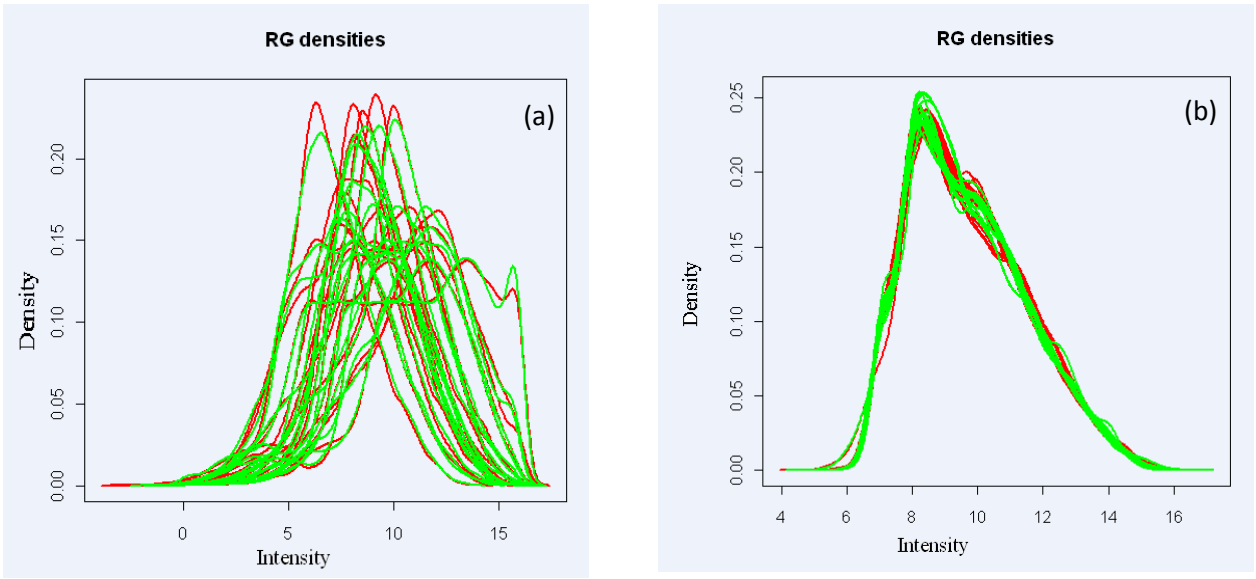



**Figure3. Red and Green densities of the whole dataset: (a) before applying quantile normalization, (b) after applying quantile normalization**

To assess the influence of preprocessing techniques on the analysis of differential gene expressions, applying all discussed methods for background correcting and normalization, we have imposed the resulting preprocessed data, represented in M-values for all samples, to diagnostic analysis. Before applying any of the preprocessing techniques, the classification attained 75% precision. The results of different combinations of

preprocessing methods, feature extraction, and classification techniques are shown in Table 1. By going through the different background correction techniques and exploring their results in previous studies [2], [8], [9]; they were ordered from low to high offset as: standard subtraction, kooperberg, Edwards, normexp, VSN, and normexp+offset. In this study and according to Figure 1, one can notice slight difference in ordering the methods from low to high offset: standard, normexp, Edward, normexp+offset, kooperberg, and VSN. The higher offsets show low classification results. The normexp, normexp+offset, and Edwards methods give high classification precision and VSN gives unexpectedly low precision at high intensities. It is also interesting to note that normexp+offset appears to give the best stabilization of the variance as a function of intensity even beating VSN, which is explicitly designed to stabilize the variance [2].

**Table1. Performance of the preprocessing algorithms**

| Background correction method | Sensitivity | Specificity | Overall accuracy |
|---|---|---|---|
| **Standard** | **80%** | **75%** | **78%** |
| **Normexp** | **80%** | **100%** | **90%** |
| **Edwards** | **80%** | **90%** | **85%** |
| **Normexp+offset** | **70%** | **100%** | **85%** |
| **Kooperberg** | **90%** | **80%** | **85%** |
| **VSN** | **80%** | **70%** | **75%** |

## 7. Conclusion

Gene expression analysis using cDNA microarrays is a multi-step procedure. Preprocessing is an important part of this procedure. Its purpose is to establish a single, potentially background corrected, and normalized expression value per gene on the same chip and from other chips of the experiment. Considering the output results most of the background correction methods used namely normexp, normexp+offset, Edwards, and Kooperberg noticeably improved the classification precision of the data. One should also take notice of the effectiveness of the quantile normalization in making the overall intensities distributions almost the same across all arrays.

## 8. References

[1] Vasyl Pihur, Susmita Datta and Somnath Datta: "RankAggreg, an R package for weighted rank aggregation". *BMC Bioinformatics* , doi:10.1186/1471-2105-10-62, 2009.
[2] Matthew E. Ritchie, Jeremy Silver, Alicia Oshlack, Melissa Holmes, Dileepa Diyagama, Andrew Holloway and Gordon K. Smyth: "A comparison of background correction methods for two colour microarrays". *Bioinformatics*, Vol. 23 no. 20, pp. 2700– 2707, 2007.
[3] Kooperberg C, Fazzio T, Delrow J, Tsukiyama T: "Improved background correction for spotted DNA microarrays". *Journal of Computational Biology*, 9, pp. 55– 66, 2002.
[4] Rafael A. Irizarry, Benjamin M. Bolstad, Francois Collin, Leslie M. Cope, Bridget Hobbs and Terence P. Speed: "Summaries of Affymetrix GeneChip probe level data". *Nucleic Acids Research*, Vol. 31 no. 4 e15, 2003.
[5] Wolfgang Huber, Anja Von Hydebrecke, Holger Sultmann, Annemarie Poustka and Martin Vingron: "Variance stabilization applied to microarray data calibration and to the quantification of differential expression". *Bioinformatics*, Vol. 18, pp. S96–S104, 2002.

[6] David Edwards: "Non-linear normalization and background correction in one-channel cDNA microarray studies". *Bioinformatics*, Vol. 19 no. 7, pp. 825–833, 2003.
[7] Johannes M. Freudenberg: "Comparison of background correction and normalization procedures for high-density oligonucleotide microarrays". *Leipzig Bioinformatics Working Paper*, Interdisciplinary Centre for Bioinformatics, No. 3, ISSN 1860 – 2746, 2005.
[8] Anders Bengtsson and Henrik Bengtsson: "Microarray image analysis: background estimation using quantile and morphological filters". *BMC Bioinformatics*, pp.7 – 96, 2006.
[9] Wotao Yin, Terrence Chen, Xiang Sean Zhou and Amit Chakraborty: "Background correction for cDNA microarray images using the TV+L1 model." *Bioinformatics*, 2005.