# Choice Impact of Soft Analysis Tools in Genes Selection

O.Kaissi[1], A.Moussa[1], B. Vannier[2] and A.Ghacham[1]
[1]LTI Laboratory, ENSA, University Abdelmalek Essaadi ,Morocco
[2]IPBC, University of Poitiers, France
amoussa@uae.ac.ma

**Abstract.**

**Motivation:** In the analysis of experiments that involves the high density of oligonucleotide chips, it is important to generate list of genes or 'targets' from the genomewide data set that contains a lot of information. Gene selection is a process that seeks to identify the most significant genes which reveal large expression changes between the baseline experiments and the conditions. Even though, several algortithms like T-test and other derived statistical algorithms were used for that selection process, the suitable Pvalue Cutoff remains difficult to choose. Therefore, one solution consists of using a False Discovery Rate (FDR) control. The Significance Analysis of Microarray (SAM) and the T-test Benjamini & Hochberg (BH) algorithms have been successfully used in such way. However, the reproductivity of results and their impact on the genes and/or experiments classification, while using different soft tools remain a subject of discussion.
**Method:** we use two Affymetrix data sets, when we look for identifying list of genes under SAM and T-test-BH algorithms with FDR control runing under R/Bioconductor project and Bioinformtics ToolBox of Mathworks.
**Results:** The list of selected genes changes significantly when using the two algorithms under both R/Bioconductor project and Bioinformtics ToolBox of Mathworks. By means of data provided from publicly databases, we illustrate, that the permutation process of the multiple statistical T-test (SAM and BH) may affect results of selection process. Moreover, list of genes using the two Soft is affected by the choice of the Pvalue-CutOff for identifyng true differential expressed genes. According to this work, we present some results clarifying sensitivity and efficiency of used soft and its influence in gene selection process. Hierarchical classification of selected genes and corresponding experiences confirm the influence of both methods and tools on the outcome of gene expression data analysis.

## 1 Introduction

The preparation of manuscripts which are to be reproduced by photo-offset requires special care. Papers submitted in a technically unsuitable form will be returned for retyping, or canceled if the volume cannot otherwise be finished on time.

The technology of DNA microarrays witnesses an exceptional growth and attracts a tremendous advantage in the scientific community. This interest lies in its efficiency; speed of obtaining results, and in its ability to study simultaneously the expression of

thousands of genes [1]. The use of microarray in various fields including biology and health, allows development of several technologies grafting and in situ [2, 3]. Therefore, several computational and statistical tools were developed to store, analyze and organize data [4]. Microarray chips consist of a DNA fragment immobilized on a solid support according to an ordered arrangement. The principle is based on the chip hybridization using a probe carrying the radioactive labeling [5]. High resolution scanner is then used to measure the signal intensity of the image that quantifies the level of genes expression.

On the other side, selection process of differentially expressed genes (DEG) across multiple conditions is one of the major goals in many microarray experiments [6]. Since one cannot analyze the raw data with thousand's or more of genes, a variety of multiple-testing procedures for DEG selection have been developed [7]. A statistical test like t-test is the main procedure used when the goal is to detect significant level of genes expression; it can be generalized to multiple groups testing for identifying DEG [8]. In literature, the statistics t-test for microarray analysis are abundant [9, 10, and 11]. Some methods use 'False Discovery Rate' (FDR) control to compute the probability that a given gene is a false positive and is identified as DEG [12]. A permutation-based approximation of this method, assumes that each gene is an independent test, is implemented in the Significant Analysis of Microarray (SAM) program [11].

In microarray data analysis, a comparative study seems to be a useful tool that leads the analyst to a suitable choice of methods, algorithm and analysis software. In this context, different comparisons have been implemented such as the comparison of normalization methods for high density oligonucleotide [13], comparison of selection methods [14] and comparison of statistical clustering techniques [15]. Yet, the reproducibility of result from these algorithms and their impact on the classification, when using different development tools and different technologies of microarrays remain a stand point of debate. For that reason, this paper presents tools of comparison study that use two methods for identifying DEG and evaluate their performances on two publicly available microarray data sets. The aim is to show: on the one hand, the impact of the P-values choice on the number of detected genes. On the other hand, to discuss the performance of selection methods and their impact on the classification on the both softwares. In the second section, this paper summarizes an overview of the the use of Affymetrix technology in DEG analysis and describes tools and statistical methods used in genes selection. Results and discussions are presented in the last section.

## 2 Materials and Methods

### 2 1 DEG Analysis in Affymetrix

Affymetrix Gene Chip represents a very reliable and standardized technology for genome-wide gene expression screening [7]. In this technology; probe sets of 11–20 pairs with 25-mer oligonucleotides are used to detect a single transcript. Each oligonucleotide pair consists of a probe with perfect match to the target (PM probe) and another probe with a single base mismatch in the 13th position (MM probe) [8]

The most widely format used for analyzing data provided from Affymetrix technology is .CEL format. This last, called “the raw data”, contains the microarray feature intensity quantification, and such data are the starting point for quality assessment and expression analysis.

Several experiences in microarray data intend to compare two conditions (treated # baseline). The objective is generally to answer the question: does the expression of a transcript on a chip (treated) change significantly with respect to the other chip (baseline)? In this context, five possible distinct answers are: Increase, Decrease, Marginal Decrease, Marginal Increase and No Change. These detections calls are giving by comparing change p-values of each gene the four thresholds chosen by the analysis for Affymetrix technology.[9]. In the case of high dimensional data, for example when comparing several experiences, the detection call is a limited tool and other solution like multiple testing procedures can be used. Some of these procedures, such as the Bonferroni procedure, control the Family-Wise-Error-Rate (FWER). The other multiple-testing procedures, such as the Benjamini and Hochberg (BH) procedure, control the False Discovery Rate (FDR) [12]. Another challenging aspect of microarray data analysis is to choose appropriate test statistics for different types of responses and covariates obtained from the datasets. The commonly used statistics including the t statistic and the F-statistic were originally designed for performing a single test but are not appropriate for large-scale data analysis. This motivated the development of many new statistics that borrow information across multiple genes for identifying differentially expressed genes, including a Significance Analysis of Microarrays (SAM); offering then a random testing approach which relies on relatively weak assumptions and yet are quite powerful [11].

Several of these methods have à strong and weak point, but there is no argument over the choice of a particular tool. For this we will try to argue in crossing a comparative study using two well used tool for gene selection under different algorithms: The SAM statistical algorithm [11], and The T-test BH algorithm[12].This choice is justified by their popularity and their availability in Expander, Bioconductor and Bioinformatics Tools Box of Mathworks.

### 2 3 Soft Tools

The first used software is Bioconductor that is a collaborative project using the statistical programming language R [18]. It allows statistical analysis on the use of different packages developed between other free applications especially designed for the analysis of biological data including microarray. For the analysis of Affymetrix chips with Bioconductor, we must first ensure that the Affymetrix libraries are installed [19]. The selection of differentially expressed genesis realized by the "limma" package integrated in Bioconductor.

The second software is Bioinformatics Tool Box of Mathworks, which presents many advantages for the analysis of microarray data: it offers an efficient and natural way of dealing with large data sets, provides a comprehensive set of functions, dedicated for the microarrays analysis.

### 2.3 Data

The first data set is Latin square produced by Affymetrix chips on human (HGUAI33A). In this publicly available set, 12 yeast genes and 14 human genes are cloned. Each of the labeled genes were pooled into groups and diluted to concentrations of 0, 0.24, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024 Pm. In each microarray experiment, 14 groups of genes in 14 different concentrations were hybridized to microarray in the presence of a complex background of expressed human genes and several control genes. For this Latin square design, 14 groups of experiments with 3 replicates for each one, give a total of 42 experiments. The concentrations of the in vitro transcript (IVT) groups in the first experiments are 0, 0.25, 0.5,….., 1024 pM, their concentrations in the second experiments are 0.25, 0.5, 1024, 0 pM, and so on [20].

The second data set concern data provided from cancer projects. In fact, the variety and extent of the cancer data, gave us access to choose the second phase of data dealing a Chronic Lymphocytic Leukemia (CLL) [21].The CLL is the most common chronic leukemia with an extremely variable clinical course. Some patients survive only a few months, whereas others have stable disease for years. The identification of novel genes mutated in CLL is important for prognostic purposes, to understand the biology of the disease and identification of targets and pathways for therapeutic intervention. In this type of data, two normal B cells isolated from peripheral blood and 5 CLL specimens have been analyzed with affymetrix (HGUAI33A) microarray for expression. The main goal is to find genes that reveal a significance change between normal B cells and CLL cells.

## 3 Results and Discussions

Throughout this research, we analyzed the performance of statistical tests integrated in Soft Tools cited below using Latin square [20] and CLL data. The first demonstration aims to identify true detected spikes according to some Cutoff of Pvalues selected. The Latin Square data seems to be the best in this case because DEG are known, therefore, we can compute the sensitivity (using the true detection rate) of used tools and methods. We have conducted some experiences when we evaluated the change of the DEG based on Pvalues from 0.001 to 0.02. In this context, the two graphs plotted below illustrate the variation of the Number of genes selected according to used algorithms and tools. Whereas the Figure3 leads to test the impact of the Pvalues change on the number of detected Spike (Fig.1).

In figure 2, the number of DEG varies according to the selection methods, depending on the software and also according to Pvalues. In connection with the test of selection, the behavior of the SAM algorithm is not expected and some DEG selected by SAM are not sorted by T-test BH algorithm under the same Pvalue Cutoff. This variation is due to a large part of the random permutation of the SAM algorithm. Once observing the size of DEG selected from each method, it's clear that bioconductor sorts a small size list of DEG comparing to Bioinformtics Toolbox of Matlab. In fact, SAM is considered a very powerful test for selection especially in the case of a large sample size

like Latin Square data set. This result show also that when we use only DEG as an element of comparison, the outcomes of plied tools still similar.

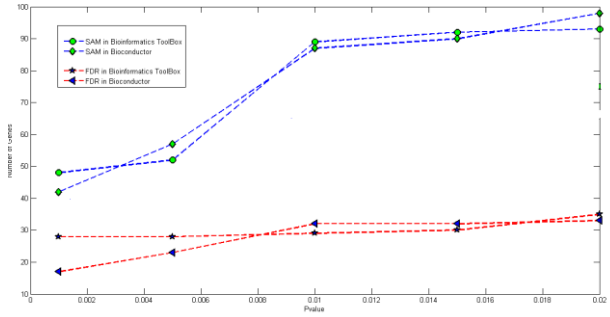


**Fig.1.**Number of genes selected according to used algorithms and tools (Latin Square)

In the same way, we use the true Detection Rate (TDR=Number of Spike Detected/Number of modulated Genes Reported) as an element of comparison. Result deduced in Figure 1 shows the variation of TDR according to the Pvalue Cutoff. To explain, the evolution of TDR with Pvalue CutOff in plotted figures demonstrates that the software plays a key role in the DEG selection. Thus, When the Pvalue increases from 0.001, the value of TDR decreases, meaning, that is preferable to usually choose a small Pvalue for DEG selection.

Another means to assess the ability of methods and tools in DEG selection is to use the hierarchical classification of selected genes and experiences. That is why we present the hierarchical classification of genes/experiences employing genes selected with Pvalues-Cutoff =0.001. This classification utilizes cancer data providing from [21] that aims to the classification of molecules.

All clusters regroup control cell in the same group. But the classification of conditions show certain change between Bioconductor and Bioinformatics ToolBox. This classification is a result of DEG selected from SAM in Matlab (Fig. 3-a), BH-T-test in Matlab (Fig. 3-b), SAM in Bioconductor (Fig. 3-c), BH-Ttest in Bioconductor (Fig. 3-d). These results confirmed the influence of the used soft on selected genes and furthermore the classification process.

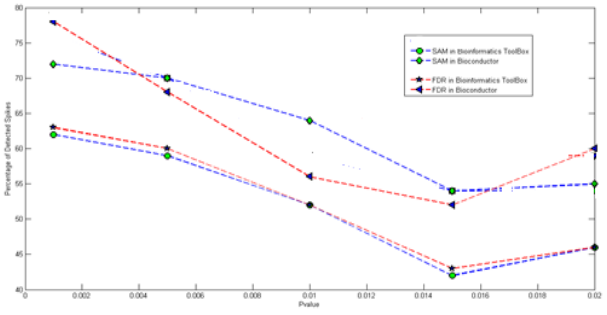


**Fig.2.** Percentage of detected spikes according to used algorithms and tools (Latin Square)

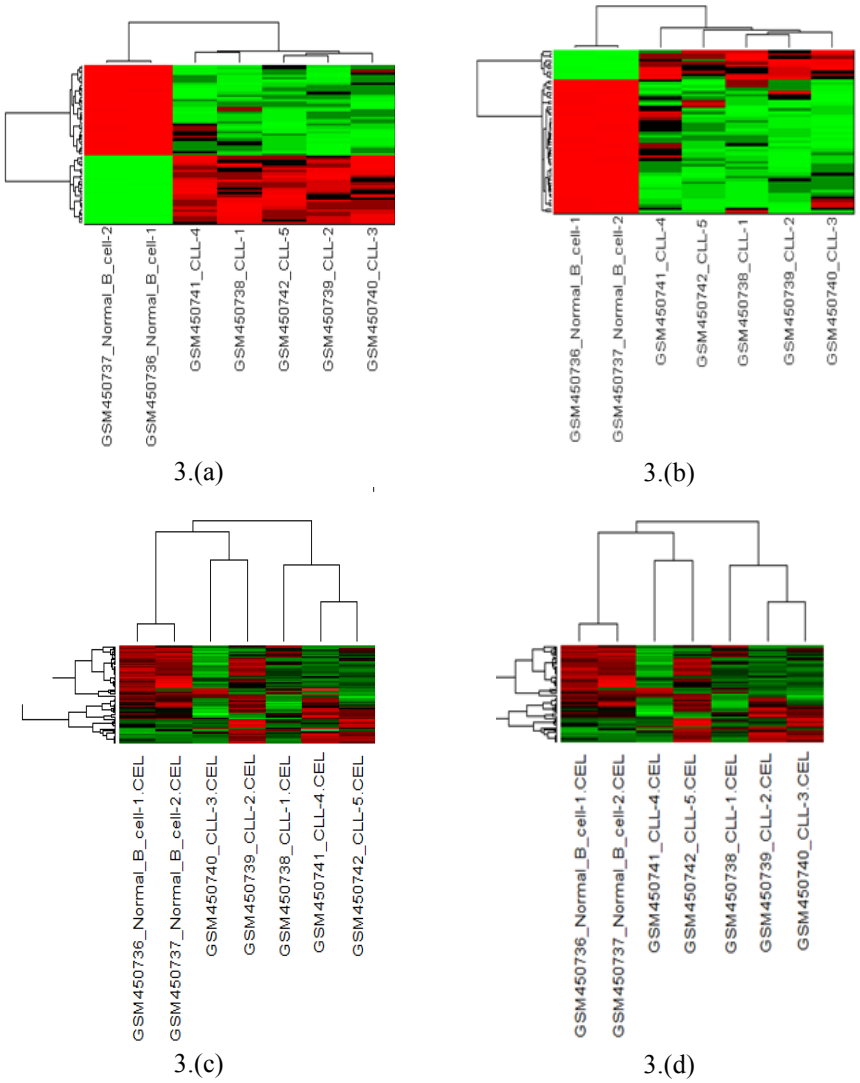


**Fig.3:** Hiearchical Biclustering of genes and condition od the second dataset

(a): Hiearchical Biclustering of genes selected by SAM in Mathlab
(b): Hiearchical Biclustering of genes selected by BH-Ttest in Mathlab
(c): Hiearchical Biclustering of genes selected by SAM in R/Bioconductor
(d): Hiearchical Biclustering of genes selected by BH-Ttest in R/Bioconductor

Finally, this comparative study shows that genes selected in microarray experiments' may depend both on methods and tools affecting thus the hierarchical biclustering of genes and conditions. In a word, we suggest the microarray data analyst to validate results by confirming the reproducibility of selected genes using various methods and tools.

## References


1. Gomases,V. S., Tagore and K.V. Kale :Microarray an approach for current Drug targets. Current Drug Metabolism, vol. 9 (2008) 221-31.
2. Leung Y. F., D. Cavalieri : Fundamentals of cDNA microarray data Analysis. Trends Genet. vol.19 (2003) 649-659.
3. Lockhar D. J t, H. Dong, M. C. Byrne, and al: Expression monitoring by hybridization to high-density oligonucleotide arrays.Nat.Biotechnol.vol.14 (1996) 1675-1680.
4. Duggan D. J., M. Bittner, Y. Chen,P. Meltzer and J. M. Trent :Expression profiling using cDNAMicroarrays. Nat.Genet. vol.21 (1999) 10-14.
5. Frouin V., and X.Gidrol : Analyse des données d'expression issues des puces à ADN. Biofutur. vol.252 (2005) 22 – 26.
6. Efron BJ : Gene association network . *Stat Assoc* (2004) **99**-96
7. Chu F. and L. Wang : Applications of support vector machines to cancer classification with microarray data. International Journal of Neural Systems, vol. 15 (2005) 475-484.
8. Faller D., H. U. Voss, J. Timmer and U. Hobohm: Normalization of DNA-microarray data by nonlinear correlation Maximization .J.Comput.Biol. vol.10 (2003) 751-762.
9. Tusher .V. G., R. Tibshirani, G. Chu : Significance analysis of microarrays applied to the ionizing radiation response.Proc. Nat. Acad. Sci.vol 98 (2001) 5116–5121.
10. Golub T. R., et al : Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science vol. 286 (1999) 531–537.
11. Model F., P. Adorjan, A. Olek, C. Piepenbrock : Feature selection for DNA methylation based cancer classification Bioinformatics vol. 17 (2001) 157–164.
12. Benjamini Y., Y. Hochberg : Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Roy. Stat. Soc. vol. B 57 (1995) 289–300.
13. Bolstad B. M., R. A. Irizarry,M. Astrand ,T. P. Speed: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.(2002)43-46.
14. Yuande T. Yin L: Comparison of methods for identifying differentially expressed genes across multiple conditions from microarray data, Vol 30 (2011) 21.
15. Susmita D, Somnath D : Comparisons and validation of statistical clustering techniques for microarray gene expression data. Vol 18 (2002) 20.
16. Lipschutz et al : Affymetrix Statistical algorithms reference guide, Technical report, Affymetrix (2001).
17. mLiu W., R. Mei, X. Di, T. Ryder, E. Hubbel, S. Dee, T. Webster, C.Harrington, M. Ho, J. Baid and S. Smeekens Analysis of High Density Expression Microarray with Signed-Rank Calls Algorithmes Bioinformatics, vol.12.(2002)1593-1599.
18. www.r-project.org
19. Bolstad B. M., R. Irizarry, M. Astrand, and T. Speed: A Comparison of normalization methods for high density oligonucleotide array data based on variance and bias'. Bioinformatics, vol.19 (2003) 185-193.
20. mLiu, W. R. Mei, X. Di, T. Ryder, E. Hubbel, S. Dee, T. Webster, C.Harrington, M. Ho, J. Baid and S. Smeekens Analysis of High Density Expression Microarray with Signed-Rank Calls Algorithmes. Bioinformatics, vol.12 ( 2002)1593-1599.
21. Sanjai S,.Alan L: Aberrant splicing of the E-cadherin transcript is a novel mechanism of gene silencing in chronic lymphocytic leukemia cells, Blood, (2009) 4179-4185.