A PROPOSED METHOD TO BALANCE EFFECT SIZES AND STATISTICAL SIGNIFICANCE IN SMALL SAMPLE SIZE TRANSCRIPTOMIC EXPERIMENTS

D. R. Schultz[1,2], I.S. Frydas[1,2], T. Papageorgiou[1,2], S. Karakitsios[1,2], D. Sarigiannis[1,2.3,4*]

[1] Environmental Engineering Laboratory, Department of Chemical Engineering, Aristotle University of Thessaloniki, Greece
[2] HERACLES Research Center – CIRI, Aristotle University of Thessaloniki, Greece
[3] Environmental Health Engineering, School for Advanced Study IUSS, Pavia, Italy
[4] National Hellenic Research Foundation, Athens, Greece

(* sarigiannis@auth.gr)

ABSTRACT

In transcriptomic studies with limited sample sizes, balancing effect size and statistical significance is a crucial step that should be considered before downstream analysis is completed. Unfortunately, and owing to the high costs of many high throughput, omics experiments, small sample sizes are common although they can lead to reduced statistical power, making effect size assessment, often represented as log-fold change (logFC), essential to accurately identify biological outcomes. Even without statistical significance below the arbitrary 0.05 cut-off, a large effect size can hold meaningful insights into biologically significant mechanisms that underly somewhat elusive physiological outcomes. Moreover, p-values, used for significance evaluation, may be less reliable in small sample size datasets due to a lack of power and hence, the sole reliance on statistical significance can be misleading. In order to address the issues associated with these aforementioned scenarios, a scoring system was set up to evaluate and consider both statistical and effect size measurements in the weighting of important genes. To do so, a sample dataset generated from microarray analysis of different chemicals in different concentrations was analysed using the in-house R-based pipeline. More specifically, raw microarray scan results underwent quality control, background correction, and normalisation, followed by batch effect correction, using the Bioconductor package limma. In order to identify the differentially expressed genes (DEGs), statistical analysis was performed on weighted array values using a moderated t-test with a Benjamin Hochberg FDR multiple testing correction. Instead of applying the generic cutoffs, the scoring method used the absolute value of the logFC and multiplied it by the log10 of the adjusted p-value. In this way, genes that were borderline statistically significant but had very large effect sizes, as well as those with moderate logFC and highly statistical significance, were weighted highly. In the areas of biomarker discovery, this may be a more meaningful method to uncover important but nuanced trends in the data versus traditional cutoffs. However, transparent reporting is crucial and, as with any statistical analysis, warrants cautious interpretation and validation as well as awareness of false positives.

KEYWORDS: transcriptomics, statistical analysis, biological significance