WB-DEGS: Within and Between Group Comparisons for Differentially Expressed Gene Selection

Kaissi O, Nimpaye E, Mishra P, Singla M, Vannier B, Ibrahimi A, Singh TR, Ghacham A, Moussa A

**Abstract**—The

**Index Terms**—

—————————— ◆ ——————————

# 1 INTRODUCTION

s

———————————————

1. O.Kaissi. LTI Laboratory, ENSA, Adbelmalek Essaadi University, Tangier, Morocco. E-mail: kaissi.ouafae@gmail.com
2. E.Nimpaye. LabTIC Laboratory, ENSA, Adbelmalek Essaadi University, Tangier, Morocco. E-mail: [eric.nimpaye@gmail.com](mailto:eric.nimpaye@gmail.com)
3. *P. Mishra, Department of Biotechnology and Bioinformatics, Jaypee University of Information and Technology, Solan, H.P., India. E-mail : prashant24.juit@gmail.com*
4. *M.Singla, Department of Biotechnology and Bioinformatics, Jaypee University of Information and Technology, Solan, H.P., India. E-mail: mohitsingla420@gmail.com*
5. A.Ghacham. LTI Laboratory, ENSA, Adbelmalek Essaadi University, Tangier, Morocco. E-mail: [aghacham@uae.ac.ma](mailto:aghacham@uae.ac.ma)
6. B.Vannier. IPBC, University of Poitiers, France.E-mail: [brigitte.vannier@univ-poitiers.fr](mailto:brigitte.vannier@univ-poitiers.fr)
7. A.Ibrahimi. 5MedBiotech Lab, Rabat Medical Pharmacy School, University of Mohammed V Souissi, Rabat, Morocco. E-mail: [a.ibrahimi@um5s.net.ma](mailto:a.ibrahimi@um5s.net.ma)
8. T.R.Singh. Department of Biotechnology and Bioinformatics, Jaypee University of Information and Technology, Solan, H.P., India. E-mail: [tiratharaj@gmail.com](mailto:tiratharaj@gmail.com)
9. A.Moussa. LabTIC Laboratory, ENSA, Adbelmalek Essaadi University, Tangier, Morocco. E-mail: [amoussa@uae.ac.ma](mailto:amoussa@uae.ac.ma)

election of DEGs across multiple conditions is one of the major goals in many microarray experiments [1]. Since one cannot analyze the raw data, with thousands or more genes, a variety of testing procedures have been developed for microarray data analysis [2]. In the early days the method of Fold Change (FC) is used. This method is the most initiative approach to finding genes that differentially regulated. Typically, an arbitrary threshold is chosen and the difference is considered as significant if it larger than the threshold [3]. Another possible approach to gene selection is to use statistical tests, like T-test. This approach essentially uses the classical hypothesis testing methodology [4]. The analysis of variance (ANOVA) is a particular interesting approach to microarray data; the idea behind ANOVA is to build an explicit model of all sources of variance that affects the measurements and the use of the data to estimate the variance [5].

For microarray studies focusing on finding sets of predictive genes, a simple method proposed by Benjamini and Hochberg [6] computes the False Discovery Rate (FDR) that means the probability of a given gene identified as a false positive. The Significant Analysis of Microarray algorithm (SAM) is a statistical method for multiple testing that uses permutations to estimate the FDR [7]. Several other methods have been used for the DEGs selection and the related literature is abundant [8], [9], [10].

Despite the wealth of a literature concerning DEGS selection methods, the majority of these methods have been widely criticized for their high rate of false positive. The algorithm implemented here use conjointly statistics provided between conditions and between replicates to address this problem [11].

In fact, replicates can be used to measure experiment variations so that statistical tests can be applied to evaluate differences. The algorithm allows two types of comparisons: between group (BG) comparisons that concern chips providing from two samples, and within group (WG) comparison that concern chip inside biological or technical replicates. For each comparison we compute the statistics parameters for each gene, precisely the FC and P-values detection [12].

This paper proposes a first implementation of this algorithm and describes different steps of its application. It also shows the performance of our algorithm compared to several DEGs selection methods.

# 2 Methodology

In order to assess the performance of DEGs selection methods, we need a set of criteria able to qualify the outcome of the selection process. In our case, the performance of DEGs selection methods has been calculated in terms of specificity, sensitivity and accuracy. In a binary decision situation, like changed or unchanged, the results can always be divided into four categories: truly changed that are reported as changed (True Positives: TP), unchanged that are reported as changed (False Positive: FP), truly changed that are reported as changed (False Positive: FP) truly changed that are reported as such (True Negatives: TN). Based on these parameters, one can define the sensitivity, specificity and accuracy that qualify the reproductivity of DEGs methods and used tools. (N is the total number of instances N=TN+TP+FN+FP)

## 2.1 Selected Algorithm and Dataset

## 2.2 Usage

1. (fig. 1-e).

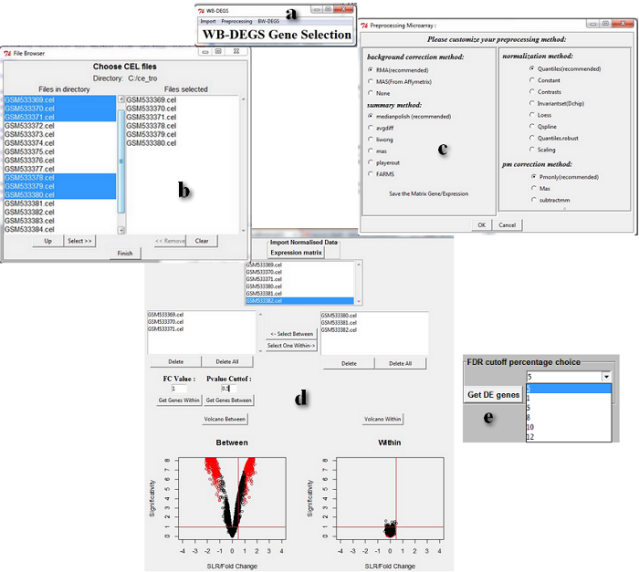


Fig. 1. Different stages of a WB-DEGS session

# 3 Results and Discussion

Fig. 2. Accuracy curve for the gene selection methods for the Platinum spike-in dataset. Fig.2 represents the accuracy curve using WB-DEGS for the Platinum spike-in dataset. This figure shows that WB-DEGS report differentially expressed genes in a stricter manner than other DEGs selection method used.

Fig. 3. A Roc curve for the gene selection methods for the Platinum spike-in dataset. Fig. 3 represents the ROC curve using WB-DEGS for the platinum spike-in dataset, this figure shows that the proposed algorithm is favored than six other methods used in the DEGs selection.

# Conclusion

**5 Availability and Requirements**

**6 Acknowledgments**

We acknowledge support from Volubilis (French-Moroccan Grant) to AI AM. This work was supported b y a grant from the NIH for H3Africa BioNet to AM & AI.

**7 REFERENCES**

1. R.Golub, et al, “Molecular Classification of Cancer,” *Science*, vol.286, no. 5439, pp. 531-537.1999.
2. O. Kaissi, “Comparative Analysis Between Soft and Tools in Microarray Data Processing” *Biomedical Conf*, pp. 791, 2013. (Conference proceedings)
3. Y. Chen, E. R. Doughert and M. L. Bittner, “Ratio-based Decisions and the Quantitative Analysis of cDNA Microarray Images, *J.Biomed*”, vol. 2, no. 4, pp. 364, 1997.
4. P.Baldi and AD. Long, “A Bayesian Framework for the Analysis of Microarray Expression Data: Regularized t -test and Statistical Inferences of Gene Changes,” *J. Bioinformatics*, vol. 17, no. 6, pp. 509, 2001.
5. F.Model et al, “Feature Selection for DNA Methylation based Cancer Classification,” *J.Bioinformaticas*, vol. 17, no. 1, pp. 157-164, 2001.
6. Y. Benjamini and Y. Hochberg, “Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing,” *Journal of the Roy.Stat.soc*, vol. 57, no. 1, pp. 289-300, 1995.
7. V. Tusher, and R. Tibshirani, “Significance Analysis of Microarray Applied to the Ionizing Radiation Response,” *J. Proc.Nat.Acad.Sci*, vol. 3, no. 2, pp. 5116-5121, 2001.
8. S. Scheid et al, “A Stochastic Downhill Search Algorithm for Estimating the Local False Discovery Rate,” *J. Computational Biology*, vol. 1, no. 3, pp. 98-108, 2004.
9. H. Fangxin , R. Breitling, W. McEntee , S Ben., L. J. Nemhauser and Joanne C, “RankProd: a BioConductor Package for Detecting Differentially Expressed Genes in Meta-analysis,”*J. Bioinformatiocs*, vol. 22, no. 22, pp. 2825-2827, 2006.
10. O. Kaissi, E. Nimpaye, TR. Singh, B. Vannier, A. Ibrahimi , A. Ghacham and A. Moussa, “Gene Selection Comparative Study in Microarray Data Analysis ,”*J. Bionformation*, vol. 9, no. 20, 2013.
11. A. Moussa , B. Vannier and Maouene M, “New Algorithm for Gene Selection in Microarray Data Analysis Workflow,” *J. Computer* *Technology & Application*, vol. 3, no. 2, pp.169-175, 2012
12. N. Jeanette, Mc. C.  Howard, J. Edenberg, “Effects of Filtering by Present Call on Analysis of Microarray Experiments,” J. *BMC Bioinformatics*, vol. 7, pp. 49, 2006.
13. I.B. Jeffery et al,” Comparison and Evaluation of Methods for Generating Differentially Expressed Gene Lists from Microarray Data,” *J. BMC Bioinformatics*, vol. 26, pp. 7:359, 2006.
14. B. Efron, R. Tibshirani, J.D. Storey, and V. Tusher, “Empirical Bayes Analysis of a Microarray Experiment,” *J. Am. Statistical Assoc.,* vol. 96, no. 456, pp. 1151-1160, 2001.
15. I. Sabaouni, A. Moussa, B. Vannier, O. Semlali, T. A. Pietka, N. A. Abumrad, A.Ibrahimi ,”[The Whole Genome Expression Analysis using Two Microarray Technologies to Identify Gene Networks That Mediate the Myocardial Phenotype of CD36 Deficiency](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3819569/),” *J. Bioinformation,* vol.9 ,no. 17,pp. 849, 2013.
16. M. N .McCall et al, “Consolidated Strategy for the Analysis of Microarray Spike-in Data,”*J. Nucleic Acids Research*, vol. 36, no. 17, pp. 108, 2008.
17. http://www.affymetrix.com/support/technical/sample\_data/datasets.affx
18. R. A.rizarry et al, “Summaries of Affymetrix GeneChip probe level data, ”J. Nucleic Acids Research, vol.15, no.3, pp. 31, 2003.
19. Q. Zhu, C. Jeffrey and S.M. Halfon “Preferred Analysis Methods for Affymetrix GeneChips Revealed by a Wholly Defined Control Dataset,”*J. Genome Biology*, vol. 6, no. 2, pp. 15, 2005.
20. R Core Team. “R: A language and environment for statistical computing.” R Foundation for Statistical Computing, Vienna, Austria, 2013. URL: http://www.R-project.org/