Journal of Biotechnology and Biomaterials

ExtendedAbstract Open Access

De novo Exocarp Transcriptome Assembly of the Organ Pipe Cactus *Stenocereusthurberi* Fruit: A First Glance

García-Coronado, Heriberto; Tafolla-Arellano, Julio-César; Hernández-Oñate, Miguel-Angel and Tiznado-Hernández, Martín-Ernesto

[1] Coordinación de Tecnología de Alimentos de Origen Vegetal. Centro de Investigación en Alimentación y Desarrollo, A.C. Carretera Gustavo Enrique Astiazarán Rosas, México. Email: tiznado@ciad.mx

Abstract

Mexico is one of the most important producers and exporters of fruits and vegetables of the world. This activity brings important revenues for the country. In this context, the postharvest fruit losses reduce the economic benefit of the country. These losses are in part due toseveral abiotic stresses present during the postharvest shelf life. Several experimental evidences strongly suggest that the cuticle plays a role in the length of the postharvest shelf life of fruits. Nevertheless, the cuticle responses to abiotic stresses such as hydric stress, high temperatures and osmotic stress is still largely unknown. Because of that, the elucidation of the molecular mechanism of cuticle biosynthesis will allow the development of strategies to design fruit with a large postharvest shelf life and more resistant to abiotic stresses. Cactus species are adapted to environments with high temperatures, high solar irradiation, and low water availability which suggests that most likely cactus species cuticle have special characteristics. According to the above, the objective of the present work was to analyze the exocarp transcriptome of organ pipe cactus *Stenocereusthurberi* fruit generated by a next generation sequencing approach.*S. thurberi* fruits are known as pitayas. Pitaya fruit with different stages of development were collected in a Sonora desert region located close to Carbo, Sonora, México. Total RNA was isolated from the exocarp and four libraries were sequenced in paired-end mode2x150 using theNextSeq 500 Iluminaplatform. A total of288,199,704 short reads were obtained, which correspond to 21.95 Gbof information. Fast QC software was used for quality analysis. Short reads with poor quality were trimmed and/or eliminated with Trimmomatic with a trailing and leading of 25, a slidingwindow of 4:25 and a minimum length of read of 80 bp. After that, 243,194,888 reads with at least 25 quality score in the Phredscale were used to carry out the *de novo* assembly by Trinity software. For the assembly, four strategies were tested: 50 percent of normalization of reads with a minimalkmer coverage of 1 (Assembly1), 50 percent of normalization of reads with a minimal kmer coverage of 5 (Assembly2), without normalization of reads with a minimal kmer coverage of 1 (Assembly3) and without normalization of reads with a minimal kmer coverage of 5 (Assembly4). Assemblies were tested through TransRate software to know the percentage of transcripts with a good number of short reads alignment, classified as "good contigs". The transcriptome completeness was determined with BUSCO using the Embryophyte database. Removal of DNA contaminating sequences from other species and rRNA was carried out through SeqClean. Assemblies redundancy showing more than 90% of identity were eliminated by using CDHits software. Besides, to identify the assembly that contain more hits with homologous sequences in publics data bases, an alignment was carried out to SwissProt database through BLASTx considering a minimal e-value of $1x10^{-5}$.Based in the statistics of the different assembled transcript collection, Assembly1 and Assembly2 were selected to continue with the analysis.After removal of the redundancy and the contaminating sequences, Assembly1 was found to have 186,687 transcripts with a N50 value of 2,103, mean length of transcripts of 1,155 bp and 86% of "good contigs", whileAssembly2 showed 113,808 transcripts with a N50 value of 2001, mean length of transcripts of 1,199 bp and 82% of "good contigs". Assembly1 showed 85.3% of complete transcripts, including 48.1% of single and 37.2% of duplicated, 10.8% fragmented and 3.9% missing, while Assembly2 showed 89% of complete transcripts, including 60.8% single and 28.2% duplicated, 6.9% fragmented and 4.1% missing.According to BLASTx alignment, 72,683 and 49,817 transcripts from Assembly1 and Assembly2, respectively, have homologous sequences in SwissProt database. The coverage distribution of these transcripts where similar between both assemblies.Nevertheless, Assembly1 showed 3,262 more hits corresponding to homologous sequences in SwissProt data base than Assembly2. Based on the data mentioned, Assembly1 was selected to further analysis.After the elimination of expression transcripts with less than 0.01 transcripts per million of reads (TPM),Assembly1 transcriptome was reduced from186,687 to 174,449 transcripts. This final version of the transcriptome showed a N50 value of 2,110 with a contig size average of 1,198.69. BUSCO results showed that 85.4% of the orthologous genes are complete. Out of these, 48.2% are single and 37.2% duplicated, 10.7% are fragmented and 3.9% are missing. Because of the lack of knowledge about *Stenocereusthurberi*specie, currently we are looking for genes to be utilized as reference genes to carry out the gene expression analysis by using quantitative real time polymerase chain reaction to validate the results generated *in silico*. Additionally, functional annotation will be performed with the transcriptome generated and genes related to cuticle biosynthesis will be identified. This information will help in the elucidation of the molecular mechanism and regulation of cuticle biosynthesis of the specie *S.thurberi*.