Preliminary association of putative GBS-based SNPs with brown rust resistance phenotypes in a sugarcane map population using GWAS and machine learning methods

Alexandre Hild Aono[1], James Shiniti Nagai[1], Estela Araujo Costa[1], Hugo Rody Vianna Silva[1], Fernanda Raquel Camilo dos Santos[2], Luciana Rossini Pinto[2], Anete Pereira de Souza[3], Reginaldo Massanobu Kuroshu[1],

1 Universidade Federal de São Paulo
2 Instituto Agronômico de Campinas
3 Universidade Estadual de Campinas

Abstract

Brazil is the world's largest sugarcane producer and its production is an important source of income, arising mostly from sugar and ethanol. Currently, an approach for generating population data for this complex polyploidy species is genotyping-by-sequencing (GBS), which allows the detection of genomic variants without a reference genome. Several diseases limit sugarcane yield and the brown rust is considered one of the most important fungal diseases of sugarcane. Herein, putative GBS-based SNPs in a sugarcane map population were associated with brown-rust phenotype differences, being genetically linked to its causative factor. A set of 182 full-sibs derived from a sugarcane commercial cross between IACSP96-3018 and IACSP-3046 were generated and genotyped by GBS. Performing a comparative alignment (BWA version 0.7.15) of 831 million reads against methyl-filtered (MF) genome, we selected the correspondences with MF Coding-DNA sequences (CDSs). The data was pre-processed using Genome Analysis Toolkit (GATK) pipeline and the variants were called using HaplotypeCaller, implemented in GATK v3.7, and SAMtools v 1.4. After a series of stringent filters, we obtained a total of 8,345 SNPs and 290 indels, identified by both callers. In order to detect putative markers associated with brown rust resistance, a Genome Wide Association Study (GWAS) and machine learning approaches for feature selection were performed in the identified variants. Using an evaluation of the severity of brown rust, we classified the cultivars in two categories: most resistant and most susceptible. As a first step in the GWAS, the principal component analysis (PCA) was performed followed by the calculation of squared Euclidean distances between isolates and hierarchical clustering with complete linkage. The correction of population structure was based on the Discriminant Analysis of Principal Component (DAPC) implemented in the R adegenet package 2.0.0. The method used for association was the multivariate DAPC-based approach. In addition, three machine learning methods of feature selection were applied to the same data, improving the results by focusing on predictivity. The mutual information (MI) was obtained to measure the dependency between the loci and the phenotype. A model of Logistic Regression (LR) was built to select the loci related to the coefficients with high levels of influence under the model and Support Vector Machine (SVM) was used as a nonlinear classifier, where the attributes with most importance under the model were retained. Using the results of the four different approaches, a

total of 241 potential variants were identified in 195 different MF scaffolds. All the consensus loci of these scaffolds were compared to the Sorghum bicolor, Oryza sativa and Zea mays CDSs genomes and a Gene Ontology enrichment analysis was performed. The results show some enriched categories involved in functions of known genes associated with rust resistance (e.g. 'protein phosphorylation', 'protein kinase activity' and 'ATP binding'). With these preliminary analyses we identified a set of putative SNPs that can be used as candidates for the development of functional specific markers in brown rust resistance.

Funding: CAPES