Proceedings of the LXIV SIGA Annual Congress
Online, 14/16 September, 2021
ISBN: **978-88-944843-2-8**

Poster Communication Abstract – 7.35

IDENTIFICATION OF CANDIDATE GENES IN MEDICAGO THROUGH CORRELATION ANALYSIS OF MICROARRAY DATA AFTER CLEANING AND NORMALIZATION

MARZORATI F.*, WANG C.**, PAVESI G.***, MIZZI L.***, MORANDINI P.*

*) Dip. di Scienze e Politiche Ambientali, Università di Milano, Via Celoria 26, 20133 Milano, Italy
**) AEB (Shanghai) Trading Co., LTD., Shanghai 200241, China
***) Dip. di Bioscienze, Università di Milano, Via Celoria 26, 20133 Milano, Italy

microarray, correlation analysis, metabolic pathway, transcript, unwanted variation

We recently improved the quality of data of the *Medicago truncatula* Gene Expression atlas (MtGEA), a public microarray database, by discarding 15% of the data on the basis of purely statistical criteria (Marzorati et al., *Plants* 2021, 10, 1240). The reduced dataset shows great improvements in the consistency of the data, as measured through correlation analysis of transcripts and the enrichment of GO terms in the relative gene lists of top correlators. Several genes show good correlation in the original datasets but are actually poorly correlated after cleaning of the database and they are therefore identified as false positives. Examples of this behaviour are for instance, many transcription factors of the bHLH, WRKY, MYB, ERF and NAC gene families. On the contrary, some genes are identifiable as false negatives, as they substantially improve their correlation value or change the spectrum of correlators upon cleaning of the database. We present evidence for both types of improvements (a reduction in both false positives and false negatives) for several genes and provide an explanation for the origin of the noise. Using the improved dataset, new gene candidates for specific processes can be proposed, such as for instance, transcription factor bHLH 74 (Medtr8g065740 corresponding to the Affymetrix probe Mtr.34810.1.S1_at), which is predicted to have a role in chromosome maintenance/stability and DNA repair, because half of the best 20 correlators in the analysis belong to this category. We also identified several new genes potentially involved in the saponin pathway. Several predictions were confirmed by analysing an independent set of microarray data after renormalization, strengthening the validity of the approach. These improved datasets thus allow the identification of strong candidate genes for wet experimental approaches.