**COMPARISON OF THREE HIGH-LEVEL MICROARRAY STATISTICAL ANALYSIS METHODS FOR DISEASE MECHANISM IDENTIFICATION**

**D.R. Schultz1,2\*, I.S. Frydas 1,2, S. Karakitsios1,2,4, D.A. Sarigiannis1,2,3,4**

1 HERACLES Health and Exposome Research Centre, Centre for Interdisciplinary Research and Innovation, Aristotle University of Thessaloniki, Greece

2 Environmental Engineering Laboratory, Department of Chemical Engineering, Aristotle University of Thessaloniki, Greece

3 Environmental Health Engineering, School for Advanced Study IUSS, Pavia, Italy

4 ENVE.X, Thessaloniki, Greece https://enve-x.com/

*\* daynaraeschultz@gmail.com*

**ABSTRACT**

Transcriptomics is a powerful tool in disease discovery that has undergone large advancements in recent years; available technologies allow for the generation of datasets with tens of thousands of outputs. This requires large computational power and advanced statistical analysis for datasets. Unfortunately, there is debate over the ideal methods to process raw data, analyze and identify differentially expressed genes (DEGs), and generate robust results while limiting Type I and Type II errors.

We compared three statistical method pipelines using Linear Models for Microarray Data (LIMMA; R software), moderated t-test (Agilent GeneSpring™ software), and Significant Analysis for Microarray (SAM; R software) for DEG discovery and functionality comparisons. R is an open-source software that allows extensive control of statistical pipelines and parameters and the ability to handle complex statistical designs, although it requires basic knowledge of programming languages. GeneSpring™ is a commercially available, user-friendly analysis software that is efficient for basic statistical analyses with the benefit of freely available technical support.

We analyzed Agilent™ microarray data generated from three real datasets from experiments from within the lab group (One-Color, SurePrint Zebrafish Gene Expression v3 4 x 44k Microarray, design ID: 026437) that aim to detect the molecular mechanisms involved in metabolic disorders associated with environmental contaminants. 3-day post fertilization (dpf) zebrafish larvae (n = ~17 per replicate, 4 replicates) were exposed for 48 hours (sampled at 5-dpf) to two concentrations of the plasticizer bis(2-ethylhexyl) phthalate (DEHP; 25 nM and10 µM), positive control (amiodarone; 1 µM), and a carrier control [dimethyl sulfoxide (DMSO); 0.1%]. Raw data was exported and analyzed using the above methods.

Results found that there was high accordance between LIMMA- and SAM-detected DEGs, which suggests these two tests are better than MTT in detecting true positives, which is in accordance with previous reports (Chrominski and Tkacz 2015). However, combining multiple methods may improve reliability of results. Overall, this work is instrumental in future efforts to generate statistically robust and reliable computational models and systems biology for the prediction of a host of metabolic diseases.

**KEYWORDS**: Systems biology, statistical analysis, computational biology, data science

**REFERENCES**

[1] Bandyopadhyay, S., Mallik, S., & Mukhopadhyay, A. (2014). A Survey and Comparative Study of Statistical Tests for Identifying Differential Expression from Microarray Data. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 11(1), 20. <https://doi.org/10.1109/TCBB.2013.147>

[2] Dudoit, S., Shaffer, J. P., & Boldrick, J. C. (2003). Multiple Hypothesis Testing in Microarray Experiments. Https://Doi.Org/10.1214/Ss/1056397487, 18(1), 71–103. <https://doi.org/10.1214/SS/1056397487>

[3] Pandey A, Priyanka H. Tripathi, Ankita H. Tripathi, Satish Chandra Pandey, Saurabh Gangola. 2019. Chapter 2 - Omics technology to study bioremediation and respective enzymes. Academic Press, Pages 23-43, https://doi.org/10.1016/B978-0-12-818307-6.00002-0.

[4] Smyth GK (2004) "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments," Statistical Applications in Genetics and Molecular Biology: Vol. 3: Iss. 1, Article 3. (Pertains to the Moderated T-Test.)

[5] Tan, S., Begley, P., Mullard, G. et al. Introduction to metabolomics and its applications in ophthalmology. Eye 30, 773–783 (2016). https://doi.org/10.1038/eye.2016.37.