Genome-Wide Association Mapping of Sugarcane Yellow Leaf Virus Resistance

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

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

[4]Plant Health Centre, Biological Institute (IB/APTA), São Paulo, Brazil

[5]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, GBS, GWAS.

Abstract

The breeding of sugarcane, a leading sugar and energy crop, is complicated by the extremely complex sugarcane genome, which burdens research in the area and delays the development of new cultivars. One of the main viral diseases that affect this crop is sugarcane yellow leaf (SCYL), which is caused by the sugarcane yellow leaf virus (SCYLV). The most common symptom of SCYL is the yellowing of leaf midribs and blades, but asymptomatic cases are frequent. Regardless of the manifestation of SCYL, 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 widely explored or explained. In this context, genome-wide association studies (GWASs) have arisen as promising tools for the identification of molecular markers associated with SCYLV resistance that can be employed in marker-assisted selection. In the present work, we performed a GWAS on sugarcane using codominant markers and genotypes of interest for breeding. A panel of 97 sugarcane genotypes inoculated with SCYLV was analyzed for SCYL symptom severity, and viral titer was estimated by reverse transcription quantitative PCR (RT-qPCR). A genotyping-by-sequencing (GBS) library was constructed for 94 individuals of this population, enabling the identification of 38,710 SNPs and 32,178 indels with information on allele proportion (AP) and position on the *Saccharum spontaneum* genome. For association analyses, several combinations of population structure and kinship were tested to reduce model inflation, and diverse marker-trait association mixed models were employed. We identified 35 markers significantly associated with SCYL symptom severity and 22 markers strongly associated with SCYLV titer that can be applied in breeding programs upon validation. By aligning the sequences flanking these markers with their

coding sequences in several plant species, we annotated the functions of 7 genes. The possible involvement of these candidates in the response to SCYLV infection is discussed.

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 is high in sugar content but susceptible to drought and 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 impacts sugarcane research and breeding; for instance, until recently it prevented the use of codominance information in marker-assisted breeding methods for this crop, leading to a loss of information during performance of these approaches and therefore limiting them (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 the 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. Other visible manifestations are the 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 parameters (Gonçalves et al., 2005; Lehrer et al., 2010); this impairs plant development and eventually translates to losses in sugarcane productivity (Vega et al., 1997; Grisham et al., 2001; Vasconcelos et al., 2009; Viswanathan et al., 2014). Many observable 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. The diagnosis of infection by molecular methods is, therefore, of great assistance. Immunological assays were the first methods developed to detect and quantify this virus (Scagliusi and Lockhart., 2000), but more recent assessments by reverse transcription followed by reverse transcription quantitative polymerase chain reaction (RT-qPCR) are much more sensitive and precise in estimating the levels of viral infection (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 *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 through meristem micropropagation (Chatenet et al., 2001; Fitch et al., 2001), which is considerably time-consuming and requires specialized

This is a provisional file, not the final typeset article

infrastructure and personnel. These characteristics make plant resistance to SCYLV the resource that can be most efficiently exploited to prevent damage provoked by this virus. Sugarcane resistance to SCYLV has been investigated 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 mapping (Debibakas 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 poor reliability of phenotyping, which added to 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 (Debibakas et al., 2014; Gouy et al., 2015).

In this study, we performed a genome-wide association study (GWAS) 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 had their viral titer estimated via relative and absolute RT-qPCR quantification methods. We constructed a genotyping-by-sequencing (GBS) library for 94 genotypes of this population and used it to discover 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 performed a GWAS in which allele proportions (APs) in each locus were used to establish genotypic classes and estimate additivity and dominance effects. Finally, by aligning neighboring sequences of the genomes of plants closely related to sugarcane, we annotated genes containing markers associated with phenotypes. We discuss the possible participation of these genes in mechanisms underlying SCYLV resistance.

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 (2019). 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 “Centro Avançado da Pesquisa Tecnológica do Agronegócio de Cana” 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 *M. sacchari* were released in the field weekly for five months starting from November 2017.

2.2 Phenotyping

Plants were phenotyped in two crop seasons, as plant cane in June 2018 (13 months after planting) (Villavicencio, 2019) and as 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 with four levels. The levels were (1) green leaf 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, 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 an integrity check. The samples were next diluted to 250 ng/µL and treated with RNase-Free RQ1 DNase (Promega, Madison, USA); after this step, the RNA was requantified and diluted again to 54.5 ng/µL 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 RT-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 the 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 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 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 RT-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:

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

This is a provisional file, not the final typeset article

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 then performed using the BLUPs to check the correlation between the three traits, and correlation distributions were plotted using the GGally R package (Schloerke et al., 2011). We also performed hierarchical clustering on principal components (HCPC) using the factoMineR R package (Lê et al., 2008) to investigate the separation of genotypes according to SCYL symptom severity and SCYLV quantification. The factoextra R package (Kassambara and Mundt, 2017) was used to plot graphs associated with this analysis. Statistical differences between the phenotypes of the clusters identified in the 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

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 PstI and MseI restriction enzymes. For operational reasons, only 94 genotypes of the panel were included in the library, which did not include genotypes 87, 88 and 95 (see Supplementary Table 1). After construction, 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 DNA was 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).

Sequencing quality was evaluated with FastQC (Andrews, 2010) and visualized with MultiQC (Ewels et al., 2016). Stacks software version 1.42 (Catchen et al., 2013) was used 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" alleles and unassembled scaffolds. After variant calling, VCFtools version 0.1.13 (Danecek et al., 2011) was used to retain biallelic markers with a minimum allelic frequency (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 (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 Association Analyses

Association analyses were performed using mixed linear model approaches in the GWASpoly R package (Rosyara et al., 2016). Three methods were used to evaluate population structure and build Q matrices, which were included in the models as a fixed effect: (I) a discriminant analysis on principal components (DAPC), performed in the adegenet R package (Jombart, 2008). For this analysis, the maximum number of groups was defined as k = 10. (II) A principal component analysis (PCA) in which the optimal number of clusters was evaluated using the elbow and silhouette methods in the factoextra package, followed by clustering with the k-means method. Three principal components (PCs) from this analysis were used for the construction of the matrix. (III) A Bayesian clustering of individuals into predetermined numbers of clusters (K) assuming an admixture model with correlated allelic frequencies between populations. Ten independent runs were implemented for K = 1-10. Estimates of probabilities of values of K in each run were taken following 100,000 generations as burn-in and 100,000 generations sampled in a Monte Carlo Markov chain. 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 (Pritchard et al., 2000); the output was interpreted in STRUCTURE HARVESTER software version 0.6.94 (Earl and vonHoldt, 2012). Clumpak software (Kopelman et al., 2015) was used to average the admixture proportions of runs and to estimate cluster membership coefficients for individuals.

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) and built with the AGHmatrix R package (Amadeu et al., 2016); and (III) a pseudodiploid matrix based on Slater et al. (2016) and 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 "quantile-quantile" (Q-Q) plot of the $-\log_{10}(p)$ values of markers was generated, from which the inflation factor λ (Devlin and Roeder, 1999) 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 used to establish the significance threshold for associations.

2.6 Marker Annotation

The sequences located in the 2000-bp window adjacent to markers significantly associated with phenotypes were aligned against the coding sequences (CDS) 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*, which are available in the Phytozome database

This is a provisional file, not the final typeset article

(https://phytozome.jgi.doe.gov/pz/portal.html). For this, BLASTn (Altschul et al., 1990) was used with an E-value of 1e-30. 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 samples from two 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 Mantel 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.

An HCPC performed with BLUP values indicated the division of the panel into three phenotypic clusters (Supplementary Figure 2); a factor map wherein these groups are plotted onto the first two dimensions of the HCPC is shown in Figure 2, and the attribution of genotypes to each cluster is available in Supplementary Table 2. Boxplots of the distribution of data among the three clusters (Supplementary Figure 3) clearly show that this division resulted in a separation of groups with contrasting phenotypes. Cluster 1 (red) was composed of 16 accessions that showed the least severe SCYL symptoms and presented low titers of SCYLV, as indicated by the two quantification methods. Cluster 2 (green) was formed by 74 accessions that also displayed mild disease symptoms but presented high viral titers. Finally, cluster 3 (blue) included 7 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 SCYL in each cluster was significantly different from those of the others and that cluster 1 showed significantly lower SCYLV titers as estimated by both quantification methods than the other two clusters (Supplementary Table 3).

3.2 GBS and Variant Calling

The sequencing of the GBS library including 94 sugarcane genotypes generated 119,117,758 raw reads, 87.88% of which had 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 4. 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 5). However, knowing that a high-quality genome assembly can greatly improve the results of GWASs (Benevenuto et al., 2019), 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 position information. Therefore, the monoploid chromosome set isolated from the *S. spontaneum* genome, which allowed the identification of 38,710 SNPs and 32,178 indels, was chosen as a reference.

3.3 Panel Genetic Structure

The genetic structure of the panel was investigated using three different approaches. Supplementary Table 6 summarizes the allocation of genotypes to the clusters identified in each analysis. First, a DAPC was performed with retention of 60 PCs that together explained ~90% of the observed variance (Supplementary Figure 4A); the Bayesian information criterion (BIC) curve indicated k = 3 as the most likely number of clusters in the panel (Supplementary Figure 4B). Eight genotypes were grouped into one cluster (D1), 79 genotypes were grouped into a second cluster (D2), and the remaining five genotypes were allocated to a third cluster (D3). The plotting of the panel elements onto the first two linear discriminants (LDs) from this analysis is shown in Figure 3A. The second method, a PCA followed by k-means clustering analysis, was executed assuming the existence of two clusters, as indicated by elbow and silhouette methods (Supplementary Figure 5A and B, respectively). Despite indicating the subdivision of the panel into two clusters, the attribution of genotypes to these groups shared similarities with the DAPC clustering. Thirteen genotypes were attributed to one cluster (K1), which corresponded to D2, and the 79 remaining genotypes were grouped into another cluster (K2), which was equivalent to D1 and D3 combined. Figure 3B depicts a projection of these groups onto the first two PCs, which accounted for ~33% of the variation in the genotypic data. Based on the third analysis, a Bayesian clustering implemented on STRUCTURE software, both ΔK and LnP(D) methods clearly indicated K = 2 as the most likely number of clusters in the panel (Supplementary Figure 6A and B, respectively). Therefore, we present here a plot of the estimated association coefficients of individuals for these two clusters (Figure 3C). Eight genotypes presented over 50% identification with cluster S2, and the remaining genotypes showed over 50% identification with cluster S1. Clustering by this method was consistent with clustering by the other two methods, as S1 corresponded to the union of D2 and D3, and S2 correspondend to D1. Curiously, the results of the three genotype-based clustering methods showed little correspondence with the phenotype-based HCPC. Although most accessions in HCPC clusters 1 and 3 were attributed to the same genotypic cluster in the three population structure analyses, accessions in cluster 2 were randomly distributed among genotypic groups identified by each genetic analysis.

3.4 Genome-Wide Association Analyses

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 the inflation factors (λ) of each of the Q + K models can be seen in Figure 4. A Mantel 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 its theoretical value under the null hypothesis (λ = 1); this 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 the most straightforward approach. Q-Q plots of the association analyses for SCYL symptom severity and

This is a provisional file, not the final typeset article

SCYLV relative and absolute quantifications can be found in Supplementary Figures 7-9, respectively. 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) as significantly associated with SCYL symptom severity (Supplementary Figure 5). 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 (Supplementary Figure 10) and 9 markers (6 SNPs and 3 indels) associated with SCYLV titers determined by absolute quantification (Supplementary Figure 11). One of these markers (S3_5029059) was common to both viral quantification methods; this could be an indication of the consistency of the association or simply reflect the correlation between the two measurements. Supplementary Table 7 supplies information on all marker-trait associations. 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.

3.5 Marker Annotation

Out of the 57 nonredundant markers associated with the three phenotypes, 17 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 8 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, seven genes are representative of all the best alignments. All hits occurred with the CDSs from the genome of *S. bicolor*, the phylogenetically closest species among those used for alignment.

Five annotated genes contained markers associated with SCYL symptom severity. The first (SbRio.01G540100) encodes an expressed transcription factor containing a GATA zinc finger domain. The second (Sobic.001G417400) encodes a putative expressed protein containing an RNA recognition motif. The third (SbRio.03G158900) encodes OsFBX103, an expressed protein containing an F-box domain. The fourth (SbRio.03G120000) encodes GRP6, a putative expressed glycine- and cysteine-rich protein precursor. The last gene (Sobic.001G200200) encodes a putative expressed protein containing a tetratricopeptide repeat. One marker associated with each SCYLV quantification method could be annotated. The gene containing the marker associated with the relative quantification method (Sobic.001G053100) encodes an unknown expressed protein, and the annotated gene containing the marker associated with the absolute quantification method (Sobic.009G121100) encodes a protein related to plant calmodulin binding.

4 Discussion

In the present study, we evaluated SCYL symptoms and SCYLV titers in a panel of 97 sugarcane accessions. Prior to phenotyping, plants were subjected to strong SCYLV inoculum pressure, an innovation over previous mapping studies, which relied on natural infection under field conditions (Costet et al., 2012; Debibakas et al., 2014; Gouy et al., 2015; Islam et al., 2018; Yang et al., 2019a).

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 (2019), 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, which shows 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 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.

Further exploring the BLUPs, we performed an HCPC analysis to discriminate accessions based on their response to SCYLV infection, which led to the separation of three clusters with considerable phenotypic differences. 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 ones (which exhibit little or no symptoms). In view of this nomenclature, we propose that the phenotypic clusters identified in the HCPC be described by the following three designations: (1) resistant, for sugarcane genotypes distinguished by a low SCYLV titer and mild SCYL symptoms; (2) tolerant, for genotypes that, despite exhibiting higher viral titers, presented fewer or no disease symptoms; and (3) susceptible, for genotypes with the most severe symptoms and presenting a high viral titer. This classification alone is of great use in sugarcane breeding programs, as it distinguishes promising clones that are sources of both SCYLV resistance and tolerance.

The main objective of this work was, however, to identify markers associated with SCYLV resistance using codominant markers and mixed modeling. This is the first study employing a two-enzyme GBS protocol in sugarcane, and we evaluated the impact of using genomic references from various backgrounds in variant calling. 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 the GBS library preparation reduced 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,

This is a provisional file, not the final typeset article

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 isolated from this genome, which provided a large number of markers (38,710 SNPs and 32,178 indels) with reliable position information.

Using these markers, we tested several approaches to account for population structuring and kinship in the panel; our results indicated that these variations did not imply significant differences in the inflation of the models used for association analyses. In many of the models employed, the inflation patterns observed in the Q-Q plots generated were similar to those of previous studies applying similar genotyping and association mapping methodologies (Yang et al., 2019a, 2019b). Thus, we opted to perform ensuing association analysis using the $Q_{PCA} + K_{MM}$ model, which presented an adequate inflation observed in Q-Q plots, apart from resulting in the median value of λ closest to 1 – its theoretical value under the null hypothesis. The 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).

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, which employed laborious inoculation and quantification techniques. 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.

Seventeen out of the associated markers were annotated to seven genes. Some of these loci do not allow extensive discussion of their involvement in SCYLV resistance; this is the case for genes known to simply encode expressed proteins with unknown function (Sobic.001G053100) or proteins containing an RNA recognition site (Sobic.001G417400). Remarkably, two annotated genes with markers associated with SCYL symptom severity 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, which encodes a GATA zinc finger protein, and Sobic.001G200200, which encodes a protein containing a tetratricopeptide repeat.

Other annotated genes presented characteristics more relatable to virus resistance in plants. This is the case for SbRio.03G120000, which is associated with SCYL severity and encodes a precursor of a glycine- and cysteine-rich protein. Glycine-rich proteins have been widely shown to participate in plant stress responses, but little is known about the mechanisms and processes in which they are involved (Czolpinska and Rurek, 2018). They are downregulated under tobacco mosaic virus infection (Lee et al., 2012) and affect the movement of a tobamovirus by enhancing vascular callose deposition (Ueki and Citovsky, 2002), for instance. A cysteine-rich protein, on the other hand, was shown to potentiate the cell-to-cell and systemic movement of a potyvirus by interacting with a viral

protein (Dunoyer et al., 2004). These reports indicate that this protein might be involved in SCYLV resistance and in the expression of SCYL symptoms.

Another annotation related to symptom severity that drew attention is SbRio.03G158900, which encodes a protein with an F-box domain. 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 viral 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), a process directly involved in virus resistance (Morel et al., 2002; Chen et al., 2015). Even more intriguing is the fact that one of the proteins encoded in the SCYLV genome, P0, contains an F-box-like domain and mediates the destabilization of AGO1 and leads to the suppression of host posttranscriptional gene silencing (Mangwende et al., 2009). Whether the F-box protein identified here plays an active role in the mechanisms involved in virus replication in sugarcane and in the manifestation of symptoms remains a question to be investigated by further studies.

Finally, Sobic.009G121100, which harbors a marker strongly associated with the SCYLV titer determined by absolute quantification, encodes a plant protein related to calmodulin binding. Calmodulins are notorious calcium transducers that regulate the activity of various proteins with diverse functions (Bouché et al., 2005). Calmodulins and calmodulin-like and calmodulin-binding proteins have been widely implicated in viral resistance in plants, often playing roles in RNA interference mechanisms (Heo et al., 1999; Nakahara et al., 2012; Li et al., 2014). Therefore, we also consider this a promising candidate gene in the regulation of SCYLV replication and spreading *in planta*.

In conclusion, this genome-wide association analysis performed with GWASpoly using multiple marker-effect models and population structure information derived from a PCA 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, which can employ the results in marker-assisted strategies for the early selection of clones. Additionally, the annotation of genes wherein these markers are located provides insights into the unexplored mechanisms possibly involved in this crop's response to infection by SCYLV.

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

MCG, LRP, MGAL and APS conceived and designed the experiments. RJGP, RCBV, CCS and IAA conducted the experiments and collected data. 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 Secretaria Nacional de Ciencia

This is a provisional file, not the final typeset article

y Tecnología (SENESYT). RJGP received MSc fellowships from CAPES (grant 88887.177386/2018-00) and from FAPESP (grant 2018/18588-8). 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

ANOVA: analysis of variance

AP: allele proportion

BLUP: best linear unbiased predictor

CDS: coding sequence

DAPC: discriminant analysis of principal components

FDR: false discovery rate

GBS: genotyping-by-sequencing

GWAS: genome-wide association study

HCPC: hierarchical clustering on principal components

LD: linear discriminant

PC: principal component

PCA: principal component analysis

QTL: quantitative trait locus

RT-qPCR: reverse transcription quantitative polymerase chain reaction

SCYL: sugarcane yellow leaf

SCYLV: sugarcane yellow leaf virus

SNP: single nucleotide polymorphism

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

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

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

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

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

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

This is a provisional file, not the final typeset article

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

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

Czolpinska, M., and Rurek, M. (2018). Plant glycine-rich proteins in stress response: an emerging, still prospective story. Front. Plant Sci. 9, 302. doi: 10.3389/fpls.2018.00302

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

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.

Debibakas, 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

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

Dunoyer, P., Thomas, C., Harrison, S., Revers, F., and Maule, A. (2004). A cysteine-rich plant protein potentiates Potyvirus movement through an interaction with the virus genome-linked protein VPg. J. Virol. 78, 2301-2309. doi: 10.1128/jvi.78.5.2301-2309.2004

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

This is a provisional file, not the final typeset article

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

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics (Oxford, England) 32, 3047-3048. doi: 10.1093/bioinformatics/btw354

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

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

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

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

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. (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

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

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

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.

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

This is a provisional file, not the final typeset article

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

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.

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

Lee, H. J., Kim, J. S., Yoo, S. J., Kang, E. Y., Han, S. H., Yang, K.-Y., et al. (2012). Different roles of glycine-rich RNA-binding protein7 in plant defense against Pectobacterium carotovorum, Botrytis cinerea, and tobacco mosaic viruses. Plant Physiol. Biochem. 60, 46-52. doi: 10.1016/j.plaphy.2012.07.020

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, 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, 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

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

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

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

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

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

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].

This is a provisional file, not the final typeset article

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

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

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.

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

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

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

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

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

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

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

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

Ueki, S., and Citovsky, V. (2002). The systemic movement of a tobamovirus is inhibited by a cadmium-ion-induced glycine-rich protein. Nat. Cell Biol. 4, 478-486. doi: 10.1038/ncb806

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. (2019) "Identification of *Saccharum* spp. genotypes resistant to yellow leaf disease (sugarcane yellow leaf virus) and mosaic disease (sugarcane mosaic virus) and association to molecular markers" PhD diss., State University of São Paulo.

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

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

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, 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

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

This is a provisional file, not the final typeset article

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. The correlation between the SCYLV quantities determined by the two methods was strong and significant (*), in contrast to the correlations of these two traits with SCYL symptom severity.

Figure 2. Factorial map generated in the hierarchical clustering on principal components (HCPC) analysis. A division into three clusters was considered.

Figure 3. Graphical outputs of the three clustering analyses employed to evaluate the genetic structures for 92 genotypes in the panel. (A) Plotting of the panel onto the first two linear discriminants (LDs) from a discriminant analysis of principal components (DAPC) using 70,888 single-nucleotide polymorphisms (SNPs) and 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 (K1 and K2) as indicated by a 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 (S1 in blue and S2 in orange). The genotypes are ordered according to ID number (see Supplementary Table 1).

Figure 4. Genomic inflation factor (λ) of twelve association models for the three SCYLV resistance traits analyzed, including different population structure (Q) and kinship (K) matrices as covariates. Abbreviations: Auto, complete autopolyploid model; DAPC, discriminant analysis of principal components; MM, MM^T realized relationship model; PCA, principal components analysis; Pseudo, pseudodiploid model; Str, STRUCTURE Bayesian clustering.

This is a provisional file, not the final typeset article

Figures

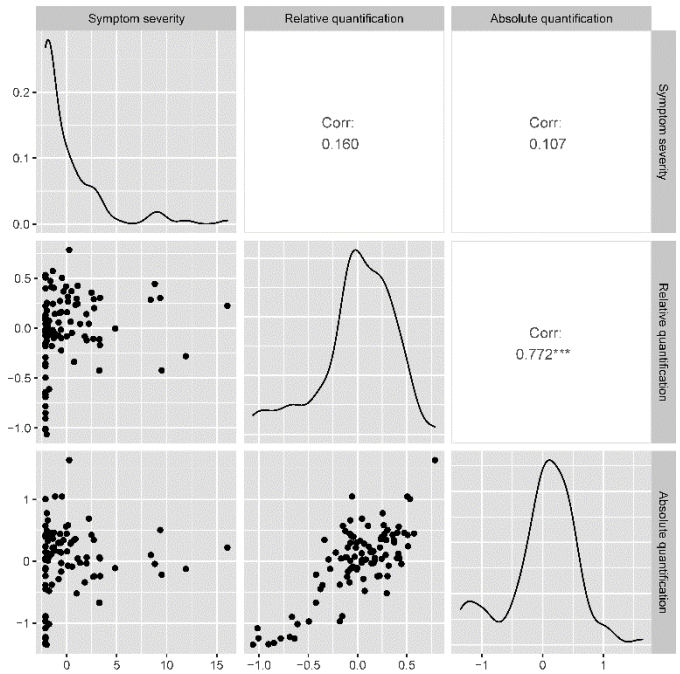

Figure 1

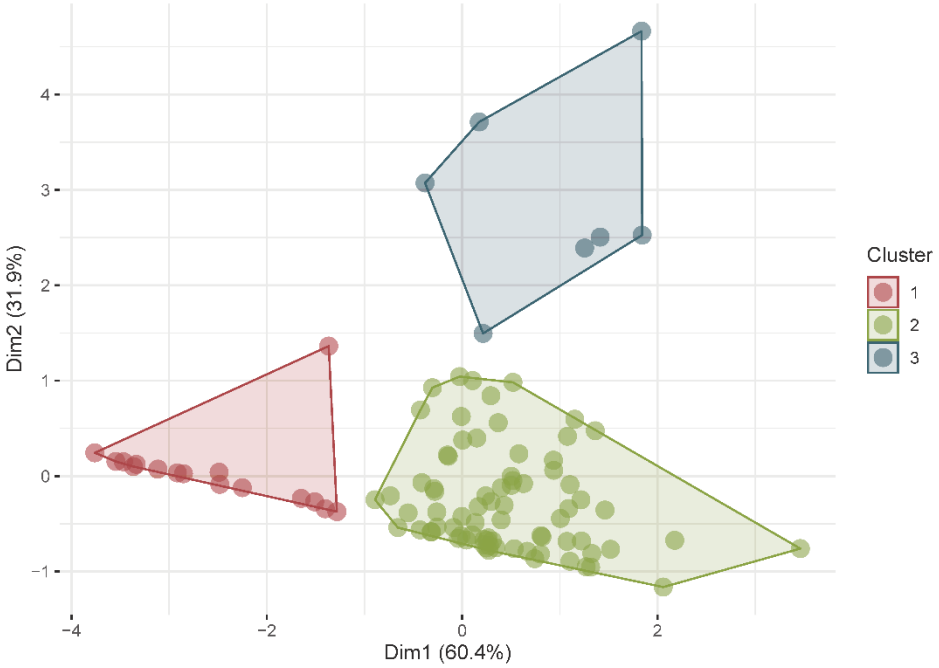

Figure 2

This is a provisional file, not the final typeset article

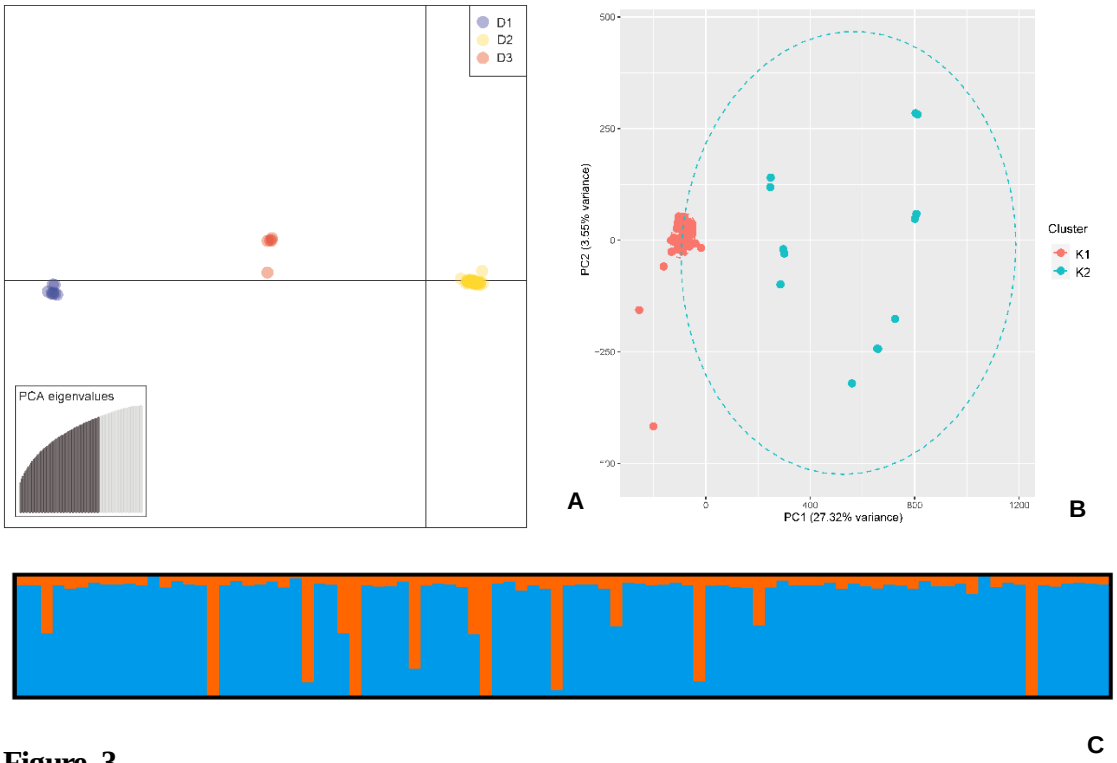
D1
D2
D3
PCA eigenvalues
A
500
250
0
−250
−500
PC2 (3.55% variance)
0
400
800
1200
PC1 (27.32% variance)
Cluster
K1
K2
B
C

Figure 3

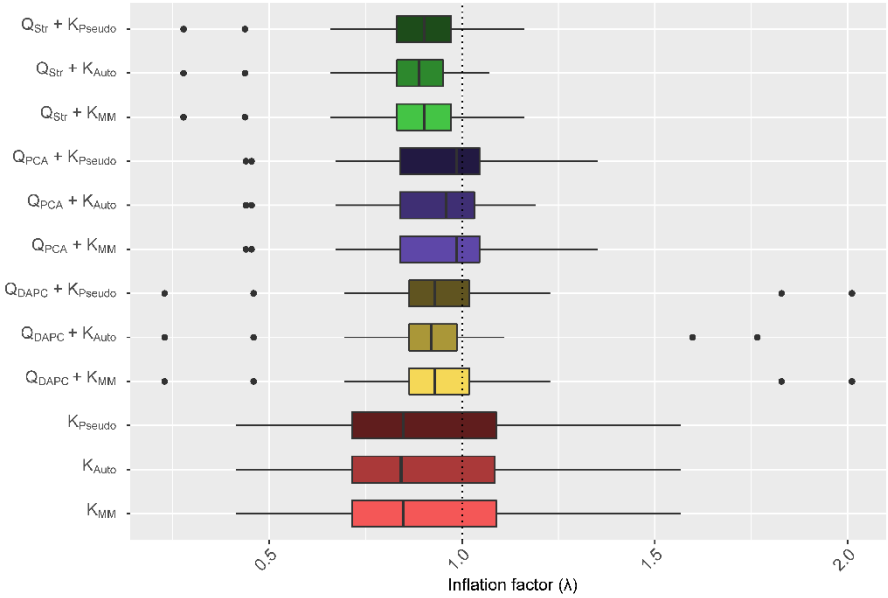

Figure 4

This is a provisional file, not the final typeset article