Genome-wide approaches for the identification of markers and genes associated with sugarcane yellow leaf virus resistance

Ricardo José Gonzaga Pimenta[1], Alexandre Hild Aono[1], Roberto Carlos Burbano Villavicencio[2], Alisson Esdras Coutinho[3], Carla Cristina da Silva[1], Ivan Antônio dos Anjos[4], Dilermando Perecin[3], Marcos Guimarães de Andrade Landell[4], Marcos Cesar Gonçalves[5], Luciana Rossini Pinto[4], Anete Pereira de Souza[1,6*]

[1]Centre of Molecular Biology and Genetic Engineering (CBMEG), University of Campinas (UNICAMP), Campinas, Brazil

[2]Gustavo Galindo Velasco Campus, Littoral Polytechnic Superior School (ESPOL), Guayaquil, Ecuador

[3]São Paulo State University (UNESP), Jaboticabal, Brazil

[4]Advanced Centre for Technological Research in Sugarcane Agribusiness, Agronomic Institute of Campinas (IAC/APTA), Ribeirão Preto, Brazil

[5]Plant Protection Research Centre, Biological Institute (IB/APTA), São Paulo, Brazil

[6]Department of Plant Biology, Institute of Biology, University of Campinas (UNICAMP), Campinas, Brazil

*** Correspondence**
Anete Pereira de Souza
anete@unicamp.br

Keywords: *Saccharum*, SCYLV, RT-qPCR, AFLP, SSR, GBS, GWAS, Machine learning.

Abstract

The breeding of sugarcane, a leading sugar and energy crop, is particularly hampered by the extremely complex sugarcane genome. One of the main diseases that affect this crop is sugarcane yellow leaf (SCYL), caused by the sugarcane yellow leaf virus (SCYLV). Regardless of the manifestation of SCYL symptoms, infection by SCYLV can lead to substantial yield losses, making resistance to this virus highly relevant to sugarcane breeding. However, the genetic basis of this trait has not been unraveled or widely explored. In this context, several methodologies have arisen as promising tools for the identification of molecular markers associated with SCYLV resistance, which are crucial for marker-assisted selection and can contribute to the understanding of the mechanisms involved in the response to this virus. Here, we investigated the genetic basis of SCYLV resistance in sugarcane using dominant and codominant markers and genotypes of interest for breeding. A sugarcane panel inoculated with SCYLV was analyzed for SCYL symptom severity, and viral titer was estimated by reverse transcription quantitative PCR. This panel was genotyped with 662 dominant markers, as well as by genotyping-by-sequencing, which enabled the identification of 70,888 SNPs and indels with information on allele proportion and position on the *Saccharum spontaneum* genome. We then performed polyploid-adapted genome-wide association analyses and used machine-learning algorithms coupled with feature selection methods to establish marker-trait associations. While each approach identified a unique set of markers associated with the phenotypes, we also observed convergences between them, demonstrating their complementarity. Finally, by aligning the sequences neighboring these markers to the genomic coding sequences of closely related

plant species, we annotated genes harboring 176 of these markers. With this procedure, we identified genes encoding emblematic participants in virus resistance mechanisms, as well as other candidates whose involvement in the responses to viruses has not been reported before. Our approach has the potential to accelerate sugarcane breeding targeting SCYLV resistance and is a powerful tool for studies on the biological processes leading to this trait.

1 Introduction

Sugarcane is one of the world's most important crops, ranking first in production quantity and sixth in net production value in 2016 (FAO, 2020). It is by far the most relevant sugar crop, accounting for approximately 80% of the world's sugar production (FAO, 2020; ISO, 2020) and is also a prominent energy crop. It is estimated that by 2022, sugarcane will account for 40% of the world's first-generation biofuels (Lalman et al., 2016), and biomass-rich energy cane is a promising source of second-generation ethanol (Carvalho-Netto et al., 2014). However, this crop has an extremely complex genome; modern cultivars are the product of a few crosses between two autopolyploid species conducted at the end of the 19th century (Bremer, 1961). *Saccharum spontaneum* ($2n$ = 5x = 40 to 16x = 128; x = 8) (Panje and Babu, 1960), a wild stress-resistant but low-sugar species, was hybridized with *Saccharum officinarum* ($2n$ = 8x = 80, x = 10) (D'Hont et al., 1998), which has a high sugar content but is sensitive to drought and susceptible to diseases. This was followed by several backcrosses with *S. officinarum* to restore sugar productivity in a process known as nobilization (Roach, 1972). These procedures gave origin to plants with very large (ca. 10 Gb) (D'Hont et al., 1998), highly polyploid (D'Hont and Glaszmann, 2001), aneuploid (Sforça et al., 2019) and remarkably repetitive (Mancini et al., 2018) genomes. This complexity directly affects sugarcane research and breeding; for instance, until recently, it prevented the use of codominance information in marker-assisted breeding strategies for this crop, leading to loss of information during the performance of these approaches and therefore limiting their use (Premachandran et al., 2011; Bourke et al., 2018).

One of the diseases that affect this crop is sugarcane yellow leaf (SCYL), which is caused by sugarcane yellow leaf virus (SCYLV), a positive-sense ssRNA belonging to the *Polerovirus* genus (Moonan et al., 2000; Smith et al., 2000). The expression of SCYL symptoms is complex and usually occurs in late stages of plant development. The main symptom of these infections is the intense yellowing of midribs in the abaxial surface of leaves, and other visible manifestations are necrosis of leaf tips, leaf chlorosis and plant stunting (Scagliusi and Lockhart, 2000; ElSayed et al., 2015). The virus is known to induce alterations in the metabolism and transport of sucrose and in photosynthetic efficiency (Gonçalves et al., 2005; Lehrer et al., 2010); this impairs plant development and eventually reflects losses in sugarcane productivity (Vega et al., 1997; Grisham et al., 2001; Vasconcelos et al., 2009; Viswanathan et al., 2014). Many symptoms of SCYL may, however, be caused by other biotic and abiotic stresses or plant senescence (Vega et al., 1997; Aljanabi et al., 2001; ElSayed et al., 2015), making SCYL identification troublesome. Therefore, the diagnosis of infection by molecular methods is of great assistance. Immunological assays were the first methods developed to detect and quantify this virus (Scagliusi and Lockhart., 2000), but later assessments by reverse transcription followed by quantitative polymerase chain reaction (RT-qPCR) are much more sensitive and precise in estimating the levels of viral infection (Gonçalves et al., 2002; Korimbocus et al., 2002; Zhu et al., 2010).

Due to SCYL's elusive symptomatology, the spread of its causal agent is silent; SCYLV is considered endemic in all sugarcane-producing regions of the world (Gonçalves, 2007). The virus is disseminated mostly during sugarcane vegetative propagation by setts but is also transmitted by

aphids, mainly the white sugarcane aphid *Melanaphis sacchari* (Zehntner, 1897) (Scagliusi and Lockhart, 2000). Unlike other pathogens, SCYLV is not efficiently eliminated by thermal treatments (Delage et al., 1999); the only way to thoroughly remove the virus from plants is by meristem micropropagation (Chatenet et al., 2001; Fitch et al., 2001), which is time-consuming and requires specialized infrastructure and personnel. These features make varietal resistance to SCYLV the resource that can be most efficiently exploited to prevent damage and losses caused by this virus. Sugarcane resistance to SCYLV has been explored in breeding programs, but research on SCYL genetics is not exempt from the difficulties generated by the complexity of this crop's genome (Gonçalves et al., 2012).

The identification of loci linked to resistance to SCYLV has been investigated in a few studies, which involved either quantitative trait locus (QTL) mapping in biparental progenies (Costet et al., 2012; Islam et al., 2018) or genome-wide association studies (GWASs) (Débibakas et al., 2014; Gouy et al., 2015; Yang et al., 2019a). Due to sugarcane's polyploid nature, most of these studies employed molecular markers that were treated as dominant, which implies a great loss of genetic information (Garcia et al., 2013). Additionally, these works employed immunological methods to phenotype SCYLV resistance. The usage of dominant markers and the poor reliability of phenotyping, allied with flaws in the assembly of association panels, were listed as factors that limited the power of the detection of QTLs and markers associated with disease resistance (Débibakas et al., 2014; Gouy et al., 2015).

In this study, we evaluated the efficacy of several genome-wide approaches to identify markers and genes associated with SCYLV resistance. We analyzed a panel of 97 *Saccharum* accessions inoculated with SCYLV, which were graded for the severity of SCYL symptoms, and their viral titration was estimated by relative and absolute RT-qPCR. This panel was genotyped with amplified fragment length polymorphisms (AFLPs) and simple sequence repeats (SSRs), which were treated as dominant. We also constructed a genotyping-by-sequencing (GBS) library for 94 genotypes of this population, which was used for the discovery of single nucleotide polymorphisms (SNPs) and insertions and deletions (indels) using a monoploid set of chromosomes of *S. spontaneum* (Zhang et al., 2018) as a genomic reference. We then employed three distinct methodologies to detect marker-trait associations: the fixed and random model circulating probability unification (FarmCPU) method using dominant markers; mixed linear modeling using SNPs and indels, in which allele proportions (APs) in each locus were employed to establish genotypic classes and estimate additive and dominant effects; and several machine learning (ML) methods coupled with feature selection (FS) techniques, using all markers to predict genotype attribution to phenotypic clusters identified by clustering analyses. Finally, by aligning neighboring sequences of the genomes of plant species closely related to sugarcane, we annotated genes containing markers associated with phenotypes. We discuss the putative participation of these genes in the mechanisms underlying resistance to SCYLV.

2 Material and Methods

2.1 Plant Material and Inoculation

The plant material and inoculation methods employed in the present study are described by Villavicencio et al. (in press). The experimental population consisted of a panel of 97 sugarcane genotypes that comprised wild germplasm accessions of *S. officinarum*, *S. spontaneum* and *Saccharum robustum*; traditional sugarcane and energy cane clones; and commercial cultivars originating from Brazilian breeding programs. The accession names and pedigree information are available in Supplementary Table 1. To ensure plant infection with SCYLV, a field nursery was

established in March 2016 at the Advanced Centre for Technological Research in Sugarcane Agribusiness located in Ribeirão Preto, São Paulo, Brazil (4°52'34" W, 21°12'50" S). Seedlings from sprouted setts of each genotype were planted in 1-meter plots with an interplot spacing of 1.5 meters. The cultivar SP71-6163, which is highly susceptible to SCYLV (Vega et al., 1997), was interspersed with the panel genotypes. *M. sacchari* aphids were reared on detached SP71-6163 leaves under laboratory conditions and released weekly in the field nursery in July 2016. Setts obtained from this nursery were used to install a field experiment following a randomized complete block design with three blocks in May 2017. Plants were grown in 1-meter-long three-row plots with row-to-row and interplot spacings of 1.5 and 2 meters, respectively. Each plot contained two plants, totaling six plants of each genotype per plot. To further assist infection by SCYLV, the cultivar SP71-6163 was planted in the borders and between blocks, and *M. sacchari* aphids were released in the field weekly for five months, starting from November 2017.

2.2 Phenotyping

Plants were phenotyped in two crop seasons: plant cane in June 2018 (13 months after planting) and ratoon cane in July 2019 (9 months after the first harvest). The severity of SCYL symptoms was assessed by three independent evaluators, who classified the top visible dewlap leaves (TVDLs) of each plot using a diagrammatic scale established by Villavicencio et al. (in press). The scale consisted of four levels, as follows: (1) green leaves with no symptoms; (2) slight yellowing of the midrib abaxial surface, with discrete or no progression to the leaf blade; (3) intense yellowing of the midrib abaxial surface and a more pronounced yellowing of the leaf blade; and (4) intense yellowing of both the midrib abaxial surface and leaf blade (Supplementary Figure 1).

In the same week as symptom evaluation was performed, fragments from the median region of at least one TVDL per plot were collected and stored at -80°C until processing. Total RNA was extracted from this tissue using TRIzol (Invitrogen, Carlsbad, USA). Samples were subjected to an additional purification process consisting of three steps: (i) mixing equal volumes of RNA extract and chloroform, (ii) precipitating the RNA overnight with 2.5 volumes of 100% ethanol and (iii) a conventional cleaning step with 70% ethanol. RNA was then quantified on a NanoDrop 2000 spectrophotometer (Thermo Scientific, Waltham, USA) and subjected to electrophoresis on a 1% agarose gel stained with ethidium bromide for integrity checks. The samples were next diluted and treated with RNase-Free RQ1 DNase (Promega, Madison, USA); after this step, the RNA was quantified and diluted again for standardization. The RNA was then converted to cDNA using the ImProm-II Reverse Transcription System kit (Promega, Madison, USA).

The SCYLV titer in each sample was determined by qPCR using GoTaq qPCR Master Mix (Promega, Madison, USA) on a Bio-Rad CFX384 Touch detection system (Bio-Rad, Philadelphia, USA). Two viral quantification methodologies were employed – one relative and one absolute – using primers and conditions as described by Chinnaraja and Viswanathan (2015). For both methods, a set of primers was used to amplify a 181-bp fragment from SCYLV ORF3 (YLSRT). For the relative quantification, an additional set of primers was used to amplify a 156-bp fragment of the 25S subunit of sugarcane ribosomal RNA (25SrRNA), which was used as an internal control. The $2^{-\Delta\Delta CT}$ method (Livak and Schmittgen, 2001) was used to correct cycle threshold (CT) values; the sample with the highest CT and a melting temperature of 82.5 ± 0.5°C for the YLSRT primers was used as a control for phenotyping in each year. The absolute quantification followed the methodology described by Chinnaraja et al. (2014). A pGEM-T Easy vector (Promega, Madison, USA) cloned with a 450-bp fragment from SCYLV ORF3 previously amplified by RT-PCR was used to construct

a serial dilution curve with six points and tenfold dilutions between points; this curve was amplified on all qPCR plates. All reactions were performed using three technical replicates.

2.3 Phenotypic Data Analyses

The normality of phenotypic data was assessed by Shapiro-Wilk tests, and normalization was carried out using the bestNormalize package (Peterson, 2017) in R software (R Core Team, 2011). The best linear unbiased predictors (BLUPs) were estimated for each trait with the breedR R package (Muñoz and Sanchez, 2015) using a mixed model as follows:

$$\boldsymbol{Y_{ijm} = \mu + B_j + Y_m + BY_{jm} + G_{i(jm)} + e_{ijm}}$$

where *Yijm* is the phenotype of the ith genotype considering the jth block and the mth year of phenotyping. The trait mean is represented by μ; fixed effects were modeled to estimate the contributions of the jth block (B_j), the mth year (Y_m) and the interaction between block and year (B_{Yjm}). Random effects included the genotype (G) and the residual error (e), representing nongenetic effects.

Pearson's correlation tests were performed using the BLUPs to check the correlation between the three traits (symptom severity, relative quantification and absolute quantification), and correlation distributions were plotted using the GGally R package (Schloerke et al., 2011). To investigate the separation of genotypes according to phenotypes, we performed two hierarchical clustering on principal components (HCPC) analyses with the factoMineR package (Lê et al., 2008) – first using only viral quantification and then employing the three analyzed traits. The factoextra R package (Kassambara and Mundt, 2017) was used to plot graphs associated with these analyses. Statistical differences between the phenotypes of the clusters identified in each HCPC were assessed by Kruskal-Wallis tests or analyses of variance (ANOVAs), depending on the distribution of the data. Post hoc Dunn's tests using the Bonferroni correction were performed with the R package dunn.test (Dinno, 2017) to verify pairwise differences between clusters.

2.4 Genotyping

2.4.1 Dominant Markers

Total DNA was extracted from leaves of each genotype following the method described by Aljanabi et al. (1999). AFLPs were developed using *Eco*RI and *Msp*I restriction enzymes (New England BioLabs). Digestion reactions were prepared in a final volume of 20 µL containing 300 ng DNA, 2.5 U of each restriction enzyme in 1X RL Buffer (New England BioLabs) and incubated in a thermocycler for 3 hours at 37°C and for 5 min at 70°C. Adapter ligation was conducted in a final volume of 40 µL containing 20 µL of the digestion reaction, 5 X buffer (40 mM Tris pH 8.4, 100 mM KCl), 0.5 µM *Eco*RI adaptor, 5 µM *Msp*I adaptor, 1 mM ATP and 0.85 U of T4 DNA ligase (67 U/µL) (BioLabs). Ligation was performed at 37°C for 2 hours and 16°C for 16 hours in a thermocycler. Preamplification was conducted with primers complementary to restriction enzyme adaptors and devoid of selective nucleotides at the 3' end (*Eco*RI+0 and *Msp*I+0 primers) and using a 6X dilution of the digestion/ligation product. This reaction was performed in a final volume of 15 µL containing 2 µL of the 6X dilution digestion/ligation product, 1X PCR buffer (20 mM Tris pH 8.4, 50 mM KCl), 3.3 µM *Eco*RI+0 and *Msp*I+0 primers, 0.17 mM dNTPs, 2 mM $MgCl_2$ and 0.07 U Taq DNA polymerase. The cycling conditions were as follows: 29 cycles at 94°C for 30 seconds, 56°C for 1 minute and 72°C for 1 minute. Preamplification reactions were diluted 10X and used for selective amplification reactions using combinations of *Eco*RI/*Msp*I primers with three selective nucleotides at the 3`end and the *Eco*RI primer labeled with fluorophores IRDye700 or IRDye800.

Thirty-five selective primer combinations were used (Supplementary Table 2). The reaction was performed in a final volume of 10 µL containing 2.5 µL of the 10X diluted preamplification, 1X PCR buffer (20 mM Tris pH 8.4, 50 mM KCl), 0.05 µM of selective Eco700 labeled primer (or 0.07 µM Eco800 primer), 0.25 µM for Msp selective primer, 0.25 µM dNTPs, 2 mM $MgCl_2$, 0.5 U of Taq DNA polymerase. The cycling conditions were as follows: one cycle at 94°C for 30 seconds, 65°C for 30 seconds and 72°C for 1 minute followed by 12 cycles at 94°C for 30 seconds, 65°C for 30 seconds (decreasing 0.7°C/cycle) and 72°C for 1 minute, followed by 23 cycles of 94°C for 30 seconds, 56°C for 30 seconds and 72°C for 1 minute. Final amplicons were separated on a 6% denaturing polyacrylamide gel and visualized with a LI-COR 4300 DNA Analyzer (LI-COR, Lincoln, NE, USA).

Twelve SSR loci previously isolated from the sugarcane expressed sequence tag database (Maccheroni et al., 2007; Oliveira et al., 2007, 2009; Marconi et al., 2011) were used for SSR genotyping; details on each SSR locus can be found in Supplementary Table 3. PCR mixes were prepared, and amplifications were conducted in a Bio-Rad MyCycler thermal cycler (Bio-Rad, Philadelphia, USA) following the conditions previously established by Oliveira et al. (2009) and Marconi et al. (2011); primers were labeled with fluorescent dyes IRDye700 and IRDye800 to allow band visualization. Amplicons were separated on a 6% denaturing polyacrylamide gel and visualized with a LI-COR 4300 DNA Analyzer. Due to sugarcane's polyploidy, both AFLPs and SSRs were treated as dominant and scored based on the presence (1) or absence (0) of bands. After genotyping, genotypes and markers with over 10% missing data were removed, as well as markers with a minor allele frequency (MAF) below 10%.

2.4.2 Genotyping-by-sequencing

Genomic DNA was extracted from young leaves using the GenElute Plant Genomic DNA Miniprep Kit (Sigma-Aldrich, St. Louis, USA). The integrity of the DNA was verified by electrophoresis on a 1% agarose gel stained with ethidium bromide, and its concentration was determined using a Qubit 3.0 fluorometer (Thermo Scientific, Wilmington, USA). The construction of the GBS library was based on a protocol by Poland et al. (2012) and used a combination of *Pst*I and *Mse*I restriction enzymes. For operational reasons, 94 out of the 97 genotypes of the panel were included in the library, which did not include genotypes 87, 88 and 95 (see Supplementary Table 1). The library was subjected to a purification step using polyethylene glycol as described by Lundin et al. (2010) with slight modifications. It was then validated with a Fragment Analyzer (Agilent Technologies, Santa Clara, USA) and quantified by RT-qPCR in a Bio-Rad CFX384 Touch detection system using the KAPPA KK4824 kit (Kapa Biosystems, Wilmington, USA). Two 150-bp single-end sequencing libraries were prepared using the NextSeq 500/550 High Output Kit (Illumina, San Diego, USA) and sequenced on a NextSeq 500 (Illumina, San Diego, USA).

After checking sequencing quality with FastQC (Andrews, 2010), we used Stacks software version 1.42 (Catchen et al., 2013) for demultiplexing and checking the amount of data generated for each sample. The TASSEL4-POLY pipeline (Pereira et al., 2018), developed from TASSEL-GBS (Glaubitz et al., 2014), was used for variant calling. Most parameters were set at their standard values; exceptions were the use of the "inclGaps" argument in the "DiscoverySNPCaller" plugin, the "misMat" argument with a value of 0.3 and the "callHets" argument in the "MergeDuplicateSNPs" plugin. Rather than aligning raw reads to a reference genome, the TASSEL-GBS pipeline first generates "tags" – unique sequences representing redundant reads – to reduce computation time (Glaubitz et al., 2014). We tested the mapping of tags against nine genomic references using two aligners: BWA version 0.7.2 (Li and Durbin, 2009) and Bowtie2 version 2.2.5 (Langmead and

Salzberg, 2012). The genomic references used were as follows: the *Sorghum bicolor* genome (Paterson et al., 2009), the methyl-filtered genome of the sugarcane cultivar SP70-1143 (Grativol et al., 2014), a sugarcane RNA-Seq assembly (Cardoso-Silva et al., 2014), a de novo assembly generated from GBS data following the GBS-SNP-CROP pipeline (Melo et al., 2016), a draft genome of the sugarcane cultivar SP80-3280 (Riaño-Pachón and Mattiello, 2017), a sugarcane transcriptome generated by Iso-Seq (Hoang et al., 2017), the mosaic monoploid genome of the sugarcane cultivar R570 (Garsmeur et al., 2018), the *S. spontaneum* genome (Zhang et al., 2018) and a monoploid chromosomic set obtained from this same reference that included the "A" haplotype and unassembled scaffolds. After variant calling, VCFtools version 0.1.13 (Danecek et al., 2011) was used to retain biallelic markers with a MAF of 0.1, no missing data and a minimum sequencing depth of 50 reads. The most appropriate reference was chosen, and adopting the method proposed by Yang et al. (2019b), the ratio between alleles (allele proportions, APs) of each variant was transformed into genotypes with a fixed ploidy of 12; this conversion was performed using the vcfR R package (Knaus and Grünwald, 2017).

2.5 Population Structure and Linkage Disequilibrium Analyses

Three procedures were used to evaluate genetic structuring in the panel, employing dominant and codominant markers separately; for all analyses, the maximum number of clusters in the panel was set to 10. The first method was a discriminant analysis on principal components (DAPC), performed in the adegenet R package (Jombart et al., 2008). The second was a principal component analysis (PCA) followed by K-means, for which missing data were imputed with the pcaMethods package (Stacklies et al., 2007) and for which the optimal number of clusters was evaluated using the elbow, silhouette and gap statistic methods in the factoextra package. The last was a Bayesian clustering of genotypes into predetermined numbers of clusters (K) performed on STRUCTURE software (Pritchard et al., 2000), assuming an admixture model with correlated allelic frequencies between populations. Ten independent runs were implemented for each K, and for dominant markers, estimates of probabilities of values of K in each run were taken following 100,000 generations as burn-in and 200,000 generations sampled in a Monte Carlo Markov Chain (MCMC). For Bayesian clustering using SNPs and indels, we used a subset of 7,000 markers randomly sampled from the total dataset, parallelized STRUCTURE with StrAuto software (Chhatre and Emerson, 2016) and sampled 100,000 generations in the MCMC. In both cases, the most likely number of genetic clusters was determined by the ad hoc statistics ΔK (Evanno et al., 2005) and the LnP(D) probability logarithm; the output was interpreted in STRUCTURE HARVESTER software version 0.6.94 (Earl et al., 2012). Clumpak software (Kopelman et al., 2015) was used to average the admixture proportions of runs and to estimate cluster membership coefficients for genotypes.

For SNPs and indels, we also measured linkage disequilibrium (LD) on the ldsep R package (Gerard, 2020) by calculating the squared correlation coefficient (r^2) between pairs of markers on the same chromosome. The decay of LD over physical distance was investigated by pooling all chromosomes, plotting pairwise r^2 values against the distance between markers and fitting a curve using the equation proposed by Hill and Weir (1988). The critical r^2 for LD decay was set to 0.1, the most commonly used threshold for determining the existence of LD (Vos et al., 2017). Only comparisons with $p < 0.05$ were used in this analysis.

2.6 Association Analyses

2.6.1 FarmCPU

Association analyses with dominant markers were performed with the FarmCPU method in R (Liu et al., 2016). For these analyses, markers were recoded to indicate the presence (0) and absence (2) of bands. We tested FarmCPU using no covariates and including matrices obtained from the three genetic structure analyses described in the previous section as such. In each case, a "quantile-quantile" (Q-Q) plot of the $-\log_{10}(p)$ values of markers was generated, and the genomic inflation factor λ (Devlin and Roeder, 1999) was calculated. The average λ from the analyses of the three traits employing each covariate matrix was calculated and used to select the model that best controlled inflation. The Bonferroni correction with $\alpha = 0.05$ was used to establish the significance threshold for associations. The phenotypic variance explained by each marker was estimated for significant marker-trait associations using a linear model in R software.

2.6.2 Mixed Modeling in GWASpoly

Association analyses using SNPs and indels were performed using mixed linear model approaches in the GWASpoly R package (Rosyara et al., 2016). The output of the three genetic structure analyses previously described was used to build Q matrices, which were included in the models as fixed effects. Similarly, three different genetic kinship matrices (K) of the panel were computed and included as random effects: (I) a realized relationship model (MM^T) matrix (VanRaden, 2008), built on GWASpoly; (II) a complete autopolyploid matrix based on Slater et al. (2016), built with the AGHmatrix R package (Amadeu et al., 2016); and (III) a pseudodiploid matrix based on Slater et al. (2016), also built with AGHmatrix. We tested twelve Q + K combinations and, for each of them, six marker-effect models were used: general, additive, simplex dominant reference, simplex dominant alternative, diploidized general and diploidized additive. For each model, a Q-Q plot of the $-\log_{10}(p)$ values of markers was generated, and λ was calculated. The average λ of all traits and models employing each Q + K combination was calculated and used to select the best set of matrices. Once this combination was chosen, Manhattan plots were generated for all models and traits. The Bonferroni and false discovery rate (FDR) correction methods with $\alpha = 0.05$ were assessed to establish the significance threshold for associations.

2.6.3 Machine Learning Coupled with Feature Selection

Finally, we assessed the capacity of ML strategies to predict the attribution of genotypes to the phenotypic groups identified in the HCPC analyses based on all markers, following the genomic prediction approach proposed by Aono et al. (2020). For this approach, we selected accessions successfully genotyped with both SNPs/indels and AFLPs/SSRs; missing data in dominant markers were imputed as the means. We evaluated the accuracy of eight ML algorithms: adaptive boosting (AB) (Freund and Schapire, 1997), decision tree (DT) (Quinlan, 1986), Gaussian naive Bayes (GNB) (Friedman et al., 1997), Gaussian process (GP) (Rasmussem, 2003), K-nearest neighbor (KNN) (Cover and Hart, 1967), multilayer perceptron (MLP) neural network (Popescu et al., 2009), random forest (RF) (Breiman, 2001) and support vector machine (SVM) (Cristianini and Shawe-Taylor, 2000), all implemented in the scikit-learn Python 3 module (Pedregosa et al., 2011). As a cross-validation strategy, we used a stratified K-fold (k=5) repeated 100 times for different data configurations.

We then tested five FS techniques to obtain feature importance and create subsets of marker data: gradient tree boosting (FS1) (Chen and Guestrin, 2016), L1-based FS through a linear support vector

classification system (FS2) (Cristianini and Shawe-Taylor, 2000), extremely randomized trees (FS3) (Geurts et al., 2006), univariate FS using ANOVA (FS4) and RF (FS5) (Breiman, 2001); all FS approaches were implemented in the scikit-learn Python 3 module. We tested the differences in the accuracy between the selected FS methods using ANOVAs and multiple comparisons by Tukey's tests implemented in the agricolae R package (De Mendiburu and de Mendiburu, 2020). We also evaluated the intersection of these datasets. Finally, the area under the receiver operating characteristic (ROC) curve (AUC) was calculated for the best ML-FS combination and plotted using the Matplotlib library89 with Python 3.

2.6 Marker Mapping and Annotation

The distribution of markers identified by all analyses along *S. spontaneum* "A" chromosomes was visualized using MapChart (Voorips, 2002). Markers previously associated with SCYLV resistance by QTL mapping (Costet et al., 2012; Islam et al., 2018) and GWAS (Débibakas et al., 2014; Yang et al., 2019a) were also retrieved and included in the map. Finally, the sequences of associated markers were annotated by aligning SSR flanking sequences or the 2000-bp window adjacent to SNPs and indels against the coding sequences (CDSs) of the genomes of 14 Poaceae species (*Brachypodium distachyon*, *Brachypodium hybridum*, *Brachypodium silvaticum*, *Hordeum vulgare*, *Oryza sativa*, *Oropetium thomaeum*, *Panicum hallii*, *Panicum virgatum*, *S. bicolor*, *Setaria italica*, *Setaria viridis*, *Triticum aestivum*, *Thinopyrum intermedium* and *Zea mays*) and *Arabidopsis thaliana* retrieved from the Phytozome database (https://phytozome.jgi.doe.gov/pz/portal.html). For this, BLASTn (Altschul et al., 1990) was used with an E-value of 1e-30, and the best alignment of each sequence was kept for analysis.

3 Results

3.1 Phenotypic Data Analyses

A total of 97 sugarcane accessions inoculated with SCYLV were evaluated for the severity of SCYL symptoms and for viral titer by relative and absolute RT-qPCR quantification of individual samples in two consecutive years. We observed a skew in the distribution of symptom severity data towards the absence of symptoms, and despite normalization procedures, these data did not follow a normal distribution, as indicated by a Shapiro-Wilk test ($p = 2.2e-16$). This was not the case for the SCYLV quantification data; both the relative ($p = 0.999$) and absolute ($p = 0.930$) quantification data were successfully normalized. Plots of the distribution of the BLUPs calculated for the three traits analyzed and of the correlations between them are depicted in Figure 1. Pearson's tests indicated that symptom severity was not significantly correlated with the viral titer determined by the relative ($p = 0.117$) or absolute ($p = 0.296$) quantification. We found, however, a significant ($p < 2.2e-16$) and strong ($r^2 = 0.772$) correlation between the values obtained by the two quantification methods, indicating their reliability.

We performed two HCPC analyses to investigate the classification of genotypes according to the phenotypes associated with SCYLV resistance. The first one, using BLUP values of SCYLV titer determined by RT-qPCR, indicated the division of the panel into three clusters (Supplementary Figure 2); a factor map wherein these groups are plotted onto the first two dimensions of this HCPC is shown in Figure 2A, and the attribution of genotypes to each cluster is available in Supplementary Table 4. Boxplots of the distribution of data among these clusters (Supplementary Figure 3) indicated that each of these groups presented distinct SCYLV titers, as indicated by the two quantification methods: cluster Q1 was composed of accessions that showed the lowest titers of SCYLV; cluster Q2 was formed by accessions that presented intermediate virus titers; and cluster Q3 included genotypes

with the highest virus titers. ANOVAs indicated that these differences were significant ($p < 2e-16$), and Dunn tests showed that the virus titer determined by both quantification methods was significantly different between all pairs of clusters (Supplementary Table 5).

The second HCPC, performed with the BLUPs of the three traits analyzed, also indicated the division of the panel into three clusters (Supplementary Figure 4); a factor map wherein these groups are plotted onto the first two dimensions of this HCPC is shown in Figure 2B, and the attribution of genotypes to each cluster is available in Supplementary Table 4. Boxplots of the distribution of phenotypic data among the three clusters (Supplementary Figure 5) clearly show that this division resulted in a separation of groups with contrasting phenotypes. Cluster SQ1 was composed of accessions that showed the least severe SCYL symptoms and presented low titers of SCYLV, as indicated by the two quantification methods; interestingly, all genotypes attributed to Q1 were included in SQ1. Cluster SQ2 was formed by accessions that also displayed mild disease symptoms but presented high viral titers. Finally, cluster SQ3 included accessions with the most severe disease symptoms and high virus titers. These differences were statistically significant; a Kruskal-Wallis test indicated the existence of significant ($p = 4.078e-08$) differences in symptom severity between all clusters, and ANOVAs showed that the same significance was present for the values of relative and absolute viral quantifications ($p < 2e-16$). Dunn tests showed that the severity of SCLV symptoms in each cluster was significantly different from the others and that SQ1 showed significantly lower SCYLV titers than the other two clusters as estimated by both quantification methods (Supplementary Table 5).

3.2 Genotyping

The genotyping of the 97 accessions of the panel using 12 SSR loci and 35 AFLP selective primer combinations generated 981 polymorphic markers (167 SRR fragments and 814 AFLP fragments). Upon filtering, four genotypes (74, 93, 96 and 97) and 319 markers were removed as they presented call rate below 90% or MAF below 10%. Therefore, 93 genotypes and 662 dominant markers were retained for further analyses that involved them.

The sequencing of the GBS library including 94 sugarcane genotypes generated 119,117,758 raw reads, 87.88% of which presented a Q-value above 30. Genotypes 31 and 77 presented very low amounts of generated data (0.19 and 1.5 KB, respectively) in contrast with the library average (598 KB) and were therefore excluded from subsequent analyses to avoid biasing. The results from tag mapping against each genomic reference using BWA and Bowtie2 aligners can be found in Supplementary Table 6. To avoid sampling of duplicated regions – which are abundant in the sugarcane genome and could bias the calculation of APs – we did not include tags with multiple alignments in the ensuing analyses. BWA presented slightly superior results in terms of uniquely mapped tags and was therefore chosen as the aligning tool. Regarding the nine genomic references tested for variant calling, the methyl-filtered genome of the cultivar SP70-1143 resulted in the largest number of isolated markers among all references, yielding 60,173 SNPs and 43,054 indels after rigorous filtering (Supplementary Table 7). However, knowing that a high-quality genome assembly can greatly improve the results of GWASs (Benevenuto et al., 2019) and also allow the estimation of long-distance LD, we opted to perform ensuing analyses with markers obtained from a reference that not only yielded a large number of markers but also provided reliable information on their position. Thus, we chose to use the monoploid chromosome set isolated from the *S. spontaneum* genome as a reference; despite not yielding the highest number of markers (38,710 SNPs and 32,178 indels), this genome had a performance superior to most of the other references tested here and, unlike many others, provided markers with information of position at chromosome level.

3.3 Panel Genetic Structure and Linkage Disequilibrium

The genetic structure of the panel was investigated using two separate sets of markers – AFLPs and SSRs scored as dominant and codominant SNPs and indels with AP information – and three different approaches – a DAPC, a PCA followed by k-means and a Bayesian clustering implemented in STRUCTURE software. Supplementary Table 8 summarizes the allocation of genotypes to the clusters identified in each analysis. First, the three methods were performed with dominant markers. The DAPC was performed retaining 70 principal components (PCs) that together explained ~90% of the observed variance (Supplementary Figure 6A); the Bayesian information criterion (BIC) curve indicated k = 3 as the most likely number of clusters in the panel (Supplementary Figure 6B). Therefore, genotypes were separated into three clusters (A1-3); the plotting of the panel elements onto the first two linear discriminants from this analysis is shown in Figure 3A. Despite resulting in the subdivision of the panel into four clusters (B1-4), as indicated by two out of three statistical methods performed to assess the optimal number of clusters in the panel (Supplementary Figures 7A-C), the clustering by PCA followed by k-means yielded results that shared similarities with the DAPC. Genotypes forming B1 and B4, for instance, were entirely represented by A3 and A1, respectively. Figure 3B depicts a projection of groups identified by k-means onto the PCA's first two PCs, which accounted for ~18% of the variation in the genotypic data. Last, based on the Bayesian clustering, both ΔK and LnP(D) methods clearly indicated K = 2 as the most likely number of clusters in the panel (Supplementary Figure 8A and B, respectively). Therefore, we present here a plot of the estimated association coefficients of individuals for these two clusters (Figure 3C), named C1 and C2. Clustering by this method was highly consistent with clustering by k-means, as C1 corresponded to the union of B1 and B2, and C2 corresponded to the union of B3 and B4.

These same three analyses were next performed with codominant markers. The DAPC was performed retaining 60 PCs explaining ~90% of the observed variance (Supplementary Figure 9A); the BIC curve indicated k = 3 as the most likely number of clusters in the panel (Supplementary Figure 9B). Thus, the panel was separated into three groups (D1-3); the plotting of the panel elements onto the first two linear discriminants from this analysis is shown in Figure 4A. The PCA followed by k-means was executed assuming the existence of two clusters (E1 and E2), as clearly indicated by all three testing methods employed (Supplementary Figure 10A-C). Despite indicating the subdivision of the panel into two clusters, the attribution of genotypes to these groups shared similarities with the DAPC clustering. E1 corresponded to D2, and E2 was equivalent to D1 and D3 combined. Figure 4B depicts a projection of these groups onto the PCA's first two PCs, which accounted for ~33% of the variation in the genotypic data. Finally, based on Bayesian clustering, both ΔK and LnP(D) methods indicated K = 2 as the most likely number of clusters in the panel (Supplementary Figure 11A and B, respectively). Thus, we present here a plot of the estimated association coefficients of individuals for these two clusters, named F1 and F2 (Figure 4C). Clustering by this procedure was also consistent with clustering by the other two methods, as F1 corresponded to the union of D2 and D3, and F2 corresponded to D1.

Despite observing extensive correspondences in clustering results between methods when the same genotypic data were used, we found that using dominant or codominant markers yielded noticeably different outcomes. Some overlap was observed between clusters identified by the analyses using each set of markers (e.g., D2 corresponded roughly to the union of A1 and A2), but overall, groups identified by these analyses shared little resemblance. Additionally, the results from these methods did not present correspondences with the phenotype-based HCPCs. Although most accessions in the HCPC group SQ3 were attributed to the same genotypic cluster in each population structure analysis, such a pattern was not observed for the other phenotypic groups.

We also estimated pairwise LD between codominant markers and analyzed its decay over distance. We observed high r^2 values (~0.4) between closely distanced markers, which dropped to 0.1 at approximately 2 Mb (Figure 5).

3.4 Association Analyses

3.4.1 FarmCPU

For FarmCPU analyses, we tested matrices obtained from each genetic structure analysis as covariates and ran the models with no covariates. Supplementary Figure 12 depicts a boxplot of the distribution of λ in each case; these data followed a normal distribution ($p = 0.975$), and an ANOVA indicated that no significant differences ($p = 0.084$) existed between models. Thus, we chose to conduct association analyses using no covariates, as this resulted in the median value of λ closest to its theoretical value under the null hypothesis ($\lambda = 1$). Appropriate profiles of inflation of p-values were obtained for the three traits, as shown in the Q-Q plots generated by FarmCPU (Supplementary Figure 13). Using a Bonferroni-corrected threshold of 0.05, one marker-trait association was detected for symptom severity, and five associations were detected for the viral titer estimated by each quantification method – with one marker (MSP_TTG-ECO_ACA-700-0133) being mutually associated with both (Supplementary Table 9); this could be an indication of the consistency of the association or simply reflect the correlation between the two measurements. All markers detected to be trait-associated were AFLP fragments, and the percentage of phenotypic variance explained by each marker ranged from 9 to 30%.

3.4.2 Mixed Modeling

Twelve combinations of population structure (Q) and kinship (K) matrices were tested as effects in the association models of the three traits analyzed. A boxplot of the distribution of λ in each Q + K model is shown in Supplementary Figure 14. A Shapiro-Wilk test indicated that the data did not follow a normal distribution ($p = 3.253e-06$), and a Kruskal-Wallis test indicated that there were no significant differences ($p = 0.869$) between the inflation of p-values of the models. Thus, we chose to conduct the following analyses with a Q + K combination that resulted in the median value of λ closest to 1, which was obtained with the combination of the first three PCs from a PCA (Q_{PCA}) with both the realized relationship (K_{MM}) and pseudodiploid (K_{Pseudo}) kinship matrices. Taking into account that the MM^T matrix can be computed by the GWASpoly package, we considered the use of the $Q_{PCA} + K_{MM}$ combination to be the most straightforward approach. Q-Q plots of the association analyses for SCYL symptom severity and SCYLV relative and absolute quantifications can be found in Supplementary Figure 15. In general, all models showed an appropriate profile of inflation of p-values.

A stringent significance threshold ($p < 0.05$ corrected by the Bonferroni method) was used to identify 35 nonredundant markers (19 SNPs and 16 indels) that were significantly associated with SCYL symptom severity (Figure 6). Using this correction, no markers were significantly associated with SCYLV titer. In an attempt to establish a less conservative threshold for association analyses of these two traits, we employed the FDR for the correction of p-values, which resulted in very low significance thresholds and the identification of thousands of associations as significant (data not shown). Therefore, we ultimately opted to use an arbitrary threshold of $p < 0.0001$ ($-\log_{10}(p) > 4$) to determine markers strongly associated with the two quantification traits. This threshold resulted in 13 markers (4 SNPs and 9 indels) associated with SCYLV titers determined by relative quantification and 9 markers (6 SNPs and 3 indels) associated with SCYLV titers determined by absolute quantification (Figure 6); one marker (S3_5029059) was common to both viral quantification

methods. Supplementary Table 10 supplies information on all marker-trait associations identified by this approach. For each trait analyzed, we observed a redundancy between markers identified as significant by the six different marker-effect models employed; this observation was particularly common between the simplex dominant alternative and the two diploidized models. No markers were identified as significantly associated with phenotypes using the simplex dominant reference model.

3.4.3 Machine Learning Coupled with Feature Selection

As a last marker-trait association method, we tested eight ML algorithms for predicting the attribution of genotypes to the phenotypic clusters identified in the two HCPC analyses. When assessing their potential in this task using the full marker dataset, the predictive accuracies varied greatly depending on the method and the phenotypic groups under analysis. They were lower for the prediction of clusters associated with viral titer (Q), ranging from 39.2% with DT to 49.6% with SVM, with an average of 44.5% (Supplementary Figure 16A). For the clusters identified including symptom severity data (SQ), accuracies were overall higher, albeit varying even more and being still unsatisfactory; they then ranged from 7.9% with GNB to 73.9% with SVM (Supplementary Figure 16B) and had an average of 58%. Therefore, we tested applying five different FS methods to reduce the marker dataset, keeping markers that were more explanatory of the phenotypic clustering results. Furthermore, we constructed three additional reduced marker datasets consisting of intersections between FS methods: markers selected by at least two of the five methods (Inter1); markers selected by at least two of the three best methods (Inter2); and markers selected by all three best methods (Inter3). We then reassessed the predictive accuracies of all ML algorithms using these new marker datasets.

After these procedures, we could achieve considerably higher accuracies in predicting Q and SQ clusters. An ANOVA followed by Tukey's test showed that three of the FS methods tested (FS1, FS2 and FS4) presented notably superior effects in both cases, while the other two (FS3 and FS5) did not result in substantial increases in accuracy (Supplementary Figure 17). In the two scenarios, the most accurate model-FS combination was MLP coupled with FS2, which was composed of 232 markers for Q and 170 markers for SQ. This combination resulted in average accuracies of 97.6% and 96.5% for the prediction of Q and SQ clusters, respectively (Supplementary Tables 11 and 12). However, in both scenarios, MLP also achieved the second-best results when using Inter2 datasets, which were composed of 190 markers for Q and 120 markers for SQ. With this strategy, we could achieve average accuracies of 95.7% for Q and 95.4% for SQ. As the usage of Inter2 represented an even further reduction in the number of markers with minimal decreases in the predictive accuracies, we considered the MLP-Inter2 combination as the most efficient approach to predict clusters associated with SCYLV resistance. To further evaluate the performance of MLP, we produced ROC curves and calculated their respective AUCs (Supplementary Figures 18 and 19). Prior to FS, MLP did not present satisfactory results, with ROC curves very close to the chance level and AUCs of 0.45-0.61 for Q and 0.40-0.56 for SQ (Supplementary Figures 18A and 19A). When Inter2 was used, the ROC curves showed a much better performance of the model, with AUCs of 1.00 for Q and of 0.98-1.00 for SQ (Supplementary Figures 18B and 19B). These results confirm that the markers present in Inter2 are in fact associated with SCYLV resistance and that MLP is an appropriate model to predict clustering based on this marker dataset.

The markers representing the reduced datasets associated with Q and SQ clusters can be found in Supplementary Tables 13 and 14, respectively. We observed twelve marker overlaps between the two datasets, i.e., two AFLP fragments, seven SNPs and three indels were present in both of them. Interestingly, several of these markers were also identified as associated with phenotypes in the

FarmCPU and mixed modeling analyses. This was the case, for instance, of MSP_TTG-ECO_ACA-700-0133, an AFLP fragment associated with SCYLV titer estimated by relative and absolute quantification and with Q and SQ clustering, and of S3_5029059, an indel associated with relative and absolute SCYLV quantification by GWAS that was also present in both the Q and SQ Inter2 datasets.

3.5 Marker Mapping and Annotation

For a better visualization of the physical location of all markers associated with SCYLV resistance, we constructed a map of their distribution along *S. spontaneum*'s "A" chromosomes (Figure 7), in which we also included markers identified as associated with SCYLV resistance in previous mapping studies. Overall, markers were considerably spread along chromosomes; however, we observed regions of dense concentration of markers identified by various methods, such as the long arms of chromosomes 1 and 3. We also verified the proximity between several markers identified in the present work and by other authors, indicating their convergence and the reliability of the methods employed here.

Out of the 362 nonredundant markers associated with all phenotypes, 176 were located in genic regions and could be annotated by aligning their 2000-bp neighboring regions with the CDSs of 14 Poaceae species and *A. thaliana* genomes; Supplementary Table 15 contains data on the alignment with the highest percentage of identity for each marker. In some cases, where two or more markers were closely located, coincident alignments and annotations were obtained; consequently, 148 genes were representative of all the best alignments. The large majority of top-scoring alignments (117) occurred with CDSs of *S. bicolor*, the phylogenetically closest species among those used for alignment; fewer alignments also occurred with the CDSs of *Z. mays* (19), *P. hallii* (4), *H. vulgare* (2), *Oropetium thomaeum* (2), *P. virgatum* (2), *S. viridis* (2), *B. sylvaticum* (1), *S. italica* (1) and *T. intermedium* (1). All of the annotated markers but one SSR fragment (SCB381.09) consisted of SNPs or indels. Several of the annotated genes could be associated with plant resistance to viruses, as detailed in the discussion.

4 Discussion

In the present study, we evaluated the severity of SCYL symptoms and SCYLV titer in a panel of 97 sugarcane accessions. These two traits are of great concern to breeding, as the exhibition of SCYL symptoms and high viral titers have both been associated with higher yield losses in SCYLV-infected sugarcane plants (Lehrer et al., 2007; Zhu et al., 2010; Viswanathan et al., 2014). Prior to phenotyping, plants were subjected to high SCYLV inoculum pressure, an innovation over all previous mapping studies (Costet et al., 2012; Débibakas et al., 2014; Gouy et al., 2015; Islam et al., 2018; Yang et al., 2019a), which relied on natural infection under field conditions. Using RT-qPCR, currently regarded as the most precise method for SCYLV quantification (Zhu et al., 2010), we assessed the viral titer in these genotypes; together with the study by Villavicencio et al. (in press), this is also the first time RT-qPCR was used for this purpose in a large number of sugarcane genotypes. We found a strong and positive correlation between the BLUPs calculated for the SCYLV titers obtained by the two quantification methods employed, showing the consistency of the data. The absence of a perfect correlation might have arisen from intrinsic differences between the two methods, which have been responsible for disparities in viral quantification by RT-qPCR in other plant-virus interactions (Jarošová et al., 2009).

However, we observed no significant correlation between the BLUPs calculated for the severity of disease symptoms and viral titers. This finding corroborates a growing body of evidence suggesting

that these traits are not strongly or necessarily correlated, i.e., high SCYLV titers are not a guarantee of more severe yellowing or of its development at all (Comstock et al., 1998; Lehrer and Komor, 2008; Chinnaraja et al., 2014). Although Zhu et al. (2010) claim to have observed a correlation between these two parameters, this was not statistically shown. Our results, on the other hand, demonstrate that no quantitative correlation between the severity of SCYL symptoms and SCYLV titers exists across the sugarcane genotypes analyzed. This reinforces the importance of SCYLV molecular screening of sugarcane clones by breeding programs; this should be done to avoid the employment of genotypes that accumulate high viral loads asymptomatically but may inconspicuously suffer yield losses, in addition to serving as a virus reservoir for vector transmission to other susceptible genotypes.

To further explore this issue, we performed two HCPC analyses to discriminate accessions based on their response to SCYLV infection, which led to the separation of clusters with considerable phenotypic differences. In the first HCPC, using only data from the two quantification methods, we could discern groups with significant variation in viral titers; in the second analysis, which also included symptom severity data, clusters with even more contrasting responses to SCYLV could be discriminated. In 1983, Cooper and Jones proposed a new terminology addressing plant responses to viral infections that is still employed by several authors in this area (Beoni et al., 2016; Foresman et al., 2016; Mansilla-Córdova et al., 2018). According to their proposal, once infected, plants present differences in the ability to restrict viral replication and invasion; at the extremes of a spectrum of behaviors are plants termed susceptible (in which the virus rapidly replicates and disperses) and resistant (in which viral replication and dispersal is slow). In addition, plants may also respond differently to the infection: another spectrum exists, at the extremes of which are sensitive plants (which exhibit severe symptoms) and tolerant plants (which exhibit little or no symptoms). In view of this nomenclature, we propose that the clusters identified in this second HCPC be described by the three following designations: (SQ1) resistant, for sugarcane genotypes distinguished by low SCYLV titer and mild or no SCYL symptoms; (SQ2) tolerant, for genotypes that, despite exhibiting higher viral titers, presented fewer or no disease symptoms; and (SQ3) susceptible, for genotypes with the most severe symptoms and presenting high viral titers. This classification per se is of great use in sugarcane breeding, as it distinguishes not only sources of tolerance to SCYLV but also an exceptionally promising group of genotypes that are truly resistant to this virus.

The main objective of this work was, however, to identify markers associated with SCYLV resistance in a broader sense. With this aim, we genotyped the plants in the panel with a combination of dominant and codominant markers, which has never been described for sugarcane to date; this is also the first study employing a two-enzyme GBS protocol in this crop. We evaluated the impact of using genomic references from various backgrounds in variant calling. In previous sugarcane GWASs, variant calling was performed using the genome of *S. bicolor* (Fickett et al., 2019; Yang et al., 2019a, 2019b, 2020), and until recently, this was sugarcane's closest relative with a well-assembled and annotated genome; however, in our analyses, this reference yielded a number of markers considerably inferior to other references, which discouraged its use. The reference that yielded the largest number of markers was the methyl-filtered genome of the SP70-1143 cultivar, in agreement with the results of a previous study employing GBS (Balsalobre et al., 2017). This is a plausible outcome, as the protocol used for library preparation reduces genome complexity by avoiding sampling of methylated regions (Elshire et al., 2011), which were also filtered out for assembly of the SP70-1143 genome (Grativol et al., 2014). However, to choose the best reference for further analyses, we also considered the quality of the assembly, which has been shown to greatly affect the results of GWASs in polyploids (Benevenuto et al., 2019). The best-assembled sugarcane genome available to date is the allele-defined genome of a haploid *S. spontaneum* accession (Zhang

et al., 2018). Despite presenting one of the highest total tag alignment rates, this reference also gave a very high rate of multiple alignments, which led to the identification of a relatively low number of markers. This was probably due to the alignment of tags to homeologous regions of different alleles rather than simply to the duplicated regions that we intended to avoid. To circumvent this situation, we opted to conduct our analyses with markers isolated using a monoploid chromosome set obtained from this genome, which provided a large number of markers (38,710 SNPs and 32,178 indels) with reliable position information.

Using these codominant markers, we analyzed the decay of LD over distance. LD has long been hypothesized to be high in sugarcane due to the short breeding history and narrow genetic basis of this crop; many studies using dominant markers have estimated it to be especially high at 5-10 cM (Jannoo et al., 1999; Raboin et al., 2008, Wei et al., 2010; Singh et al., 2016; Barreto et al., 2019). The first study to use SNPs for this task and estimate LD decay in bp (Yang et al., 2019c) indicated that LD was extremely long lasting, with the average r^2 decaying to 0.2 at 3.5 Mb in hybrids. Our results further confirm the persistence of LD at long distances in sugarcane, albeit indicating that it decayed more quickly – with r^2 dropping to 0.2 at less than 1 Mb and to 0.1 at 2 Mb. These results impact mapping studies, as a high LD implies that a low density of markers might be needed for accurate mapping of quantitative traits.

With the two distinct marker datasets generated – including dominant and codominant markers with AP information – we tested several approaches to evaluate population structure in the panel; these assessments yielded remarkably different results. Some of the few existing studies contrasting the usage of dominant and codominant markers in plants have also shown discrepancies in measures of genetic structure and diversity (Gaudeul et al., 2004; Fang et al., 2006; Roncallo et al., 2019). However, this sort of comparison has never been performed including markers with AD or AP information in polyploids, let alone in sugarcane, which has one the most complex genomes known. In this crop, the most relatable findings available are those reported by Creste et al. (2010), who compared the effect of using dominant AFLP, SSR and target region amplification polymorphism markers in the estimation of genetic relationships in sugarcane. Alike we demonstrate now, these authors showed that using different types of markers can bias genetic analyses in this crop and that thus the choice of marker to be used must be guided by the specific goal of each study. For genome-wide mapping of quantitative traits – for which a high density of markers is necessary – codominant SNPs and indels are currently more cost-effective, as they can be easily identified in much larger numbers. Additionally, in the case of polyploids, these markers offer the possibility of estimating allele dosages or APs, which are highly informative for genetic studies (de Bem Oliveira, 2019; Matias et al., 2019; Aono et al., 2020). Hence, we believe the results we obtained with codominant SNPs and indels are more reliable, as they rely on much larger amounts of genetic information.

In contrast with the differences arisen from the type of marker used, we observed little divergence between the results of the methods used for each marker dataset; the separation of genotypes into groups was overall similar, with consistent correspondences between groups identified in each analysis. Additionally, eventual discrepancies between methods did not result in significant differences in the inflation of the models used for both FarmCPU and codominant mixed modeling analyses; the inflation patterns observed in Q-Q plots were similar to those of previous studies applying similar genotyping and marker-trait association methodologies (Barreto et al., 2019; Yang et al., 2019a, 2019b, 2020). For both association mapping strategies, we opted to perform ensuing association analyses using the covariates that resulted in the value of λ closest to 1. For FarmCPU, this corresponded to the "naive" model with no covariates; for codominant mixed modeling analysis, this was the Q_{PCA} + K_{MM} combination. K_{MM} is the usual choice of relationship matrix in polyploid

association mapping (Rosyara et al., 2016; Berdugo-Cely et al., 2017; Byrne et al., 2020), as Q matrices obtained from PCA are commonly used to control population structure in GWASs (Nimmakayala et al., 2016; Su et al., 2016; Ferrão et al., 2018).

FarmCPU analyses using dominant markers identified one AFLP fragment significantly associated with symptom severity, which explained a small part of the phenotypic variation ($r^2 = 0.116$). Eight out of the nine markers found to be associated with viral titer explained larger parts of the variation in the phenotypes (21-30%). These results are more promising than those obtained in a previous GWAS employing dominant markers in the investigation of SCYLV resistance, which found markers explaining smaller percentages (9-14%) of the variation observed (Débibakas et al., 2014). Albeit low, values in this range are, in fact, the rule rather than the exception in sugarcane association studies. Evidence indicates that almost all of this crop's traits are highly quantitative, with the notable exception of brown rust resistance, for which two major QTLs are known (Daugrois et al., 1996; Raboin et al., 2006). For other relevant traits, such as yield, sugar or fiber content and resistance to other diseases, it is common to find the most associated markers explaining ≤ 10% of the phenotypic variation (Gouy et al., 2013; Barreto et al., 2019; Fickett et al., 2019).

A few authors have suggested that these suboptimal results could be improved with the usage of markers with dosage; this was also performed here, using SNPs and indels with AP information for codominant mixed modeling analyses. Although we successfully identified markers associated with SCYL symptom severity using the Bonferroni correction, the same was not observed for SCYLV quantification data obtained by RT-qPCR. This was probably influenced by the limited size of the panel, a factor that restricts the power of GWASs (Wang et al., 2012; Yang et al., 2014). As pointed out by Racedo et al. (2016), assembling and phenotyping large sugarcane association panels is a challenging task, and this was especially burdensome in our study, in which laborious inoculation and quantification techniques were employed. Furthermore, the Bonferroni method is notorious for its conservative nature, which contributes to a low capacity for controlling false negatives (Gao et al., 2010; Johnson et al., 2010; Fadista et al., 2016). This led us to establish an arbitrary threshold ($p < 0.0001$) to select markers strongly associated with SCYLV titer for further investigation. Using this methodology, we identified 57 nonredundant markers associated with the three phenotypes.

Finally, as a last approach to identify marker-trait associations, we tested several ML algorithms coupled with FS methods to predict genotype attribution to phenotypic clusters identified by HCPC analyses. We could achieve very high accuracies of prediction (95%) with considerably reduced marker datasets comprising 120-190 SNPs and indels. These results are very similar to what was obtained for predicting sugarcane brown rust resistance groups using the same methods, where an accuracy of 95% was obtained using 131 SNPs (Aono et al., 2020). Marker datasets selected by ML have rarely been employed in genetic association studies in plants, but the few existing examples show their power to identify genes associated with phenotypes of interest (Steinfath et al., 2009; Heer et al., 2018; Zhou and Bellis, 2019).

Here, we annotated 176 markers associated with SCYLV resistance to 148 genes. Many of these candidates do not allow extensive discussion on their involvement in resistance to this viral disease, as they either have very generic descriptions (e.g., "expressed protein") or have not been previously linked to plant virus resistance. Other proteins, such as helix-loop-helix proteins or glycosyltransferases, have occasionally been associated with responses to viruses but are members of very large gene families with extremely diverse biological roles. Thus, we will focus on discussing genes whose involvement in disease resistance is better established.

Remarkably, a few candidates harboring markers associated with SCYL resistance encode proteins previously associated with the response to SCYLV infection. This was the case for SbRio.10G317500.1, encoding a peroxidase precursor and associated with SQ clustering. Peroxidases are long known to be activated in response to pathogens, but most notably, there is evidence that a guaiacol peroxidase is more active in sugarcane plants exhibiting SCYL symptoms than in uninfected or asymptomatic plants (Scagliusi et al., 2009). Our results provide further evidence that these enzymes are in fact involved in the response to SCYLV and in the development of yellowing symptoms. Three other candidates harboring markers associated with SCYL symptom severity and SQ clustering encode proteins with motifs that, despite not being commonly linked to defense against viruses, were also found to be associated with SCYLV resistance in a previous GWAS (Yang et al., 2019a). These are Sobic.001G023900, encoding a GATA zinc finger protein, and Sobic.001G200200 and Zm00001d037864_T030, which encode proteins containing tetratricopeptide repeats.

Other annotations included classic participants in more general disease resistance mechanisms, such as several genes encoding proteins with leucine-rich repeat (LRR) motifs. These structures are an important part of nucleotide-binding LRR (NBS-LRR) proteins, receptors that detect pathogen-associated proteins and elicit effector-triggered immunity in plants (DeYoung and Innes, 2006). NBS-LRR proteins represent over 80% of all plant disease resistance (R) genes known (Dubey and Singh, 2018) and have been shown to determine resistance to viruses in several crops (Dinesh-Kumar et al., 2000; Seo et al., 2006; Xun et al., 2019). In our analyses, we found one LRR protein (Sobic.008G156600.1) associated with SCYLV titer determined by relative quantification, one disease resistance NBS-LRR (Sobic.007G085400.1) and one N-terminal leucine zipper NBS-LRR resistance gene analog (Sobic.005G203500.1) associated with Q clustering, and an additional LRR protein (Sobic.001G452600.1) associated with SQ clustering. Furthermore, we annotated one gene (Sobic.009G204800.1) associated with Q clustering that encodes a precursor of a receptor-like serine/threonine-protein kinase, i.e., the family to which LRR proteins belong. This is similar to a finding by Yang et al. (2019), who identified a serine/threonine-protein kinase associated with SCYLV resistance. We consider these proteins to be highly promising candidates to be involved in the recognition of infection by SCYLV, which could trigger response mechanisms leading to the restriction of the virus. Further virus-host interaction studies involving these proteins might help confirm this hypothesis, which would represent a major breakthrough in understanding resistance to SCYLV.

Two other annotated genes associated with Q and SQ clustering were readily identified as involved in plant disease resistance mechanisms. Sobic.010G131300.2 contains a Bric-a-Brac, Tramtrack, Broad Complex/Pox virus and Zinc finger (BTB/POZ) domain, while Sobic.007G198400.1 contains two BTB domains, as well as ankyrin repeat regions. These domains are present in and are essential for the function of NONEXPRESSOR OF PATHOGENESIS-RELATED GENES 1 (NPR1), a central player in plant disease responses (Cao et al., 1997; Rochon et al., 2006). This family of transcription factors is involved in establishing both systemic acquired resistance and induced systemic resistance (Pieterse et al., 1998), mediating the crosstalk between salicylic acid and jasmonic acid/ethylene responses (Spoel et al., 2003). Correspondingly, NPR1 has been widely shown to be involved in resistance to viruses (Liu et al., 2002; Lin et al., 2004); therefore, it is reasonable to suggest its participation in the response to infection by SCYLV.

We also found a few candidates with putative roles in the RNA interference mechanism, one of the most prominent processes that contribute to resistance against viruses in plants. This is the case for Sobic.001G214000.1, associated with SQ clustering, which encodes a Dicer. Dicers are part of a

mechanism known as RNA silencing, where they are responsible for the recognition and cleavage of long double-stranded RNA molecules into mature small RNAs, which in turn guide the site-directed cleavage of viral mRNAs that disrupt virus replication (Llave, 2010); accordingly, these nucleases have been linked to resistance to viruses in several plant species (Deleris, 2006; Alam et al., 2019). Another gene possibly involved in RNA interference is Sobic.009G121100, which harbors a marker associated with SCYLV titer determined by absolute quantification. This gene encodes a protein related to calmodulin binding – a calcium transducer that regulates the activity of various proteins with diverse functions (Bouché et al., 2005) and has been widely implicated in viral resistance in plants, often playing roles in RNA interference (Heo et al., 1999; Nakahara et al., 2012; Li et al., 2014). Therefore, we consider these genes promising candidates in the regulation of SCYLV replication and spreading *in planta*, as well as in the development of SCYL symptoms.

Two additional annotations linked to the mechanism of RNA interference are those of genes encoding proteins with F-box domains, SbRio.03G158900 and Sobic.002G019750.1. These are motifs with ca. 50 amino acids that occur in large numbers in plant genomes (Correa et al., 2013). F-box proteins are involved in virus resistance in several plant species (Cao et al., 2008; Thiel et al., 2012); a particularly interesting case is FBW2 from *A. thaliana*, which regulates AGO1, an Argonaute protein with a central role in RNA silencing (Earley et al., 2010) and posttranscriptional repression of target viral RNAs through a sequence homology-dependent mechanism (Morel et al., 2002; Chen et al., 2015; Yang and Li, 2018). Even more intriguing is the fact that one of the proteins encoded by the SCYLV genome, P0, contains an F-box-like domain and mediates the destabilization of AGO1, leading to the suppression of host posttranscriptional gene silencing (Mangwende et al., 2009). Whether the F-box proteins identified here play active roles in the mechanisms involved in posttranscriptional gene silencing of SCYLV replication in sugarcane and in the manifestation of symptoms remains a question to be investigated by further studies.

Several other annotated genes may represent host factors involved in various steps of plant-virus interactions. For instance, Sobic.010G160500.4, associated with Q clustering, encodes an RNA helicase with a DEAD-box domain; these host proteins are often coopted by viruses to promote viral translation or replication, playing important roles in regulating infection (Huang et al., 2010; Kovalev et al., 2012; Li et al., 2016). Similarly, soluble N-ethylmaleimide-sensitive-factor attachment protein receptor (SNARE) proteins, characterized by the motif present in Q clustering-associated Sobic.001G528000.1, are essential in the biogenesis and fusion of vesicles involved in intercellular replication and movement of several plant viruses (Wei et al., 2013; Cabanillas et al., 2018; Sasvari et al., 2018; Ibrahim et al., 2020). We also found one gene encoding a myosin (Sobic.002G108000.1) associated with Q clustering and two genes related to kinesin (Sobic.001G346600.1 and Sobic.001G399200.2) associated with SQ clustering. These are two classes of filament-associated motor proteins that influence numerous biological processes due to their role in the transport of organelles (Heinlein, 2015). In a few cases, both myosins (Harries et al., 2009; Amari et al., 2011, 2014) and kinesins (Abdelkhalek et al., 2019) have been shown to be involved in viral intercellular movement, but the mechanisms are currently unclear. One last interesting annotation was Sobic.003G101500.1, a protein with a DNAJ domain associated with Q clustering. DNAJs have been shown to interact with proteins of various plant viruses and to be associated with resistance, sometimes being crucial for virus infection and spread (Hofius et al., 2007; Lu et al., 2009; Shimizu et al., 2009; Liu et al., 2013). We consider these genes to be promising candidates as host cofactors in the response to SCYLV infection.

In conclusion, this array of genome-wide analyses allowed us to detect markers significantly associated with SCYLV resistance in sugarcane. If validated, these markers represent an especially

valuable resource for sugarcane breeding programs, as the results can be directly employed in marker-assisted strategies for the early selection of clones. The annotation of several genes wherein these markers are located revealed many candidates with long-established and pivotal roles in viral disease resistance, further demonstrating the efficiency of the methods employed for this purpose. Additionally, this annotation provides valuable insights into the unexplored mechanisms possibly involved in sugarcane's response to infection by SCYLV, introducing new candidates whose role in this process can be further investigated in future studies.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

DP, MGAL, MCG, LRP and APS conceived the project and designed the experiments. RJGP, RCBV, CCS, IAA and LRP performed phenotyping. RJGP, RCBV and AEC performed genotyping. RJGP and AHA analyzed the data and interpreted the results. RJGP wrote the manuscript. All authors read and approved the manuscript.

Funding

This work was supported by grants from the São Paulo Research Foundation (FAPESP), the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ, grant 424050/2016-1), the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Computational Biology Program), the Littoral Polytechnic Superior School (ESPOL) and the Secretaría Nacional de Ciencia y Tecnología (SENESYT). RJGP received an MSc fellowship from CAPES (grant 88887.177386/2018-00) and MSc and PhD fellowships from FAPESP (grants 2018/18588-8 and 2019/21682-9). AHA received a PhD fellowship from FAPESP (grant 2019/03232-6). RCBV received a PhD fellowship from PAEDEx-AUIP. CCS received a PD fellowship from FAPESP (grant 2015/24346-9).

Abbreviations

AB: adaptive boosting

AFLP: amplified fragment length polymorphism

ANOVA: analysis of variance

AP: allele proportion

AUC: area under the receiver operating characteristic curve

BIC: Bayesian information criterion

BLUP: best linear unbiased predictor

BTB/POZ: Broad Complex/Pox virus and Zinc finger

CDS: coding sequence

CT: cycle threshold

DAPC: discriminant analysis of principal components

DT: decision tree

FarmCPU: fixed and random model circulating probability unification

FDR: false discovery rate

FS: feature selection

GBS: genotyping-by-sequencing

GNB: Gaussian naive Bayes

GP: Gaussian process

GWAS: genome-wide association study

HCPC: hierarchical clustering on principal components

Indel: insertion and deletion

KNN: K-nearest neighbor

LD: linkage disequilibrium

LRR: leucine-rich repeat

MAF: minimum allele frequency

ML: machine learning

MLP: multilayer perceptron

NBS: nucleotide-binding site

NPR1: NONEXPRESSOR OF PATHOGENESIS-RELATED GENES 1

PC: principal component

PCA: principal component analysis

Q-Q: quantile-quantile

QTL: quantitative trait locus

RF: random forest

ROC: receiver operating characteristic

RT-qPCR: reverse transcription quantitative polymerase chain reaction

SCYL: sugarcane yellow leaf

SCYLV: sugarcane yellow leaf virus

SNARE: soluble N-ethylmaleimide-sensitive-factor attachment protein receptor

SNP: single nucleotide polymorphism

SSR: simple sequence repeat

SVM: support vector machine

TVDL: top visible dewlap leaf

Acknowledgments

We thank Aline C. L. Moraes for assistance in constructing and sequencing the GBS library and Maicon Volpin for assistance in fieldwork.

References

Abdelkhalek, A., Ismail, I. A., Dessoky, E. S., El-Hallous, E. I., & Hafez, E. (2019). A tomato kinesin-like protein is associated with Tobacco mosaic virus infection. Biotechnology & Biotechnological Equipment, 33(1), 1424-1433. doi: 10.1080/13102818.2019.1673207

Alam, C. M., Jain, G., Kausar, A., Singh, A. K., Mandal, B., Varma, A.... & Chakraborty, S. (2019). Dicer 1 of Candida albicans cleaves plant viral dsRNA in vitro and provides tolerance in plants against virus infection. Virus Disease, 30(2), 237-244. doi: 10.1007/s13337-019-00520-x

Aljanabi, S. M., Forget, L., & Dookun, A. (1999). An improved and rapid protocol for the isolation of polysaccharide- and polyphenol-free sugarcane DNA. Plant Molecular Biology Reporter, 17(3), 281-282. doi: 10.1023/A:1007692929505

Aljanabi, S. M., Parmessur, Y., Moutia, Y., Saumtally, S., and Dookun, A. (2001). Further evidence of the association of a phytoplasma and a virus with yellow leaf syndrome in sugarcane. Plant Pathol. 50, 628-636. doi: 10.1046/j.1365-3059.2001.00604.x

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. J. Mol. Biol. 215, 403-410. doi: 10.1016/s0022-2836(05)80360-2

Amadeu, R. R., Cellon, C., Olmstead, J. W., Garcia, A. A. F., Resende, M. F. R., and Muñoz, P. R. (2016). AGHmatrix: R package to construct relationship matrices for autotetraploid and diploid species: a blueberry example. Plant Genome 9, 1-10. doi: 10.3835/plantgenome2016.01.0009

Amari, K., Lerich, A., Schmitt-Keichinger, C., Dolja, V. V., & Ritzenthaler, C. (2011). Tubule-guided cell-to-cell movement of a plant virus requires class XI myosin motors. PLoS Pathogens, 7(10), e1002327. doi: 10.1371/journal.ppat.1002327

Amari, K., Di Donato, M., Dolja, V. V., & Heinlein, M. (2014). Myosins VIII and XI play distinct roles in reproduction and transport of tobacco mosaic virus. PLoS Pathogens, 10(10), e1004448. doi: 10.1371/journal.ppat.1004448

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data. Babraham: Babraham Bioinformatics.

Aono, A. H., Costa, E. A., Rody, H. V. S., Nagai, J. S., Pimenta, R. J. G., Mancini, M. C., dos Santos, F. R. C., Pinto, L. R., Landell, M. G. de A., de Souza, A. P., & Kuroshu, R. M. (2020). Machine learning approaches reveal genomic regions associated with sugarcane brown rust resistance. Scientific Reports, 10. doi:10.1038/s41598-020-77063-5

Balsalobre, T. W. A., da Silva Pereira, G., Margarido, G. R. A., Gazaffi, R., Barreto, F. Z., Anoni, C. O.... and de Souza, A. P. (2017). GBS-based single dosage markers for linkage and QTL mapping allow gene mining for yield-related traits in sugarcane. BMC Genomics, 18, 72. doi: 10.1186/s12864-016-3383-x

Barreto, F. Z., Rosa, J. R. B. F., Balsalobre, T. W. A., Pastina, M. M., Silva, R. R., Hoffmann, H. P.... & Carneiro, M. S. (2019). A genome-wide association study identified loci for yield

component traits in sugarcane (*Saccharum* spp.). PloS One, 14(7), e0219843. doi: 10.1371/journal.pone.0219843

Benevenuto, J., Ferrão, L. F. V., Amadeu, R. R., and Munoz, P. (2019). How can a high-quality genome assembly help plant breeders? GigaScience 8, giz068. doi: 10.1093/gigascience/giz068

Beoni, E., Chrpová, J., Jarošová, J., and Kundu, J. K. (2016). Survey of Barley yellow dwarf virus incidence in winter cereal crops, and assessment of wheat and barley resistance to the virus. Crop Pasture Sci. 67, 1054-1063. doi: 10.1071/cp16167

Berdugo-Cely, J., Valbuena, R. I., Sánchez-Betancourt, E., Barrero, L. S., and Yockteng, R. (2017). Genetic diversity and association mapping in the Colombian Central Collection of *Solanum tuberosum* L. Andigenum group using SNPs markers. PLoS One 12, e0173039. doi: 10.1371/journal.pone.0173039

Bouché, N., Yellin, A., Snedden, W. A., and Fromm, H. (2005). Plant-specific calmodulin-binding proteins. Annu. Rev. Plant Biol. 56, 435-466. doi: 10.1146/annurev.arplant.56.032604.144224

Bourke, P. M., Voorrips, R. E., Visser, R. G. F., and Maliepaard, C. (2018). Tools for genetic studies in experimental populations of polyploids. Front. Plant Sci. 9, 513. doi: 10.3389/fpls.2018.00513

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. doi: 10.1023/A:1010933404324

Bremer, G. (1961). Problems in breeding and cytology of sugar cane. Euphytica 10, 59-78. doi: 10.1007/bf00037206

Byrne, S., Meade, F., Mesiti, F., Griffin, D., Kennedy, C., and Milbourne, D. (2020). Genome-wide association and genomic prediction for fry color in potato. Agronomy 10, 90. doi: 10.3390/agronomy10010090

Cabanillas, D. G., Jiang, J., Movahed, N., Germain, H., Yamaji, Y., Zheng, H., & Laliberté, J. F. (2018). Turnip mosaic virus uses the SNARE protein VTI11 in an unconventional route for replication vesicle trafficking. The Plant Cell, 30(10), 2594-2615. doi: 10.1105/tpc.18.00281

Cao, H., Glazebrook, J., Clarke, J. D., Volko, S., & Dong, X. (1997). The *Arabidopsis* NPR1 gene that controls systemic acquired resistance encodes a novel protein containing ankyrin repeats. Cell, 88(1), 57-63. doi: 10.1016/S0092-8674(00)81858-9

Cao, Y., Yang, Y., Zhang, H., Li, D., Zheng, Z., and Song, F. (2008). Overexpression of a rice defense-related F-box protein geneOsDRF1in tobacco improves disease resistance through potentiation of defense gene expression. Physiol. Plant. 134, 440-452. doi: 10.1111/j.1399-3054.2008.01149.x

Cardoso-Silva, C. B., Costa, E. A., Mancini, M. C., Balsalobre, T. W. A., Canesin, L. E. C., Pinto, L. R., et al. (2014). De novo assembly and transcriptome analysis of contrasting sugarcane varieties. PLoS One 9, e88462. doi: 10.1371/journal.pone.0088462

Carvalho-Netto, O. V., Bressiani, J. A., Soriano, H. L., Fiori, C. S., Santos, J. M., Barbosa, G. V. S., et al. (2014). The potential of the energy cane as the main biomass crop for the cellulosic industry. Chem. Biol. Technol. Agric. 1, 20. doi: 10.1186/s40538-014-0020-2

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., and Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. Mol. Ecol. 22, 3124-3140. doi: 10.1111/mec.12354

Chatenet, M., Delage, C., Ripolles, M., Irey, M., Lockhart, B. E. L., and Rott, P. (2001). Detection of sugarcane yellow leaf virus in quarantine and production of virus-free sugarcane by apical meristem culture. Plant Dis. 85, 1177-1180. doi: 10.1094/pdis.2001.85.11.1177

Chen, H., Zhang, L., Yu, K., and Wang, A. (2015). Pathogenesis of Soybean mosaic virus in soybean carrying Rsv1 gene is associated with miRNA and siRNA pathways, and breakdown of AGO1 homeostasis. Virology 476, 395-404. doi: 10.1016/j.virol.2014.12.034

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). doi: 10.1145/2939672.2939785

Chhatre, V. E., & Emerson, K. J. (2017). StrAuto: automation and parallelization of STRUCTURE analysis. BMC Bioinformatics, 18, 1-5. doi: 10.1186/s12859-017-1593-0

Chinnaraja, C., and Viswanathan, R. (2015). Quantification of sugarcane yellow leaf virus in sugarcane following transmission through aphid vector, *Melanaphis sacchari*. Virus Disease 26, 237-242. doi: 10.1007/s13337-015-0267-7

Chinnaraja, C., Viswanathan, R., Sathyabhama, M., Parameswari, B., Bagyalakshmi, K., Malathi, P., et al. (2014). Quantification of sugarcane yellow leaf virus in in vitro plantlets and asymptomatic plants of sugarcane by RT-qPCR. Curr. Sci. 106, 729-734

Comstock, J. C., Irey, M. S., Lockhart, B. E. L., and Wang, Z. K. (1998). Incidence of yellow leaf syndrome in CP cultivars based on polymerase chain reaction and serological techniques. Sugar Cane 4, 21-24

Cooper, J. I., and Jones, A. T. (1983). Responses of plants to viruses: proposals for the use of terms. Phytopathology 73, 127-128. doi: 10.1094/phyto-73-127

Correa, R. L., Bruckner, F. P., de Souza Cascardo, R., and Alfenas-Zerbini, P. (2013). The role of F-box proteins during viral infection. International journal of molecular sciences, 14, 4030-4049. doi: 10.3390/ijms14024030

Costet, L., Raboin, L.-M., Payet, M., D'Hont, A., and Nibouche, S. (2012). A major quantitative trait allele for resistance to the sugarcane yellow leaf virus (Luteoviridae). Plant Breed. 131, 637-640. doi: 10.1111/j.1439-0523.2012.02003.x

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), 21-27. doi: 10.1109/TIT.1967.1053964

Creste, S., Sansoli, D. M., Tardiani, A. C. S., Silva, D. N., Gonçalves, F. K., Fávero, T. M.... & Pinto, L. R. (2010). Comparison of AFLP, TRAP and SSRs in the estimation of genetic relationships in sugarcane. Sugar Tech, 12(2), 150-154. doi: 10.1007/s12355-010-0029-1

Cristianini, N., & Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press. doi: 10.1017/CBO9780511801389

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A.... and McVean, G. (2011). The variant call format and VCFtools. Bioinformatics, 27, 2156-2158. doi: 10.1093/bioinformatics/btr330

Daugrois, J. H., Grivet, L., Roques, D., Hoarau, J. Y., Lombard, H., Glaszmann, J. C., & D'Hont, A. (1996). A putative major gene for rust resistance linked with a RFLP marker in sugarcane cultivar 'R570'. Theoretical and Applied Genetics, 92(8), 1059-1064. doi: 10.1007/BF00224049

Deleris, A., Gallego-Bartolome, J., Bao, J., Kasschau, K. D., Carrington, J. C., & Voinnet, O. (2006). Hierarchical action and inhibition of plant Dicer-like proteins in antiviral defense. Science, 313(5783), 68-71. doi: 10.1126/science.1128214

DeYoung, B. J., & Innes, R. W. (2006). Plant NBS-LRR proteins in pathogen sensing and host defense. Nature Immunology, 7(12), 1243-1249. doi: 10.1038/ni1410

D'Hont, A., Ison, D., Alix, K., Roux, C., and Glaszmann, J. C. (1998). Determination of basic chromosome numbers in the genus Saccharum by physical mapping of ribosomal RNA genes. Genome 41, 221-225. doi: 10.1139/g98-023

D'Hont, A., and Glaszmann, J. C. (2001). Sugarcane genome analysis with molecular markers: a first decade of research, in Proceedings of the International Society of Sugarcane Technologists XXVI Congress, ed. D. M. Hogarth (Brisbane, Australia: The Australian Society of Sugar Cane Technologists), 556-559.

De Bem Oliveira, I., Resende, M. F., Ferrão, L. F. V., Amadeu, R. R., Endelman, J. B., Kirst, M.... & Munoz, P. R. (2019). Genomic prediction of autotetraploids; influence of relationship matrices, allele dosage, and continuous genotyping calls in phenotype prediction. G3: Genes, Genomes, Genetics, 9(4), 1189-1198. doi: 10.1534/g3.119.400059

De Mendiburu, F., & De Mendiburu, M. F. (2020). Package 'agricolae'. R package version, 1-2.

Débibakas, S., Rocher, S., Garsmeur, O., Toubi, L., Roques, D., D'Hont, A., et al. (2014). Prospecting sugarcane resistance to sugarcane yellow leaf virus by genome-wide association. Theor. Appl. Genet. 127, 1719-1732. doi: 10.1007/s00122-014-2334-7

Delage, C., Rippoles, M., Chatenet, M., Irey, M., and Rott, P. (1999). Elimination of sugarcane yellow leaf virus from sugarcane by meristem tip culture, in Proceedings of the XXIII Congress of the International Society of Sugar Cane Technologists. New Delhi, India: ISSCT Congress (Vol. 404).

Devlin, B., and Roeder, K. (1999). Genomic control for association studies. Biometrics 55, 997-1004. doi: 10.1111/j.0006-341x.1999.00997.x

Dinesh-Kumar, S. P., Tham, W. H., & Baker, B. J. (2000). Structure–function analysis of the tobacco mosaic virus resistance gene N. Proceedings of the National Academy of Sciences, 97(26), 14789-14794. doi: 10.1073/pnas.97.26.14789

Dinno, A. (2017). dunn.test: Dunn's test of multiple comparisons using rank sums. R package version, 1(4).

Dubey, N., & Singh, K. (2018). Role of NBS-LRR proteins in plant defense. In Molecular Aspects of Plant-Pathogen Interaction (pp. 115-138). Springer, Singapore. doi: 10.1007/978-981-10-7371-7_5

Earl, D. A., and vonHoldt, B. M. (2012). Structure harvester: a website and program for visualizing structure output and implementing the Evanno method. Conserv. Genet. Resour. 4, 359-361. doi: 10.1007/s12686-011-9548-7

Earley, K., Smith, M., Weber, R., Gregory, B., and Poethig, R. (2010). An endogenous F-box protein regulates ARGONAUTE1 in *Arabidopsis thaliana*. Silence 1, 15. doi: 10.1186/1758-907X-1-15

ElSayed, A. I., Komor, E., Boulila, M., Viswanathan, R., and Odero, D. C. (2015). Biology and management of sugarcane yellow leaf virus: an historical overview. Arch. Virol. 160, 2921-2934. doi: 10.1007/s00705-015-2618-5

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., and Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS one, 6(5), e19379. doi: 10.1371/journal.pone.0019379

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: a simulation study. Mol. Ecol. 14, 2611-2620. doi: 10.1111/j.1365-294x.2005.02553.x

Fadista, J., Manning, A. K., Florez, J. C., and Groop, L. (2016). The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. Eur. J. Hum. Genet. 24, 1202-1205. doi: 10.1038/ejhg.2015.269

Fang, J., Twito, T., Zhang, Z., & Chao, C. T. (2006). Genetic relationships among fruiting-mei (*Prunus mume* Sieb. et Zucc.) cultivars evaluated with AFLP and SNP markers. Genome, 49(10), 1256-1264. doi: 10.1139/g06-097

FAO (2020). FAOSTAT: Production Sheet. Rome: FAO.

Ferrão, L. F. V., Benevenuto, J., Oliveira, I. D. B., Cellon, C., Olmstead, J., Kirst, M., et al. (2018). Insights into the genetic basis of blueberry fruit-related traits using diploid and polyploid models in a GWAS context. Front. Ecol. Evol. 6, 107. doi: 10.3389/fevo.2018.00107

Fickett, N., Gutierrez, A., Verma, M., Pontif, M., Hale, A., Kimbeng, C., & Baisakh, N. (2019). Genome-wide association mapping identifies markers associated with cane yield components

and sucrose traits in the Louisiana sugarcane core collection. Genomics, 111(6), 1794-1801. doi: 10.1016/j.ygeno.2018.12.002

Fitch, M. M. M., Lehrer, A. T., Komor, E., and Moore, P. H. (2001). Elimination of sugarcane yellow leaf virus from infected sugarcane plants by meristem tip culture visualized by tissue blot immunoassay. Plant Pathol. 50, 676-680. doi: 10.1046/j.1365-3059.2001.00639.x

Foresman, B. J., Oliver, R. E., Jackson, E. W., Chao, S., Arruda, M. P., and Kolb, F. L. (2016). Genome-wide association mapping of barley yellow dwarf virus tolerance in spring oat (*Avena sativ*a L.). PLoS One 11, e0155376. doi: 10.1371/journal.pone.0155376

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences, 55(1), 119-139. doi: 10.1006/jcss.1997.1504

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. Machine Learning, 29(2-3), 131-163. doi: 10.1023/A:1007465528199

Gerard, D. (2020). Pairwise Linkage Disequilibrium Estimation for Polyploids. bioRxiv. doi: 10.1101/2020.08.03.234476.

Gao, X., Becker, L. C., Becker, D. M., Starmer, J. D., and Province, M. A. (2010). Avoiding the high Bonferroni penalty in genome-wide association studies. Genet. Epidemiol. 34, 100-105. doi: 10.1002/gepi.20430

Garcia, A. A. F., Mollinari, M., Marconi, T. G., Serang, O. R., Silva, R. R., Vieira, M. L. C., et al. (2013). SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. Sci. Rep. 3, 3399. doi: 10.1038/srep03399

Garsmeur, O., Droc, G., Antonise, R., Grimwood, J., Potier, B., Aitken, K., et al. (2018). A mosaic monoploid reference sequence for the highly complex genome of sugarcane. Nat. Commun. 9, 2638. doi: 10.1038/s41467-018-05051-5

Gaudeul, M., Till-Bottraud, I., Barjon, F., & Manel, S. (2004). Genetic diversity and differentiation in *Eryngium alpinum* L. (Apiaceae): comparison of AFLP and microsatellite markers. Heredity, 92(6), 508-518. doi: 10.1038/sj.hdy.6800443

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine Learning, 63(1), 3-42. doi: 10.1007/s10994-006-6226-1

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., et al. (2014). TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS One 9, e90346. doi: 10.1371/journal.pone.0090346

Gonçalves, M.C.; Klerks, M.M.; Verbeek, M.; Vega, J.; Van Den Heuvel, J.F.J.M. (2002). The use of molecular beacons combined with NASBA for the sensitive detection of Sugarcane yellow leaf virus. Eur. J. Plant Pathol. 08, 401-407. doi: 10.1023/a:1016040314260.

Gonçalves, M. C. (2007). Doenças causadas por vírus, in Cana-de-açúcar, eds. L. L. Dinardo-Miranda, A. C. M. Vasconcelos, and M. G. A. Landell. (Campinas: Instituto Agronômico de Campinas), 150-165.

Gonçalves, M. C., Pinto, L. R., Souza, S. C., and Landell, M. G. A. (2012). Virus diseases of sugarcane: a constant challenge to sugarcane breeding in Brazil. Funct. Plant Sci. Biotechnol. 6, 108-116.

Gonçalves, M. C., Vega, J., Oliveira, J. G., and Gomes, M. M. A. (2005). Sugarcane yellow leaf virus infection leads to alterations in photosynthetic efficiency and carbohydrate accumulation in sugarcane leaves. Fitopatol. Bras. 30, 10-16. doi: 10.1590/s0100-41582005000100002

Gouy, M., Rousselle, Y., Chane, A. T., Anglade, A., Royaert, S., Nibouche, S., et al. (2015). Genome wide association mapping of agro-morphological and disease resistance traits in sugarcane. Euphytica 202, 269-284. doi: 10.1007/s10681-014-1294-y

Grativol, C., Regulski, M., Bertalan, M., McCombie, W. R., da Silva, F. R., Zerlotini Neto, A., et al. (2014). Sugarcane genome sequencing by methylation filtration provides tools for genomic research in the genus *Saccharum*. Plant J. 79, 162-172. doi: 10.1111/tpj.12539

Grisham, M., Pan, Y., Legendre, B., Godshall, M., and Eggleston, G. (2001). Effect of sugarcane yellow leaf virus on sugarcane yield and juice quality. Proc. Int. Soc. Sugar Cane Technol. 24, 434-438

Harries, P. A., Park, J. W., Sasaki, N., Ballard, K. D., Maule, A. J., & Nelson, R. S. (2009). Differing requirements for actin and myosin by plant viruses for sustained intercellular movement. Proceedings of the National Academy of Sciences, 106(41), 17594-17599. doi: 10.1073/pnas.0909239106

Heer, K., Behringer, D., Piermattei, A., Bässler, C., Brandl, R., Fady, B.... & Vendramin, G. G. (2018). Linking dendroecology and association genetics in natural populations: Stress responses archived in tree rings associate with SNP genotypes in silver fir (*Abies alba* Mill.). Molecular Ecology, 27(6), 1428-1438. doi: 10.1111/mec.14538

Heinlein, M. (2015). Plant virus replication and movement. Virology, 479, 657-671. doi: 10.1016/j.virol.2015.01.025

Heo, W. D., Lee, S. H., Kim, M. C., Kim, J. C., Chung, W. S., Chun, H. J., et al. (1999). Involvement of specific calmodulin isoforms in salicylic acid-independent activation of plant disease resistance responses. Proc. Natl. Acad. Sci. U. S. A. 96, 766-771. doi: 10.1073/pnas.96.2.766

Hill, W. G., & Weir, B. S. (1988). Variances and covariances of squared linkage disequilibria in finite populations. Theoretical Population Biology, 33(1), 54-78. doi: 10.1016/0040-5809(88)90004-4

Hoang, N. V., Furtado, A., Mason, P. J., Marquardt, A., Kasirajan, L., Thirugnanasambandam, P. P., et al. (2017). A survey of the complex transcriptome from the highly polyploid sugarcane genome using full-length isoform sequencing and de novo assembly from short read sequencing. BMC Genom. 18, 395-395. doi: 10.1186/s12864-017-3757-8

Hofius, D., Maier, A. T., Dietrich, C., Jungkunz, I., Börnke, F., Maiss, E., & Sonnewald, U. (2007). Capsid protein-mediated recruitment of host DnaJ-like proteins is required for Potato virus Y infection in tobacco plants. Journal of Virology, 81(21), 11870-11880. doi: 10.1128/JVI.01525-07

Huang, T. S., Wei, T., Laliberté, J. F., and Wang, A. (2010). A host RNA helicase-like protein, AtRH8, interacts with the potyviral genome-linked protein, VPg, associates with the virus accumulation complex, and is essential for infection. Plant Physiology, 152(1), 255-266. doi: 10.1104/pp.109.147983

Ibrahim, A., Yang, X., Liu, C., Cooper, K. D., Bishop, B. A., Zhu, M.... & Nelson, R. S. (2020). Plant SNAREs SYP22 and SYP23 interact with Tobacco mosaic virus 126 kDa protein and SYP2s are required for normal local virus accumulation and spread. Virology, 547, 57-71. doi: 10.1016/j.virol.2020.04.002

Islam, M. S., Yang, X., Sood, S., Comstock, J. C., and Wang, J. (2018). Molecular characterization of genetic basis of sugarcane yellow leaf virus (SCYLV) resistance in *Saccharum* spp. hybrid. Plant Breed. 137, 598-604. doi: 10.1111/pbr.12614

ISO (2020). International Sugar Organization. London: ISO.

Jannoo, N., Grivet, L., Dookun, A., D'Hont, A., & Glaszmann, J. C. (1999). Linkage disequilibrium among modern sugarcane cultivars. Theoretical and Applied Genetics, 99(6), 1053-1060. doi: 10.1007/s001220051414

Jarošová, J., Gadiou, S., and Kumar, J. K. (2009). Real-time RT-PCR quantitative analysis of plant viruses in stone fruit tissues. Julius-Kühn-Archiv, 61

Johnson, R. C., Nelson, G. W., Troyer, J. L., Lautenberger, J. A., Kessing, B. D., Winkler, C. A., et al. (2010). Accounting for multiple comparisons in a genome-wide association study (GWAS). BMC Genom. 11, 724. doi: 10.1186/1471-2164-11-724

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24, 1403-1405. doi: 10.1093/bioinformatics/btn129

Kassambara, A., and Mundt, F. (2017). Factoextra: extract and visualize the results of multivariate data analyses. R Package Version 1, 2017

Knaus, B. J., and Grünwald, N. J. (2017). vcfr: a package to manipulate and visualize variant call format data in R. Mol. Ecol. Resour. 17, 44-53. doi: 10.1111/1755-0998.12549

Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., and Mayrose, I. (2015). Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Mol. Ecol. Resour. 15, 1179-1191. doi: 10.1111/1755-0998.12387

Korimbocus, J., Coates, D., Barker, I., and Boonham, N. (2002). Improved detection of sugarcane yellow leaf virus using a real-time fluorescent (TaqMan) RT-PCR assay. J. Virol. Methods 103, 109-120. doi: 10.1016/s0166-0934(01)00406-2

Kovalev, N., Pogany, J., and Nagy, P. D. (2012). A co-opted DEAD-box RNA helicase enhances tombusvirus plus-strand synthesis. PLoS Pathogens, 8(2), e1002537. doi: 10.1371/journal.ppat.1002537

Lalman, J. A., Shewa, W. A., Gallagher, J., and Ravella, S. (2016). Biofuels production from renewable feedstocks, in Quality Living Through Chemurgy and Green Chemistry, ed. P. C. K. Lau. (Berlin, Heidelberg: Springer), 193-220. doi: 10.1007/978-3-662-53704-6_8

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357-359. doi: 10.1038/nmeth.1923

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: an R package for multivariate analysis. J. Stat. Softw. 25, 1-18. doi: 10.18637/jss.v025.i01

Lehrer, A. T., Moore, P. H., and Komor, E. (2007). Impact of sugarcane yellow leaf virus (SCYLV) on the carbohydrate status of sugarcane: comparison of virus-free plants with symptomatic and asymptomatic virus-infected plants. Physiological and Molecular Plant Pathology, 70(4-6), 180-188. doi: 10.1016/j.pmpp.2007.09.005

Lehrer, A., Yan, S.-L., Fontaniella, B., ElSayed, A., and Komor, E. (2010). Carbohydrate composition of sugarcane cultivars that are resistant or susceptible to sugarcane yellow leaf virus. J. Gen. Plant Pathol. 76, 62-68. doi: 10.1007/s10327-009-0210-0

Lehrer, A. T., and Komor, E. (2008). Symptom expression of yellow leaf disease in sugarcane cultivars with different degrees of infection by sugarcane yellow leaf virus. Plant Pathol. 57, 178-189. doi: 10.1111/j.1365-3059.2007.01696.x

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England) 25, 1754-1760. doi: 10.1093/bioinformatics/btp324

Li, F., Huang, C., Li, Z., and Zhou, X. (2014). Suppression of RNA silencing by a plant DNA virus satellite requires a host calmodulin-like protein to repress RDR6 expression. PLoS Pathog. 10, e1003921. doi: 10.1371/journal.ppat.1003921

Li, Y., Xiong, R., Bernards, M., and Wang, A. (2016). Recruitment of *Arabidopsis* RNA helicase AtRH9 to the viral replication complex by viral replicase to promote turnip mosaic virus replication. Scientific Reports, 6, 30297. doi: 10.1038/srep30297

Lin, W. C., Lu, C. F., Wu, J. W., Cheng, M. L., Lin, Y. M., Yang, N. S.... & Cheng, C. P. (2004). Transgenic tomato plants expressing the Arabidopsis NPR1 gene display enhanced resistance to a spectrum of fungal and bacterial diseases. Transgenic Research, 13(6), 567-581. doi: 10.1007/s11248-004-2375-9

Liu, Y., Schiff, M., Marathe, R., & Dinesh-Kumar, S. P. (2002). Tobacco Rar1, EDS1 and NPR1/NIM1 like genes are required for N-mediated resistance to tobacco mosaic virus. The Plant Journal, 30(4), 415-429. doi: 10.1046/j.1365-313x.2002.01297.x

Liu, J. Z., & Whitham, S. A. (2013). Overexpression of a soybean nuclear localized type–III DnaJ domain-containing HSP40 reveals its roles in cell death and disease resistance. The Plant Journal, 74(1), 110-121. doi: 10.1111/tpj.12108

Liu, X., Huang, M., Fan, B., Buckler, E. S., & Zhang, Z. (2016). Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. PLoS Genetics, 12(2), e1005767. doi: 10.1371/journal.pgen.1005767

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. Methods 25, 402-408. doi: 10.1006/meth.2001.1262

Llave, C. (2010). Virus-derived small interfering RNAs at the core of plant–virus interactions. Trends in Plant Science, 15(12), 701-707. doi: 10.1016/j.tplants.2010.09.001

Lu, L., Du, Z., Qin, M., Wang, P., Lan, H., Niu, X.... & Wu, Z. (2009). Pc4, a putative movement protein of Rice stripe virus, interacts with a type I DnaJ protein and a small Hsp of rice. Virus Genes, 38(2), 320-327. doi: 10.1007/s11262-008-0324-z

Lundin, S., Stranneheim, H., Pettersson, E., Klevebring, D., and Lundeberg, J. (2010). Increased throughput by parallelization of library preparation for massive sequencing. PLoS One 5, e10029. doi: 10.1371/journal.pone.0010029

Maccheroni, W., Jordão, H., Degaspari, R., & Matsuoka, S. (2007). Development of a dependable microsatellite-based fingerprinting system for sugarcane. Proceedings of the International Society of Sugar Cane Technologistics, Durban.

Mancini, M. C., Cardoso-Silva, C. B., Sforça, D. A., and Pereira de Souza, A. (2018). "Targeted sequencing by gene synteny," a new strategy for polyploid species: sequencing and physical structure of a complex sugarcane region. Front. Plant Sci. 9, 397. doi: 10.3389/fpls.2018.00397

Mangwende, T., Wang, M.-L., Borth, W., Hu, J., Moore, P. H., Mirkov, T. E., et al. (2009). The P0 gene of sugarcane yellow leaf virus encodes an RNA silencing suppressor with unique activities. Virology 384, 38-50. doi: 10.1016/j.virol.2008.10.034

Mansilla-Córdova, P. J., Bampi, D., Rondinel-Mendoza, N. V., Melo, P. C. T., Lourenção, A. L., and Rezende, J. A. M. (2018). Screening tomato genotypes for resistance and tolerance to Tomato chlorosis virus. Plant Pathol. 67, 1231-1237. doi: 10.1111/ppa.12826

Marconi, T. G., Costa, E. A., Miranda, H. R., Mancini, M. C., Cardoso-Silva, C. B., Oliveira, K. M.... & Souza, A. P. (2011). Functional markers for gene mapping and genetic diversity studies in sugarcane. BMC Research Notes, 4(1), 264. doi: 10.1186/1756-0500-4-264

Matias, F. I., Alves, F. C., Meireles, K. G. X., Barrios, S. C. L., do Valle, C. B., Endelman, J. B., & Fritsche-Neto, R. (2019). On the accuracy of genomic prediction models considering multi-trait and allele dosage in *Urochloa* spp. interspecific tetraploid hybrids. Molecular Breeding, 39(7), 100. doi: 10.1007/s11032-019-1002-7

Melo, A. T. O., Bartaula, R., and Hale, I. (2016). GBS-SNP-CROP: a reference-optional pipeline for SNP discovery and plant germplasm characterization using variable length, paired-end genotyping-by-sequencing data. BMC Bioinform. 17, 29. doi: 10.1186/s12859-016-0879-y

Moonan, F., Molina, J., and Mirkov, T. E. (2000). Sugarcane yellow leaf virus: an emerging virus that has evolved by recombination between luteoviral and poleroviral ancestors. Virology 269, 156-171. doi: 10.1006/viro.1999.0162

Morel, J.-B., Godon, C., Mourrain, P., Béclin, C., Boutet, S., Feuerbach, F., et al. (2002). Fertile hypomorphic ARGONAUTE (ago1) mutants impaired in post-transcriptional gene silencing and virus resistance. Plant Cell 14, 629-639. doi: 10.1105/tpc.010358

Muñoz, F., and Sanchez, L.S. (2014). breedR: statistical methods for forest genetic resources analysis. In Trees for the future: plant material in a changing climate.

Nakahara, K. S., Masuta, C., Yamada, S., Shimura, H., Kashihara, Y., Wada, T. S., et al. (2012). Tobacco calmodulin-like protein provides secondary defense by binding to and directing degradation of virus RNA silencing suppressors. Proc. Natl. Acad. Sci. U. S. A. 109, 10113-10118. doi: 10.1073/pnas.1201628109

Nimmakayala, P., Tomason, Y. R., Abburi, V. L., Alvarado, A., Saminathan, T., Vajja, V. G., et al. (2016). Genome-wide differentiation of various melon horticultural groups for use in GWAS for fruit firmness and construction of a high resolution genetic map. Front. Plant Sci. 7, 1437. doi: 10.3389/fpls.2016.01437

Oliveira, K. M., Pinto, L. R., Marconi, T. G., Margarido, G. R., Pastina, M. M., Teixeira, L. H. M.... & Souza, A. P. (2007). Functional integrated genetic linkage map based on EST-markers for a sugarcane (*Saccharum* spp.) commercial cross. Molecular Breeding, 20(3), 189-208. doi: 10.1007/s11032-007-9082-1

Oliveira, K. M., Pinto, L. R., Marconi, T. G., Mollinari, M., Ulian, E. C., Chabregas, S. M.... & Souza, A. P. (2009). Characterization of new polymorphic functional markers for sugarcane. Genome, 52(2), 191-209. doi: 10.1139/g08-105

Panje, R. R., and Babu, C. N. (1960). Studies in Saccharum spontaneum distribution and geographical association of chromosome numbers. Cytologia 25, 152-172. doi: 10.1508/cytologia.25.152

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., et al. (2009). The *Sorghum bicolor* genome and the diversification of grasses. Nature 457, 551-556. doi: 10.1038/nature07723

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O.... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research, 12, 2825-2830.

Pereira, G. S., Garcia, A. A. F., and Margarido, G. R. A. (2018). A fully automated pipeline for quantitative genotype calling from next generation sequencing data in autopolyploids. BMC Bioinform. 19, 398. doi: 10.1186/s12859-018-2433-6

Peterson, R. (2017). Estimating normalization transformations with bestNormalize. https://github.com/petersonR/bestNormalize. [Accessed September 2, 2020].

Pieterse, C. M., Van Wees, S. C., Van Pelt, J. A., Knoester, M., Laan, R., Gerrits, H.... & Van Loon, L. C. (1998). A novel signaling pathway controlling induced systemic resistance in Arabidopsis. The Plant Cell, 10(9), 1571-1580. doi: 10.1105/tpc.10.9.1571

Poland, J. A., Brown, P. J., Sorrells, M. E., and Jannink, J.-L. (2012). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7, e32253. doi: 10.1371/journal.pone.0032253

Popescu, M. C., Balas, V. E., Perescu-Popescu, L., & Mastorakis, N. (2009). Multilayer perceptron and neural networks. WSEAS Transactions on Circuits and Systems, 8(7), 579-588.

Premachandran, M. N., Prathima, P. T., and Maya, L. (2011). Sugarcane and polyploidy - a review. J. Sugarcane Res. 1, 1-15

Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81-106. doi: 10.1007/BF00116251

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics 155, 945-959

R Core Team (2011). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Raboin, L. M., Oliveira, K. M., Lecunff, L., Telismart, H., Roques, D., Butterfield, M.... & D‘Hont, A. (2006). Genetic mapping in sugarcane, a high polyploid, using bi-parental progeny: identification of a gene controlling stalk colour and a new rust resistance gene. Theoretical and Applied Genetics, 112(7), 1382-1391. doi: 10.1007/s00122-006-0240-3

Raboin, L. M., Pauquet, J., Butterfield, M., D’Hont, A., & Glaszmann, J. C. (2008). Analysis of genome-wide linkage disequilibrium in the highly polyploid sugarcane. Theoretical and Applied Genetics, 116(5), 701-714. doi: 10.1371/journal.pbio.0030196

Racedo, J., Gutiérrez, L., Perera, M. F., Ostengo, S., Pardo, E. M., Cuenya, M. I., ... and Castagnaro, A. P. (2016). Genome-wide association mapping of quantitative traits in a breeding population of sugarcane. BMC Plant Biology, 16(1), 142. doi: 10.1186/s12870-016-0829-x

Rasmussen, C. E. (2003). Gaussian processes in machine learning. In Summer School on Machine Learning (pp. 63-71). Springer, Berlin, Heidelberg. doi: 10.1007/978-3-540-28650-9_4

Riaño-Pachón, D. M., and Mattiello, L. (2017). Draft genome sequencing of the sugarcane hybrid SP80-3280. F1000Research 6, 861. doi: 10.12688/f1000research.11859.2

Roach, B. T. (1972). Nobilisation of sugarcane. Proc. Int. Soc. Sugar Cane Technol. 14, 206-216

Rochon, A., Boyle, P., Wignes, T., Fobert, P. R., & Després, C. (2006). The coactivator function of Arabidopsis NPR1 requires the core of its BTB/POZ domain and the oxidation of C-terminal cysteines. The Plant Cell, 18(12), 3670-3685. doi: 10.1105/tpc.106.046953

Roncallo, P. F., Beaufort, V., Larsen, A. O., Dreisigacker, S., & Echenique, V. (2019). Genetic diversity and linkage disequilibrium using SNP (KASP) and AFLP markers in a worldwide durum wheat (*Triticum turgidum* L. var *durum*) collection. PloS One, 14(6), e0218562. doi: 10.1371/journal.pone.0218562

Rosyara, U. R., De Jong, W. S., Douches, D. S., and Endelman, J. B. (2016). Software for genome-wide association studies in autopolyploids and its application to potato. Plant Genome 9, 1-10. doi: 10.3835/plantgenome2015.08.0073

Sasvari, Z., Kovalev, N., Gonzalez, P. A., Xu, K., & Nagy, P. D. (2018). Assembly-hub function of ER-localized SNARE proteins in biogenesis of tombusvirus replication compartment. PLoS Pathogens, 14(5), e1007028. doi: 10.1371/journal.ppat.1007028

Scagliusi, S. M., and Lockhart, B. E. L. (2000). Transmission, characterization, and serology of a luteovirus associated with yellow leaf syndrome of sugarcane. Phytopathology 90, 120-124. doi: 10.1094/phyto.2000.90.2.120

Scagliusi, S. M., Basu, S. K., de Gouvea, J. A., & Vega, J. (2009). Physiological Alterations in Brazilian Sugarcane Varieties Infected by Sugarcane yellow leaf virus (ScYLV). Functional Plant Science and Biotechnology, 3(1), 19-25.

Schloerke, B., Crowley, J., Cook, D., Hofmann, H., Wickham, H., Briatte, F., et al. (2011). Ggally: Extension to ggplot2.

Seo, Y. S., Rojas, M. R., Lee, J. Y., Lee, S. W., Jeon, J. S., Ronald, P.... & Gilbertson, R. L. (2006). A viral resistance gene from common bean functions across plant families and is up-regulated in a non-virus-specific manner. Proceedings of the National Academy of Sciences, 103(32), 11856-11861. doi: 10.1073/pnas.0604815103

Sforça, D. A., Vautrin, S., Cardoso-Silva, C. B., Mancini, M. C., Romero-da Cruz, M. V., Pereira, G. d. S., et al. (2019). Gene duplication in the sugarcane genome: a case study of allele interactions and evolutionary patterns in two genic regions. Front. Plant Sci. 10, 553. doi: 10.3389/fpls.2019.00553

Shimizu, T., Yoshii, A., Sakurai, K., Hamada, K., Yamaji, Y., Suzuki, M.... & Hibi, T. (2009). Identification of a novel tobacco DnaJ-like protein that interacts with the movement protein of tobacco mosaic virus. Archives of Virology, 154(6), 959-967. doi: 10.1007/s00705-009-0397-6

Singh, R. K., Banerjee, N., Khan, M. S., Yadav, S., Kumar, S., Duttamajumder, S. K.... and Paterson, A. H. (2016). Identification of putative candidate genes for red rot resistance in sugarcane (*Saccharum* species hybrid) using LD-based association mapping. Molecular Genetics and Genomics, 291(3), 1363-1377. doi: 10.1007/s00438-016-1190-3

Slater, A. T., Cogan, N. O. I., Forster, J. W., Hayes, B. J., and Daetwyler, H. D. (2016). Improving genetic gain with genomic selection in autotetraploid potato. Plant Genome 9, 1-15. doi: 10.3835/plantgenome2016.02.0021

Smith, G. R., Borg, Z., Braithwaite, K. S., Lockhart, B. E. L., and Gibbs, M. J. (2000). Sugarcane yellow leaf virus: a novel member of the Luteoviridae that probably arose by inter-species recombination. J. Gen. Virol. 81, 1865-1869. doi: 10.1099/0022-1317-81-7-1865

Spoel, S. H., Koornneef, A., Claessens, S. M., Korzelius, J. P., Van Pelt, J. A., Mueller, M. J.... & Van Loon, L. C. (2003). NPR1 modulates cross-talk between salicylate-and jasmonate-dependent defense pathways through a novel function in the cytosol. The Plant Cell, 15(3), 760-770. doi: 10.1105/tpc.009159

Stacklies, W., Redestig, H., Scholz, M., Walther, D., & Selbig, J. (2007). pcaMethods – a bioconductor package providing PCA methods for incomplete data. Bioinformatics, 23(9), 1164-1167. doi: 10.1093/bioinformatics/btm069

Steinfath, M., Gärtner, T., Lisec, J., Meyer, R. C., Altmann, T., Willmitzer, L., & Selbig, J. (2010). Prediction of hybrid biomass in *Arabidopsis thaliana* by selected parental SNP and metabolic markers. Theoretical and Applied Genetics, 120(2), 239-247. doi: 10.1007/s00122-009-1191-2

Su, J., Fan, S., Li, L., Wei, H., Wang, C., Wang, H., et al. (2016). Detection of favorable QTL alleles and candidate genes for lint percentage by GWAS in Chinese Upland cotton. Front. Plant Sci. 7, 1576. doi: 10.3389/fpls.2016.01576

Thiel, H., Hleibieh, K., Gilmer, D., and Varrelmann, M. (2012). The P25 pathogenicity factor of beet necrotic yellow vein virus targets the sugar beet 26S proteasome involved in the induction of a hypersensitive resistance response via interaction with an F-box protein. Mol. Plant-Microbe Interact. 25, 1058-1072. doi: 10.1094/mpmi-03-12-0057-r

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. J. Dairy Sci. 91, 4414-4423. doi: 10.3168/jds.2007-0980

Vasconcelos, A., Gonçalves, M. C., Pinto, L. R., Landell, M. G., and Perecin, D. (2009). Effects of sugarcane yellow leaf virus on sugarcane yield and root system development. Funct. Plant Sci. Biotechnol. 3, 31-35

Vega, J., Scagliusi, S. M. M., and Ulian, E. C. (1997). Sugarcane yellow leaf disease in Brazil: evidence of association with a luteovirus. Plant Dis. 81, 21-26. doi: 10.1094/pdis.1997.81.1.21

Villavicencio, R. C. B., Gonçalves, M. C., Nóbile, P. M., Anjos, I. A., Silva, M. F., Perecin, D., Anjos, L. S., Fernandes, V. B., Pinto, L. R. R. (in press) Screening of *Saccharum* spp. genotypes for sugarcane yellow leaf virus resistance by combining symptom phenotyping and highly precise virus titration. Crop Protection.

Viswanathan, R., Chinnaraja, C., Malathi, P., Gomathi, R., Rakkiyappan, P., Neelamathi, D., et al. (2014). Impact of sugarcane yellow leaf virus (SCYLV) infection on physiological efficiency and growth parameters of sugarcane under tropical climatic conditions in India. Acta Physiol. Plant. 36, 1805-1822. doi: 10.1007/s11738-014-1554-4

Voorrips, R. E. (2002). MapChart: software for the graphical presentation of linkage maps and QTLs. Journal of Heredity, 93(1), 77-78. doi: 10.1093/jhered/93.1.77

Vos, P. G., Paulo, M. J., Voorrips, R. E., Visser, R. G., van Eck, H. J., & van Eeuwijk, F. A. (2017). Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. Theor. App. Genet. 130, 123-135. doi: 10.1007/s00122-016-2798-8

Wang, H., Smith, K. P., Combs, E., Blake, T., Horsley, R. D., and Muehlbauer, G. J. (2012). Effect of population size and unbalanced data sets on QTL detection using genome-wide association mapping in barley breeding germplasm. Theor. Appl. Genet. 124, 111-124. doi: 10.1007/s00122-011-1691-8

Wei, X., Jackson, P. A., Hermann, S., Kilian, A., Heller-Uszynska, K., and Deomano, E. (2010). Simultaneously accounting for population structure, genotype by environment interaction, and spatial variation in marker–trait associations in sugarcane. Genome, 53(11), 973-981. doi: 10.1139/G10-050

Wei, T., Zhang, C., Hou, X., Sanfaçon, H., & Wang, A. (2013). The SNARE protein Syp71 is essential for turnip mosaic virus infection by mediating fusion of virus-induced vesicles with chloroplasts. PLoS Pathogens, 9(5), e1003378. doi: 10.1371/journal.ppat.1003378

Xun, H., Yang, X., He, H., Wang, M., Guo, P., Wang, Y.... & Liu, B. (2019). Over-expression of GmKR3, a TIR–NBS–LRR type R gene, confers resistance to multiple viruses in soybean. Plant Molecular Biology, 99(1-2), 95-111. doi: 10.1007/s11103-018-0804-z

Yang, N., Lu, Y., Yang, X., Huang, J., Zhou, Y., Ali, F., et al. (2014). Genome wide association studies using a new nonparametric model reveal the genetic architecture of 17 agronomic traits in an enlarged maize association panel. PLoS Genet. 10, e1004573-e1004573. doi: 10.1371/journal.pgen.1004573

Yang, Z., Li, Y. (2018). Dissection of RNAi-based antiviral immunity in plants. Current Opinion in Virology, 32, 88-99. doi: 10.1016/j.coviro.2018.08.003

Yang, X., Sood, S., Luo, Z., Todd, J., and Wang, J. (2019a). Genome-wide association studies identified resistance loci to orange rust and yellow leaf virus diseases in sugarcane (*Saccharum* spp.). Phytopathology 109, 623-631. doi: 10.1094/phyto-08-18-0282-r

Yang, X., Todd, J., Arundale, R., Binder, J. B., Luo, Z., Islam, M. S., et al. (2019b). Identifying loci controlling fiber composition in polyploid sugarcane (*Saccharum* spp.) through genome-wide association study. Ind. Crops Prod. 130, 598-605. doi: 10.1016/j.indcrop.2019.01.023

Yang, X., Song, J., Todd, J., Peng, Z., Paudel, D., Luo, Z.... and Zhao, Y. (2019c). Target enrichment sequencing of 307 germplasm accessions identified ancestry of ancient and modern hybrids and signatures of adaptation and selection in sugarcane (*Saccharum* spp.), a 'sweet'crop with 'bitter'genomes. Plant Biotechnology Journal, 17(2), 488-498. doi: 10.1111/pbi.12992

Yang, X., Luo, Z., Todd, J., Sood, S., & Wang, J. (2020). Genome-wide association study of multiple yield traits in a diversity panel of polyploid sugarcane (*Saccharum* spp.). The Plant Genome, e20006. doi: 10.1002/tpg2.20006

Zhang, J., Zhang, X., Tang, H., Zhang, Q., Hua, X., Ma, X., et al. (2018). Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. Nat. Genet. 50, 1565-1573. doi: 10.1038/s41588-018-0237-2

Zhou, W., & Bellis, E. S. (2019). Minor QTLs mining through the combination of GWAS and machine learning feature selection. BioXriv. doi: 10.1101/702761

Zhu, Y. J., Lim, S. T. S., Schenck, S., Arcinas, A., and Komor, E. (2010). RT-PCR and quantitative real-time RT-PCR detection of sugarcane yellow leaf virus (SCYLV) in symptomatic and asymptomatic plants of Hawaiian sugarcane cultivars and the correlation of SCYLV titre to yield. Eur. J. Plant Pathol. 127, 263-273. doi: 10.1007/s10658-010-9591-3

Figure Legends

Figure 1. Distributions of BLUP values and correlations between the three traits analyzed. Asterisks (***) indicate statistical significance ($p < 0.0005$).

Figure 2. Factorial maps generated in the two hierarchical clustering on principal components (HPCP) analyses using BLUP values. (A) Factorial map of HCPCs performed using the SCYLV titer determined by RT-qPCR. A division into three clusters (Q1, Q2 and Q3) was considered. **(B)** Factorial map of HCPC performed using SCYL symptom severity and SCYLV titer determined by RT-qPCR. A division into three clusters (SQ1, SQ2 and SQ3) was considered.

Figure 3: Graphical outputs of the three clustering analyses employed to evaluate the genetic structure for 93 genotypes of the panel using 662 amplified fragment length polymorphisms (AFLPs) and simple sequence repeats (SSRs). (A) Plotting of the panel onto the first two linear discriminants from a discriminant analysis of principal components (DAPC). A division into three clusters (A1, A2 and A3) was considered. **(B)** Projection of the panel onto the first two principal components (PCs) from a principal component analysis (PCA). Genotypes are divided into four clusters (B1, B2, B3 and B4) as indicated by a k-means clustering analysis. Ellipses represent a 95% confidence interval. **(C)** Bayesian clustering performed on STRUCTURE. Each genotype is represented by a vertical bar split in accordance with the admixture coefficients for two clusters (C1 in dark blue and C2 in red). Genotypes are ordered according to ID number (see Supplementary Table 1).

Figure 4. Graphical outputs of the three clustering analyses employed to evaluate the genetic structures for 92 genotypes in the panel using codominant markers. (A) Plotting of the panel onto the first two linear discriminants from a discriminant analysis of principal components (DAPC) using 70,888 single-nucleotide polymorphisms (SNPs) and insertions and deletions (indels). A division into three clusters (D1, D2 and D3) was considered. **(B)** Projection of the panel onto the first two principal components (PCs) from a principal component analysis (PCA) using 70,888 SNPs and indels. The genotypes were divided into two clusters (E1 and E2) as indicated by k-means clustering analysis. Ellipses represent a 95% confidence interval. **(C)** Bayesian clustering performed on STRUCTURE using 7,000 SNP and indel markers. Each genotype is represented by a vertical bar that is split in accordance with the admixture coefficients for two clusters (F1 in blue and F2 in orange). The genotypes are ordered according to ID number (see Supplementary Table 1).

Figure 5. Decay of linkage disequilibrium (r^2) as a function of physical distance (bp) between pairs of 67,007 single nucleotide polymorphisms (SNPs) and insertions and deletions (indels) located on *Saccharum spontaneum* chromosomes 1A-8A. Only r^2 values with $P < 0.05$ are included.

Figure 6. Manhattan plots generated in association analyses using the best linear unbiased predictor (BLUP) values of the three traits analyzed. Six different models were tested: general, additive, simplex dominant reference (1-dom-ref), simplex dominant alternative (1-dom-alt), diploidized general (diplo-general) and diploidized additive (diplo-additive). On the x-axis, S represents scaffolds not associated with any of the *S. spontaneum* chromosomes.

Figure 7. Distribution of markers associated with SCYLV resistance along *Saccharum spontaneum* "A" chromosomes. In each chromosome, marker positions are shown on the left, and marker names are indicated on the right, labeled and colored according to the method employed for their identification. Markers identified by previous mapping studies are colored in gray.

Figures

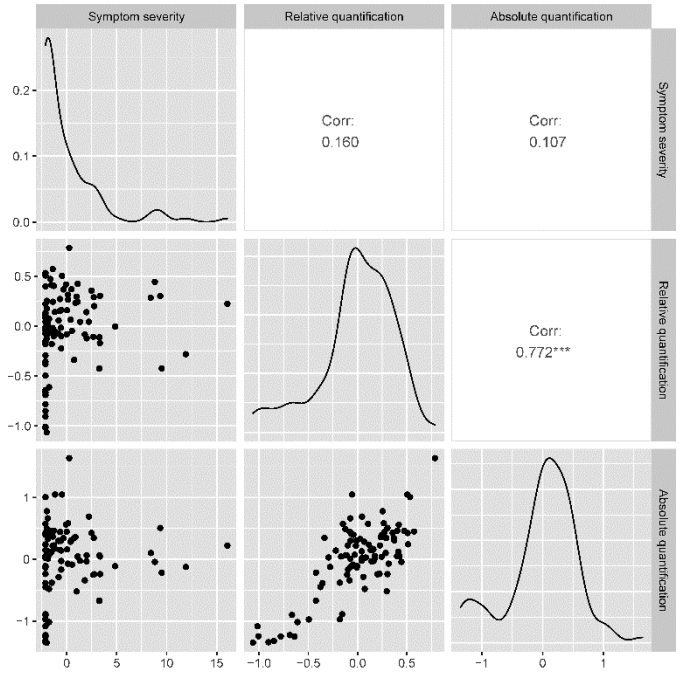

Figure 1

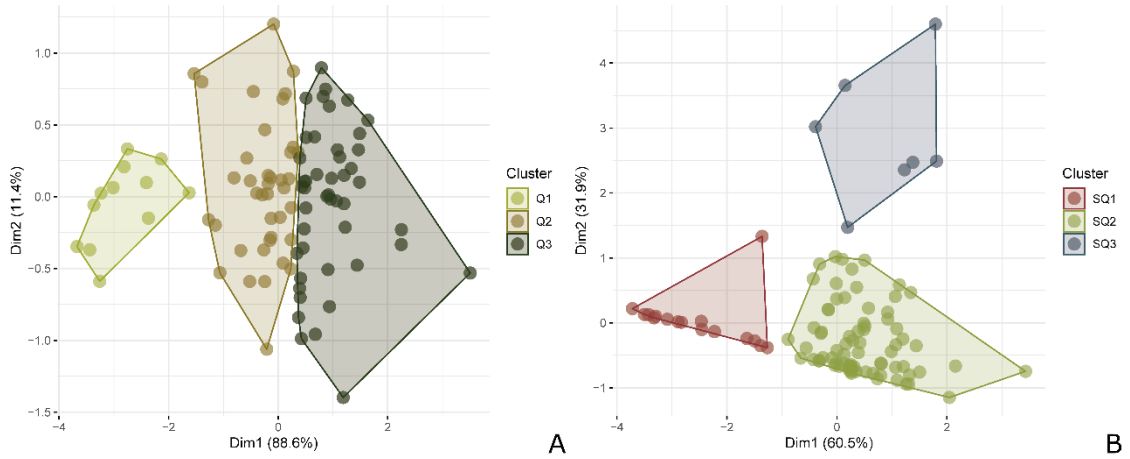
Dim2 (11.4%)
Dim1 (88.6%)
Cluster
Q1
Q2
Q3
A
Dim2 (31.9%)
Dim1 (60.5%)
Cluster
SQ1
SQ2
SQ3
B

Figure 2

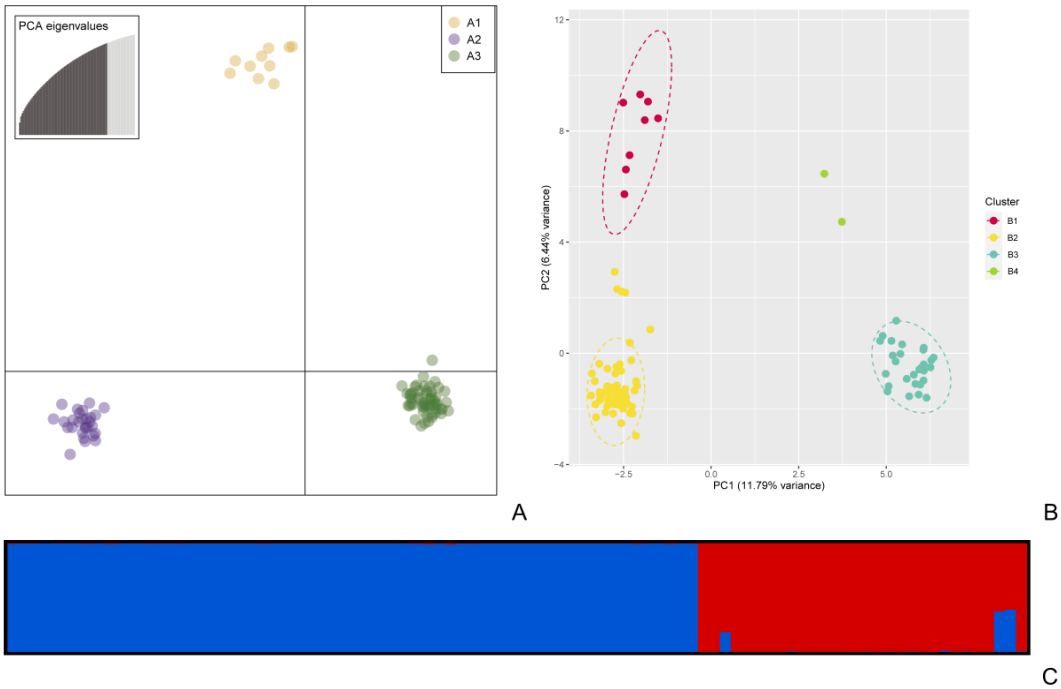
PCA eigenvalues
A1
A2
A3
A
PC2 (6.44% variance)
PC1 (11.79% variance)
Cluster
B1
B2
B3
B4
B
C

Figure 3

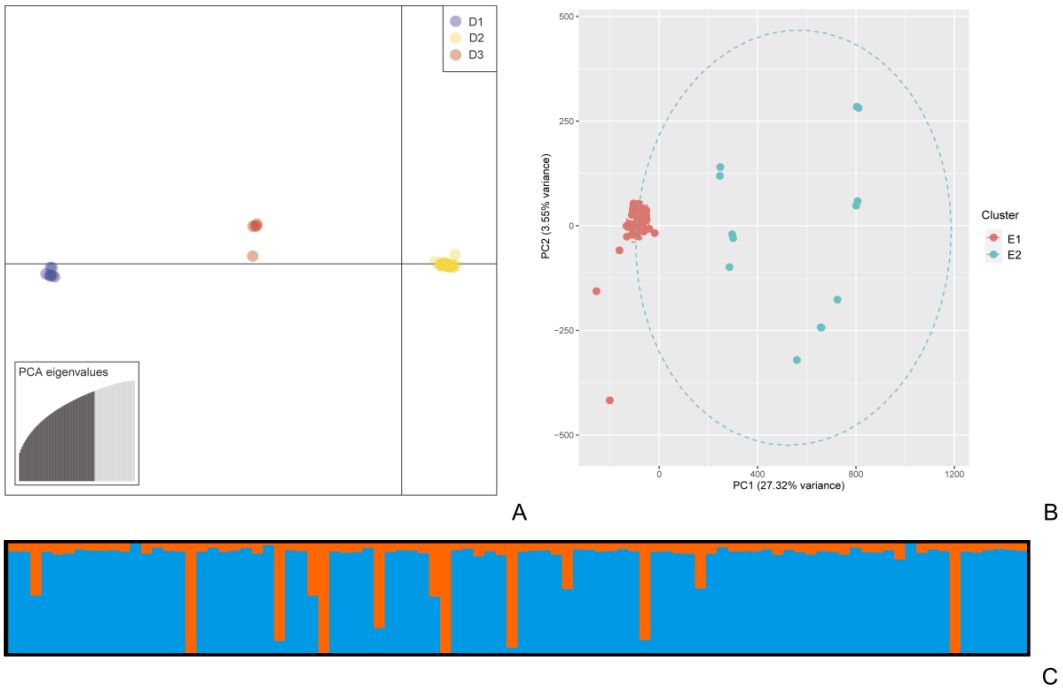
D1
D2
D3
PCA eigenvalues
A
500
250
0
−250
−500
PC2 (3.55% variance)
0
400
800
1200
PC1 (27.32% variance)
Cluster
E1
E2
B
C

Figure 4

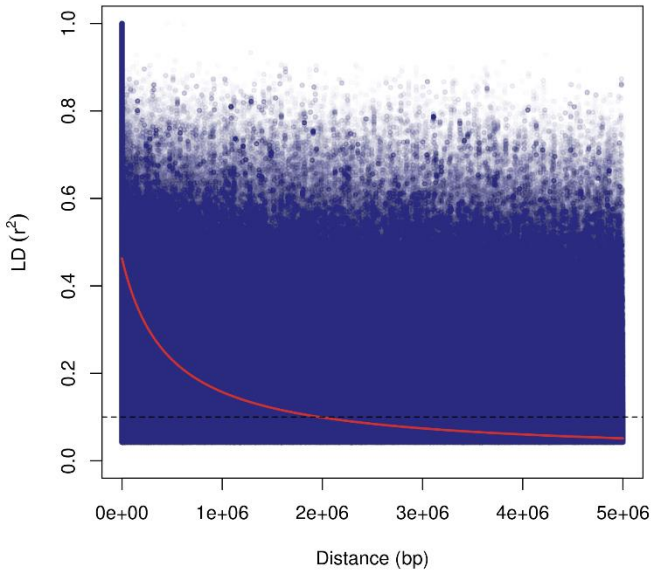
LD (r^2)
0.0
0.2
0.4
0.6
0.8
1.0
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
Distance (bp)

Figure 5

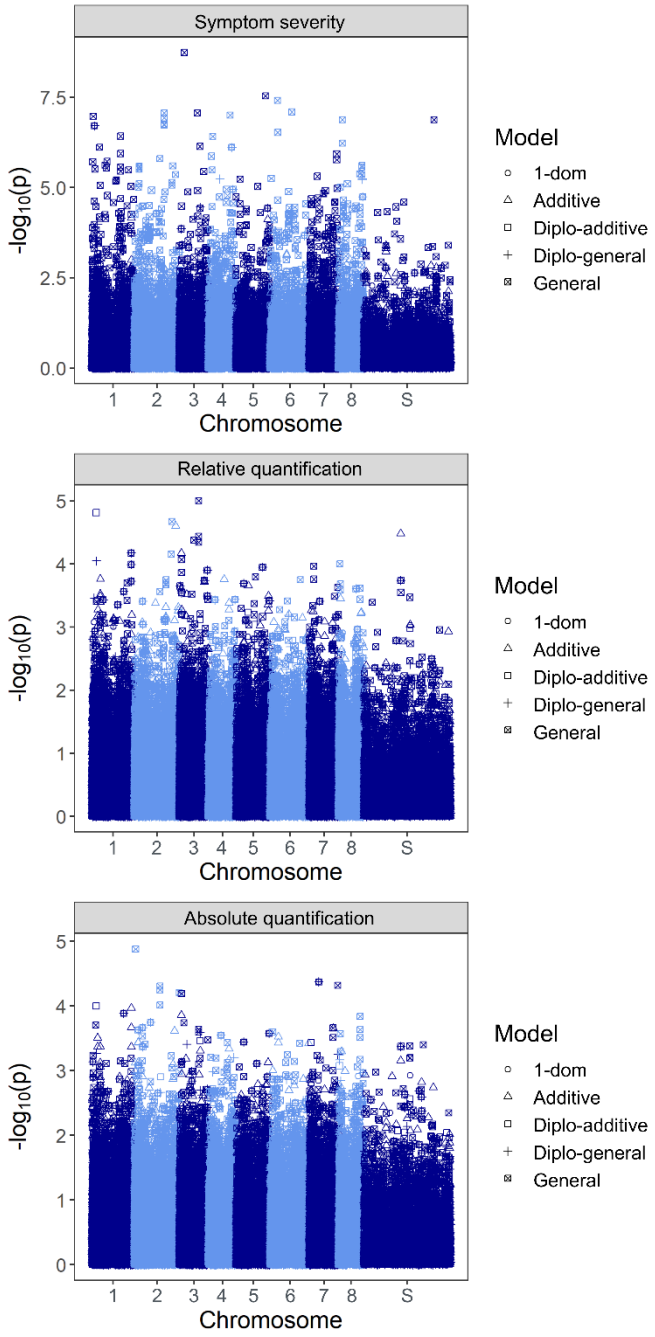
Symptom severity
Relative quantification
Absolute quantification
-log10(p)
Chromosome
1 2 3 4 5 6 7 8 S
Model
1-dom
Additive
Diplo-additive
Diplo-general
General

Figure 6

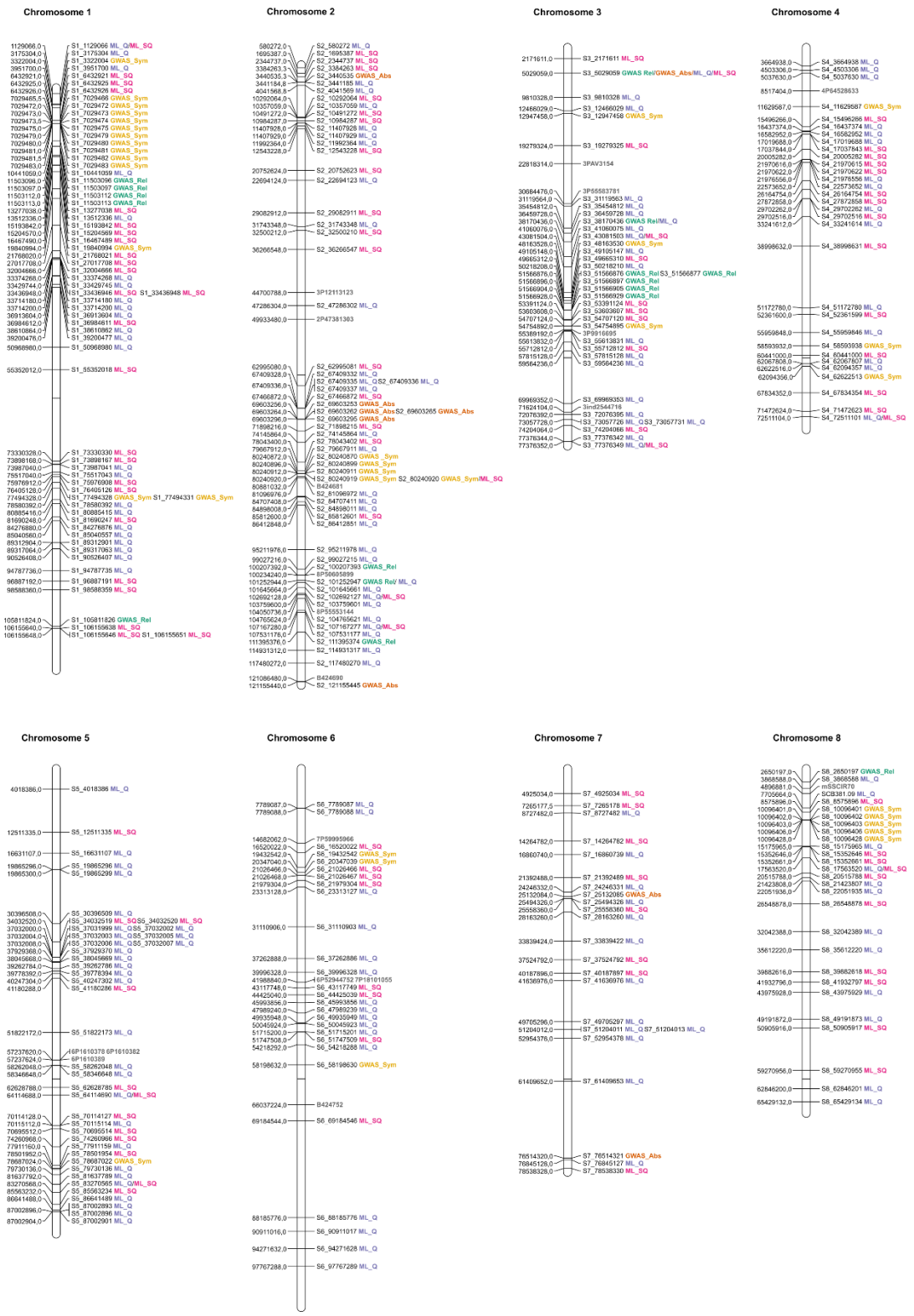
Chromosome 1
Chromosome 2
Chromosome 3
Chromosome 4
Chromosome 5
Chromosome 6
Chromosome 7
Chromosome 8

Figure 7