Multiomic investigation of sugarcane mosaic virus resistance in sugarcane

Ricardo José Gonzaga Pimenta[1], Alexandre Hild Aono[1], Roberto Carlos Villavicencio Burbano[2], Marcel Fernando da Silva[3], 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 for 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 Protection Research Centre, Biological Institute (IB/APTA), São Paulo, Brazil

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

***Corresponding author**

Anete Pereira de Souza

E-mail: anete@unicamp.br

Phone: +55 19 35211132

Abstract

Sugarcane mosaic virus (SCMV) is the main etiological agent of sugarcane mosaic disease, which affects sugarcane, maize and other economically important grass species. Despite the extensive characterization of quantitative trait loci controlling resistance to SCMV in maize, the genetic basis of this trait is largely unexplored in sugarcane. Here, a genome-wide association study was performed and machine learning coupled to feature selection was used for the genomic prediction of resistance to SCMV in a diverse panel of sugarcane accessions. This ultimately led to the identification of nine single nucleotide polymorphisms (SNPs) explaining up to 29.9% of the phenotypic variance and a 73-SNP set that predicted resistance with high accuracy, precision, recall, and F1 scores. Both marker sets were validated in additional sugarcane genotypes, in which the SNPs explained up to 23.6% of the phenotypic variation and predicted resistance with a maximum accuracy of 69.1%. Synteny analyses showed that the gene responsible for the major SCMV resistance in maize is probably absent in sugarcane, explaining why such a major resistance source is thus far unknown in this crop. Lastly, using sugarcane RNA sequencing data, markers associated with the resistance to SCMV in sugarcane were annotated and a gene coexpression network was constructed to identify the predicted biological processes involved in SCMV resistance. This allowed the identification of candidate resistance genes and confirmed the involvement of stress responses, photosynthesis and regulation of transcription and translation in the resistance to this virus. These results provide a viable marker-assisted breeding approach for sugarcane and identify target genes for future molecular studies on resistance to SCMV.

Keywords: *Saccharum*, SCMV, GWAS, Machine Learning, Feature Selection, QTL, RNA-Seq, Coexpression Networks

1. Introduction

Sugarcane (*Saccharum* spp.) is highly economically important in tropical regions worldwide, as this plant is not only the world's most important sugar-producing crop but also an important source of renewable energy obtained from its juice and bagasse (Carvalho-Netto *et al.*, 2014; ISO, 2022). Brazil has been the leader of the global cultivation of sugarcane for many years and is currently responsible for approximately 40% of global production (FAO, 2022). However, sugarcane yield is threatened by several diseases, with sugarcane mosaic being of the most important at the global scale (Wu *et al.*, 2012). In addition to the characteristic mosaic pattern displayed on the leaves, other symptoms of this disease include dwarfing, striping and streaking of culms, and shortening of internodes in highly susceptible genotypes (Gonçalves *et al.*, 2007). In Brazil, this disease emerged in the beginning of the 20th century; it led to massive yield losses and drove the sugarcane industry to the brink of collapse in 1920-30. Damage caused by sugarcane mosaic disease has since been controlled with the employment of resistant cultivars and the adoption of several practices, such as the planting of healthy setts and roguing of nurseries and commercial fields. However, this disease is still a threat to sugarcane production, and resistance to it is a primary concern in breeding programs (Gonçalves *et al.*, 2012).

Three viruses of the Potyviridae family are currently recognized as etiological agents of this disease in sugarcane: sugarcane mosaic virus (SCMV), sorghum mosaic virus, and sugarcane streak mosaic virus (Hall *et al.*, 1998). SCMV, belonging to the *Potyvirus* genus, is a widespread species and the only one of these viruses found to naturally infect sugarcane in Brazil (Gonçalves *et al.*, 2004, 2007, 2011). SCMV has been reported to cause sugarcane yield losses of up to 40-50% (Costa and Muller, 1982; Smith *et al.*, 1992) while also reducing juice quality (Singh *et al.*, 2003), sett germination and plant photosynthetic activity (Viswanathan and Balamuralikrishnan,

2005; Gonçalves *et al.*, 2007). High yield losses arising from infection by this virus have led to the discontinuation of several sugarcane cultivars (Singh *et al.*, 1997).

SCMV also infects many other closely related Poaceae species, including maize (*Zea mays*); this virus is responsible for extensive losses in maize yields, especially in Europe and China (Wu *et al.*, 2012). As a result of being transmitted by various aphid species in a nonpersistent manner (Hassan *et al.*, 2003), SCMV is very hard to control in the field, making host resistance an important resource to avoid damage caused by this virus. Thus, numerous quantitative trait locus (QTL) mapping studies have been performed to investigate the resistance of maize to SCMV; these studies resulted in the identification of two major and three minor QTLs controlling this trait in this species (Melchinger *et al.*, 1998; Xia *et al.*, 1999; Xu *et al.*, 1999; Dußle *et al.*, 2000, 2003; Zhang *et al.*, 2003; Wu *et al.*, 2007; Liu *et al.*, 2009; Soldanova *et al.*, 2012). Together, the major loci, named *Scmv1* and *Scmv2*, usually explain up to ~60-70% of the phenotypic variance observed for resistance (Xia *et al.*, 1999; Dußle *et al.*, 2000; Soldanova *et al.*, 2012). Recently, researchers have finely mapped the location of these QTLs and identified the causal genes responsible for resistance in maize (Ding *et al.*, 2012; Tao *et al.*, 2013; Li *et al.*, 2016; Liu *et al.*, 2017).

However, with respect to sugarcane, data on resistance to this virus are scarce. A few phenotypic studies have been performed in Brazil: specifically, researchers have evaluated the genotypic correlation of this disease incidence in sugarcane families (Xavier *et al.*, 2013) and screened diverse genotypes for resistance (da Silva *et al.*, 2015a, b). In addition, three marker–trait association studies have been carried out targeting SCMV resistance in this crop (Barnes *et al.*, 1997; Pinto *et al.*, 2013; Burbano *et al.*, 2022); however, most included very few genotypes (≤50), and all employed dominantly scored markers. This apparent disparity between the

information on SCMV resistance available for sugarcane and maize can be partially explained by the larger economic importance of the latter species in several countries; however, another important factor, i.e., sugarcane's genomic complexity, has an effect. Modern cultivars are derived from a few crosses between two highly autopolyploid species, *Saccharum spontaneum* ($2n$ = 5x = 40 to 16x = 128; x = 8) (Panje and Babu, 1960) and *Saccharum officinarum* ($2n$ = 8x = 80; x = 10) (D'Hont *et al.*, 1998). These hybrids have large (D'Hont *et al.*, 1998), highly polyploid (D'Hont and Glaszmann, 2001), aneuploid (Sforça *et al.*, 2019) and duplicated (Aono et al., 2021) genomes that hinder sugarcane breeding research. Additionally, studies suggest that the majority of sugarcane traits are controlled by many small-effect loci (Gouy *et al.*, 2015; Fickett *et al.*, 2019; Pimenta *et al.*, 2021). However, given the existence of *Scmv1* and *Scmv2* in maize, it is odd that no major loci controlling SCMV resistance in sugarcane have been identified.

In view of this crop's complex genome and the high impact of SCMV on its yield, the exploration of novel methodologies is required for the investigation of sugarcane's resistance to this virus. This study aimed to identify markers associated with SCMV resistance and provide insights into its molecular basis through the use of state-of-the-art genomic and transcriptomic approaches. To achieve this, a panel of *Saccharum* accessions was assessed by phenotyping for SCMV resistance in the field and was genotyped via genotyping by sequencing (GBS), enabling the discovery of single nucleotide polymorphisms (SNPs) with information on allele proportion (AP) and position in a monoploid set of chromosomes of *S. spontaneum*. These data were used to perform a genome-wide association study (GWAS) to identify markers associated with SCMV resistance and to predict genotype attribution to resistant or susceptible groups by the use of machine learning (ML) coupled with feature selection (FS). Associated markers were genotyped

on additional accessions previously assessed for SCMV resistance for validation and subsequently annotated by the use of a newly assembled sugarcane transcriptome. This allowed the incorporation of SCMV-associated genes into a coexpression network and thus a broader investigation of the molecular basis underlying sugarcane resistance to this virus.

2. Results

2.1. Panel phenotyping and genotyping

Ninety-seven sugarcane accessions were evaluated for the presence and severity of SCMV symptoms in two consecutive years. A skew in the distribution of the data toward the absence of symptoms was observed; despite normalization procedures, these data did not follow a normal distribution as indicated by a Shapiro–Wilk test ($p = 2.2e-16$). Based on the occurrence of SCMV symptoms, the panel could be divided into two groups: 62 resistant genotypes, which did not present symptoms in any block or year, and 35 susceptible genotypes, which presented symptoms on at least one occasion.

Following the construction and sequencing of a GBS library, 3,747,524, 3,152,409, and 569,360 biallelic SNPs were identified using FreeBayes, SAMtools and the TASSEL-4-POLY pipeline, respectively. After filtering procedures were performed and examining the intersection between tools, 37,001 of these markers were found to be called by TASSEL4-POLY and at least one of the other tools; thus, these markers constituted the final set of reliable SNPs.

2.2. Association analyses

2.2.1. Mixed modeling

Data from 92 accessions of the panel were subjected to mixed modeling genome-wide association analyses on GWASpoly, and six different marker-effect models were used. Q-Q plots generated by these analyses can be found in Figure S2. In general, most models showed an appropriate profile of inflation of p values; exceptions disregarded for further analyses included the general model, which presented insufficient control of inflation, and the simplex dominant alternative model, which presented deflation. A stringent significance threshold ($p < 0.05$ corrected by the Bonferroni method) was used to establish 20 significant marker–trait associations, some of which were highly significant (Figure 1); the r^2 values of associations ranged from 0.017 to 0.299 (Table S3). Several markers were associated with SCMV resistance according to more than one model, and nine nonredundant markers were representative of all associations.

2.2.2. ML coupled with FS

Eight ML algorithms for predicting the attribution of sugarcane genotypes to SCMV-resistant or SCMV-susceptible groups based on genotypic data were tested. When assessing their potential for this task when the full marker dataset was used, the predictive accuracies ranged from 52.8 (DT) to 66.9% (RF), with a mean of 60.3% (Table 1 and Figure S3). The remaining metrics evaluated showed much inferior results, with means of 21%, 26.8% and 20.7% found for precision, recall and F1 score, respectively. GP performed particularly poorly, with the mean of measurements equal to zero for these three metrics (Table 1 and Figures S4-S6).

Therefore, three FS methods were used to reduce the marker dataset and improve model performance, and the SNPs identified by at least two of these methods were selected. This enabled the identification of a 73-SNP dataset, which led to considerable increases in all the metrics of all the models. With the reduced dataset, a mean accuracy of 90.2%, with a maximum of 99.7% using MLP, was obtained (Table 1 and Figure S3). Even more pronounced increases were observed for the other metrics: the mean precision was 79.6%, with a maximum of 100% with MLP (Table 1 and Figure S4); the mean recall was 91.4%, with a maximum of 100% with KNN and SVM (Table 1 and Figure S5); and the mean F1 score was 83.6%, with a maximum of 99.6 with MLP (Table 1 and Figure S6). ROC curves and their AUCs supported the promising results of FS in the predictive task. When all the markers were used, all the models presented ROC curves rather close to the level associated with chance alone, with AUCs ranging between 0.46 and 0.57 (Figure 2A). However, when markers selected by FS were used, most ROC curves indicated much better model performance, with AUCs of up to 0.99 (Figure 2B). Only DT and GP did not show appreciable increases in the AUC; thus, these models were excluded as appropriate methods for genomic prediction in this case.

2.3. Marker validation

Two groups of sugarcane genotypes previously assessed for SCMV resistance were genotyped via the MonsterPlex technology to validate markers identified in the association panel. The sequencing of the MonsterPlex library generated a total of 38,581,797 single-end reads, 99.8% of which presented a mean Q-value greater than 30; these values remained consistently high for the first 100 bases of the reads (Figure S7A). These data encompassed 81 of the 92 samples sent for analysis; DNA from genotypes 4, 5, 18, 68, 69, 70, 71, 73, 76, 77, and 79 (see Table S2) did not amplify well, and consequently, these samples were absent in the sequencing results.

Interestingly, ten out of these samples were represented by *S. officinarum* F1 accessions, with only genotype 18 representing a hybrid variety. After trimming was performed, 38,574,693 reads were retained, 99.9% of which had a mean Q-value greater than 30 (Figure S7B). Using SAMtools and FreeBayes, 53 out of the 82 target SNPs (64.6%) could be called.

Seven of these SNPs were identified by GWAS as being significantly associated with SCMV resistance. When associations involving these markers were tested by linear models, the r^2 values were overall lower than those of the original panel and were frequently close to zero, especially in the small wild accession panel. However, these values remained positive and reached as high as 0.236 (Table S4). The remaining 46 SNPs identified belonged to the reduced 73-SNP dataset identified by FS. These markers were applied to the eight ML models tested, which resulted in a mean accuracy of 61.6%, with a maximum of 69.1% by the RF model. This model was also among the top-ranking ones in terms of precision (68.6%), recall (94.1%) and F1 score (79.3%) for the identification of resistant genotypes (Table 2).

2.4. Synteny analyses

To assess the presence of the two major SCMV resistance QTLs from maize, *Scmv1* and *Scmv2*, in sugarcane, the CDSs of the causal genes at these loci were aligned against the *S. spontaneum* genomic sequence employed for SNP calling. Despite the close phylogenetic relation of the two species, no hits were found for *Scmv1*, indicating that this gene is likely absent from the *S. spontaneum* genome. Complementary searches in the genomes of an additional six sugarcane accessions also revealed no matches for this gene. The sequence of *Scmv2*, on the other hand, resulted in a 373-bp alignment with 88.7% identity and an E-value of 4.92e-127, corresponding to the Sspon.02G0027920-1A gene, which, like the causal gene at *Scmv2*, encodes an auxin-

binding protein. This gene is located on chromosome 2A and is 1.7 kb away from the marker Chr2A_103190628, which was identified as being associated with SCMV resistance by FS (Figure 3).

2.5. Coexpression network construction and marker annotation

To assemble a de novo sugarcane transcriptome for marker annotation and expression analysis, more than two billion (2,477,287,294) sugarcane RNA sequencing (RNA-Seq) reads were retrieved from the SRA, 76% of which (1.9 billion) were retained after trimming. The transcriptomic reference assembled by Trinity comprised 611,480 transcripts with an N50 of 1,233 bp, represented by 212,076 longest isoforms (henceforth referred to as "genes") with an N50 of 2,561 bp. The complete assembly contained 83.8% of conserved orthologs from green plants, as reported by BUSCO (Table S5). After quantification with Salmon, 131,615 genes were discarded for exhibiting very low expression. The remaining genes were used to construct a GWGCN. Using the UPGMA method, 64 functional modules were defined in this network, with sizes ranging from 58 to 32,980 genes and a mean size of 1,257.

To annotate the markers identified as associated with SCMV resistance through GWAS and FS, the transcriptome assembly was aligned against the *S. spontaneum* genome used for SNP calling, and the closest genes aligned upstream and downstream of each marker were retrieved. This enabled the association of 69 markers with 220 isoforms representing 84 genes. Thirty-five of these genes were located in 26 modules in the coexpression network; a summary of the alignment results is provided in Table S6. Among the annotated genes, a disease resistance protein associated with the Chr1A_90316612 marker was particularly interesting. To obtain better visualization of the biological processes associated with all the annotated genes, their GO-

associated terms were retrieved and used for constructing a network using the REVIGO tool (Figure 4). The most prominent terms identified were linked to the regulation of transcription and translation, stress responses, and organismal development. The “modulation by virus of host process” term, which is associated with a peroxisomal oxidase and the marker Chr6A_86163774, was also displayed.

Additionally, the sequence of *Scmv2* from maize was aligned against the transcriptome, which resulted in several hits, the vast majority of which with TRINITY_DN5998_c0_g1. This gene was annotated as an auxin-binding protein and located within functional module 5 in the coexpression network—in which four genes close to SCMV resistance SNPs were also located. Among the hits with TRINITY_DN5998_c0_g1, the one with the highest E-value (1.07e-176) occurred in isoform 9, representing a 524-bp alignment with 88.1% identity. This and all isoforms of the gene also presented high-scoring alignments with the region containing Sspon.02G0027920-1A.

As a last strategy to investigate the biological processes involved in SCMV resistance, the GO terms of the 14,732 genes present in the 26 modules containing the genes associated with resistance were determined. Because these genes were initially associated with a very large number of GO terms (3,859 terms in the biological process category), a GO enrichment analysis using Fisher’s test with Bonferroni correction was performed before further procedures. The 117 terms resulting from this analysis were used to construct a TreeMap on REVIGO (Figure S8). The main terms found were distinct from those identified based only on the associated genes and included “response to salt stress”, “DNA integration”, “regulation of multicellular organismal process”, “photosynthesis”, and “seed germination”.

3. Discussion

Despite recent advances in sequencing technologies, genomic studies in sugarcane remain considerably hindered by the complexity of sugarcane's genome (Thirugnanasambandam *et al.*, 2018). Studies focusing on resistance to viruses are particularly limited due to sugarcane plants' large size and vegetative propagation, which limit the size of controlled experiments and the number of genotypes that can be evaluated (da Silva *et al.*, 2015; Pimenta *et al.*, 2021). Moreover, the genotypes to be used in such assays can be made virus free only by tissue culture (Chatenet *et al.*, 2001; Dewanti *et al.*, 2016), a hard and time-consuming process. However, given the extensively reported economic impacts of SCMV infection on this crop's yield (Costa and Muller, 1982; Bailey and Fox, 1987; Smith *et al.*, 1992; Cronje *et al.*, 1994; Singh *et al.*, 2003), it is remarkable that this work represents the first genome-wide study targeting resistance to this pathogen in sugarcane.

Although our GWAS analysis did not reveal major loci controlling resistance to SCMV in sugarcane, it led to the identification of nine SNPs significantly associated with this trait, explaining a small (1.7%) to moderate (29.9%) percentage of the phenotypic variation. Compared to *Scmv1*, the major SCMV resistance QTL identified in maize, which explains 54-56% of the variation alone (Xia *et al.*, 1999; Soldanova *et al.*, 2012), these findings might seem modest. However, they fit in the upper range of those of other mapping studies in sugarcane; with the exceptions of resistance to brown and orange rust (Daugrois *et al.*, 1996; Yang *et al.*, 2018), markers explaining 10% or less of the phenotypic variation in traits of agronomic importance are common in this crop (Gouy *et al.*, 2015; Fickett *et al.*, 2019). Specifically, for resistance to SCMV, previous studies identified markers explaining 5-14% of the variance observed individually (Pinto *et al.*, 2013; Burbano *et al.*, 2022) or up to ~40% together (Barnes

et al., 1997). The apparent quantitative nature of resistance to SCMV in sugarcane attests to the limitations of traditional marker-assisted breeding in this crop and demonstrates the need for the deployment of high-throughput genotyping and specific methodologies for association analyses in sugarcane.

One finding from our study that contributes to the understanding of this panorama is the absence of a gene that is highly homologous to *ZmTrxh*, the causal gene at *Scmv1*, in the *S. spontaneum* genome. The expression of *ZmTrxh*, which encodes an atypical thioredoxin that acts as a molecular chaperone, is necessary to disrupt infection by SCMV (Liu *et al.*, 2017). This phenomenon could involve SCMV's RNA silencing suppressor helper-component proteinase (HC-Pro), which has been shown to interact with a maize ferredoxin (Cheng *et al.*, 2008). These proteins are part of the ferredoxin–thioredoxin system, which is directly involved in photosynthesis (Buchanan, 1991) and might interfere with the suppression of silencing by SCMV HC-Pro and thus with resistance to this virus. Similar to other Poaceae (Liu *et al.*, 2017), close orthologs of *ZmTrxh* are not present in *S. spontaneum*, resulting in the lack of this specific resistance mechanism in this crop. Since *ZmTrxh* is absent even in a few maize lines, its presence in other sugarcane genotypes cannot be completely ruled out. However, BLASTn alignments of its sequence were also performed against those of all other sugarcane genomes available to date, none of which returned significant alignments with the *ZmTrxh* sequence. Because sugarcane has a very narrow genetic basis (Panje and Babu, 1960), this gene is likely to be absent in other genotypes of commercial relevance.

In addition to performing a GWAS, ML algorithms coupled with FS were employed to predict genotype resistance or susceptibility to SCMV. Similar to previous works in which this genomic prediction methodology was applied to sugarcane to evaluate resistance to brown rust

(Aono *et al.*, 2020) and sugarcane yellow leaf virus (Pimenta *et al.*, 2021), very promising results for several metrics were achieved. These results are considerably superior to those obtained by Barnes *et al.* (1997), who predicted sugarcane resistance to SCMV with an accuracy of 76% based on random amplified polymorphic DNA markers. Importantly, our results arose from a highly restricted SNP set obtained by FS, composed of only 73 markers, none of which had been identified by GWAS. A similar joint learning methodology that is based on FS and ML and combines classification and regression strategies has recently been shown to be highly suitable for the genomic prediction of several agronomic traits of sugarcane and polyploid forage grass species (Aono *et al.*, 2022).

Unlike *Scmv1*, the causal gene at the second major SCMV resistance QTL from maize (*Scmv2*) has an ortholog in the *S. spontaneum* genome. Interestingly, one marker identified through FS (Chr2A_103190628) was found to be close (1.7 kb) to this region. Linkage disequilibrium is high in sugarcane, persisting for up to 2-3.5 Mb (Yang *et al.*, 2019b; Pimenta *et al.*, 2021). Thus, it is possible that this marker is linked to Sspon.02G0027920-1A, the *S. spontaneum* gene syntenic to the auxin-binding protein gene at *Scmv2* (Ding *et al.*, 2012). This is an indication of the potential suitability of FS methodologies for the identification of QTLs, which is supported by other studies in which researchers analyzed traits controlled by many loci (Zhou *et al.*, 2019).

To apply the findings of our study to sugarcane breeding, validation of the markers associated with SCMV resistance was performed. For sugarcane, validation of individual SNPs was successfully achieved only for resistance to orange rust (Yang *et al.*, 2018; McCord *et al.*, 2019). In the context of genomic prediction, SNP validation in sugarcane test populations was recently implemented by Hayes *et al*. (2021), who employed single-dose markers obtained

through a SNP chip and achieved mean predictive accuracies of 29-47% for various agronomic traits. Due to the high cost of chip genotyping (De Donato *et al.*, 2013; Bajgain *et al.*, 2016) and the importance of including allele dose information in polyploid genetic studies (de Bem Oliveira *et al.*, 2019; Aono *et al.*, 2020), the MonsterPlex technology was chosen for the validation of our GBS-based markers. However, a few issues arose with this method, with failure in the amplification of a considerable percentage of sugarcane genotypes (~22%) and loci (~35%).

To some extent, this is expected from the technique, which does not guarantee the successful amplification and sequencing of all targets. The nonamplification of some genotypes might have been a consequence of the genomic reference used for SNP calling—the genome of *S. spontaneum*, a species different from that of almost all the genotypes for which amplification failed (*S. officinarum*). Because the majority of modern sugarcane cultivars that should be targeted by marker-assisted breeding are hybrids of these two species (Panje and Babu, 1960), failures in the amplification of whole genotypes are expected to be minimized. However, this highlights the importance of providing high-quality sequence data for genetically complex species such as sugarcane, which would certainly contribute positively to the research and breeding of this crop. Nevertheless, this low-cost targeted sequencing technology has the potential to be a viable approach for sugarcane marker-assisted breeding, especially if coupled to the ML-based genomic prediction approach used in this study, which effectively reduces the number of markers to be genotyped, contributing to the cost effectiveness of the process. The good results regarding predictive accuracy (69.1%), precision (68.6%), recall (94.1%) and F1 value (79.3%) obtained with the RF model are a strong indicator that this approach can be adopted for other traits of economic importance in sugarcane.

Another objective of the present work was to contribute to the elucidation of the molecular processes involved in sugarcane resistance to SCMV. To do so, RNA-Seq data was employed to annotate the markers identified as being associated with resistance and to construct coexpression networks to further investigate biological processes linked to them. Although RNA-Seq data from sugarcane plants infected with SCMV are available (Akbar *et al.*, 2020), they include data from only two biological replicates, which equates to a very low sample number for network modeling—the WGCNA developers recommend a minimum of 15 samples to avoid noise and biologically meaningless inferences (Langfelder and Horvath, 2017). The summarization of GO terms from genes close to markers directly associated with SCMV resistance revealed a few general processes previously associated with responses to this virus; these included stress responses and the regulation of transcription and translation (Akbar *et al.*, 2020; da Silva *et al.*, 2020).

A more detailed examination of marker annotations revealed several genes previously linked to resistance to plant viruses; in many cases, these associations were established by RNA-Seq or proteomics. This is the case for allantoinases (Vuorinen *et al.*, 2010), GLO oxidases (Varela *et al.*, 2017), alpha-galactosidases (Naqvi *et al.*, 2019), WD repeat-containing protein homolog genes (Şahin-Çevik *et al.*, 2019), and pentatricopeptide repeat-containing proteins, which have also been associated with resistance to SCMV in maize through GWASs (Abdelkhalek *et al.*, 2018; Gustafson *et al.*, 2018). Similarly, Shen *et al.* (2021) identified a ribonuclease H protein gene at a QTL responsible for potyvirus resistance in soybean and showed that the expression of this gene was upregulated in resistant cultivars and influenced viral accumulation.

However, there is much more compelling evidence of associations with virus resistance in plants for other candidates identified in our study. For instance, resistance gene analogs with nucleotide-binding site leucine-rich repeat (NBS-LRR) motifs, such as RGA5, are often involved in disease resistance by their ability to recognize pathogenic effector proteins and induce effector-triggered immunity (Sekhwal *et al.*, 2015). RGA5 specifically has been shown to bind effectors of a fungal pathogen in rice (Cesari *et al.*, 2013), but NBS-LRR proteins also participate in the recognition and resistance to potyviruses (Ma *et al.*, 2018; Xun *et al.*, 2019). The existence of preliminary evidence of associations between polymorphisms in NBS-LRR protein genes and SCMV resistance in sugarcane (Brune and Rutherford, 2005) strengthens the hypothesis that RGA5 could in fact act as a resistance protein against infection by this virus.

Several genes that may represent susceptibility factors to SCMV were also annotated. For instance, a chloroplast carbonic anhydrase has been identified as a salicylic acid-binding protein that plays a role in the hypersensitive response of tobacco (Slaymaker *et al.*, 2002). Furthermore, an *Arabidopsis* homolog of this protein was subsequently shown to interact with potyviral HC-Pro, weakening host defense responses and facilitating viral infection (Poque *et al.*, 2018). A lower abundance of carbonic anhydrase was also associated with successful infection by *Tobamovirus* (Konakalla *et al.*, 2021). In *Arabidopsis*, SCE1, a SUMO-conjugating enzyme, has been shown to interact with potyviral RNA-dependent RNA polymerase, and SCE1 knockdown resulted in increased resistance to turnip mosaic virus (Xiong and Wang, 2013). This protein also interacts with the replication initiator protein of begomoviruses and interferes with their replication (Castillo *et al.*, 2004, 2007).

Additionally, three proteins that have chaperone activity and also participate in resistance to viruses were annotated. DNAJ and DNA-like proteins such as C76 and DNAJ 10 have been

shown to interact with the coat protein of potyviruses, benefitting viral infection and replication (Hofius *et al.*, 2007; Zong *et al.*, 2020). Similarly, a heavy metal-associated isoprenylated plant protein was shown to interact with the *Pomovirus* movement protein, affecting virus long-distance movement (Cowan *et al.*, 2018). Interestingly, transcripts of two chaperones and a heavy metal-associated isoprenylated protein differentially accumulated in response to SCMV in sugarcane (da Silva et al., 2020). Another protein annotated in the present study that has been shown to interact with the potyviral movement protein P3N-PIPO is a beta-glucosidase, possibly facilitating viral spread through the plant (Song *et al.*, 2016); beta-glucosidase genes have also been found in QTLs for resistance to SCMV and other potyviruses (Gustafson *et al.*, 2018; Rubio *et al.*, 2019). Thus, it would be of great value to perform yeast two-hybrid assays including these host proteins and SCMV coat and movement proteins, the results of which could elucidate the involvement of these proteins in the replication and movement of SCMV.

Finally, the increase in the number of enriched GO terms associated with resistance through our GWGCN analysis sheds light on the complex network of biological processes involved in resistance to SCMV. The investigation of modules in coexpression networks can reveal sets of genes that are modulated together to execute specific functions; this is based on the "guilt-by-association" principle, which proposes that components (in our case, genes) with correlated biological functions tend to interact in networks such as GWGCNs (Oliver, 2000; Wolfe *et al.*, 2005). According to the results of our analysis, biological processes enriched in SCMV resistance-associated modules included stress responses, regulation of transcription and translation, and a process that has long been known to be affected by SCMV infection (Irvine, 1971) but has not been featured by the analysis of genes directly associated with resistance—photosynthesis. Recent transcriptomic and proteomic studies have shown that infection by

SCMV indeed affects the regulation of genes and proteins involved in these processes (Wu *et al.*, 2013; Chen *et al.*, 2017; Akbar *et al.*, 2020; da Silva *et al.*, 2020). Notably, the results of our coexpression network analyses indicate that the expression of genes identified in the present study as being associated with SCMV resistance are also associated with those controlling such processes; thus, those genes possibly play roles in their regulation during viral infection.

Our study indicates that resistance to SCMV in sugarcane has a more quantitative nature than in maize, which is in accordance with what has been observed for most traits in this crop. It also provides evidence that the ML-based strategy employed represents a viable approach for marker-assisted breeding in sugarcane; this strategy should therefore be assessed for its efficacy for other quantitative traits of economic importance. The annotation of identified markers via a transcriptomic assembly and analysis of gene coexpression networks showed that associated genes participate in key mechanisms of resistance to SCMV. These findings also revealed strong candidates for future investigation of resistance to this virus, which could help elucidate the molecular mechanisms involved in it.

4. Experimental procedures

4.1. Plant material

The plant material employed in the present study has been described elsewhere (Pimenta *et al.*, 2021). The experimental population consisted of a panel of 97 sugarcane genotypes comprising wild accessions of *S. officinarum*, *S. spontaneum* and *Saccharum robustum*; traditional sugarcane and energy cane clones; and commercial cultivars from Brazilian breeding programs. The accession names and pedigree information are available in Table S1. A field experiment

following a randomized complete block design with three blocks was established in May 2017 at the Advanced Center for Technological Research in Sugarcane Agribusiness located in Ribeirão Preto, São Paulo, Brazil (4°52’34” W, 21°12’50” S). 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 row contained two plants, totaling six plants of each genotype per plot. Infection by SCMV isolate RIB-2 (Burbano *et al.*, 2022) was allowed to occur under natural conditions in conjunction with high inoculum pressure and a high incidence of aphid vectors.

4.2. Phenotyping

Plants were phenotyped in two cropping seasons: plant cane in February 2018 (9 months after planting) and ratoon cane in July 2019 (9 months after the first harvest). The severity of SCMV symptoms was assessed by 2-3 independent evaluators, who classified the top visible dewlap leaves in each plot by the use of a diagrammatic scale consisting of four levels of increasing intensity of mosaic symptoms (Figure S1).

The data normality was assessed by the Shapiro–Wilk test, 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 with the breedR R package (Munoz and Rodriguez, 2014) using a mixed model, as follows:

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

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

(BY_{jm}). Random effects included the genotype (G) and the residual error (e), representing nongenetic effects.

4.3. Genotyping

The library preparation and sequencing methods used were thoroughly described by Pimenta *et al.* (2021). Briefly, genomic DNA was extracted from the leaves and used for the construction of a GBS library following the protocol by Poland *et al.* (2012). For operational reasons, 94 out of the 97 genotypes of the panel were included in the library; genotypes 87, 88 and 95 were excluded (see Table S1). Two 150-bp single-end sequencing libraries were prepared, and their contents were sequenced on a NextSeq 500 instrument (Illumina, San Diego, USA). After checking the sequencing quality, three tools were used for SNP calling: SAMtools version 1.6 (Li *et al.*, 2009), FreeBayes version 1.1.0-3 (Garrison and Marth, 2012) and the TASSEL4-POLY pipeline (Pereira *et al.*, 2018). A monoploid chromosome set obtained from the *S. spontaneum* genome (Zhang *et al.*, 2018) that included the A haplotype and unassembled scaffolds was used as a genomic reference. After variant calling, VCFtools version 0.1.13 (Danecek *et al.*, 2011) was used to retain biallelic SNPs with a minor allele frequency of 0.1, a maximum of 25% missing data and a minimum sequencing depth of 50 reads. SNPs identified by TASSEL and at least one other tool were then selected, and the ratio between alleles (allele proportions, APs) was obtained for each marker.

4.4. Association analyses

4.4.1. Mixed modeling

Association analyses were performed using mixed linear modeling in the GWASpoly R package (Rosyara *et al.*, 2016). For these analyses, APs were transformed into genotypic classes with a fixed ploidy of 12 in the vcfR R package (Knaus and Grünwald, 2017), as proposed by Yang *et al.* (2019a). A realized relationship model (MM^T) matrix (VanRaden, 2008), built in GWASpoly, was included as a random effect, and three principal components from a principal component analysis performed with genotypic data were included as fixed effects. Six marker-effect models were used for association analyses, namely, general, additive, simplex dominant reference, simplex dominant alternative, diploidized general and diploidized additive models. Q-Q plots of $-\log_{10}(p)$ values of the markers were generated for all the models, and Manhattan plots were constructed for models with appropriate inflation profiles. The Bonferroni correction method with $\alpha = 0.05$ was used to establish the significance threshold for associations. The phenotypic variance explained by each marker (r^2) significantly associated with SCMV resistance was estimated using a linear model in R.

4.4.2. ML coupled with FS

Following a genomic prediction approach previously employed for sugarcane (Aono *et al.*, 2020; Pimenta *et al.*, 2021), ML algorithms coupled with FS were used to predict the attribution of genotypes to two groups: those that presented mosaic symptoms at any block or year (susceptible) and those that did not present symptoms in any case (resistant). Eight ML algorithms implemented in the scikit-learn Python 3 module (Pedregosa *et al.*, 2011) were tested:

adaptive boosting (AB) (Freund and Schapire, 1997), decision tree (DT) (Quinlan, 1986), Gaussian naive Bayes (GNB) (Friedman *et al.*, 1997), Gaussian process (GP) (Rasmussen, 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). Three FS techniques were employed 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) and univariate FS using analysis of variance (FS3) (Geurts *et al.*, 2006), which were also implemented in scikit-learn. The markers selected by at least two of these FS methods were identified and used with the referred ML algorithms to classify genotypes as resistant or susceptible. To implement a cross-validation strategy, a stratified K-fold (k=5) repeated 100 times for different data configurations was used. The following metrics were evaluated: accuracy (proportion of correctly classified items), recall (items correctly classified as positive among the total quantity of positives), precision (items correctly classified as positive among the total items identified as positive), and the F1 score (the harmonic mean of precision and recall). The area under the receiver operating characteristic (ROC) curve (AUC) was also calculated for all the models using scikit-learn and plotted with the ggplot2 R package (Wickham, 2011).

4.5. Marker validation

Markers significantly associated with SCMV resistance were subjected to validation in two additional panels with sugarcane genotypes previously assessed for this trait. The first panel comprised 28 wild accessions, including representatives of *S. officinarum*, *S. spontaneum*, *S.*

robustum, *Saccharum barberi*, and interspecific hybrids (da Silva *et al.*, 2015a), and the second panel comprised 64 Brazilian varieties and elite clones from the three main sugarcane breeding programs in Brazil (da Silva *et al.*, 2015b). These 92 genotypes (Table S2) were used for validation using MonsterPlex Technology (Floodlight Genomics, Knoxville, USA). DNA was extracted from leaves following the methods described by Aljanabi et al. (1999) or using the GenElute Plant Genomic DNA Miniprep Kit (Sigma–Aldrich, St. Louis, USA). DNA samples and marker flanking sequences were sent to Floodlight Genomics, where multiplex PCR was used to amplify ~100-bp fragments containing markers, which were then sequenced on a HiSeq platform (Illumina, San Diego, USA). Trimmomatic version 0.39 (Bolger *et al.*, 2014) was used to trim the single-end sequencing reads using a 5-bp sliding window with a minimum average Phred quality score of 20 and removing reads shorter than 30 bp. The trimmed reads were aligned to reference flanking sequences using Bowtie2 version 2.2.5 (Langmead and Salzberg, 2012), and SNP calling was performed using SAMtools and FreeBayes. After APs/genotypic classes were obtained for each locus, linear models in R were used to estimate marker r^2 values for each panel, and ML models were used to predict resistance phenotypes as previously described.

4.6. Synteny analyses

For synteny analyses, the coding DNA sequences (CDSs) of the causal genes at *Scmv1* and *Scmv2* were retrieved from the MaizeGDB database (Portwood *et al.*, 2019) and aligned against the *S. spontaneum* genome sequence using BLASTn (Altschul *et al.*, 1990). Synteny plots were constructed using Circos software version 0.69.9 (Krzywinski *et al.*, 2009). The *Scmv1* CDS was also aligned to the genome sequences of *S. spontaneum* Np-X (Zhang et al., 2022), *S.*

officinarum LA Purple (SRA Bioproject accession PRJNA744175), and the hybrids SP70-1143 (Grativol *et al.*, 2014), R570 (Garsmeur *et al.*, 2018), SP80-3280 (Souza *et al.*, 2019), and CC01-1940 (Trujillo-Montenegro *et al.*, 2021).

4.7. Coexpression network construction and marker annotation

To annotate markers associated with SCMV and investigate their expression profile, RNA-Seq data supplied by Marquardt *et al.* (2019) was used. This study provided data from samples with five biological replicates, each made up of four to five bulked leaves, which were considered suitable for the construction of a highly robust coexpression network. Sequencing data were downloaded from the Sequence Read Archive (SRA; BioProject PRJNA474042) and trimmed with Trimmomatic version 0.39 (Bolger *et al.*, 2014), with the default parameters.

A de novo transcriptome was assembled using Trinity version 2.5.1 (Grabherr *et al.*, 2011), with the minimum contig length set to 300 bp. The completeness of the assembly was evaluated with BUSCO version 5.1.2 (Simão *et al.*, 2015) using datasets of conserved orthologs from Viridiplantae. Annotations were performed with Trinotate (Bryant *et al*., 2017) and included homology searches of sequences in the UniProt database, domain identification according to information in the Pfam database, and predictions of signal peptides with SignalP and transmembrane domains using TMHMM. Salmon version 1.1.0 software (Patro *et al.*, 2017) was used for transcript quantification, with the default parameters used. Genes with a mean of less than 5 transcripts per million (TPM) in at least one sample type were filtered out to avoid genes expressed at low levels, and genes with no variance across quantifications were excluded using the WGCNA package (Langfelder and Horvath, 2008).

A global weighted gene coexpression network (GWGCN) was constructed with WGCNA. Pairwise Pearson correlations of TPM values that considered a power function to fit a scale-free independence were used. For that, a soft threshold power beta estimation of 25, corresponding to an r^2 value of 0.85, was estimated and generated a scale-free topology model. Functional modules in the network were defined by the use of the unweighted pair group method with arithmetic mean (UPGMA) based on a topological overlap matrix and dynamic dendrogram pruning based on the dendrogram only.

To annotate markers associated with SCMV resistance and locate them in the network, the de novo transcriptome assembly was aligned against the *S. spontaneum* genomic reference used for SNP calling via BLASTn, and the closest genes upstream and downstream of each marker at a maximum distance of 2 Mb were retrieved. The following parameters were used: a minimum of 90% identity, a minimum E-value of 1e-50, and best hit algorithm overhang and edge values of 0.1. Similarly, the CDSs of the causal genes at *Scmv1* and *Scmv2* were aligned against the transcriptome assembly using BLASTn with the default parameters.

All genes present in the network modules containing genes associated with SCMV resistance were recovered and used for a Gene Ontology (GO) enrichment analysis with the topGO R package (Alexa and Rahnenfuhrer, 2010) in conjunction with Fisher's test with a Bonferroni correction with $\alpha = 0.01$. The REVIGO tool (Supek *et al.*, 2011) was used for the visualization and analysis of GO categories of the genes associated with SCMV resistance and in enriched categories associated with the genes in the network modules.

Acknowledgments

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

Conflicts 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 and APS conceived the project and designed the experiments. RJGP, RCVB, MFS, IAA and LRP performed the phenotyping. RJGP and AHA performed the genotyping, analyzed the data and interpreted the results. RJGP wrote the manuscript. All the 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), 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). RCVB

received a PhD fellowship from PAEDEx-AUIP. APS received a research fellowship from CNPq (grant 312777/2018-3).

Data statement

All datasets analyzed during the current study are available online and referenced in the corresponding papers.

References

Abdelkhalek, A., Elmorsi, A., Alshehaby, O., Sanan-Mishra, N. and Hafez, E. (2018) Identification of genes differentially expressed in onion infected with Iris yellow spot virus. *Phytopathol. Mediterr.*, **57**, 334-340.

Akbar, S., Wei, Y., Yuan, Y., Khan, M.T., Qin, L., Powell, C.A., Chen, B. and Zhang, M. (2020) Gene expression profiling of reactive oxygen species (ROS) and antioxidant defense system following *Sugarcane mosaic virus* (SCMV) infection. *BMC Plant Biol.*, **20**, 532. https://doi.org/10.1186/s12870-020-02737-1

Alexa, A. and Rahnenfuhrer, J. (2010) *topGO: Enrichment Analysis for Gene Ontology. R Package Version*.

Aljanabi, S. M., Forget, L. and Dookun, A. (1999). An improved and rapid protocol for the isolation of polysaccharide-and polyphenol-free sugarcane DNA. *Plant Mol. Biol. Rep.*, **17(3)**, 281-282.

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. https://doi.org/10.1016/s0022-2836(05)80360-2

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.D.A., de Souza, A.P. and Kuroshu, R.M. (2020) Machine learning approaches reveal genomic regions associated with sugarcane brown rust resistance. *Sci. Rep.*, **10**, 20057. https://doi.org/10.1038/s41598-020-77063-5

Aono, A.H., Pimenta, R.J.G., Garcia, A.L.B., Correr, F.H., Hosaka, G.K., Carrasco, M.M., Cardoso-Silva, C.B., Mancini, M.C., Sforça, D.A., Santos, L.B., Nagai, J.S., Pinto, L.R., Landell, M.G.A., Carneiro, M.S., Balsalobre, T.W., Quiles, M.G., Pereira, W.A., Margarido, G.R.A. and Souza, A.P. (2021) The Wild Sugarcane and Sorghum Kinomes: Insights Into Expansion, Diversification, and Expression Patterns. *Front. Plant Sci.* **12**, 668623. doi: 10.3389/fpls.2021.668623

Aono, A.H., Ferreira, R.C.U., Moraes, A.D.C.L., Lara, L.A.D.C., Pimenta, R.J.G., Costa, E.A., Pinto, L.R., Landell, M.G.D.A., Santos, M.F., Jank, L. Barrios, S.C.L., do Valle, C.B., Chiari, L., Garcia, A.A.F., Kuroshu, R.M., Lorena, A.C., Gorjanc, G. and Souza, A. P. (2022) A joint learning approach for genomic prediction in polyploid grasses. *Scientific Reports*, **12(1)**, 1-17. https://doi.org/10.1038/s41598-022-16417-7

Bailey, R.A. and Fox, P.H. (1987) A preliminary report on the effect of sugarcane mosaic virus on the yield of sugarcane varieties NCo376 and N12. In *Proceedings of the South African Sugar Technologists' Association*, pp. 1-4.

Bajgain, P., Rouse, M.N. and Anderson, J.A. (2016) Comparing genotyping-by-sequencing and single nucleotide polymorphism chip genotyping for quantitative trait loci mapping in wheat. *Crop Sci.*, **56**, 232-248. https://doi.org/10.2135/cropsci2015.06.0389

Barnes, J., Rutherford, R., Botha, F., Barnes, J. and Rutherford, R. (1997) The identification of potential genetic markers in sugarcane varieties for the prediction of

disease and pest resistance ratings. *Proc. Annu. Congr. S. Afr. Sugar Technol. Assoc.*, **71**, 57-61.

Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114-2120. https://doi.org/10.1093/bioinformatics/btu170

Breiman, L. (2001) Random forests. *Mach. Learn.*, **45**, 5-32. https://doi.org/10.1023/a:1010933404324

Brune, A.E. and Rutherford, R.S. (2005) Investigating the association of sugarcane kinase analogs and NBS-LRR resistance gene analogs with susceptibility and/or resistance to the pathogens Ustilago scitaminea (smut) and sugarcane mosaic virus (SCMV). In *Proceedings of the 79th Annual Congress of South African Sugar Technologists' Association*. Kwa-Shukela, Mount Edgecombe, South Africa: South African Sugar Technologists' Association, pp. 235-238.

Bryant, D. M., Johnson, K., DiTommaso, T., Tickle, T., Couger, M. B., Payzin-Dogru, D., Lee T. J., Leigh, N. D., Kuo T. H., Davis, F. G., Bateman, J., Bryant, S., Guzikowski, A. R., Tsai, S. L., Coyne, S., Ye, W. W., Freeman Jr, R. M., Peshkin, L., Tabin, C. J., Regev, A., Haas, B. J. and Whited, J. L. (2017). A tissue-mapped axolotl de novo transcriptome enables identification of limb regeneration factors. Cell Rep., 18(3), 762-776.

Buchanan, B.B. (1991) Regulation of CO2 assimilation in oxygenic photosynthesis: the ferredoxin/thioredoxin system. Perspective on its discovery, present status, and future development. *Arch. Biochem. Biophys.*, **288**, 1-9. https://doi.org/10.1016/0003-9861(91)90157-e

Burbano, R.C.V., da Silva, M.F., Coutinho, A.E., Gonçalves, M.C., dos Anjos, I.A., Anjos, L.O.S., Perecin, D. and Pinto, L.R. (2022) Marker-trait association for resistance to sugarcane mosaic virus (SCMV) in a Sugarcane (*Saccharum* spp.) panel. *Sugar Tech*, 1-13. https://doi.org/10.1007/s12355-022-01131-5

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

Castillo, A.G., Kong, L.J., Hanley-Bowdoin, L. and Bejarano, E.R. (2004) Interaction between a geminivirus replication protein and the plant sumoylation system. *J. Virol.*, **78**, 2758-2769. https://doi.org/10.1128/jvi.78.6.2758-2769.2004

Castillo, A.G., Morilla, G., Lozano, R., Collinet, D., Perez-Luna, A., Kashoggi, A. and Bejarano, E. (2007) Identification of plant genes involved in TYLCV replication. In *Tomato Yellow Leaf Curl Virus Disease: Management, Molecular Biology, Breeding for Resistance* (Czosnek, H., ed). Dordrecht: Springer, pp. 207-221.

Cesari, S., Thilliez, G., Ribot, C., Chalvon, V., Michel, C., Jauneau, A., Rivas, S., Alaux, L., Kanzaki, H., Okuyama, Y., Morel, J.-B., Fournier, E., Tharreau, D., Terauchi, R. and Kroj, T. (2013) The rice resistance protein pair RGA4/RGA5 recognizes the Magnaporthe oryzae effectors AVR-Pia and AVR1-CO39 by direct binding. *Plant Cell*, **25**, 1463-1481. https://doi.org/10.1105/tpc.112.107201

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. https://doi.org/10.1094/pdis.2001.85.11.1177

Chen, H., Cao, Y., Li, Y., Xia, Z., Xie, J., Carr, J.P., Wu, B., Fan, Z. and Zhou, T. (2017) Identification of differentially regulated maize proteins conditioning *Sugarcane mosaic virus* systemic infection. *New Phytol.*, **215**, 1156-1172. https://doi.org/10.1111/nph.14645

Chen, T. and Guestrin, C. (2016) XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, California, USA: Association for Computing Machinery, pp. 785-794.

Cheng, Y.Q., Liu, Z.M., Xu, J., Zhou, T., Wang, M., Chen, Y.T., Li, H.F. and Fan, Z.F. (2008) HC-Pro protein of sugar cane mosaic virus interacts specifically with maize ferredoxin-5 in vitro and in planta. *J. Gen. Virol.*, **89**, 2046-2054. https://doi.org/10.1099/vir.0.2008/001271-0

Costa, A.S. and Muller, G.W. (1982) General evaluation of the impacts of virus diseases of economic crops on the development of Latin American Countries. In *Proc. Conf. Impact of Viral Diseases in Developing Latin American and Caribbean Countries*. Rio de Janeiro, pp. 216-130.

Cover, T. and Hart, P. (1967) Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory*, **13**, 21-27. https://doi.org/10.1109/tit.1967.1053964

Cowan, G.H., Roberts, A.G., Jones, S., Kumar, P., Kalyandurg, P.B., Gil, J.F., Savenkov, E.I., Hemsley, P.A. and Torrance, L. (2018) Potato mop-top virus co-opts the stress sensor HIPP26 for long-distance movement. *Plant Physiol.*, **176**, 2052-2070. https://doi.org/10.1104/pp.17.01698

Cristianini, N. and Shawe-Taylor, J. (2000) *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods.* Cambridge, UK: Cambridge University Press.

Cronje, C.P.R., Bechet, G.R. and Bailey, R.A. (1994) Symptom expression of sugarcane mosaic virus (SCMV) and associated effects on sugarcane yield. In *Proceedings of the Annual Congress South African Sugar Technologists' Association*, pp. 8-11.

D'Hont, A. and Glaszmann, J.C. (2001) Sugarcane genome analysis with molecular markers: a first decade of research. In *International Society of Sugar Cane Technologists. Proceedings of the XXIV Congress*. Mackay, Australia: Australian Society of Sugar Cane Technologists, pp. 556-559.

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. https://doi.org/10.1139/g98-023

Da Silva, M.F., Gonçalves, M.C., Brito, M.D.S., Medeiros, C.N., Harakava, R., Landell, M.G.D.A. and Pinto, L.R. (2020) Sugarcane mosaic virus mediated changes in cytosine methylation pattern and differentially transcribed fragments in resistance-contrasting sugarcane genotypes. *PLoS One*, **15**, e0241493. https://doi.org/10.1371/journal.pone.0241493

Da Silva, M.F., Gonçalves, M.C., Melloni, M.N.G., Perecin, D., Landell, M.G.A., Xavier, M.A. and Pinto, L.R. (2015a) Screening sugarcane wild accessions for resistance to Sugarcane mosaic virus (SCMV). *Sugar Tech*, **17**, 252-257. https://doi.org/10.1007/s12355-014-0323-4

Da Silva, M.F., Gonçalves, M.C., Pinto, L.R., Perecin, D., Xavier, M.A. and Landell, M.G.A. (2015b) Evaluation of Brazilian sugarcane genotypes for resistance to *Sugarcane*

mosaic virus under greenhouse and field conditions. *Crop Prot.*, **70**, 15-20. https://doi.org/10.1016/j.cropro.2015.01.002

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R. and Genomes Project Analysis Group (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156-2158. https://doi.org/10.1093/bioinformatics/btr330

Daugrois, J.H., Grivet, L., Roques, D., Hoarau, J.Y., Lombard, H., Glaszmann, J.C. and D'Hont, A. (1996) A putative major gene for rust resistance linked with a RFLP marker in sugarcane cultivar 'R570'. *Theor. Appl. Genet.*, **92**, 1059-1064. https://doi.org/10.1007/bf00224049

de Bem Oliveira, I., Resende, M.F.R., Ferrão, L.F.V., Amadeu, R.R., Endelman, J.B., Kirst, M., Coelho, A.S.G. and Munoz, P.R. (2019) Genomic prediction of autotetraploids; influence of relationship matrices, allele dosage, and continuous genotyping calls in phenotype prediction. *G3*, **9**, 1189-1198. https://doi.org/10.1534/g3.119.400059

De Donato, M., Peters, S.O., Mitchell, S.E., Hussain, T. and Imumorin, I.G. (2013) Genotyping-by-sequencing (GBS): a novel, efficient and cost-effective genotyping method for cattle using next-generation sequencing. *PLoS One*, **8**, e62137. https://doi.org/10.1371/journal.pone.0062137

Dewanti, P., Widuri, L.I., Ainiyati, C., Okviandari, P., Maisaro and Sugiharto, B. (2016) Elimination of SCMV (*Sugarcane Mozaik Virus*) and rapid propagation of virus-free sugarcane (*Saccharum officinarum* L.) using somatic embryogenesis. *Procedia Chem.*, **18**, 96-102. https://doi.org/10.1016/j.proche.2016.01.016

Ding, J., Li, H., Wang, Y., Zhao, R., Zhang, X., Chen, J., Xia, Z. and Wu, J. (2012) Fine mapping of *Rscmv2*, a major gene for resistance to sugarcane mosaic virus in maize. *Mol. Breed.*, **30**, 1593-1600. https://doi.org/10.1007/s11032-012-9741-8

Dußle, C., Quint, M., Melchinger, A., Xu, M. and Lübberstedt, T. (2003) Saturation of two chromosome regions conferring resistance to SCMV with SSR and AFLP markers by targeted BSA. *Theor. Appl. Genet.*, **106**, 485-493. https://doi.org/10.1007/s00122-002-1107-x

Dußle, C.M., Melchinger, A.E., Kuntze, L., Stork, A. and Lubberstedt, T. (2000) Molecular mapping and gene action of Scm1 and Scm2, two major QTL contributing to SCMV resistance in maize. *Plant Breed.*, **119**, 299-303. https://doi.org/10.1046/j.1439-0523.2000.00509.x

FAO (2022) *FAOSTAT: Production Sheet.* Rome: FAO.

Fickett, N., Gutierrez, A., Verma, M., Pontif, M., Hale, A., Kimbeng, C. and 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**, 1794-1801. https://doi.org/10.1016/j.ygeno.2018.12.002

Freund, Y. and Schapire, R.E. (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, **55**, 119-139. https://doi.org/10.1006/jcss.1997.1504

Friedman, N., Geiger, D. and Goldszmidt, M. (1997) Bayesian network classifiers. *Mach. Learn.*, **29**, 131-163. https://doi.org/10.1023/a:1007465528199

Garrison, E. and Marth, G. (2012) Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*

Garsmeur, O., Droc, G., Antonise, R., Grimwood, J., Potier, B., Aitken, K., Jenkins, J., Martin, G., Charron, C., Hervouet, C., Costet, L., Yahiaoui, N., Healey, A., Sims, D., Cherukuri, Y., Sreedasyam, A., Kilian, A., Chan, A., Van Sluys, M.A., Swaminathan, K., Town, C., Bergès, H., Simmons, B., Glaszmann, J.C., van der Vossen, E., Henry, R., Schmutz, J. and D'Hont, A. (2018) A mosaic monoploid reference sequence for the highly complex genome of sugarcane. *Nat. Commun.*, **9**, 2638. https://doi.org/10.1038/s41467-018-05051-5

Geurts, P., Ernst, D. and Wehenkel, L. (2006) Extremely randomized trees. *Mach. Learn.*, **63**, 3-42. https://doi.org/10.1007/s10994-006-6226-1

Gonçalves, M.C., Galdeano, D.M., Maia, I.D.G. and Chagas, C.M. (2011) Genetic variability of Sugarcane mosaic virus causing maize mosaic in Brazil. *Pesqui. Agropecu. Bras.*, **46**, 362-369. https://doi.org/10.1590/s0100-204x2011000400004

Gonçalves, M.C., Moreira, Y.J.C.B., Maia, I.G., Santos, A., Fantin, G.M., Chaves, A.L.R., Frangione, D.S.S. and Ulian, E.C. (2004) Identificação e caracterização de isolados pertencentes ao subgrupo do *Sugarcane mosaic virus* no estado de São Paulo. *Fitopatol. Bras.*, **29**, 129.

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., Santos, A.S., Maia, I.G., Chagas, C.M. and Harakava, R. (2007) Caracterização de um isolado do Sugarcane mosaic virus que quebra a resistência de variedades comerciais de cana-de-açúcar. *Fitopatol. Bras.*, **32**, 32-39. https://doi.org/10.1590/s0100-41582007000100004

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

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N. and Regev, A. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, **29**, 644-652. https://doi.org/10.1038/nbt.1883

Grativol, C., Regulski, M., Bertalan, M., McCombie, W.R., da Silva, F.R., Neto, A.Z., Vicentini, R., Farinelli, L., Hemerly, A.S., Martienssen, R.A. and Ferreira, P.C.G. (2014) Sugarcane genome sequencing by methylation filtration provides tools for genomic research in the genus *Saccharum*. *Plant J.*, **79**, 162-172. https://doi.org/10.1111/tpj.12539

Gustafson, T.J., Leon, N., Kaeppler, S.M. and Tracy, W.F. (2018) Genetic analysis of *Sugarcane mosaic virus* resistance in the wisconsin diversity panel of maize. *Crop Sci.*, **58**, 1853-1865. https://doi.org/10.2135/cropsci2017.11.0675

Hall, J.S., Adams, B., Parsons, T.J., French, R., Lane, L.C. and Jensen, S.G. (1998) Molecular cloning, sequencing, and phylogenetic relationships of a new potyvirus: sugarcane streak mosaic virus, and a reevaluation of the classification of the Potyviridae. *Mol. Phylogenetics Evol.*, **10**, 323-332. https://doi.org/10.1006/mpev.1998.0535

Hassan, M., Sahi, G.M., Wakil, W. and Imanat, Y. (2003) Aphid transmission of Sugarcane mosaic virus (SCMV). *Pak. J. Agric. Sci.*, **40**, 74-76.

Hayes, B. J., Wei, X., Joyce, P., Atkin, F., Deomano, E., Yue, J., Nguyen, L., Ross, E. M., Cavallaro, T., Aitken K. S. and and Voss-Fels, K. P. (2021). Accuracy of genomic prediction of complex traits in sugarcane. *Theor. Appl. Genet.*, **134(5)**, 1455-1462.

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

Irvine, J.E. (1971) Photosynthesis in sugarcane varieties infected with strains of sugarcane mosaic virus. *Physiol. Plant.*, **24**, 51-54. https://doi.org/10.1111/j.1399-3054.1971.tb06714.x

ISO (2022) *International Sugar Organization.* London: ISO.

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. https://doi.org/10.1111/1755-0998.12549

Konakalla, N.C., Nitin, M., Kaldis, A., Masarapu, H., Carpentier, S. and Voloudakis, A. (2021) DsRNA molecules from the Tobacco mosaic virus p126 gene counteract TMV-induced proteome changes at an early stage of infection. *Front. Plant Sci.*, **12**, 663707. https://doi.org/10.3389/fpls.2021.663707

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.*, **19**, 1639-1645. https://doi.org/10.1101/gr.092759.109

Langfelder, P. and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform.*, **9**, 559. https://doi.org/10.1186/1471-2105-9-559

Langfelder, P. and Horvath, S. (2017) WGCNA package: frequently asked questions.

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

Li, Z., Chen, J., Han, L., Wen, J., Chen, G., Li, H., Wang, Y., Zhao, R., Zhang, X., Xia, Z., Yan, J., Wu, J. and Ding, J. (2016) Association mapping resolving the major loci Scmv2 conferring resistance to sugarcane mosaic virus in maize. *Eur. J. Plant Pathol.*, **145**, 385-391. https://doi.org/10.1007/s10658-015-0852-z

Liu, Q., Liu, H., Gong, Y., Tao, Y., Jiang, L., Zuo, W., Yang, Q., Ye, J., Lai, J., Wu, J., Lübberstedt, T. and Xu, M. (2017) An atypical thioredoxin imparts early resistance to sugarcane mosaic virus in maize. *Mol. Plant*, **10**, 483-497. https://doi.org/10.1016/j.molp.2017.02.002

Liu, X.H., Tan, Z.B. and Rong, T.Z. (2009) Molecular mapping of a major QTL conferring resistance to SCMV based on immortal RIL population in maize. *Euphytica*, **167**, 229-235. https://doi.org/10.1007/s10681-008-9874-3

Ma, F.F., Wu, M., Liu, Y.N., Feng, X.Y., Wu, X.Z., Chen, J.Q. and Wang, B. (2018) Molecular characterization of NBS–LRR genes in the soybean Rsv3 locus reveals several divergent alleles that likely confer resistance to the soybean mosaic virus. *Theor. Appl. Genet.*, **131**, 253-265. https://doi.org/10.1007/s00122-017-2999-9

Marquardt, A., Henry, R.J. and Botha, F.C. (2019) Midrib sucrose accumulation and sugar transporter gene expression in YCS-affected sugarcane leaves. *Trop. Plant Biol.*, **12**, 186-205. https://doi.org/10.1007/s12042-019-09221-7

McCord, P., Glynn, N. and Comstock, J. (2019) Identifying markers for resistance to sugarcane orange rust (*Puccinia kuehnii*) via selective genotyping and capture sequencing. *Euphytica*, **215**, 150. https://doi.org/10.1007/s10681-019-2340-6

Melchinger, A.E., Kuntze, L., Gumber, R.K., Lübberstedt, T. and Fuchs, E. (1998) Genetic basis of resistance to sugarcane mosaic virus in European maize germplasm. *Theor. Appl. Genet.*, **96**, 1151-1161. https://doi.org/10.1007/s001220050851

Munoz, F. and Rodriguez, L.S. (2014) breedR: statistical methods for forest genetic resources analysis. In *Trees for the Future: Plant Material in a Changing Climate*. Tulln, Austria, pp. 13.

Naqvi, R.Z., Zaidi, S.S.E.A., Mukhtar, M.S., Amin, I., Mishra, B., Strickler, S., Mueller, L.A., Asif, M. and Mansoor, S. (2019) Transcriptomic analysis of cultivated cotton Gossypium hirsutum provides insights into host responses upon whitefly-mediated transmission of cotton leaf curl disease. *PLoS One*, **14**, e0210011. https://doi.org/10.1371/journal.pone.0210011

Oliver, S. (2000) Guilt-by-association goes global. *Nature*, **403**, 601-602. https://doi.org/10.1038/35001165

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

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A. and Kingsford, C. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, **14**, 417-419. https://doi.org/10.1038/nmeth.4197

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. and Dubourg, V. (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, **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**, 398-398. https://doi.org/10.1186/s12859-018-2433-6

Peterson, R. (2017) Estimating normalization transformations with bestNormalize. https://github.com/petersonR/bestNormalize. [Accessed February 23, 2021].

Pimenta, R.J.G., Aono, A.H., Burbano, R.C.V., Coutinho, A.E., da Silva, C.C., Dos Anjos, I.A., Perecin, D., Landell, M.G.D.A., Gonçalves, M.C., Pinto, L.R. and de Souza, A.P. (2021) Genome-wide approaches for the identification of markers and genes associated with sugarcane yellow leaf virus resistance. *Sci. Rep.*, **11**, 15730. https://doi.org/10.1038/s41598-021-95116-1

Pinto, L.R., Gonçalves, M.C., Galdeano, D.M., Perecin, D., Medeiros, C.N.F., Gonçalves, B.S., Mancini, M.C. and Landell, M.G.D.A. (2013) Preliminary investigation of sugarcane mosaic virus resistance and marker association in a sugarcane family sample derived from a bi-parental cross. In *Proceedings of International Society of Sugar Cane Technologists*. Brisbane: International Society of Sugar Cane Technologists.

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. https://doi.org/10.1371/journal.pone.0032253

Popescu, M.C., Balas, V.E., Perescu-Popescu, L. and Mastorakis, N. (2009) Multilayer perceptron and neural networks. *WSEAS Trans. Circuits Syst.*, **8**, 579-588.

Poque, S., Wu, H.W., Huang, C.H., Cheng, H.W., Hu, W.C., Yang, J.Y., Wang, D. and Yeh, S.D. (2018) Potyviral gene-silencing suppressor HCPro interacts with salicylic acid (SA)-binding protein 3 to weaken SA-mediated defense responses. *Mol. Plant Microbe Interact.*, **31**, 86-100. https://doi.org/10.1094/mpmi-06-17-0128-fi

Portwood, J.L., Woodhouse, M.R., Cannon, E.K., Gardiner, J.M., Harper, L.C., Schaeffer, M.L., Walsh, J.R., Sen, T.Z., Cho, K.T., Schott, D.A., Braun, B.L., Dietze, M., Dunfee, B., Elsik, C.G., Manchanda, N., Coe, E., Sachs, M., Stinard, P., Tolbert, J., Zimmerman, S. and Andorf, C.M. (2019) MaizeGDB 2018: the maize multi-genome genetics and genomics database. *Nucleic Acids Res.*, **47**, D1146-D1154. https://doi.org/10.1093/nar/gky1046

Quinlan, J.R. (1986) Induction of decision trees. *Mach. Learn.*, **1**, 81-106. https://doi.org/10.1007/bf00116251

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

Rasmussen, C.E. (2003) Gaussian processes in machine learning. In *Advanced Lectures on Machine Learning. ML 2003. Lecture Notes in Computer Science* (Bousquet, O., von Luxburg, U. and Rätsch, G., eds). Berlin, Heidelberg: Springer, pp. 63-71.

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** https://doi.org/10.3835/plantgenome2015.08.0073

Rubio, B., Cosson, P., Caballero, M., Revers, F., Bergelson, J., Roux, F. and Schurdi-Levraud, V. (2019) Genome-wide association study reveals new loci involved in *Arabidopsis thaliana* and *Turnip mosaic virus* (TuMV) interactions in the field. *New Phytol.*, **221**, 2026-2038. https://doi.org/10.1111/nph.15507

Şahin-Çevik, M., Sivri, E.D. and Çevik, B. (2019) Identification and expression analysis of genes induced in response to tomato chlorosis virus infection in tomato. *Plant Pathol. J.*, **35**, 257-273. https://doi.org/10.5423/PPJ.OA.12.2018.0287

Sekhwal, M.K., Li, P., Lam, I., Wang, X., Cloutier, S. and You, F.M. (2015) Disease resistance gene analogs (RGAs) in plants. *Int. J. Mol. Sci.*, **16**, 19248-19290. https://doi.org/10.3390/ijms160819248

Sforça, D.A., Vautrin, S., Cardoso-Silva, C.B., Mancini, M.C., Romero-da Cruz, M.V., Pereira, G.D.S., Conte, M., Bellec, A., Dahmer, N., Fourment, J., Rodde, N., Van Sluys, M.-A., Vicentini, R., Garcia, A.A.F., Forni-Martins, E.R., Carneiro, M.S., Hoffmann, H.P., Pinto, L.R., Landell, M.G.D.A., Vincentz, M., Berges, H. and de Souza, A.P. (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. https://doi.org/10.3389/fpls.2019.00553

Shen, Y., Xie, L., Chen, B., Cai, H., Chen, Y., Zhi, H. and Li, K. (2021) Fine mapping of the RSC9 gene and preliminary functional analysis of candidate resistance genes in soybean (*Glycine max*). *Plant Breed.*, **141**, 49-62. https://doi.org/10.1111/pbr.12987

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210-3212. https://doi.org/10.1093/bioinformatics/btv351

Singh, S.P., Rao, G.P., Singh, J. and Singh, S.B. (1997) Effect of sugarcane mosaic potyvirus infection on metabolic activity, yield and juice quality. *Sugar Cane*, **5**, 19-23.

Singh, V., Sinha, O. and Kumar, R. (2003) Progressive decline in yield and quality of sugarcane due to sugarcane mosaic virus. *Indian Phytopathol.*, **56**, 500-502.

Slaymaker, D.H., Navarre, D.A., Clark, D., del Pozo, O., Martin, G.B. and Klessig, D.F. (2002) The tobacco salicylic acid-binding protein 3 (SABP3) is the chloroplast carbonic anhydrase, which exhibits antioxidant activity and plays a role in the hypersensitive defense response. *Proc. Natl Acad. Sci. USA*, **99**, 11640-11645. https://doi.org/10.1073/pnas.182427699

Smith, G.R., Ford, R., Frenkel, M.J., Shukla, D.D. and Dale, J.L. (1992) Transient expression of the coat protein of sugarcane mosaic virus in sugarcane protoplasts and expression inEscherichia coli. *Arch. Virol.*, **125**, 15-23. https://doi.org/10.1007/bf01309625

Soldanova, M., Cholastova, T., Polakova, M., Piakova, Z. and Hajkova, P. (2012) Molecular mapping of quantitative trait loci (QTLs) determining resistance to Sugarcane mosaic virus in maize using simple sequence repeat (SSR) markers. *Afr. J. Biotechnol.*, **11**, 3496-3501. https://doi.org/10.5897/AJB11.027

Song, P., Chen, X., Wu, B., Gao, L., Zhi, H. and Cui, X. (2016) Identification for soybean host factors interacting with P3N-PIPO protein of Soybean mosaic virus. *Acta Physiol. Plant.*, **38** https://doi.org/10.1007/s11738-016-2126-6

Souza, G.M., Van Sluys, M.-A., Lembke, C.G., Lee, H., Margarido, G.R.A., Hotta, C.T., Gaiarsa, J.W., Diniz, A.L., Oliveira, M.D.M., Ferreira, S.D.S., Nishiyama, M.Y., Ten-Caten, F., Ragagnin, G.T., Andrade, P.D.M., de Souza, R.F., Nicastro, G.G., Pandya, R., Kim, C., Guo, H., Durham, A.M., Carneiro, M.S., Zhang, J., Zhang, X., Zhang, Q., Ming, R., Schatz, M.C., Davidson, B., Paterson, A.H. and Heckerman, D. (2019) Assembly of the 373k gene space of the polyploid sugarcane genome reveals reservoirs of functional diversity in the world's leading biomass crop. *GigaScience*, **8**, giz129. https://doi.org/10.1093/gigascience/giz129

Supek, F., Bošnjak, M., Škunca, N. and Šmuc, T. (2011) REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One*, **6**, e21800. https://doi.org/10.1371/journal.pone.0021800

Tao, Y., Jiang, L., Liu, Q., Zhang, Y., Zhang, R., Ingvardsen, C.R., Frei, U.K., Wang, B., Lai, J., Lübberstedt, T. and Xu, M. (2013) Combined linkage and association mapping reveals candidates for *Scmv1*, a major locus involved in resistance to sugarcane mosaic virus (SCMV) in maize. *BMC Plant Biol.*, **13**, 162. https://doi.org/10.1186/1471-2229-13-162

Thirugnanasambandam, P.P., Hoang, N.V. and Henry, R.J. (2018) The challenge of analyzing the sugarcane genome. *Front. Plant Sci.*, **9**, 616-616. https://doi.org/10.3389/fpls.2018.00616

Trujillo-Montenegro, J.H., Cubillos, M.J.R., Loaiza, C.D., Quintero, M., Espitia-Navarro, H.F., Villareal, F.A.S., Valens, C.A.V., Barrios, A.F.G., De Vega, J., Duitama, J. and Riascos, J.J. (2021) Unraveling the genome of a high yielding colombian sugarcane hybrid. *Front. Plant Sci.*, **12**, 694859. https://doi.org/10.3389/fpls.2021.694859

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

Varela, A.L.N., Komatsu, S., Wang, X., Silva, R.G.G., Souza, P.F.N., Lobo, A.K.M., Vasconcelos, I.M., Silveira, J.A.G. and Oliveira, J.T.A. (2017) Gel-free/label-free proteomic, photosynthetic, and biochemical analysis of cowpea (*Vigna unguiculata* [L.] Walp.) resistance against Cowpea severe mosaic virus (CPSMV). *J. Proteom.*, **163**, 76-91. https://doi.org/10.1016/j.jprot.2017.05.003

Viswanathan, R. and Balamuralikrishnan, M. (2005) Impact of mosaic infection on growth and yield of sugarcane. *Sugar Tech*, **7**, 61-65. https://doi.org/10.1007/bf02942419

Vuorinen, A.L., Gammelgård, E., Auvinen, P., Somervuo, P., Dere, S. and Valkonen, J.P.T. (2010) Factors underpinning the responsiveness and higher levels of virus resistance realised in potato genotypes carrying virus-specific R genes. *Ann. Appl. Biol.*, **157**, 229-241. https://doi.org/10.1111/j.1744-7348.2010.00424.x

Wickham, H. (2011) ggplot2. *Wiley Interdiscip. Rev. Comput. Stat.*, **3**, 180-185. https://doi.org/10.1002/wics.147

Wolfe, C.J., Kohane, I.S. and Butte, A.J. (2005) Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinform.*, **6**, 227-227. https://doi.org/10.1186/1471-2105-6-227

Wu, J.Y., Ding, J.Q., Du, Y.X., Xu, Y.B. and Zhang, X.C. (2007) Genetic analysis and molecular mapping of two dominant complementary genes determining resistance to sugarcane mosaic virus in maize. *Euphytica*, **156**, 355-364. https://doi.org/10.1007/s10681-007-9384-8

Wu, L., Han, Z., Wang, S., Wang, X., Sun, A., Zu, X. and Chen, Y. (2013) Comparative proteomic analysis of the plant–virus interaction in resistant and susceptible ecotypes of maize infected with sugarcane mosaic virus. *J. Proteom.*, **89**, 124-140. https://doi.org/10.1016/j.jprot.2013.06.005

Wu, L., Zu, X., Wang, S. and Chen, Y. (2012) Sugarcane mosaic virus – long history but still a threat to industry. *Crop Prot.*, **42**, 74-78. https://doi.org/10.1016/j.cropro.2012.07.005

Xavier, M., Silva, M., Gonçalves, M., Pinto, L., Perecin, D. and Landell, M. (2013) Family selection for detection of promising crosses of sugarcane varieties for resistance to SCMV in Ribeirão Preto and Jaú. In *International Society of Sugar Cane Technologists: Proceedings of the XXVIIIth Congress*. São Paulo, Brazil: International Society of Sugar Cane Technologists.

Xia, X., Melchinger, A.E., Kuntze, L. and Lübberstedt, T. (1999) Quantitative trait loci mapping of resistance to sugarcane mosaic virus in maize. *Phytopathology*, **89**, 660-667. https://doi.org/10.1094/phyto.1999.89.8.660

Xiong, R. and Wang, A. (2013) SCE1, the SUMO-conjugating enzyme in plants that interacts with NIb, the RNA-dependent RNA polymerase of Turnip mosaic virus, is required for viral infection. *J. Virol.*, **87**, 4704-4715. https://doi.org/10.1128/JVI.02828-12

Xu, M.L., Melchinger, A.E., Xia, X.C. and Lübberstedt, T. (1999) High-resolution mapping of loci conferring resistance to sugarcane mosaic virus in maize using RFLP, SSR, and AFLP markers. *Mol. Gen. Genet.*, **261**, 574-581. https://doi.org/10.1007/s004380051003

Xun, H., Yang, X., He, H., Wang, M., Guo, P., Wang, Y., Pang, J., Dong, Y., Feng, X., Wang, S. and Liu, B. (2019) Over-expression of GmKR3, a TIR–NBS–LRR type R

gene, confers resistance to multiple viruses in soybean. *Plant Mol. Biol.*, **99**, 95-111. https://doi.org/10.1007/s11103-018-0804-z

Yang, X., Islam, M.S., Sood, S., Maya, S., Hanson, E.A., Comstock, J. and Wang, J. (2018) Identifying quantitative trait loci (QTLs) and developing diagnostic markers linked to orange rust resistance in sugarcane (*Saccharum* spp.). *Front. Plant Sci.*, **9**, 350. https://doi.org/10.3389/fpls.2018.00350

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

Yang, X., Song, J., Todd, J., Peng, Z., Paudel, D., Luo, Z., Ma, X., You, Q., Hanson, E., Zhao, Z., Zhao, Y., Zhang, J., Ming, R. and Wang, J. (2019b) 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 Biotechnol. J.*, **17**, 488-498. https://doi.org/10.1111/pbi.12992

Zhang, J., Zhang, X., Tang, H., Zhang, Q., Hua, X., Ma, X., Zhu, F., Jones, T., Zhu, X., Bowers, J., Wai, C.M., Zheng, C., Shi, Y., Chen, S., Xu, X., Yue, J., Nelson, D.R., Huang, L., Li, Z., Xu, H., Zhou, D., Wang, Y., Hu, W., Lin, J., Deng, Y., Pandey, N., Mancini, M., Zerpa, D., Nguyen, J.K., Wang, L., Yu, L., Xin, Y., Ge, L., Arro, J., Han, J.O., Chakrabarty, S., Pushko, M., Zhang, W., Ma, Y., Ma, P., Lv, M., Chen, F., Zheng, G., Xu, J., Yang, Z., Deng, F., Chen, X., Liao, Z., Zhang, X., Lin, Z., Lin, H., Yan, H., Kuang, Z., Zhong, W., Liang, P., Wang, G., Yuan, Y., Shi, J., Hou, J.,

Lin, J., Jin, J., Cao, P., Shen, Q., Jiang, Q., Zhou, P., Ma, Y., Zhang, X., Xu, R., Liu, J., Zhou, Y., Jia, H., Ma, Q., Qi, R., Zhang, Z., Fang, J., Fang, H., Song, J., Wang, M., Dong, G., Wang, G., Chen, Z., Ma, T., Liu, H., Dhungana, S.R., Huss, S.E., Yang, X., Sharma, A., Trujillo, J.H., Martinez, M.C., Hudson, M., Riascos, J.J., Schuler, M., Chen, L.-Q., Braun, D.M., Li, L., Yu, Q., Wang, J., Wang, K., Schatz, M.C., Heckerman, D., Van Sluys, M.-A., Souza, G.M., Moore, P.H., Sankoff, D., VanBuren, R., Paterson, A.H., Nagai, C. and Ming, R. (2018) Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. *Nat. Genet.*, **50**, 1565-1573. https://doi.org/10.1038/s41588-018-0237-2

Zhang, S., Li, X., Wang, Z., George, M., Jeffers, D., Wang, F., Liu, X., Li, M. and Yuan, L. (2003) QTL mapping for resistance to SCMV in Chinese maize germplasm. *Maydica*, **48**, 307-312.

Zhou, W., Bellis, E.S., Stubblefield, J., Causey, J., Qualls, J., Walker, K. and Huang, X. (2019) Minor QTLs mining through the combination of GWAS and machine learning feature selection. *BioRxiv*, 712190. https://doi.org/10.1101/712190

Zong, T., Yin, J., Jin, T., Wang, L., Luo, M., Li, K. and Zhi, H. (2020). A DnaJ protein that interacts with soybean mosaic virus coat protein serves as a key susceptibility factor for viral infection. *Virus Res*., **281**, 197870.

Tables

Table 1. Predictive ability of machine learning (ML) models for predicting SCMV resistance before and after feature selection (FS). The ML models tested were adaptive boosting (AB), decision tree (DT), Gaussian naive Bayes (GNB), Gaussian process (GP), K-nearest neighbor (KNN), multilayer perceptron neural network (MLP), random forest (RF) and support vector machine (SVM).

Model	Accuracy	Precision	Recall	F1
Before Feature Selection				
AB	60.7	23.8	36.6	27.3
DT	52.8	30.7	29.5	29.2
GNB	54.4	30.9	31.7	28.9
GP	66.3	0.00	0.00	0.00
KNN	59.6	18.1	30.9	21.6
MLP	55.3	**54.6**	38.3	**43.0**
RF	**66.9**	8.2	**38.8**	13.2
SVM	66.6	1.4	8.4	2.4
Mean	60.3	21.0	26.8	20.7
After Feature Selection				
AB	86.2	70.0	87.9	76.5
DT	66.4	50.4	52.1	49.2
GNB	96.7	95.2	95.6	94.9
GP	93.2	81.7	98.2	88.4
KNN	95.7	87.3	**100.0**	92.8
MLP	**99.7**	**100.0**	99.3	**99.6**
RF	85.6	57.5	98.6	70.7
SVM	98.1	94.5	**100.0**	96.9
Mean	90.2	79.6	91.4	83.6

Table 2. Predictive accuracy, precision, recall and F1 scores of machine learning (ML) approaches employed to predict groups associated with SCMV resistance in the validation panel. The ML models tested were adaptive boosting (AB), decision tree (DT), Gaussian naive Bayes (GNB), Gaussian process (GP), K-nearest neighbor (KNN), multilayer perceptron neural network (MLP), random forest (RF) and support vector machine (SVM).

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
AB	61.7	64.7	86.3	73.9
DT	59.3	61.8	92.2	74.0
GNB	46.9	61.8	41.2	49.4
GP	65.4	65.3	**96.1**	77.8
KNN	64.2	63.8	100.0	77.9
MLP	63.0	**69.1**	74.5	71.7
RF	**69.1**	68.6	94.1	**79.3**
SVM	63.0	54.0	79.1	64.2
Mean	61.6	63.6	82.9	71.0

Figure legends

Figure 1. Manhattan plots generated from association analyses in which the best linear unbiased predictor (BLUP) values of SCMV symptom severity were used. Four different models were used: additive, simplex dominant reference (1-dom-ref), diploidized general (diplo-general) and diploidized additive (diplo-additive) models. On the x-axis, S represents scaffolds not associated with any of the *Saccharum spontaneum* chromosomes.

Figure 2. Receiver operating characteristic (ROC) curves and area under the curve (AUC) results concerning the performance of machine learning models for predicting SCMV resistance in which the full marker dataset (A) and markers selected by feature selection (FS) (B) were used. The machine learning models tested were adaptive boosting (AB), decision tree (DT), Gaussian naive Bayes (GNB), Gaussian process (GP), K-nearest neighbor (KNN), multilayer perceptron neural network (MLP), random forest (RF) and support vector machine (SVM).

Figure 3. Synteny plot of *Scmv2* on chromosome 3 of *Zea mays* (blue) and *Saccharum spontaneum* A chromosomes (red). The red and black ticks represent markers associated with sugarcane mosaic virus (SCMV) resistance by association mapping and feature selection (FS), respectively.

Figure 4. Network of Gene Ontology (GO) biological process terms obtained from genes associated with sugarcane mosaic virus (SCMV) resistance.

Figures

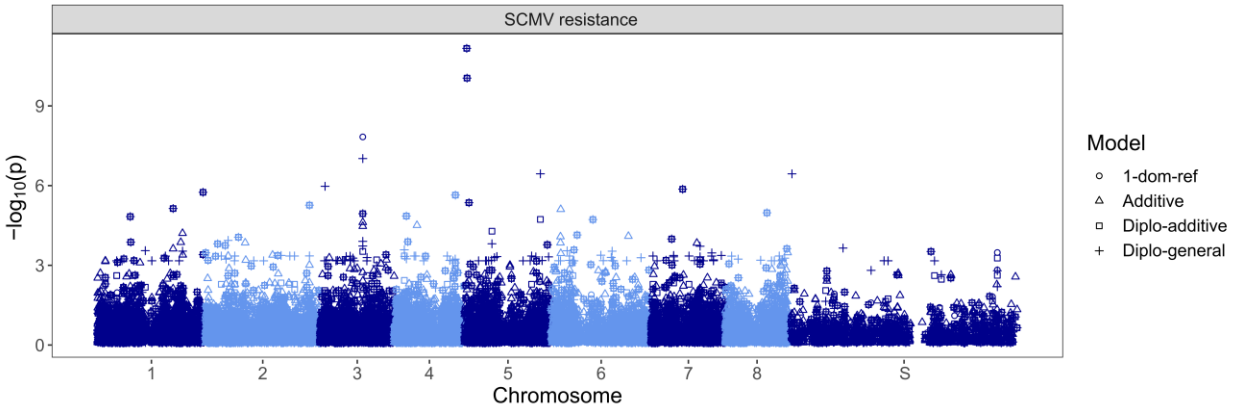

Figure 1

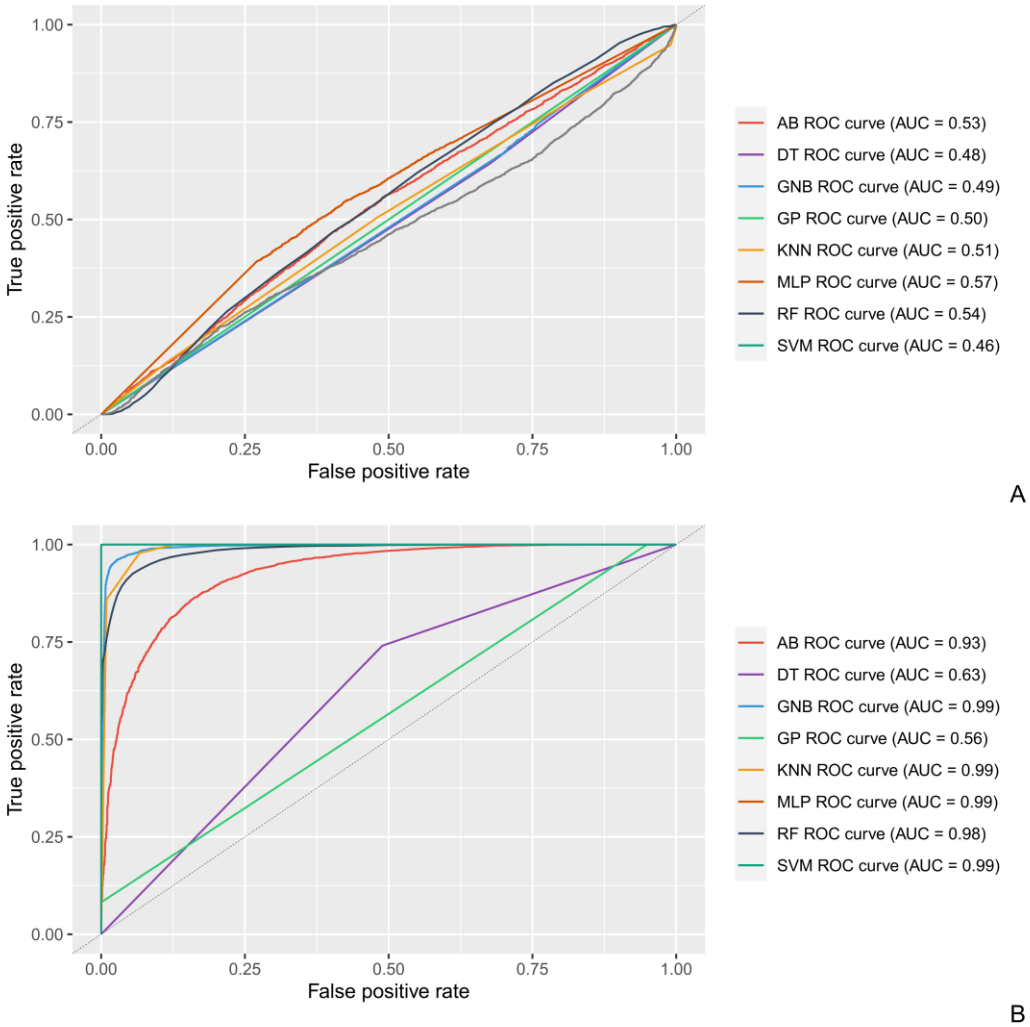
True positive rate
False positive rate
0.00
0.25
0.50
0.75
1.00
AB ROC curve (AUC = 0.53)
DT ROC curve (AUC = 0.48)
GNB ROC curve (AUC = 0.49)
GP ROC curve (AUC = 0.50)
KNN ROC curve (AUC = 0.51)
MLP ROC curve (AUC = 0.57)
RF ROC curve (AUC = 0.54)
SVM ROC curve (AUC = 0.46)
A
AB ROC curve (AUC = 0.93)
DT ROC curve (AUC = 0.63)
GNB ROC curve (AUC = 0.99)
GP ROC curve (AUC = 0.56)
KNN ROC curve (AUC = 0.99)
MLP ROC curve (AUC = 0.99)
RF ROC curve (AUC = 0.98)
SVM ROC curve (AUC = 0.99)
B

Figure 2

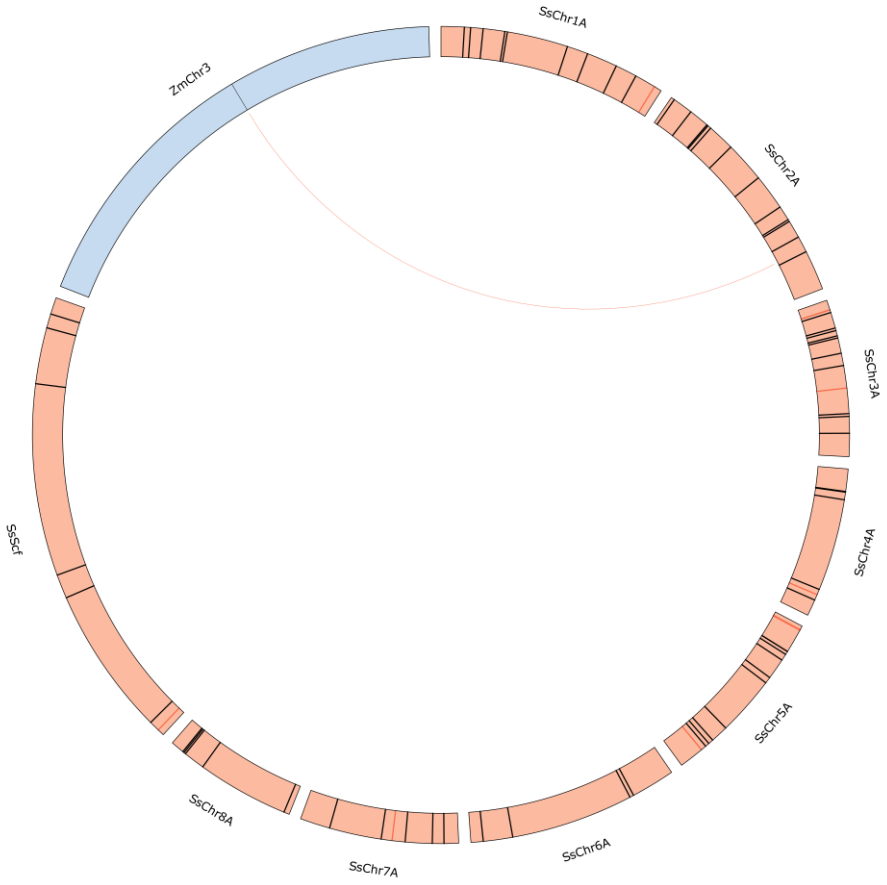
SsChr1A
SsChr2A
SsChr3A
SsChr4A
SsChr5A
SsChr6A
SsChr7A
SsChr8A
SsScf
ZmChr3

Figure 3

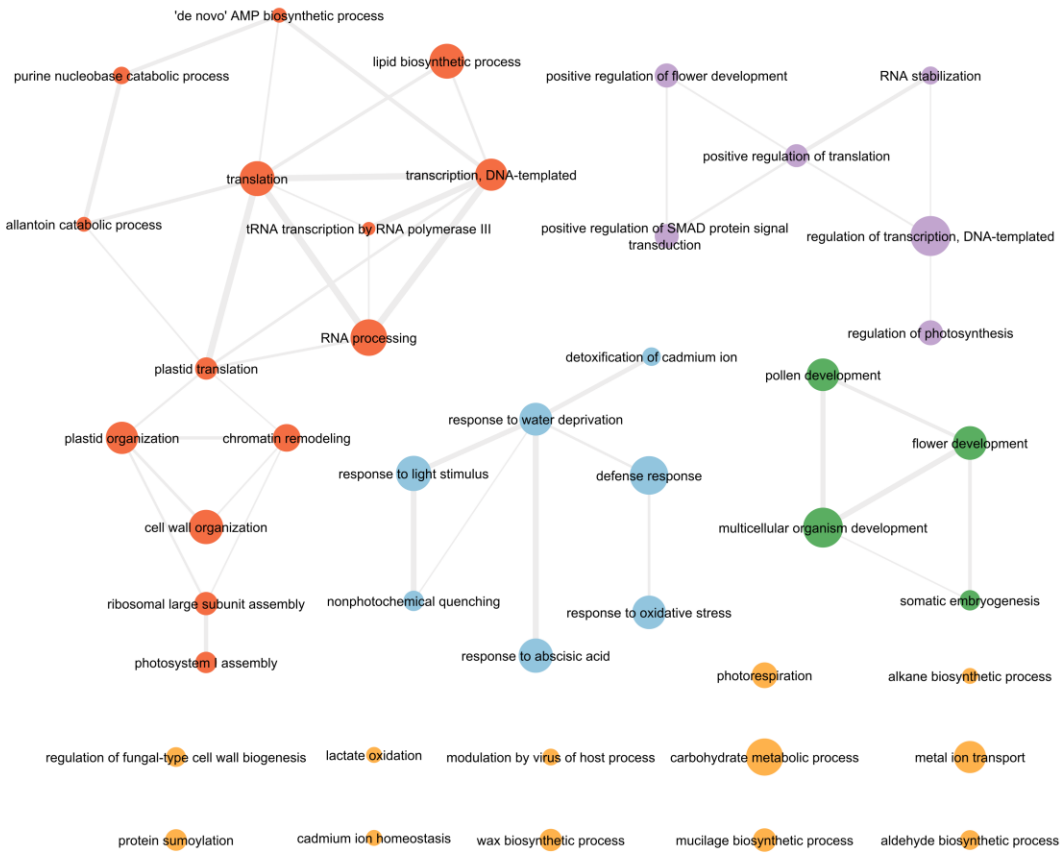
'de novo' AMP biosynthetic process
purine nucleobase catabolic process
lipid biosynthetic process
positive regulation of flower development
RNA stabilization
positive regulation of translation
translation
transcription, DNA-templated
allantoin catabolic process
tRNA transcription by RNA polymerase III
positive regulation of SMAD protein signal transduction
regulation of transcription, DNA-templated
RNA processing
regulation of photosynthesis
plastid translation
detoxification of cadmium ion
pollen development
response to water deprivation
plastid organization
chromatin remodeling
flower development
response to light stimulus
defense response
cell wall organization
multicellular organism development
ribosomal large subunit assembly
nonphotochemical quenching
response to oxidative stress
somatic embryogenesis
photosystem I assembly
response to abscisic acid
photorespiration
alkane biosynthetic process
regulation of fungal-type cell wall biogenesis
lactate oxidation
modulation by virus of host process
carbohydrate metabolic process
metal ion transport
protein sumoylation
cadmium ion homeostasis
wax biosynthetic process
mucilage biosynthetic process
aldehyde biosynthetic process

Figure 4