Preliminary analysis of functional SNPs mining from GBS data using GATK in a sugarcane map population

Alexandre H. Aono[1], Estela A. Costa[1], Hugo V. S. Rody[1], James S. Nagai[1], Anete P. de Souza[2,3], Reginaldo M. Kuroshu[1]

[1]Instituto de Ciência e Tecnologia, Universidade Federal de São Paulo, São José dos Campos, SP - Brazil, [2] Molecular Biology and Genetic Engineering Center (CBMEG) – University of Campinas (UNICAMP), Campinas, SP – Brazil, [3] Vegetal Biology Department, Institute of Biology, University of Campinas (UNICAMP), Campinas - SP, Brazil.

Sugarcane is the source of sugar in all tropical and subtropical countries and it is becoming increasingly important for bio-based fuels. However, its large (10 Gb), polyploid, complex genome has hindered genome based breeding efforts. Currently, genotyping-by-sequencing (GBS) has been the most economical approach for generating population genomic data without the need of a reference genome. Here, GBS was carried out in 182 full-sibs derived from a sugarcane commercial cross (IACSP96-3018 x IACSP93-3046) in order to establish a pipeline to identify SNPs in polyploidy species and generate informative molecular markers. After a sequencing using the platform Illumina GAIIX (1x120bp), we processed the data following the GATK pipeline, modifying it and creating in-house scripts to handle polyploidy genomes. As a first step in the pre-processing phase of the analysis, we used FASTX-Toolkit for demultiplexing and barcode processing. A comparative alignment was performed using BWA-MEM algorithm against three different references: sorghum genome (1), sugarcane RNA-seq data (2) and sugarcane methyl-filtered genome (3). Picard tools were used to mark alignment duplicates and SAMtools for controlling the process. Genotype calls were first made in gVCF format for each sample using HaplotypeCaller with stringent parameters for phred-scaled confidence threshold and ploidy level as 12. Then, all samples were joined into a VCF file as implemented in GATK 3.6 pipeline. As a result, from ~174 million reads generated, it was obtained: 94% of correspondence in (1), 33% in (2) and 41% in (3). In order to identify functional SNPs, we selected a set of aligned contigs to (2). These contigs were previously classified as part of two representative pathways: Carbon Fixation in Photosynthetic (C4 photosynthetic pathway) and the Starch and Sucrose Metabolism. From 53 selected contigs, we obtained 129 putative SNPs in 28 contigs. In C4 photosynthetic pathway, 91 putative SNPs were found and 38 in the Starch and Sucrose pathway. With these preliminary analyses we identified SNPs that can be used as candidates for the development of functional specific markers and started a process to establish a pipeline for searching SNPs in polyploidy and aneuploidy species as sugarcane. As future issues, we expect to find more SNPs using the other references for sugarcane, by mapping more and different regions; including non-coding regions.