Assembly, identification and characterisation of sugarcane transcripts

Rezende P. M., Ribeiro T. H. C., Schumacher P. V., Lima A. A., Chalfun-Junior A.

Laboratory of Plant Molecular Physiology, Federal University of Lavras

Sugarcane (*Saccharum officinarum L.*) stands out as an important crop due to its role in sugar and ethanol production, which are widely consumed worldwide. Brazil is the world leader in sugar production from sugarcane, being responsible for 38.7% (1.9 billion tons) of the world production in 2014. Unlike other important crop species, such as Maize (*Zea mays L.*), Sorghum (*Sorghum bicolor L. Moench*), and Rice (*Oryza sativa*), sugarcane haven't had its genome sequenced so far, what limits the understanding of important biological processes from the molecular point of view. Among these important processes is flowering, which should be avoided in sugarcane since it requires a great deal of energy, reducing the sugar content of the sugarcane stalks. Thus, a better understanding of the molecular mechanisms regulating sugarcane flowering induction is crucial for the reduction of flowering intensity in sugarcane and the development new cultivars less susceptible to flowering. This study aimed to perform an assembly strategy, the identification, and the characterisation of sugarcane transcripts available in public databases. Transcript assembling was carried out using the software Trinity, using the SUCEST (Sugarcane Expressed Sequence Tag) database and sugarcane reads generated from a high-throughput sequencing platform as input data. After the identification of the reads, the coding regions and the protein sequences of the candidates genes were predicted using the software TransDecoder. The predicted proteins were aligned against the protein database SWISSPROT. The alignments, a conserved domain analysis, and the construction of phylogenetic tress using flowering gene sequences from other species, allowed the identification of two important sugarcane flowering genes, FD and FT (*Flowering Locus T)*. 151389 genes and 170756 transcripts were assembled, and the N50 value for these transcripts was 929. The alignment against the SWISSPROT database showed that 137126 transcripts displayed an e-value below 10, what corresponds to 97 % of the transcripts identified. From these 137126 transcripts, 57% showed and alignment coverage higher than 80%. The conserved domain analysis indicated the presence of 14 sugarcane sequences that possess the conserved bZIP (*basic Leucine Zipper*), present in the FD gene from other species, and one of these sequences was shown to be a putative sugarcane FD based on the phylogenetic analysis. On the other hand, for the FT gene, 11 sugarcane sequences were found to possess the PEBP (*Phosphatidyl Ethanolamine-Binding Protein*) domain, which characterises FT genes, and the phylogenic analysis showed six of them were putative FT genes, although four of the sequences were found to be partial sequences. Thus, these results show that the approach used in this study can be used to identify putative genes related to important biological process of sugarcane.