GBS data from a population of modern sugarcane variety reveals duplicate genes retained by breeding process

Hugo V. S. Rody[1], James S. Nagai[1], Estela A. Costa[1], Alexandre H. Aono[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 most important crop for sugar and biofuel production. Despite all economic interest, sugarcane breeding is challenging due to its overcomplex genetics, with cultivars varying in chromosome number from 80 to 130. New sequencing methods such as genotyping-by-sequencing (GBS) have accelerated studies using genomic data from populations of non-model organisms. However, because short reads are likely to map with equal probability in multiple positions, duplicate genes have been typically filtered from GBS data. We used a pipeline to expose duplicate genes in GBS data from a population of modern sugarcane variety from the Sugarcane Breeding Program at IAC/Apta, obtained using IACSP96-3046 and IACSP95-3018 as parents. Additionally, we investigated which duplicate gene categories are consistently overrepresented across the population. After GBS raw reads manipulation, final high quality reads, with minimum of 80% of Q > 20 and 85bp long, were used for De novo assembly; performed by Stacks v.1.42. To filter young duplicate genes that typically are merged as a single locus in GBS data, we allowed maximum of two nucleotides mismatches among reads to form a putative locus. Old duplicates are expected to have accumulated enough mutations to form a new locus. Using BLASTn, all consensus loci were compared to the Sorghum Coding-DNA sequence (CDS) genome, with a cutoff e-20 and minimum alignment length equal to 60. With BLASTn result, four subsets were created, grouping the genes of sorghum by the number of individuals in the sugarcane population that harbored at least one consensus locus showing similarity to respective sorghum gene. In Subset1, sorghum genes were present in a unique individual across the sugarcane population. Subset2 was formed by sorghum genes that occurred from 2 to 50 individuals, Subset3 from 51 to 100 individuals, and Subset4 by genes that occurred above 100 individuals. Gene Ontology (GO) enrichment analysis, based on sorghum annotation, was carried out for each subset. Different subsets had different GO categories overrepresented. Subset1 is enriched by gene categories whose products likely stand alone in metabolic pathways, such as "cellular response to stress". Whereas in Subset4, most of genes overrepresented are connected genes, such as those involved with signaling. Further, essential genes for carbon fixation in C4 organisms were overrepresented in Subset4. We showed overrepresented duplicate gene categories highly and lowly distributed across the sugarcane population, suggesting how breeding process has influenced duplicate gene retention that formed the characteristics of modern sugarcane.