The effect of genetic diversity in differential gene expression analyses using RNA-Seq data

Victor Mello, Ana Letycia Basso Garcia, Fernando Henrique Correr, Guilherme Kenichi Hosaka, Amanda Ghelfi Dumit, Gabriel Rodrigues Alves Margarido

ESALQ - USP

Abstract

Differential gene expression studies focus on discovering which genes are more highly expressed in a certain condition, tissue or group of organisms. Statistical tests for differential expression using RNA-Seq data often rely on estimates of both the average expression levels and dispersion (variance) of transcript abundances for the contrasting groups. This is done by sequencing biological replicates subjected to the same experimental conditions. Researchers commonly use clones to compose these groups in many cases where it is convenient, such as in vegetatively propagated plants, in order to obtain the highest possible uniformity among the replicates. However, this approach restrains the generalization of results to only a few genotypes, and those may not be valid in a broader sense. In this work, we compare the outcomes of differential gene expression analysis when using a strategy based on clones (SBC) or on diverse genotypes (SBDG) of sugarcane as biological replicates. The samples consisted of sugarcane top internodes grouped by the soluble solids level (Brix), namely Very Low, Low, High and Very High Brix. Within each group there were three biological replicates, which included clones of the same genotype in one set and three different genotypes in the other. The 24 samples, 12 from each set, were utilized to perform a de novo transcriptome assembly, resulting in 262, 281 putative genes. We found that the common biological coefficient of variation with the SBDG was about twice higher than with the SBC. The total number of differentially expressed genes (DEG) was equal to 28, 699 and 10, 380 in the clone and diverse approaches, respectively, taking into account up and downregulated genes. A functional enrichment analysis showed more Gene Ontology (GO) enriched terms directly related to the contrasting phenotypic traits for the "Very Low Brix against others" contrast (FDR < 0.01) using the SBDG. On the other hand, for the other proposed contrasts, the SBC resulted in many more statistically significant enriched terms, although none of them was related to sugar yield or carbon partitioning. The obtained results suggest that using clones as biological replicates minimize the variance of transcript expression levels, resulting in a higher statistical power to detect DEG and GO enrichment, but both might be not representative of the phenomena of interest. These conclusions highlight the existence of bias depending on the sample and replicate choices, which should ideally present a balance between the expected statistical power and the biological meaning of the results.

Funding: FAPESP