Ploidy level analysis of functional SNPs from GBS data in a sugarcane map population

Estela Araujo Costa[1], Alexandre Hild Aono[1], Hugo Rody Vianna Silva[1], James Shiniti Nagai[1], Anete Pereira de Souza[2], Reginaldo Massanobu Kuroshu[1],

1 Universidade Federal de São Paulo
2 Universidade Estadual de Campinas

Abstract

Sugarcane is one of the most important energy source in Brazil. Because of the complexity in determining ploidy levels, studies with molecular markers have been limited by the lack of information so far. Genotyping-by-sequencing (GBS) made possible deep genetic analysis, making accessible studies with molecular markers using ploidy level from genomic data of complex polyploids. Herein, a pipeline for identifying SNPs from GBS data in poliploidy genomes was established and we used the SuperMASSA software to estimate the ploidy level of some functional SNPs. In total 831 million 100-bp single end reads were generated from GBS of 182 individuals from a sugarcane map population. After demultiplexing and barcode processing, BWA version 0.7.15 was used to align reads against sugarcane methyl-filtered (MF) genome sequence. From a total of 96% of reads aligned, we selected alignments found in 10,957 MF Coding-DNA sequences (CDSs). Variants were called using GATK and Samtools, resulting in 8,345 SNPs in the intersection of both callers. All consensus loci were compared to Sorghum bicolor, Oryza sativa and Zea mays CDSs genomes using BLASTn. An enrichment analysis was performed where 538 SNPs in 94 MF scaffolds, which correspond to some important GO categories involved in sugar transport and metabolism in the storage tissue. Those 94 scaffolds were used as input to KAAS server, resulting in 65 scaffolds with 352 SNPs representative in 64 KEGG orthology (KO) terms. During KEGG analyses, some important pathways were represented, such as "Glycolysis / Gluconeogenesis" and "Amino Sugar and Nucleotide Sugar Metabolism", with four and five SNPs respectively. In order to genotype the population, the SuperMASSA software was used to infer on the ploidy level in a range from 2 to 20, and a posterior probability was associated. In total, 510 putative SNPs were able to estimate the ploidy level using SuperMassa. The results showed a large number of ploidy level at 20 (35%) in the category GO:1901576 'organic substance biosynthetic process'. The most representative ploidy level was 20 (50%). These results suggest a high level of ploidy estimation. Possibly, because those called SNPs are in genes involved to important energetic process such as the sugar metabolism, they might have multiple copies in the sugarcane genome, what makes the estimative performed by the SuperMassa software biased towards high ploidy levels. By unveiling many putative functional SNPs, the pipeline we established brought positive results to help the understanding of the complex ploidy levels of sugarcane genome.

Funding: Capes