Gene Assembly, Prediction and Phylogenomic Analysis of Erianthus arundinaceus, a crop for biomass production

Nicholas Vinicius Silva[1], Luciana Souto Mofatto[1], Juliana José[1], Gonçalo Amarante Guimarães Pereira[2], Marcelo Falsarella Carazzolle[3],

1 UNICAMP
2 Brazilian Bioethanol Science and Technology Laboratory, Brazilian Center for Research in Energy and Materials, Biology Institute – UNICAMP
3 Biology Institute - UNICAMP, National Center for High Performance Computing

Abstract

Erianthus arundinaceus is a wild perennial C4 grass, considered closely related to Saccharum. It has a good perennial ratooning ability, excellent vigor, high fiber and low sugar content, waterlogging and diseases resistance. Due to its high biomass production and strong tolerance to environmental stresses, it is regarded as one of the most promising crops for biomass production and source of desirable traits genes for breeding programs in sugarcane. The aim of this research was providing genomic information and phylogenomic analysis of Erianthus arundinaceus, in order to understand the evolutionary relationships among this species and other crops. Genomic sequences were obtained through Illumina MiSeq platform, generating 93 million of paired-end reads. The sequences were assembled using "The Polyploid Gene Assembler (PGA) Pipeline" with reference-based genomes of Sorghum bicolor, Zea mays, Setaria italica and Panicum virgatum, resulting in 15.596, 3.610, 2.532, 1.788 assembled sequences respectively. The remaining unmapped reads were De novo assembled using TRINITY, resulting in 389.960 sequences (N50=1923bp, Larger sequence of 15.566b. All sequences were used as reference for RNA-seq reads mapping using STAR, for the gene prediction analysis. Intron-exon junctions, provided by RNA-seq mapping, were used as hints for gene prediction in Genemark, resulting in 13.053 putative genes. These genes were filtered to find the most reliable, according to Blastp similarities with proteins from related genus. The 3.016 final reliable genes were used as training and test groups in AUGUSTUS, and gene predictions were performed with and without hints. The phylogenomic analysis among E. arundinaceus and five publicly available grasses genomes with reliable protein predictions (S. bicolor, S. bicolor "Rio", Z. mays, S. italica and B. distachyon) used the concatenated protein alignments of 472 single copy-ortholog genes identified with OrthoFinder. Proteins were globally aligned using T-COFEE, and submitted to Maximum Likelihood (ML) and Bayesian Inference (BI) phylogenetic analysis. We provide the first phylogeny with a genome dataset for E. arundinaceu, with strong branch supports and evidences that this crop is closely related to S. bicolor than previously inferred in literature. The new informations we present are central for further genome investigations and gene prospection from E. arundinaceus, and also for the development of new techniques for the improvement of sugarcane breeding in the production of biofuels and bioproducts.

Funding: CNPq