scientific reports

Check for updates

OPEN

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

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

Sugarcane yellow leaf (SCYL), caused by the sugarcane yellow leaf virus (SCYLV) is a major disease affecting sugarcane, a leading sugar and energy crop. Despite damages caused by SCYLV, the genetic base of resistance to this virus remains largely unknown. Several methodologies have arisen to identify molecular markers associated with SCYLV resistance, which are crucial for marker-assisted selection and understanding response mechanisms to this virus. We investigated the genetic base of SCYLV resistance using dominant and codominant markers and genotypes of interest for sugarcane breeding. A sugarcane panel inoculated with SCYLV was analyzed for SCYL symptoms, and viral titer was estimated by RT-qPCR. This panel was genotyped with 662 dominant markers and 70,888 SNPs and indels with allele proportion information. We used polyploid-adapted genome-wide association analyses and machine-learning algorithms coupled with feature selection methods to establish marker-trait associations. While each approach identified unique marker sets associated with phenotypes, convergences were observed between them and demonstrated their complementarity. Lastly, we annotated these markers, identifying genes encoding emblematic participants in virus resistance mechanisms and previously unreported candidates involved in viral responses. Our approach could accelerate sugarcane breeding targeting SCYLV resistance and facilitate studies on biological processes leading to this trait.

Sugarcane is one of the world's most important crops, ranking first in production quantity and sixth in net production value in 2016[1]. It is by far the most relevant sugar crop, accounting for approximately 80% of the world's sugar production[1,2] and is also a prominent energy crop. However, it has an extremely complex genome; modern cultivars are the product of a few crosses between two autopolyploid species. *Saccharum spontaneum* ($2n = 5x = 40$ to $16x = 128$; $x = 8$)[3], a wild stress-resistant but low-sugar species, was hybridized and backcrossed with *Saccharum officinarum* ($2n = 8x = 80$, $x = 10$)[4], which has a high sugar content but is sensitive to drought and susceptible to diseases. These procedures gave origin to plants with very large (ca. 10 Gb), highly polyploid, aneuploid and remarkably duplicated genomes[5,6]. This complexity directly affects sugarcane research and breeding and, until recently, it also prevented the use of codominance information in marker-assisted breeding strategies for this crop, limiting such approaches[7,8].

[1]Centre of Molecular Biology and Genetic Engineering (CBMEG), University of Campinas (UNICAMP), Campinas, São Paulo, Brazil. [2]Gustavo Galindo Velasco Campus, Littoral Polytechnic Superior School (ESPOL), Guayaquil, Ecuador. [3]São Paulo State University (UNESP), Jaboticabal, São Paulo, Brazil. [4]Advanced Centre for Technological Research in Sugarcane Agribusiness, Agronomic Institute of Campinas (IAC/APTA), Ribeirão Preto, São Paulo, Brazil. [5]Plant Protection Research Centre, Biological Institute (IB/APTA), São Paulo, São Paulo, Brazil. [6]Department of Plant Biology, Institute of Biology, University of Campinas (UNICAMP), Campinas, São Paulo, Brazil. ✉email: anete@unicamp.br

One of the diseases that affect this crop is sugarcane yellow leaf (SCYL), which is caused by sugarcane yellow leaf virus (SCYLV), a positive-sense ssRNA virus belonging to the *Polerovirus* genus[9,10]. The expression of SCYL symptoms is complex and usually occurs in late stages of plant development, being mainly characterized by the intense yellowing of midribs in the abaxial surface of leaves[11,12]. SCYLV alters the metabolism and transport of sucrose and photosynthetic efficiency[13,14], impairing plant development, which eventually reflects in productivity losses[15–20]. Many SCYL symptoms may, however, be caused by other stresses or plant senescence[12,15,21], making SCYL identification troublesome. Therefore, molecular diagnosis of SCYLV infection is of great importance; this was initially performed through immunological assays[11], but more sensitive and accurate methods using reverse transcription followed by quantitative polymerase chain reaction (RT-qPCR) were later developed[18,22,23].

Due to SCYL's elusive symptomatology, SCYLV's spread is silent; it is disseminated mostly during sugarcane vegetative propagation but is also transmitted by aphids, mainly the sugarcane aphid *Melanaphis sacchari* (Zehntner, 1897)[11]. Unlike other pathogens, the virus is not efficiently eradicated by thermal treatments[24]; the only way to thoroughly eliminate it is by meristem micropropagation[25,26], which is time-consuming and requires specialized infrastructure and personnel. These features make varietal resistance to SCYLV the most efficient resource to prevent damage and losses caused by this virus. Resistance has been explored in breeding programs and by a few genetic mapping studies[27–32]. However, research on SCYL genetics is not exempt from the difficulties generated by the complexity of the sugarcane genome[33]. Due to this crop's polyploid nature, most of these works employed dominantly scored molecular markers, implying a great loss of genetic information[34]. Additionally, they employed immunological methods to phenotype SCYLV resistance. The usage of dominant markers and the poor reliability of phenotyping were listed as key factors limiting the power of these studies[28,29].

Here, we evaluated the efficacy of several genome-wide approaches to identify markers and genes associated with SCYLV resistance. We analyzed a panel of *Saccharum* accessions inoculated with SCYLV, which were graded for the severity of SCYL symptoms, and their viral titer was estimated by relative and absolute RT-qPCR. This panel was genotyped with amplified fragment length polymorphisms (AFLPs) and simple sequence repeats (SSRs), as well as single nucleotide polymorphisms (SNPs) and insertions and deletions (indels) obtained by genotyping-by-sequencing (GBS). We then employed three distinct methodologies to detect marker-trait associations: the fixed and random model circulating probability unification (FarmCPU) method using dominant AFLPs and SSRs; mixed linear modeling using SNPs and indels, in which allele proportions (APs) in each locus were employed to establish genotypic classes and estimate additive and dominant effects; and several machine learning (ML) methods coupled with feature selection (FS) techniques, using all markers to predict genotype attribution to phenotypic clusters. Finally, we annotated genes containing markers associated with phenotypes, discussing the putative participation of these genes in the mechanisms underlying resistance to SCYLV.

Results

Phenotypic data analyses. A total of 97 sugarcane accessions inoculated with SCYLV were evaluated for the severity of SCYL symptoms and for viral titer estimated by relative and absolute RT-qPCR quantification in two consecutive years, as comprehensively described in Supplementary Results. Based on best linear unbiased prediction (BLUP) estimations, symptom severity was not correlated with the viral titer determined by relative ($p = 0.117$) or absolute ($p = 0.296$) quantification. We found, however, a significant ($p < 2.2e{-}16$) and strong ($r^2 = 0.772$) correlation between the values obtained by the two quantification methods, indicating their reliability (Supplementary Fig. 2).

Using BLUP values, we performed two hierarchical clustering on principal components (HCPC) analyses to investigate the classification of genotypes according to SCYLV resistance phenotypes—the first using BLUP values of SCYLV titers determined by RT-qPCR, and the second including BLUP values of all three traits analyzed. Both analyses indicated a division of the panel into three clusters (Supplementary Figs. 3–4)—named Q1-3 for the first HCPC and SQ1-3 for the second analysis. Factor maps wherein these groups are plotted onto the first two dimensions of HCPCs are shown in Fig. 1, and the attribution of genotypes to each cluster is available in Supplementary Table 4. Each group defined in the first HCPC presented significantly different SCYLV titers as estimated by both quantification methods (Supplementary Fig. 5, Supplementary Table 5). The second HCPC also resulted in a separation of groups with contrasting phenotypes: SQ1 accessions showed the least severe SCYL symptoms and the lowest titers of SCYLV; SQ2 accessions displayed significantly more severe disease symptoms and higher viral titers; and SQ3 accessions had the most severe disease symptoms and equally higher virus titers (Supplementary Fig. 6, Supplementary Table 5).

Genotyping and genetic analyses. After genotyping and filtering procedures, 93 accessions of the panel were successfully characterized with 550 AFLP fragments and 112 SSR fragments, totaling 662 polymorphic dominant markers. The GBS library constructed allowed the successful genotyping of 92 panel accessions, as described in detail in the Supplementary Results. We performed variant calling using BWA aligner and a monoploid chromosome set isolated from the *S. spontaneum* genome as a reference. This genome allowed the discovery of a large number of markers (38,710 SNPs and 32,178 indels) with AP information after rigorous filtering (Supplementary Tables 6–7). Additionally, unlike many of the references tested, it provided markers with information of position at chromosome level, allowing the estimation of long-distance linkage disequilibrium (LD). Pairwise LD between markers located within chromosomes was obtained and its decay was analyzed over distance. We observed high r^2 values (~0.4) between closely distanced markers, which dropped to 0.1 at approximately 2 Mb (Fig. 2).

The genetic structure of the panel was investigated separately using the two marker datasets generated – AFLPs and SSRs scored as dominant and codominant SNPs and indels with AP information –, and three different approaches—a discriminant analysis of principal components (DAPC), a principal component analysis

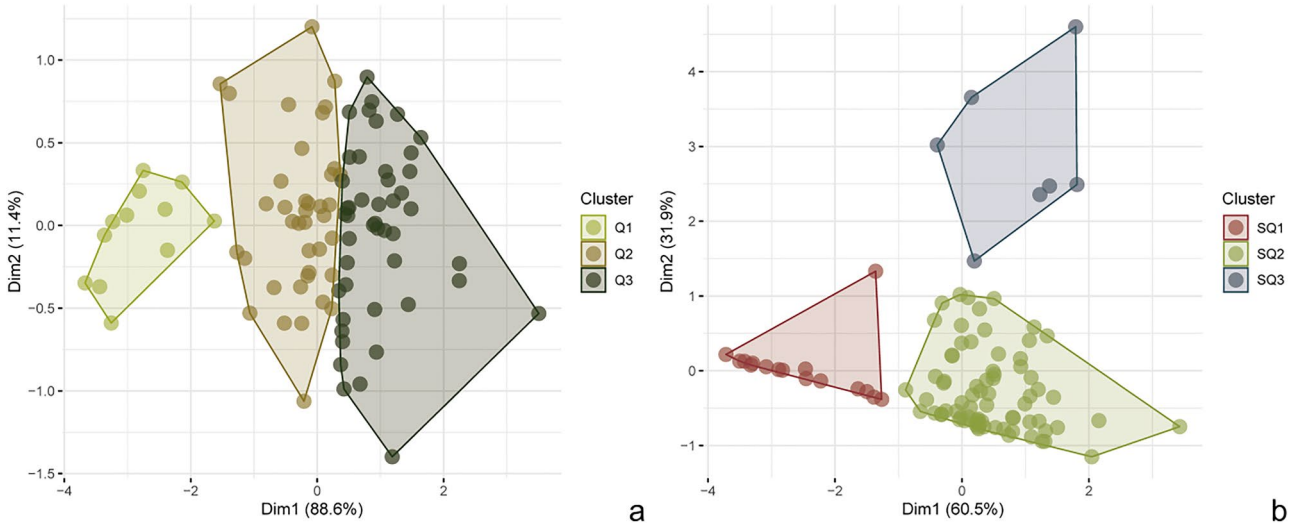

Figure 1. Factorial maps generated in the two hierarchical clustering on principal components (HPCP) analyses using BLUP values. (**a**) Factorial map of HCPC performed using the SCYLV titer determined by RT-qPCR. A division into three clusters (Q1, Q2 and Q3) was considered. (**b**) Factorial map of HCPC performed using SCYL symptom severity and SCYLV titer determined by RT-qPCR. A division into three clusters (SQ1, SQ2 and SQ3) was considered.

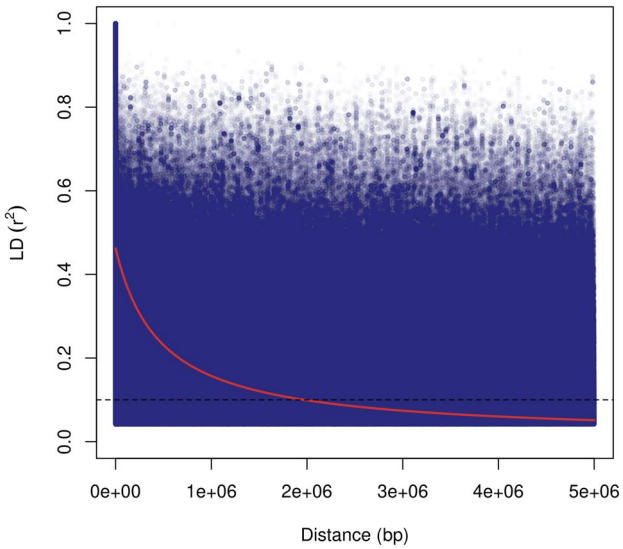

Figure 2. Decay of linkage disequilibrium (r^2) as a function of physical distance (bp) between pairs of 67,007 single nucleotide polymorphisms (SNPs) and insertions and deletions (indels) located on *Saccharum spontaneum* chromosomes 1A-8A. Only r^2 values with $p < 0.05$ are included.

(PCA) followed by k-means and a Bayesian clustering implemented in STRUCTURE. Results are thoroughly described in the Supplementary Results, and Supplementary Table 8 summarizes the allocation of genotypes to the clusters identified in each analysis. Analyses performed with dominant markers identified two to four clusters, depending on the structure analysis employed (Supplementary Figs. 7–10); however, we observed extensive similarities between the groups identified in each method. A similar pattern was observed when the same three structure analyses were performed with codominant markers. Each method resulted in a unique separation of accessions, varying between two and three groups (Supplementary Figs. 11–14), but the clustering obtained by these different analyses was overall coincident. We found, however, that using dominant or codominant markers yielded noticeably different outcomes. Some overlap was observed between clusters identified by the analyses using each set of markers but, overall, groups identified by these analyses shared little resemblance. Additionally, the results from these methods did not present correspondences with those from phenotype-based HCPCs.

Association analyses. *FarmCPU.* For FarmCPU analyses, we tested matrices obtained from each genetic structure analysis as covariates and ran the models with no covariates. The distribution of the genomic inflation factor λ (Supplementary Fig. 15) was normal ($p = 0.975$) and no significant differences ($p = 0.084$) were observed between the inflation of p values of models. Thus, we chose to conduct FarmCPU analyses using no covariates, as this resulted in the median value of λ closest to its theoretical value under the null hypothesis (λ = 1) and in appropriate profiles of inflation of p values as seen in quantile–quantile (Q–Q) plots (Supplementary Fig. 16). Using a Bonferroni-corrected threshold of 0.05, one marker-trait association was detected for symptom severity and five associations were detected for the viral titer estimated by each quantification method—with one marker being mutually associated with both. The percentage of phenotypic variance explained by each marker ranged from 9 to 30% (Supplementary Table 9).

Mixed modeling. Twelve combinations of population structure (Q) and kinship (K) matrices were tested as effects in the codominant association models. The distribution of λ in each Q + K combination (Supplementary Fig. 17) was not normal ($p = 3.253e{-}06$) and no significant differences ($p = 0.869$) were detected between models. Thus, following analyses were conducted with a Q + K combination that resulted in the median value of λ closest to 1, which was obtained with the combination of the first three PCs from a PCA with both the realized relationship (MM^T) and pseudodiploid kinship matrices. As the MM^T matrix is directly computed by the GWASpoly package, we considered the $Q_{PCA} + K_{MM}$ combination to be the most straightforward. Q–Q plots of the association analyses for SCYL symptom severity and SCYLV relative and absolute quantifications can be found in Supplementary Fig. 18; in general, all models showed appropriate inflation of p values.

A stringent significance threshold ($p < 0.05$ corrected by the Bonferroni method) was used to identify 35 nonredundant markers significantly associated with SCYL symptom severity (Fig. 3). Using this correction, no markers were significantly associated with SCYLV titer. In an attempt to establish a less conservative threshold for association analyses of these two traits, we employed the false discovery rate (FDR) for the correction of p values, which resulted in very low significance thresholds and the identification of thousands of associations as significant. Therefore, we ultimately opted to use an arbitrary threshold of $p < 0.0001$ to determine markers strongly associated with the two quantification traits. This resulted in 13 and 9 markers associated with SCYLV titer determined by relative and absolute quantifications, respectively (Fig. 3); one marker was common to both analyses. Supplementary Table 10 supplies information on all marker-trait associations identified by this approach. For each trait, we observed a redundancy between markers identified as significant by different marker-effect models; this observation was particularly common between the simplex dominant alternative and the diploidized models.

Machine learning coupled with feature selection. As a last marker-trait association method, we tested eight ML algorithms for predicting the attribution of genotypes to the phenotypic clusters identified in the HCPCs. When assessing their potential in this task using the full marker dataset, predictive accuracies varied greatly depending on the method and phenotypic groups under analysis. Accuracies were lower for the prediction of clusters associated with viral titer (Q), ranging between 39.2 and 49.6%, with an average of 44.5% (Supplementary Fig. 19a). For clusters identified including symptom severity data (SQ), accuracies were overall higher, albeit varying even more and being still unsatisfactory; they ranged between 7.9 and 73.9% (Supplementary Fig. 19b) and had an average of 58%. Therefore, we tested applying five FS methods to reduce the marker dataset, and constructed three additional reduced marker datasets consisting of intersections between FS methods.

These procedures led to considerably higher accuracies in predicting Q and SQ clusters. Three FS methods (FS1, FS2 and FS4) presented notably superior effects in increasing accuracy in both cases (Supplementary Fig. 20). In the two scenarios, the most accurate model-FS combination was a multilayer perceptron neural network (MLP) coupled with FS2, which was composed of 232 markers for Q and 170 markers for SQ. This combination resulted in average accuracies of 97.6% and 96.5% for the prediction of Q and SQ, respectively (Supplementary Tables 11 and 12). However, in both scenarios, MLP achieved the second-best results when using Inter2 datasets, composed of markers present in at least two out of the three best FS methods, which represented 190 markers for Q and 120 markers for SQ. With this strategy, we could achieve equally high accuracies (95.7% for Q and 95.4% for SQ) with further reductions in marker numbers. To farther evaluate the performance of MLP, we produced receiver operating characteristic (ROC) curves and calculated their respective area under the curves (AUCs). Prior to FS, MLP did not present satisfactory results, with ROC curves very close to the chance level and AUCs of 0.45–0.61 for Q and 0.40–0.56 for SQ (Fig. 4a). When Inter2 was used, ROC curves showed much better model performances, with AUCs of 1.00 for Q and of 0.98–1.00 for SQ (Fig. 4b). These results confirm that Inter2 markers are in fact associated with SCYLV resistance and that MLP is an appropriate model to predict clustering based on this dataset. The markers representing the reduced datasets associated with Q and SQ clusters can be found in Supplementary Tables 13 and 14, respectively. We observed twelve marker overlaps between the two datasets; interestingly, several of these markers were also identified as associated with phenotypes in the FarmCPU and mixed modeling analyses.

Marker mapping and annotation. For a better visualization of the physical location of all markers associated with SCYLV resistance, we constructed a map of their distribution along *S. spontaneum*'s "A" chromosomes (Fig. 5), in which we also included markers identified as associated with SCYLV resistance in previous mapping studies. Overall, markers were considerably spread along chromosomes; however, we observed regions of dense concentration of markers identified by various methods, such as the long arms of chromosomes 1 and 3. We also verified the proximity between several markers identified in the present work and by other authors, indicating their convergence and the reliability of the methods employed here.

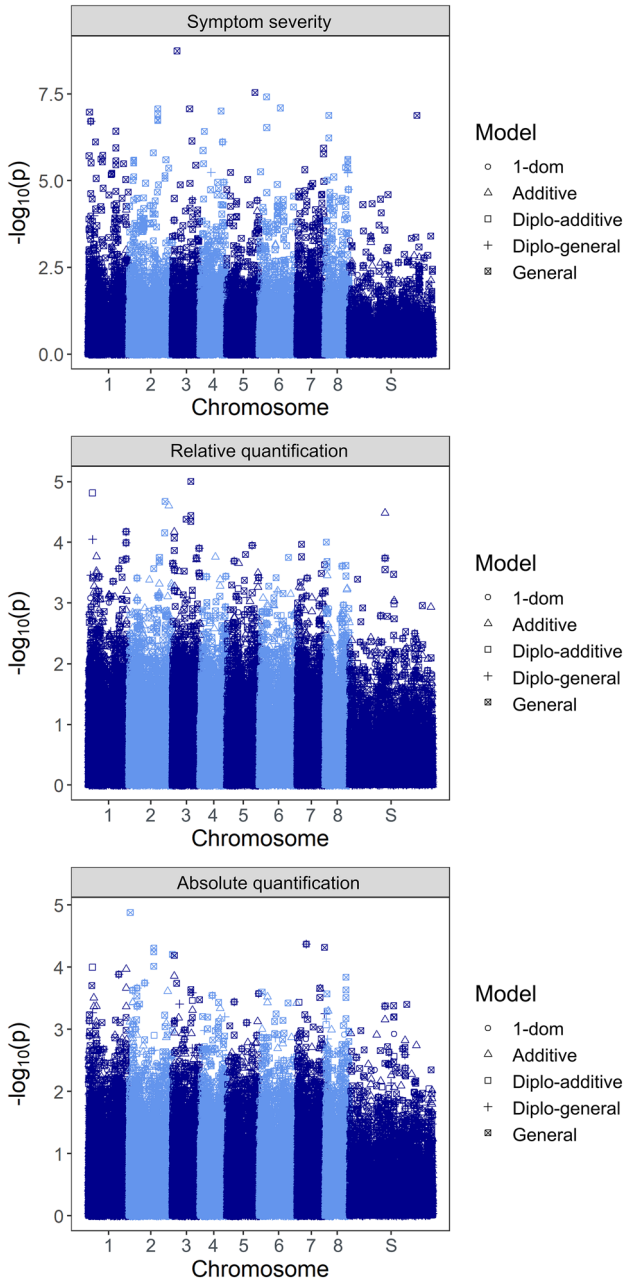

Figure 3. Manhattan plots generated in association analyses using the best linear unbiased predictor (BLUP) values of the three traits analyzed. Six different models were tested: general, additive, simplex dominant reference (1-dom-ref), simplex dominant alternative (1-dom-alt), diploidized general (diplo-general) and diploidized additive (diplo-additive). On the x-axis, S represents scaffolds not associated with any of the *S. spontaneum* chromosomes.

Out of the 362 nonredundant markers associated with all phenotypes, 176 were located in genic regions and could be annotated by aligning their 2000-bp neighboring regions with the coding sequences (CDSs) of 14 Poaceae species and *Arabidopsis thaliana* genomes; Supplementary Table 15 contains data on the alignment with the highest percentage of identity for each marker. In some cases, where two or more markers were closely located, coincident alignments and annotations were obtained; consequently, 148 genes were representative of all the best alignments. The large majority of top-scoring alignments (117) occurred with CDSs of *Sorghum bicolor*, the phylogenetically closest species among those used for alignment. Fewer alignments also occurred with the CDSs of other species. Several of the annotated genes could be associated with plant resistance to viruses, as detailed in the discussion.

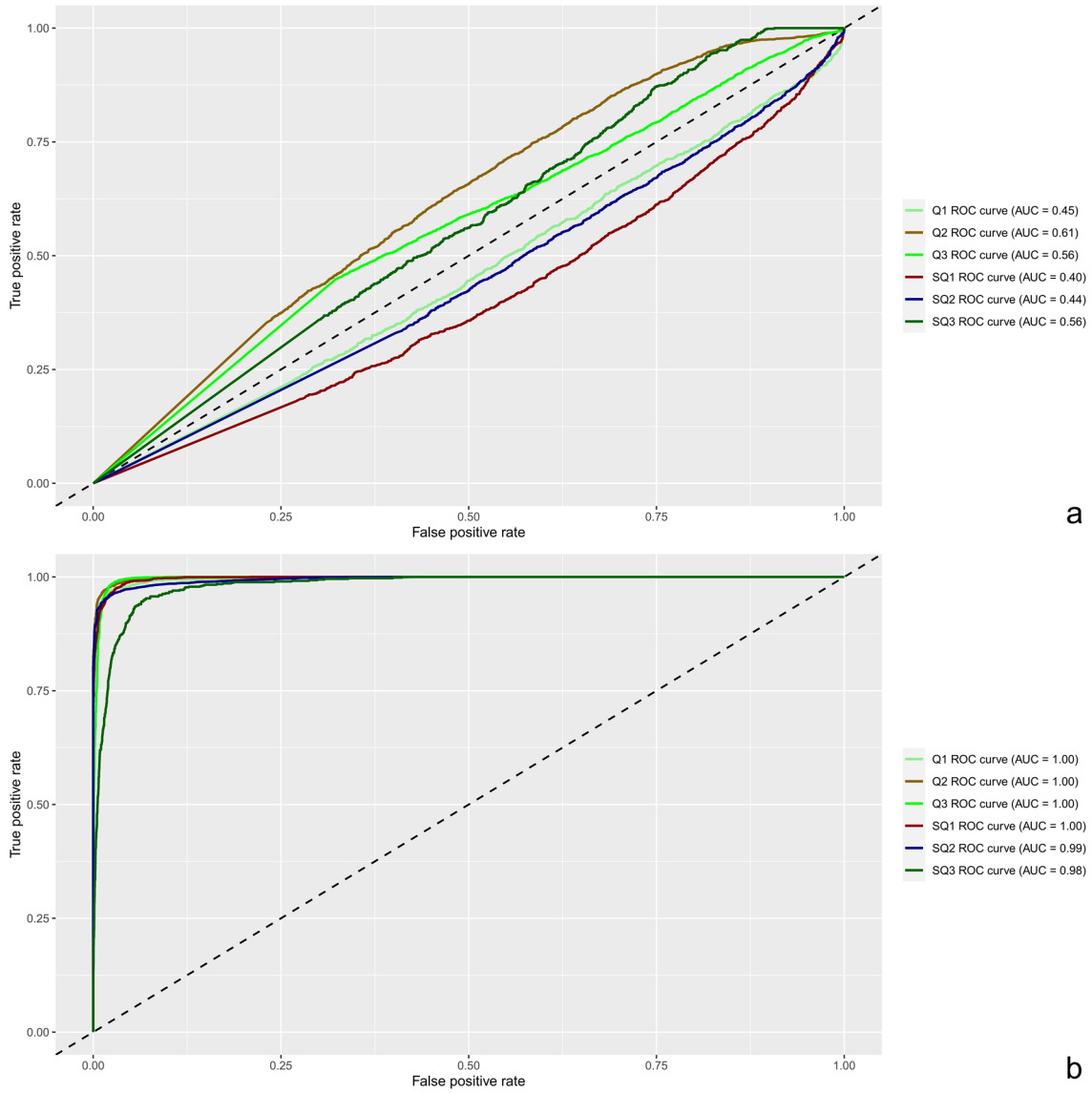

Figure 4. Receiver operating characteristic (ROC) curves and area under the curve (AUC) results regarding the performance of MLP in predicting clustering by by SCYLV titer determined by RT-qPCR (Q) and SCYLV titer determined by RT-qPCR and SCYL symptom severity (SQ). (**a**) Model performance obtained using the full marker dataset. (**b**) Model performance obtained using the marker dataset obtained from the intersection of at least two of the three best feature selection methods employed in the study (Inter2).

Discussion

We evaluated the severity of SCYL symptoms and SCYLV titer in a panel of 97 sugarcane accessions. These two traits are of great concern to breeding, as both have been associated with higher yield losses in SCYLV-infected sugarcane plants[18,22,35]. Prior to phenotyping, plants were subjected to high and uniform SCYLV inoculum pressure, an innovation over all previous SCYLV genetic mapping studies[27–31], which relied on natural infection under field conditions. Using RT-qPCR, currently regarded as the most precise method for SCYLV quantification[18], we assessed the viral titer in these genotypes. We found a strong and positive correlation between the BLUPs calculated for the SCYLV titers obtained by the two quantification methods employed, showing the consistency of the data. The absence of a perfect correlation might have arisen from intrinsic differences between methods, which have been responsible for disparities in viral quantification by RT-qPCR in other plant-virus interactions[36].

However, we observed no quantitative correlation between the severity of SCYL symptoms and SCYLV titers across the sugarcane genotypes analyzed. This finding corroborates a growing body of evidence suggesting that these traits are not strongly or necessarily correlated, i.e., high SCYLV titers are not a guarantee of more severe yellowing or of its development at all[37–39]. This reinforces the importance of SCYLV molecular screening of sugarcane clones by breeding programs, in an effort to avoid the employment of genotypes that accumulate high viral loads asymptomatically but may inconspicuously suffer yield losses as well as serving as a virus reservoir for vector transmission to other susceptible genotypes.

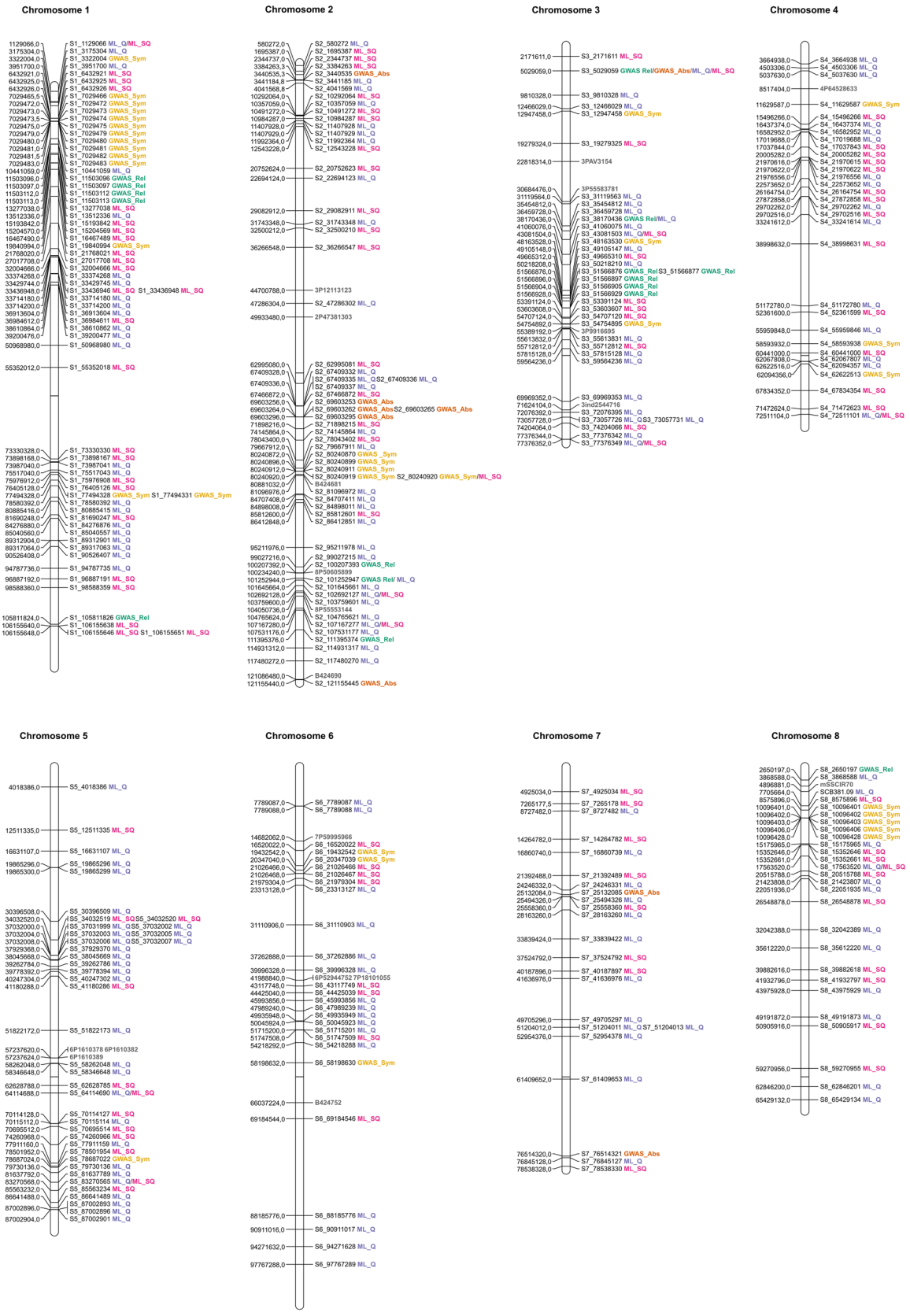

Figure 5. Distribution of markers associated with SCYLV resistance along *Saccharum spontaneum* "A" chromosomes. In each chromosome, marker positions are shown on the left, and marker names are indicated on the right, labeled and colored according to the method employed for their identification. Markers identified by previous mapping studies are colored in gray.

To further explore this issue, we performed two HCPC analyses to discriminate accessions based on their response to SCYLV, which led to the separation of clusters with considerable phenotypic differences. In the first HCPC, using only viral quantification data, we could discern groups with significant variation in viral titers. In the second analysis, which also included symptom severity data, clusters with even more contrasting responses to SCYLV could be discriminated. Cooper and Jones[40] proposed a terminology addressing plant responses to viral infections that is still employed today[41–43]. According to this proposal, once infected, plants present differences in their ability to restrict viral replication and invasion; the extremes of a spectrum of behaviors are plants termed susceptible and resistant. Additionally, they may also respond differently to the infection in terms of symptom development: another spectrum exists, at the extremes of which are sensitive and tolerant plants. In view of this nomenclature, we propose that the clusters identified in this second HCPC be described as follows: (SQ1) resistant, for sugarcane genotypes distinguished by low SCYLV titer and mild or no SCYL symptoms; (SQ2) tolerant, for genotypes that, despite exhibiting higher viral titers, presented few or no disease symptoms; and (SQ3) susceptible, for genotypes with the most severe symptoms and presenting high viral titers. This classification per se is of great use in sugarcane breeding, as it distinguishes not only sources of tolerance to SCYLV but also an exceptionally promising group of truly resistant genotypes.

Our main objective was, however, to identify markers associated with SCYLV resistance in a broader sense. With this aim, we performed genotyping with a combination of dominant and codominant markers, which has never been described for sugarcane. We evaluated the impact of using genomic references from various backgrounds in variant calling from GBS. In previous sugarcane GWASs, this was performed using the genome of *S. bicolor*[31,44–46], a close relative species with a well-assembled and annotated genome. However, in our analyses, this reference yielded a number of markers considerably inferior to other references. The methyl-filtered genome of the SP70-1143 cultivar yielded the most markers, in agreement with a previous study employing GBS[47]; this is a plausible outcome, as this method avoids sampling of methylated regions[48] which were also filtered out for this genomic assembly[49]. However, to choose the best reference for further analyses, we also considered the quality of the assembly, which greatly affects the results of GWASs in polyploids[50]. The best-assembled sugarcane genome available to date is the allele-defined genome of a haploid *S. spontaneum* accession[51]. Despite presenting one of the highest total tag alignment rates, this reference also gave a very high rate of multiple alignments, leading to the identification of relatively few markers. This was probably due to the alignment of tags to hom(e)ologous regions of different alleles rather than to the duplicated regions that we intended to avoid. To circumvent this situation, we conducted our analyses with markers isolated using a monoploid chromosome set obtained from this genome, which provided a large number of markers with reliable position information.

Using these codominant markers, we analyzed the decay of LD over distance. LD has long been hypothesized to be high in sugarcane due to the short breeding history and narrow genetic base of this crop; many studies using dominant markers have estimated it to be especially high at 5–10 cM[52–56]. The first study to use SNPs for this task and estimate LD decay in bp[57] indicated that LD was extremely long lasting, with the average r^2 decaying to 0.2 at 3.5 Mb in hybrids. Our results further confirm the persistence of LD at long distances in sugarcane, albeit indicating that it decayed more quickly—with r^2 dropping to 0.2 at less than 1 Mb and to 0.1 at 2 Mb. These results impact mapping studies, as a high LD implies that a low density of markers might be needed for accurate mapping of quantitative traits.

We tested several approaches to evaluate population structure in the panel using each distinct marker dataset generated, which yielded remarkably different results. Studies contrasting the usage of dominant and codominant markers in plants have shown discrepancies in measures of genetic structure and diversity[58–60], but this sort of comparison has never been performed including markers with dosage information in polyploids—let alone in sugarcane. In this crop, the most relatable findings available are those reported by Creste et al.[61], who showed that using different dominant markers can bias genetic analyses, and thus the choice of marker must be guided by the specific goal of each study. For GWASs—for which a high density of markers is usually necessary—SNPs and indels are currently more cost-effective, as they can be easily identified in much larger numbers, in addition to offering the possibility of estimating highly-informative allele dosages or APs[62–64]. Hence, we believe the results we obtained with codominant SNPs and indels are more reliable, as they lean on much more genetic information.

In contrast with the differences arising from the type of marker used, we observed little divergence between results of different structure methods performed with each marker dataset, and eventual discrepancies did not result in significant differences in the inflation of the association models, whose patterns were similar to those of previous studies[31,45,46,56]. Therefore, we opted to perform association analyses using the covariates that resulted in the value of λ closest to 1. For FarmCPU, this corresponded to the "naive" model with no covariates; for codominant mixed modeling analysis, this was the $Q_{PCA} + K_{MM}$ combination. K_{MM} is the usual choice of relationship matrix in polyploid association mapping[66–67], while Q matrices obtained from PCA are commonly used to control population structure in GWASs[68–70].

FarmCPU analyses using dominant markers identified one AFLP fragment significantly associated with symptom severity, which explained a small part of the phenotypic variation ($r^2 = 0.116$). Eight out of the nine markers associated with viral titer explained larger parts of the variation in the phenotypes (21–30%). These results are more promising than those obtained in a previous dominant GWAS targeting SCYLV resistance, which found r^2 ranging between 0.09 and 0.14[28]. Albeit low, values in this range are very common in sugarcane association studies. Evidence indicates that almost all of this crop's traits are highly quantitative, with the notable exception of brown rust resistance[71,72]. For other relevant traits, it is common to find most associated markers explaining ≤ 10% of the phenotypic variation[29,44,56].

A few authors have suggested that these suboptimal results could be improved with the usage of markers with dosage, which was also performed here using SNPs and indels with AP information. Although codominant mixed modeling analyses successfully identified markers associated with SCYL symptom severity using the Bonferroni correction, the same was not observed for SCYLV titer. This was probably influenced by the modest size of the

panel, a factor that restricts the power of GWASs[73,74]. As previously noted by Racedo et al.[75], assembling and phenotyping large sugarcane association panels is a challenging task. Thus, it is not uncommon for association studies of this crop to evaluate fewer than 100 genotypes[44,75–78]. Our study was particularly burdensome, as extremely laborious inoculation and quantification techniques were employed to generate highly reliable phenotypic data. Furthermore, the Bonferroni method is notorious for its conservative nature, poorly controlling false negatives[79–81]. This led us to establish an arbitrary threshold ($p < 0.0001$) to select markers strongly associated with SCYLV titer for further investigation. Using this methodology, we identified 57 nonredundant markers associated with the three phenotypes.

As a last approach to identify marker-trait associations, we tested several ML algorithms coupled with FS methods to predict genotype attribution to phenotypic clusters identified by HCPC analyses. Unlike methods built on classical statistics, these algorithms are not as heavily impacted by the sample size. We could achieve very high accuracies of prediction (up to 95%) with considerably reduced datasets comprising 120–190 markers. These results are very similar to what was obtained for predicting sugarcane brown rust resistance groups, where an accuracy of 95% was obtained using 131 SNPs[64]. Marker datasets selected by ML have rarely been employed in genetic association studies in plants, but the few existing examples show their power to identify genes associated with phenotypes of interest[82–84].

We annotated 176 markers associated with SCYLV resistance to 148 genes. Many candidates do not allow extensive discussion on their involvement in resistance to this disease, as they either have very generic descriptions or have not been previously linked to plant virus resistance. Other proteins have occasionally been associated with responses to viruses but are members of very large gene families with extremely diverse biological roles and will not be discussed. Remarkably, few candidates encode proteins previously associated with the response to SCYLV infection. This was the case for SbRio.10G317500.1, encoding a peroxidase precursor. Peroxidases are long known to be activated in response to pathogens, but most notably, a guaiacol peroxidase has been shown to be more active in sugarcane plants exhibiting SCYL symptoms than in uninfected or asymptomatic plants[85]. Our results provide further evidence that these enzymes are in fact involved in the response to SCYLV. Other candidates harboring markers associated with SCYLV resistance encode proteins with motifs previously associated with SCYLV resistance[31]: Sobic.001G023900, encoding a GATA zinc finger protein, and Sobic.001G200200 and Zm00001d037864_T030, both of which encode proteins containing tetratricopeptide repeats.

Other annotations included classic participants in more general disease resistance mechanisms, such as several genes encoding proteins with leucine-rich repeat (LRR) motifs. These structures are part of nucleotide-binding LRR (NBS-LRR) proteins, receptors that detect pathogen-associated proteins and elicit effector-triggered immunity[86]. Hence, NBS-LRRs have been widely shown to determine resistance to viruses in plants[87–89]. We found two LRR proteins (Sobic.008G156600.1 and Sobic.001G452600.1), one disease resistance NBS-LRR (Sobic.007G085400.1) and one N-terminal leucine zipper NBS-LRR resistance gene analog (Sobic.005G203500.1) associated with SCYLV resistance. Furthermore, we annotated one gene (Sobic.009G204800.1) that encodes a precursor of a receptor-like serine/threonine–protein kinase within the family to which LRR proteins belong. Yang et al.[31] also identified a serine/threonine-protein kinase associated with SCYLV resistance. We consider these proteins highly promising candidates to be involved in the recognition of infection by SCYLV, which could trigger response mechanisms leading to the restriction of the virus. Further virus–host interaction studies involving these proteins might help confirm this hypothesis, which would represent a major breakthrough in understanding resistance to SCYLV.

Two other annotated genes were readily identified as involved in plant disease resistance mechanisms. Sobic.010G131300.2 contains a Bric-a-Brac, Tramtrack, Broad Complex/Pox virus and Zinc finger (BTB/POZ) domain, while Sobic.007G198400.1 contains two BTB domains, as well as ankyrin repeat regions. These domains are present in and are essential for the function of NONEXPRESSOR OF PATHOGENESIS-RELATED GENES 1 (NPR1), a central player in plant disease responses[90,91]. This family of transcription factors is involved in establishing both systemic acquired resistance and induced systemic resistance[92], mediating the crosstalk between salicylic acid and jasmonic acid/ethylene responses[93]. Correspondingly, NPR1 has been widely shown to be involved in resistance to viruses[94,95], and it is therefore reasonable to suggest its participation in the response to infection by SCYLV.

We also found a few candidates with putative roles in the RNA interference mechanism, one of the most prominent processes that contribute to resistance against viruses in plants. This is the case for Sobic.001G214000.1, which encodes a Dicer. Dicers are part of a mechanism known as RNA silencing, recognizing and cleaving long double-stranded RNA molecules into mature small RNAs that guide the cleavage of viral mRNAs and disrupt virus replication[96]; accordingly, they have been linked to resistance to viruses in several plant species[97,98]. Another gene possibly involved in RNA interference is Sobic.009G121100, encoding a protein related to calmodulin binding—a calcium transducer that regulates the activity of various proteins with diverse functions[99] and has been widely implicated in viral resistance in plants, often playing roles in RNA interference[100–102]. Consequently, we consider these genes promising candidates in the regulation of SCYLV replication and spread *in planta*, as well as in the development of SCYL symptoms.

Two additional annotations linked to the mechanism of RNA interference are those of genes encoding proteins with F-box domains, SbRio.03G158900 and Sobic.002G019750.1. F-box proteins are involved in virus resistance in several plant species[103,104]. A particularly interesting case is FBW2 from *Arabidopsis thaliana*, which regulates AGO1, an Argonaute protein with a central role in RNA silencing[105] and repression of target viral RNAs[106–108]. Even more intriguing is the fact that one of the proteins encoded by the SCYLV genome, P0, contains an F-box-like domain and mediates the destabilization of AGO1, leading to the suppression of host gene silencing[109]. Whether the F-box proteins identified here play active roles in silencing of SCYLV remains a question to be investigated by further studies.

Other annotated genes may represent host factors involved in various steps of plant–virus interactions. For instance, Sobic.010G160500.4 encodes an RNA helicase with a DEAD-box domain, which is often coopted by viruses to promote viral translation or replication, thus playing important roles in regulating infection[110–112]. Similarly, soluble N-ethylmaleimide-sensitive-factor attachment protein receptor (SNARE) proteins such as Sobic.001G528000.1 are essential in the biogenesis and fusion of vesicles of several plant viruses[113–116]. We also found one gene encoding a myosin (Sobic.002G108000.1) and two genes related to kinesin (Sobic.001G346600.1 and Sobic.001G399200.2), all filament-associated motor proteins involved in the transport of organelles[117]. In a few cases, both myosins[118–120] and kinesins[121] have been shown to be involved in viral intercellular movement through poorly understood mechanisms. One last interesting annotation was Sobic.003G101500.1, a protein with a DNAJ domain. DNAJs have been shown to interact with proteins of various plant viruses and to be associated with resistance, sometimes being crucial for virus infection and spread[122–125]. We consider these genes to be promising candidates as host cofactors in the response to SCYLV infection.

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

Material and methods

Plant material and inoculation. The plant material and inoculation methods employed in the present study are described by Burbano et al.[126] and are in compliance with local and national regulations. The experimental population consisted of a panel of 97 sugarcane genotypes comprising wild germplasm accessions of *S. officinarum*, *S. spontaneum* and *Saccharum robustum*; traditional sugarcane and energy cane clones; and commercial cultivars originating from Brazilian breeding programs (Supplementary Table 1). To ensure plant infection with SCYLV, a field nursery was established in March 2016 at the Advanced Centre for Technological Research in Sugarcane Agribusiness located in Ribeirão Preto, São Paulo, Brazil (4°52′34″ W, 21°12′50″ S). Seedlings from sprouted setts of each genotype were planted in 1-m plots with an interplot spacing of 1.5 m. The cultivar SP71-6163, which is highly susceptible to SCYLV[15], was interspersed with the panel genotypes. *M. sacchari* vector aphids were reared on RT-PCR tested SCYLV-infected SP71-6163 plants. After an acquisition access period of at least 48 h, aphids were released weekly in the field nursery in July 2016. After plant growth, setts obtained from this nursery were used to install a field experiment following a randomized complete block design with three blocks in May 2017. Plants were grown in 1-m-long three-row plots with row-to-row and interplot spacings of 1.5 and 2 m, respectively. Each row contained two plants, totaling six plants of each genotype per plot. To further assist infection by SCYLV, the cultivar SP71-6163 was planted in the borders and between blocks, and *M. sacchari* aphids were again released in the field weekly for 5 months, starting from November 2017.

Phenotyping. Plants were phenotyped in two crop seasons: plant cane in June 2018 and ratoon cane in July 2019. The severity of SCYL symptoms was assessed by three independent evaluators, who classified the top visible dewlap leaves (TVDLs) of each plot using a diagrammatic scale established by Burbano et al.[126], as shown in Supplementary Fig. 1. In the same week as symptom evaluation was performed, fragments from the median region of at least one TVDL per plot were collected and stored at − 80 °C until processing. Total RNA was extracted from this tissue using TRIzol (Invitrogen, Carlsbad, USA). Samples were subjected to an additional purification process consisting of three steps: (1) mixing equal volumes of RNA extract and chloroform, (2) precipitating the RNA overnight with 2.5 volumes of 100% ethanol and (3) a conventional cleaning step with 70% ethanol. RNA was then quantified on a NanoDrop 2000 spectrophotometer (Thermo Scientific, Waltham, USA) and subjected to electrophoresis on a 1% agarose gel stained with ethidium bromide for integrity checks. Samples were next diluted, treated with RNase-Free RQ1 DNase (Promega, Madison, USA), quantified and diluted again for standardization, and converted to cDNA using the ImProm-II Reverse Transcription System kit (Promega, Madison, USA).

The SCYLV titer in each sample was determined by qPCR using GoTaq qPCR Master Mix (Promega, Madison, USA) on a Bio-Rad CFX384 Touch detection system (Bio-Rad, Philadelphia, USA). Two viral quantification methodologies were employed—one relative and one absolute—using primers and conditions as described by Chinnaraja and Viswanathan[127]. For both methods, a set of primers was used to amplify a 181-bp fragment from SCYLV ORF3 (YLSRT). For the relative quantification, an additional set of primers was used to amplify a 156-bp fragment of the 25S subunit of sugarcane ribosomal RNA (25SrRNA), used as an internal control. The $2^{-\Delta\Delta CT}$ method[128] was used to correct cycle threshold (CT) values; the sample with the highest CT and a melting temperature of 82.5 ± 0.5 °C for the YLSRT primers was used as a control for phenotyping in each year. The absolute quantification followed the methodology described by Chinnaraja et al.[39]. A pGEM-T Easy vector (Promega, Madison, USA) cloned with a 450-bp fragment from SCYLV ORF3 previously amplified by RT-PCR was used to construct a serial dilution curve with six points and tenfold dilutions between points, which were amplified on qPCR plates. All reactions were performed using three technical replicates.

Phenotypic data analyses. The normality of phenotypic data was assessed by Shapiro–Wilk tests, and normalization was carried out using the bestNormalize package[129] in R software[130]. BLUPs were estimated for each trait with the breedR R package[131] using a mixed model as follows:

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

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

Pearson's correlation tests were performed using the BLUPs to check the correlation between traits, and correlation distributions were plotted using the GGally R package[132]. To investigate the separation of genotypes according to phenotypes, we performed two HCPC analyses with the factoMineR package[133]—first using only viral quantification and then employing the three analyzed traits. The factoextra R package[134] was used to plot graphs associated with these analyses. Statistical differences between the phenotypes of the clusters identified in each HCPC were assessed by Kruskal–Wallis tests or analyses of variance (ANOVAs), depending on the distribution of the data. Post hoc Dunn's tests using the Bonferroni correction were performed with the R package dunn.test[135] to verify pairwise differences between clusters.

Genotyping. *Dominant markers.* Total DNA was extracted from leaves of each genotype following the method described by Aljanabi et al.[136]. AFLPs were developed using *Eco*RI and *Msp*I restriction enzymes (New England BioLabs). Digestion reactions were prepared in a final volume of 20 µL containing 300 ng DNA, 2.5 U of each restriction enzyme in 1X RL Buffer (New England BioLabs) and incubated for 3 h at 37 °C and for 5 min at 70 °C. Adapter ligation was conducted in a final volume of 40 µL containing 20 µL of the digestion reaction, 5× buffer (40 mM Tris pH 8.4, 100 mM KCl), 0.5 µM *Eco*RI adaptor, 5 µM *Msp*I adaptor, 1 mM ATP and 0.85 U of T4 DNA ligase (67 U/µL) (New England BioLabs). Ligation was performed at 37 °C for 2 h and 16 °C for 16 h. Preamplification was conducted with primers complementary to restriction enzyme adaptors and devoid of selective nucleotides at the 3′ end (*Eco*RI + 0 and *Msp*I + 0 primers) and using a 6× dilution of the digestion/ligation product. This reaction was performed in a final volume of 15 µL containing 2 µL of the 6× dilution digestion/ligation product, 1× PCR buffer (20 mM Tris pH 8.4, 50 mM KCl), 3.3 µM *Eco*RI + 0 and *Msp*I + 0 primers, 0.17 mM dNTPs, 2 mM $MgCl_2$ and 0.07 U Taq DNA polymerase. The cycling conditions were as follows: 29 cycles at 94 °C for 30 s, 56 °C for 1 min and 72 °C for 1 min. Preamplification reactions were diluted 10X and used for selective amplification reactions using combinations of *Eco*RI/*Msp*I primers with three selective nucleotides at the 3`end and the *Eco*RI primer labeled with fluorophores IRDye700 or IRDye800. Thirty-five selective primer combinations were used (Supplementary Table 2). The reaction was performed in a final volume of 10 µL containing 2.5 µL of the 10× diluted preamplification, 1× PCR buffer (20 mM Tris pH 8.4, 50 mM KCl), 0.05 µM of selective Eco700 labeled primer (or 0.07 µM Eco800 primer), 0.25 µM for Msp selective primer, 0.25 µM dNTPs, 2 mM $MgCl_2$, 0.5 U of Taq DNA polymerase. Cycling conditions were as follows: 94 °C for 30 s, 65 °C for 30 s and 72 °C for 1 min followed by 12 cycles at 94 °C for 30 s, 65 °C for 30 s (decreasing 0.7 °C/cycle) and 72 °C for 1 min, followed by 23 cycles of 94 °C for 30 s, 56 °C for 30 s and 72 °C for 1 min. Final amplicons were separated on a 6% denaturing polyacrylamide gel and visualized with a LI-COR 4300 DNA Analyzer (LI-COR, Lincoln, NE, USA).

Twelve SSR loci previously isolated from the sugarcane expressed sequence tag database[137–140] were used for SSR genotyping (Supplementary Table 3). PCR mixes were prepared and amplifications were conducted in a Bio-Rad MyCycler thermocycler (Bio-Rad, Philadelphia, USA) following the conditions previously established by Oliveira et al.[139] and Marconi et al.[140]; primers were labeled with fluorescent dyes IRDye700 and IRDye800 to allow band visualization. Amplicons were separated on a 6% denaturing polyacrylamide gel and visualized with a LI-COR 4300 DNA Analyzer. Due to sugarcane polyploidy, both AFLPs and SSRs were treated as dominant and scored based on the presence (1) or absence (0) of bands. After genotyping, genotypes and markers with over 10% missing data were removed, as well as markers with a MAF below 10%.

Genotyping-by-sequencing. Genomic DNA was extracted from leaves using the GenElute Plant Genomic DNA Miniprep Kit (Sigma-Aldrich, St. Louis, USA). The integrity of the DNA was verified by electrophoresis on a 1% agarose gel stained with ethidium bromide, and its concentration was determined using a Qubit 3.0 fluorometer (Thermo Scientific, Wilmington, USA). The construction of the GBS library was based on a protocol by Poland et al.[141] and used a combination of *Pst*I and *Mse*I restriction enzymes. For operational reasons, 94 out of the 97 genotypes of the panel were included in the library, which did not include genotypes 87, 88 and 95 (see Supplementary Table 1). The library was subjected to a purification step using polyethylene glycol as described by Lundin et al.[142] with slight modifications. It was then validated with a Fragment Analyzer (Agilent Technologies, Santa Clara, USA) and quantified by RT-qPCR in a Bio-Rad CFX384 Touch detection system using the KAPPA KK4824 kit (Kapa Biosystems, Wilmington, USA). Two 150-bp single-end sequencing libraries were prepared using the NextSeq 500/550 High Output Kit (Illumina, San Diego, USA) and sequenced on a NextSeq 500 (Illumina, San Diego, USA).

After checking sequencing quality with FastQC[143], we used Stacks software version 1.42[144] for demultiplexing and checking the amount of data generated for each sample. The TASSEL4-POLY pipeline[145], developed from TASSEL-GBS[146], was used for variant calling. Most parameters were set at their standard values; exceptions were the use of the "inclGaps" argument in the "DiscoverySNPCaller" plugin, the "misMat" argument with a value of 0.3 and the "callHets" argument in the "MergeDuplicateSNPs" plugin. Rather than aligning raw reads to a reference genome, the TASSEL-GBS pipeline first generates "tags"—unique sequences representing redundant reads—to reduce computation time[145]. We tested mapping tags against nine genomic references using two aligners: BWA version 0.7.2[147] and Bowtie2 version 2.2.5[148]. The genomic references used were as follows: the *S. bicolor* genome[149], the methyl-filtered genome of the sugarcane cultivar SP70-1143[49], a sugarcane RNA-Seq

assembly[150], a de novo assembly generated from GBS data following the GBS-SNP-CROP pipeline[151], a draft genome of the sugarcane cultivar SP80-3280[152], a sugarcane transcriptome generated by Iso-Seq[153], the mosaic monoploid genome of the sugarcane cultivar R570[154], the *S. spontaneum* genome[51] and a monoploid chromosomic set obtained from this same reference that included the "A" haplotype and unassembled scaffolds. To avoid sampling of duplicated regions, we did not include tags with multiple alignments in the ensuing analyses. After variant calling, VCFtools version 0.1.13[155] was used to retain biallelic markers with an MAF of 0.1, no missing data and a minimum sequencing depth of 50 reads. The most appropriate reference was chosen, and adopting the method proposed by Yang et al.[45], the ratio between alleles (allele proportions, APs) of each variant was transformed into genotypes with a fixed ploidy of 12 using the vcfR R package[156].

Linkage disequilibrium and population structure analyses. For SNPs and indels, we measured LD on the ldsep R package[157] by calculating the squared correlation coefficient (r^2) between pairs of markers on the same chromosomes. The decay of LD over physical distance was investigated by pooling all chromosomes, plotting pairwise r^2 values against the distance between markers and fitting a curve using the equation proposed by Hill and Weir[158]. The critical r^2 for LD decay was set to 0.1, the most commonly used threshold for determining the existence of LD[159]. Only comparisons with $p < 0.05$ were used in this analysis.

Three procedures were used to evaluate genetic structuring in the panel, employing dominant and codominant markers separately; for all analyses, the maximum number of clusters in the panel was set to 10. The first method was a DAPC, performed in the adegenet R package[160]. The second was PCA followed by K-means, for which missing data were imputed with the nonlinear estimation by iterative partial least squares method in the pcaMethods package[161] and for which the optimal number of clusters was evaluated using the elbow, silhouette and gap statistic methods in the factoextra package. The last was a Bayesian clustering of genotypes into predetermined numbers of clusters (K) performed on STRUCTURE software[162], assuming an admixture model with correlated allelic frequencies between populations. Ten independent runs were implemented for each K, and for dominant markers, estimates of probabilities of values of K in each run were taken following 100,000 generations as burn-in and 200,000 generations sampled in a Monte Carlo Markov Chain (MCMC). For Bayesian clustering using SNPs and indels, we used a subset of 7,000 markers randomly sampled from the total dataset, parallelized STRUCTURE with StrAuto software[163] and sampled 100,000 generations in the MCMC. In both cases, the most likely number of genetic clusters was determined by the ad hoc statistics ΔK[164] and the LnP(D) probability logarithm; the output was interpreted in STRUCTURE HARVESTER software version 0.6.94[165]. Clumpak software[166] was used to average the admixture proportions of runs and to estimate cluster membership coefficients for genotypes.

Association analyses. *FarmCPU.* Association analyses with dominant markers were performed with the FarmCPU[167] method in R. For these analyses, markers were recoded to indicate the presence (0) and absence (2) of bands. We tested FarmCPU using no covariates and including matrices obtained from the three genetic structure analyses described in the previous section as such. In each case, a Q–Q plot of the $-\log_{10}(p)$ values of markers was generated, and the genomic inflation factor λ[168] was calculated. The average λ from analyses employing each covariate matrix was calculated and used to select the model that best controlled inflation. The Bonferroni correction with $\alpha = 0.05$ was used to establish the significance threshold for associations, and the phenotypic variance explained by each marker was estimated for significant marker-trait associations using a linear model in R software.

Mixed modeling in GWASpoly. Association analyses using SNPs and indels were performed using mixed linear model approaches in the GWASpoly R package[65]. The output of the three genetic structure analyses previously described was used to build Q matrices, which were included in the models as fixed effects. Similarly, three different genetic kinship matrices (K) of the panel were computed and included as random effects: (I) a MM^T matrix[169], built on GWASpoly; (II) a complete autopolyploid matrix based on Slater et al.[170], built with the AGHmatrix R package[171]; and (III) a pseudodiploid matrix based on Slater et al.[170], also built with AGHmatrix. We tested twelve Q + K combinations, and for each of them, six marker-effect models were used: general, additive, simplex dominant reference, simplex dominant alternative, diploidized general and diploidized additive. For each model, a Q–Q plot of the $-\log_{10}(p)$ values of markers was generated, and λ was calculated. The average λ of all traits and models employing each Q + K combination was calculated and used to select the best set of matrices. Once this combination was chosen, Manhattan plots were generated for all models and traits. The Bonferroni and FDR correction methods with $\alpha = 0.05$ were assessed to establish the significance threshold for associations.

Machine learning coupled with feature selection. Finally, we assessed the capacity of ML strategies to predict the attribution of genotypes to the phenotypic groups identified in the HCPC analyses based on all markers, following the genomic prediction approach proposed by Aono et al.[64]. For this approach, we selected accessions successfully genotyped with both SNPs/indels and AFLPs/SSRs; missing data in dominant markers were imputed as the means. We evaluated the accuracy of eight ML algorithms: adaptive boosting (AB)[172], decision tree (DT)[173], Gaussian naive Bayes (GNB)[174], Gaussian process (GP)[175], K-nearest neighbor (KNN)[176], MLP[177], random forest (RF)[178] and support vector machine (SVM)[179], all implemented in the scikit-learn Python 3 module[180]. As a cross-validation strategy, we used a stratified K-fold (k = 5) repeated 100 times for different data configurations.

We then tested five FS techniques to obtain feature importance and create subsets of marker data: gradient tree boosting (FS1)[181], L1-based FS through a linear support vector classification system (FS2)[179], extremely randomized trees (FS3)[182], univariate FS using ANOVA (FS4) and RF (FS5)[178]. All FS approaches were implemented in the scikit-learn Python 3 module. We tested the differences in the accuracy between the selected FS

methods using ANOVAs and multiple comparisons by Tukey's tests implemented in the agricolae R package[183]. We also evaluated intersections between these datasets: markers selected by at least two of the five methods (Inter1); markers selected by at least two of the three best methods (Inter2); and markers selected by all three best methods (Inter3). Finally, the area under ROC curves was calculated for the best ML-FS combination and plotted using the Matplotlib library89 with Python 3.

Marker mapping and annotation. The distribution of markers identified by all analyses along *S. spontaneum* "A" chromosomes was visualized using MapChart[184]. Markers previously associated with SCYLV resistance by QTL mapping[27,30] and GWAS[28,31] were also retrieved and included in the map. Finally, the sequences of associated markers were annotated by aligning SSR flanking sequences or the 2000-bp window adjacent to SNPs and indels against a database comprising CDSs of the genomes of 14 Poaceae species and *A. thaliana*[64]. For this, BLASTn[185] was used with an E-value of 1e−30, and the best alignment of each sequence was kept for analysis.

Data availability

The raw sequencing data used in this article have been submitted to the SRA/NCBI under BioProject PRJNA702641.

Received: 12 April 2021; Accepted: 19 July 2021
Published online: 03 August 2021

References

1. FAO. *FAOSTAT: Production Sheet* (FAO, 2020).
2. ISO. *International Sugar Organization* (ISO, 2020).
3. Panje, R. R. & Babu, C. N. Studies in *Saccharum spontaneum* distribution and geographical association of chromosome numbers. *Cytologia* **25**, 152–172. https://doi.org/10.1508/cytologia.25.152 (1960).
4. D'Hont, A., Ison, D., Alix, K., Roux, C. & Glaszmann, J. C. 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 (1998).
5. Dhont, A. & Glaszmann, J. C. Sugarcane genome analysis with molecular markers: A first decade of research. In *Proceedings of the International Society of Sugarcane Technologists XXVI Congress* (ed. Hogarth, D. M.) 556–559 (The Australian Society of Sugar Cane Technologists, Brisbane, 2001).
6. Sforça, D. A. *et al.* 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 (2019).
7. Premachandran, M. N., Prathima, P. T. & Maya, L. Sugarcane and polyploidy—A review. *J. Sugarcane Res.* **1**, 1–15 (2011).
8. Bourke, P. M., Voorrips, R. E., Visser, R. G. F. & Maliepaard, C. Tools for genetic studies in experimental populations of polyploids. *Front. Plant Sci.* **9**, 513. https://doi.org/10.3389/fpls.2018.00513 (2018).
9. Moonan, F., Molina, J. & Mirkov, T. E. Sugarcane yellow leaf virus: An emerging virus that has evolved by recombination between luteoviral and poleroviral ancestors. *Virology* **269**, 156–171. https://doi.org/10.1006/viro.1999.0162 (2000).
10. Smith, G. R., Borg, Z., Braithwaite, K. S., Lockhart, B. E. L. & Gibbs, M. J. Sugarcane yellow leaf virus: A novel member of the Luteoviridae that probably arose by inter-species recombination. *J. Gen. Virol.* **81**, 1865–1869. https://doi.org/10.1099/0022-1317-81-7-1865 (2000).
11. Scagliusi, S. M. & Lockhart, B. E. L. Transmission, characterization, and serology of a luteovirus associated with yellow leaf syndrome of sugarcane. *Phytopathology* **90**, 120–124. https://doi.org/10.1094/phyto.2000.90.2.120 (2000).
12. ElSayed, A. I., Komor, E., Boulila, M., Viswanathan, R. & Odero, D. C. Biology and management of sugarcane yellow leaf virus: An historical overview. *Arch. Virol.* **160**, 2921–2934. https://doi.org/10.1007/s00705-015-2618-5 (2015).
13. Gonçalves, M. C., Vega, J., Oliveira, J. G. & Gomes, M. M. A. Sugarcane yellow leaf virus infection leads to alterations in photosynthetic efficiency and carbohydrate accumulation in sugarcane leaves. *Fitopatol. Bras.* **30**, 10–16. https://doi.org/10.1590/s0100-41582005000100002 (2005).
14. Lehrer, A., Yan, S.-L., Fontaniella, B., ElSayed, A. & Komor, E. Carbohydrate composition of sugarcane cultivars that are resistant or susceptible to sugarcane yellow leaf virus. *J. Gen. Plant Pathol.* **76**, 62–68. https://doi.org/10.1007/s10327-009-0210-0 (2010).
15. Vega, J., Scagliusi, S. M. M. & Ulian, E. C. Sugarcane yellow leaf disease in Brazil: Evidence of association with a luteovirus. *Plant Dis.* **81**, 21–26. https://doi.org/10.1094/pdis.1997.81.1.21 (1997).
16. Grisham, M., Pan, Y., Legendre, B., Godshall, M. & Eggleston, G. Effect of sugarcane yellow leaf virus on sugarcane yield and juice quality. *Proc. Int. Soc. Sugar Cane Technol.* **24**, 434–438 (2001).
17. Vasconcelos, A., Gonçalves, M. C., Pinto, L. R., Landell, M. G. & Perecin, D. Effects of sugarcane yellow leaf virus on sugarcane yield and root system development. *Funct. Plant Sci. Biotechnol.* **3**, 31–35 (2009).
18. Zhu, Y. J., Lim, S. T. S., Schenck, S., Arcinas, A. & Komor, E. RT-PCR and quantitative real-time RT-PCR detection of Sugarcane Yellow Leaf Virus (SCYLV) in symptomatic and asymptomatic plants of Hawaiian sugarcane cultivars and the correlation of SCYLV titre to yield. *Eur. J. Plant Pathol.* **127**, 263–273. https://doi.org/10.1007/s10658-010-9591-3 (2010).
19. Viswanathan, R. *et al.* Impact of Sugarcane yellow leaf virus (ScYLV) infection on physiological efficiency and growth parameters of sugarcane under tropical climatic conditions in India. *Acta Physiol. Plant* **36**, 1805–1822. https://doi.org/10.1007/s11738-014-1554-4 (2014).
20. Boukari, W. *et al.* Field infection of virus-free sugarcane by Sugarcane yellow leaf virus and effect of yellow leaf on sugarcane grown on organic and on mineral soils in Florida. *Plant Dis.* **103**(9), 2367–2373. https://doi.org/10.1094/PDIS-01-19-0199-RE (2019).
21. Aljanabi, S. M., Parmessur, Y., Moutia, Y., Saumtally, S. & Dookun, A. Further evidence of the association of a phytoplasma and a virus with yellow leaf syndrome in sugarcane. *Plant Pathol.* **50**, 628–636. https://doi.org/10.1046/j.1365-3059.2001.00604.x (2001).
22. Gonçalves, M. C., Klerks, M. M., Verbeek, M., Vega, J. & van den Heuvel, J. F. J. M. The use of molecular beacons combined with NASBA for the sensitive detection of sugarcane yellow leaf virus. *Eur. J. Plant Pathol.* **108**, 401–407. https://doi.org/10.1023/A:1016040314260 (2002).
23. Korimbocus, J., Coates, D., Barker, I. & Boonham, N. Improved detection of Sugarcane yellow leaf virus using a real-time fluorescent (TaqMan) RT-PCR assay. *J. Virol. Methods* **103**, 109–120. https://doi.org/10.1016/s0166-0934(01)00406-2 (2002).
24. Delage, C., Rippolesn, M., Chatenet, M., Irey, M. & Rott, P. Elimination of sugarcane yellow leaf virus from sugarcane by meristem tip culture. In *Proceedings of the XXIII Congress of the International Society Of Sugar Cane Technologists* (eds Singh, V. & Kumar, V.) (ISSCT Congress, New Delhi, 1999).

25. Chatenet, M. *et al.* 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 (2001).
26. Fitch, M. M. M., Lehrer, A. T., Komor, E. & Moore, P. H. Elimination of sugarcane yellow leaf virus from infected sugarcane plants by meristem tip culture visualized by tissue blot immunoassay. *Plant Pathol.* **50**, 676–680. https://doi.org/10.1046/j.1365-3059.2001.00639.x (2001).
27. Costet, L., Raboin, L.-M., Payet, M., D'Hont, A. & Nibouche, S. A major quantitative trait allele for resistance to the Sugarcane yellow leaf virus (Luteoviridae). *Plant Breed.* **131**, 637–640. https://doi.org/10.1111/j.1439-0523.2012.02003.x (2012).
28. Debibakas, S. *et al.* Prospecting sugarcane resistance to Sugarcane yellow leaf virus by genome-wide association. *Theor. Appl. Genet.* **127**, 1719–1732. https://doi.org/10.1007/s00122-014-2334-7 (2014).
29. Gouy, M. *et al.* 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 (2015).
30. Islam, M. S., Yang, X., Sood, S., Comstock, J. C. & Wang, J. Molecular characterization of genetic basis of sugarcane yellow leaf virus (SCYLV) resistance in *Saccharum* spp. hybrid. *Plant Breed.* **137**, 598–604. https://doi.org/10.1111/pbr.12614 (2018).
31. Yang, X., Sood, S., Luo, Z., Todd, J. & Wang, J. Genome-wide association studies identified resistance loci to orange rust and yellow leaf virus diseases in Sugarcane (*Saccharum* spp.). *Phytopathology* **109**, 623–631. https://doi.org/10.1094/phyto-08-18-0282-r (2019).
32. You, Q., Yang, X., Peng, Z., Islam, M. S., Sood, S., Luo, Z., et al. Development of an Axiom Sugarcane100K SNP array for genetic map construction and QTL identification. *Theor. Appl. Genet.* **132**, 2829–2845.https://doi.org/10.1007/s00122-019-03391-4 (2019).
33. Gonçalves, M. C., Pinto, L. R., Souza, S. C. & Landell, M. G. A. Virus diseases of Sugarcane: A constant challenge to sugarcane breeding in Brazil. *Funct. Plant Sci. Biotechnol.* **6**, 108–116 (2012).
34. Garcia, A. A. F. *et al.* SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. *Sci .Rep.* **3**, 3399. https://doi.org/10.1038/srep03399 (2013).
35. Lehrer, A. T., Moore, P. H. & Komor, E. Impact of sugarcane yellow leaf virus (ScYLV) on the carbohydrate status of sugarcane: Comparison of virus-free plants with symptomatic and asymptomatic virus-infected plants. *Physiol. Mol. Plant Pathol.* **70**, 180–188. https://doi.org/10.1016/j.pmpp.2007.09.005 (2007).
36. Jarošová, J., Gadiou, S. & Kumar, J. K. Real-time RT-PCR quantitative analysis of plant viruses in stone fruit tissues. *Julius-Kühn-Archiv* **61**, 1–437 (2009).
37. Comstock, J. C., Irey, M. S., Lockhart, B. E. L. & Wang, Z. K. Incidence of yellow leaf syndrome in CP cultivars based on polymerase chain reaction and serological techniques. *Sugar Cane* **4**, 21–24 (1998).
38. Lehrer, A. T. & Komor, E. Symptom expression of yellow leaf disease in sugarcane cultivars with different degrees of infection by Sugarcane yellow leaf virus. *Plant Pathol.* **57**, 178–189. https://doi.org/10.1111/j.1365-3059.2007.01696.x (2008).
39. Chinnaraja, C. *et al.* Quantification of sugarcane yellow leaf virus in *in vitro* plantlets and asymptomatic plants of sugarcane by RT-qPCR. *Curr. Sci.* **106**, 729–734 (2014).
40. Cooper, J. I. & Jones, A. T. Responses of plants to viruses: Proposals for the use of terms. *Phytopathology* **73**, 127–128. https://doi.org/10.1094/phyto-73-127 (1983).
41. Beoni, E., Chrpová, J., Jarošová, J. & Kundu, J. K. Survey of Barley yellow dwarf virus incidence in winter cereal crops, and assessment of wheat and barley resistance to the virus. *Crop Pasture Sci.* **67**, 1054–1063. https://doi.org/10.1071/cp16167 (2016).
42. Foresman, B. J. *et al.* Genome-wide association mapping of barley yellow dwarf virus tolerance in spring oat (*Avena sativa* L.). *PLoS ONE* **11**, e0155376. https://doi.org/10.1371/journal.pone.0155376 (2016).
43. Mansilla-Córdova, P. J. *et al.* Screening tomato genotypes for resistance and tolerance to *Tomato chlorosis virus*. *Plant Pathol.* **67**, 1231–1237. https://doi.org/10.1111/ppa.12826 (2018).
44. Fickett, N. *et al.* 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 (2019).
45. Yang, X. *et al.* 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 (2019).
46. Yang, X., Luo, Z., Todd, J., Sood, S. & Wang, J. Genome-wide association study of multiple yield traits in a diversity panel of polyploid sugarcane (*Saccharum* spp.). *Plant Genome* **13**, e20006. https://doi.org/10.1002/tpg2.20006 (2020).
47. Balsalobre, T. W. A. *et al.* GBS-based single dosage markers for linkage and QTL mapping allow gene mining for yield-related traits in sugarcane. *BMC Genom.* **18**, 72. https://doi.org/10.1186/s12864-016-3383-x (2017).
48. Elshire, R. J. *et al.* A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* **6**, e19379. https://doi.org/10.1371/journal.pone.0019379 (2011).
49. Grativol, C. *et al.* Sugarcane genome sequencing by methylation filtration provides tools for genomic research in the genus *Saccharum*. *Plant J.* **79**, 162–172 (2014).
50. Benevenuto, J., Ferrão, L. F. V., Amadeu, R. R. & Munoz, P. How can a high-quality genome assembly help plant breeders?. *GigaScience* **8**, giz068. https://doi.org/10.1093/gigascience/giz068 (2019).
51. Zhang, J. *et al.* Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. *Nat. Genet.* **50**, 1565–1573. https://doi.org/10.1038/s41588-018-0237-2 (2018).
52. Jannoo, N., Grivet, L., Dookun, A., D'Hont, A. & Glaszmann, J. C. Linkage disequilibrium among modern sugarcane cultivars. *Theor. Appl. Genet.* **99**, 1053–1060. https://doi.org/10.1007/s001220051414 (1999).
53. Raboin, L. M., Pauquet, J., Butterfield, M., D'Hont, A. & Glaszmann, J. C. Analysis of genome-wide linkage disequilibrium in the highly polyploid sugarcane. *Theor. Appl. Genet.* **116**, 701–714. https://doi.org/10.1007/s00122-007-0703-1 (2008).
54. Wei, X. *et al.* Simultaneously accounting for population structure, genotype by environment interaction, and spatial variation in marker-trait associations in sugarcane. *Genome* **53**, 973–981. https://doi.org/10.1139/g10-050 (2010).
55. Singh, R. K. *et al.* Identification of putative candidate genes for red rot resistance in sugarcane (*Saccharum* species hybrid) using LD-based association mapping. *Mol. Genet. Genom.* **291**, 1363–1377. https://doi.org/10.1007/s00438-016-1190-3 (2016).
56. Barreto, F. Z. *et al.* A genome-wide association study identified loci for yield component traits in sugarcane (*Saccharum* spp.). *PLoS ONE* **14**, e0219843. https://doi.org/10.1371/journal.pone.0219843 (2019).
57. Yang, X. *et al.* 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 (2019).
58. Gaudeul, M., Till-Bottraud, I., Barjon, F. & Manel, S. Genetic diversity and differentiation in *Eryngium alpinum* L. (Apiaceae): Comparison of AFLP and microsatellite markers. *Heredity (Edinb)* **92**, 508–518. https://doi.org/10.1038/sj.hdy.6800443 (2004).
59. Fang, J., Twito, T., Zhang, Z. & Chao, C. T. Genetic relationships among fruiting-mei (*Prunus mume* Sieb. et Zucc.) cultivars evaluated with AFLP and SNP markers. *Genome* **49**, 1256–1264. https://doi.org/10.1139/g06-097 (2006).
60. Roncallo, P. F., Beaufort, V., Larsen, A. O., Dreisigacker, S. & Echenique, V. Genetic diversity and linkage disequilibrium using SNP (KASP) and AFLP markers in a worldwide durum wheat (*Triticum turgidum* L. var durum) collection. *PLoS ONE* **14**, e0218562. https://doi.org/10.1371/journal.pone.0218562 (2019).
61. Creste, S. *et al.* Comparison of AFLP, TRAP and SSRs in the estimation of genetic relationships in sugarcane. *Sugar Tech* **12**, 150–154. https://doi.org/10.1007/s12355-010-0029-1 (2010).

62. de Bem Oliveira, I. *et al.* Genomic prediction of autotetraploids; influence of relationship matrices, allele dosage, and continuous genotyping calls in phenotype prediction. *G3 (Bethesda)* **9**, 1189–1198. https://doi.org/10.1534/g3.119.400059 (2019).
63. Matias, F. I. *et al.* On the accuracy of genomic prediction models considering multi-trait and allele dosage in *Urochloa* spp. interspecific tetraploid hybrids. *Mol Breed.* **39**, 100. https://doi.org/10.1007/s11032-019-1002-7 (2019).
64. Aono, A. H. *et al.* 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 (2020).
65. Rosyara, U. R., De Jong, W. S., Douches, D. S. & Endelman, J. B. Software for genome-wide association studies in autopolyploids and its application to potato. *Plant Genome* **9**, 1–10. https://doi.org/10.3835/plantgenome2015.08.0073 (2016).
66. Berdugo-Cely, J., Valbuena, R. I., Sánchez-Betancourt, E., Barrero, L. S. & Yockteng, R. Genetic diversity and association mapping in the colombian central collection of *Solanum tuberosum* L. Andigenum group using SNPs markers. *PLoS ONE* **12**, e0173039. https://doi.org/10.1371/journal.pone.0173039 (2017).
67. Byrne, S. *et al.* Genome-wide association and genomic prediction for fry color in potato. *Agronomy* **10**, 90. https://doi.org/10.3390/agronomy10010090 (2020).
68. Nimmakayala, P. *et al.* Genome-wide differentiation of various melon horticultural groups for use in GWAS for fruit firmness and construction of a high resolution genetic map. *Front. Plant Sci.* **7**, 1437. https://doi.org/10.3389/fpls.2016.01437 (2016).
69. Su, J. *et al.* Detection of favorable QTL alleles and candidate genes for lint percentage by GWAS in Chinese Upland cotton. *Front. Plant Sci.* **7**, 1576. https://doi.org/10.3389/fpls.2016.01576 (2016).
70. Ferrão, L. F. V. *et al.* Insights into the genetic basis of blueberry fruit-related traits using diploid and polyploid models in a GWAS context. *Front. Ecol. Evol.* **6**, 107. https://doi.org/10.3389/fevo.2018.00107 (2018).
71. Daugrois, J. H. *et al.* 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 (1996).
72. Raboin, L. M. *et al.* Genetic mapping in sugarcane, a high polyploid, using bi-parental progeny: Identification of a gene controlling stalk colour and a new rust resistance gene. *Theor. Appl. Genet.* **112**, 1382–1391. https://doi.org/10.1007/s00122-006-0240-3 (2006).
73. Wang, H. *et al.* Effect of population size and unbalanced data sets on QTL detection using genome-wide association mapping in barley breeding germplasm. *Theor. Appl. Genet.* **124**, 111–124. https://doi.org/10.1007/s00122-011-1691-8 (2012).
74. Yang, N. *et al.* Genome wide association studies using a new nonparametric model reveal the genetic architecture of 17 agronomic traits in an enlarged maize association panel. *PLoS Genet.* **10**, e1004573–e1004573. https://doi.org/10.1371/journal.pgen.1004573 (2014).
75. Racedo, J. *et al.* Genome-wide association mapping of quantitative traits in a breeding population of sugarcane. *BMC Plant Biol.* **16**, 142. https://doi.org/10.1186/s12870-016-0829-x (2016).
76. Barnes, J., Rutherford, R. & Botha, F. 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 (1997).
77. Diola, V., Barbosa, M. H. P., Veiga, C. F. M. & Fernandes, E. C. Molecular markers EST-SSRs for genotype-phenotype association in sugarcane. *Sugar Tech* **16**, 241–249. https://doi.org/10.1007/s12355-013-0268-z (2014).
78. Siraree, A. *et al.* Identification of marker-trait associations for morphological descriptors and yield component traits in sugarcane. *Physiol. Mol. Biol. Plants* **23**, 185–196. https://doi.org/10.1007/s12298-016-0403-x (2017).
79. Gao, X., Becker, L. C., Becker, D. M., Starmer, J. D. & Province, M. A. Avoiding the high Bonferroni penalty in genome-wide association studies. *Genet. Epidemiol.* **34**, 100–105. https://doi.org/10.1002/gepi.20430 (2010).
80. Johnson, R. C. *et al.* Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genom.* **11**, 724. https://doi.org/10.1186/1471-2164-11-724 (2010).
81. Fadista, J., Manning, A. K., Florez, J. C. & Groop, L. The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *Eur. J. Hum. Genet.* **24**, 1202–1205. https://doi.org/10.1038/ejhg.2015.269 (2016).
82. Steinfath, M. *et al.* Prediction of hybrid biomass in *Arabidopsis thaliana* by selected parental SNP and metabolic markers. *Theor. Appl. Genet.* **120**, 239–247. https://doi.org/10.1007/s00122-009-1191-2 (2010).
83. Heer, K. *et al.* Linking dendroecology and association genetics in natural populations: Stress responses archived in tree rings associate with SNP genotypes in silver fir (*Abies alba* Mill.). *Mol. Ecol.* **27**, 1428–1438. https://doi.org/10.1111/mec.14538 (2018).
84. Zhou, W. *et al.* Minor QTLs mining through the combination of GWAS and machine learning feature selection. *bioRxiv* https://doi.org/10.1101/702761 (2019).
85. Scagliusi, S. M., Basu, S. K., de Gouvea, J. A. & Vega, J. physiological alterations in Brazilian sugarcane varieties infected by Sugarcane yellow leaf virus (ScYLV). *Funct. Plant Sci. Biotechnol.* **3**, 19–25 (2009).
86. DeYoung, B. J. & Innes, R. W. Plant NBS-LRR proteins in pathogen sensing and host defense. *Nat. Immunol.* **7**, 1243–1249. https://doi.org/10.1038/ni1410 (2006).
87. Dinesh-Kumar, S. P., Tham, W. H. & Baker, B. J. Structure-function analysis of the tobacco mosaic virus resistance gene N. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 14789–14794. https://doi.org/10.1073/pnas.97.26.14789 (2000).
88. Seo, Y. S. *et al.* A viral resistance gene from common bean functions across plant families and is up-regulated in a non-virus-specific manner. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 11856–11861. https://doi.org/10.1073/pnas.0604815103 (2006).
89. Xun, H. *et al.* 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 (2019).
90. Cao, H., Glazebrook, J., Clarke, J. D., Volko, S. & Dong, X. The Arabidopsis NPR1 gene that controls systemic acquired resistance encodes a novel protein containing ankyrin repeats. *Cell* **88**, 57–63. https://doi.org/10.1016/s0092-8674(00)81858-9 (1997).
91. Rochon, A., Boyle, P., Wignes, T., Fobert, P. R. & Després, C. The coactivator function of Arabidopsis NPR1 requires the core of its BTB/POZ domain and the oxidation of C-terminal cysteines. *Plant Cell* **18**, 3670–3685. https://doi.org/10.1105/tpc.106.046953 (2006).
92. Pieterse, C. M. *et al.* A novel signaling pathway controlling induced systemic resistance in Arabidopsis. *Plant Cell* **10**, 1571–1580. https://doi.org/10.1105/tpc.10.9.1571 (1998).
93. Spoel, S. H. *et al.* NPR1 modulates cross-talk between salicylate- and jasmonate-dependent defense pathways through a novel function in the cytosol. *Plant Cell* **15**, 760–770. https://doi.org/10.1105/tpc.009159 (2003).
94. Liu, Y., Schiff, M., Marathe, R. & Dinesh-Kumar, S. P. Tobacco Rar1, EDS1 and NPR1/NIM1 like genes are required for N-mediated resistance to tobacco mosaic virus. *Plant J.* **30**, 415–429. https://doi.org/10.1046/j.1365-313x.2002.01297.x (2002).
95. Lin, W. C. *et al.* Transgenic tomato plants expressing the Arabidopsis NPR1 gene display enhanced resistance to a spectrum of fungal and bacterial diseases. *Transgenic Res.* **13**, 567–581. https://doi.org/10.1007/s11248-004-2375-9 (2004).
96. Llave, C. Virus-derived small interfering RNAs at the core of plant-virus interactions. *Trends Plant Sci.* **15**, 701–707. https://doi.org/10.1016/j.tplants.2010.09.001 (2010).
97. Deleris, A. *et al.* Hierarchical action and inhibition of plant Dicer-like proteins in antiviral defense. *Science* **313**, 68–71. https://doi.org/10.1126/science.1128214 (2006).
98. Alam, C. M. *et al.* Dicer 1 of *Candida albicans* cleaves plant viral dsRNA in vitro and provides tolerance in plants against virus infection. *Virusdisease* **30**, 237–244. https://doi.org/10.1007/s13337-019-00520-x (2019).
99. Bouché, N., Yellin, A., Snedden, W. A. & Fromm, H. Plant-specific calmodulin-binding proteins. *Annu. Rev. Plant Biol.* **56**, 435–466. https://doi.org/10.1146/annurev.arplant.56.032604.144224 (2005).

100. Heo, W. D. *et al.* Involvement of specific calmodulin isoforms in salicylic acid-independent activation of plant disease resistance responses. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 766–771. https://doi.org/10.1073/pnas.96.2.766 (1999).
101. Nakahara, K. S. *et al.* Tobacco calmodulin-like protein provides secondary defense by binding to and directing degradation of virus RNA silencing suppressors. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 10113–10118. https://doi.org/10.1073/pnas.1201628109 (2012).
102. Li, F., Huang, C., Li, Z. & Zhou, X. Suppression of RNA silencing by a plant DNA virus satellite requires a host calmodulin-like protein to repress RDR6 expression. *PLoS Pathog.* **10**, e1003921. https://doi.org/10.1371/journal.ppat.1003921 (2014).
103. Cao, Y. *et al.* Overexpression of a rice defense-related F-box protein geneOsDRF1in tobacco improves disease resistance through potentiation of defense gene expression. *Physiol. Plant* **134**, 440–452. https://doi.org/10.1111/j.1399-3054.2008.01149.x (2008).
104. Thiel, H., Hleibieh, K., Gilmer, D. & Varrelmann, M. The P25 pathogenicity factor of beet necrotic yellow vein virus targets the sugar beet 26S proteasome involved in the induction of a hypersensitive resistance response via interaction with an F-box protein. *Mol. Plant-Microbe Interact.* **25**, 1058–1072. https://doi.org/10.1094/mpmi-03-12-0057-r (2012).
105. Earley, K., Smith, M., Weber, R., Gregory, B. & Poethig, R. An endogenous F-box protein regulates ARGONAUTE1 in *Arabidopsis thaliana*. *Silence* **1**, 15. https://doi.org/10.1186/1758-907X-1-15 (2010).
106. Morel, J.-B. *et al.* Fertile hypomorphic ARGONAUTE (ago1) mutants impaired in post-transcriptional gene silencing and virus resistance. *Plant Cell* **14**, 629–639. https://doi.org/10.1105/tpc.010358 (2002).
107. Chen, H., Zhang, L., Yu, K. & Wang, A. Pathogenesis of Soybean mosaic virus in soybean carrying Rsv1 gene is associated with miRNA and siRNA pathways, and breakdown of AGO1 homeostasis. *Virology* **476**, 395–404. https://doi.org/10.1016/j.virol.2014.12.034 (2015).
108. Yang, Z. & Li, Y. Dissection of RNAi-based antiviral immunity in plants. *Curr. Opin. Virol.* **32**, 88–99. https://doi.org/10.1016/j.coviro.2018.08.003 (2018).
109. Mangwende, T. *et al.* The P0 gene of Sugarcane yellow leaf virus encodes an RNA silencing suppressor with unique activities. *Virology* **384**, 38–50. https://doi.org/10.1016/j.virol.2008.10.034 (2009).
110. Huang, T. S., Wei, T., Laliberté, J. F. & Wang, A. A host RNA helicase-like protein, AtRH8, interacts with the potyviral genome-linked protein, VPg, associates with the virus accumulation complex, and is essential for infection. *Plant Physiol.* **152**, 255–266. https://doi.org/10.1104/pp.109.147983 (2010).
111. Kovalev, N., Pogany, J. & Nagy, P. D. A Co-Opted DEAD-Box RNA helicase enhances tombusvirus plus-strand synthesis. *PLoS Pathog.* **8**, e1002537. https://doi.org/10.1371/journal.ppat.1002537 (2012).
112. Li, Y., Xiong, R., Bernards, M. & Wang, A. Recruitment of *Arabidopsis* RNA helicase AtRH9 to the viral replication complex by viral replicase to promote turnip mosaic virus replication. *Sci. Rep.* **6**, 30297–30297. https://doi.org/10.1038/srep30297 (2016).
113. Wei, T., Zhang, C., Hou, X., Sanfaçon, H. & Wang, A. The SNARE protein Syp71 is essential for turnip mosaic virus infection by mediating fusion of virus-induced vesicles with chloroplasts. *PLoS Pathog.* **9**, e1003378. https://doi.org/10.1371/journal.ppat.1003378 (2013).
114. Cabanillas, D. G. *et al.* Turnip mosaic virus uses the SNARE protein VTI11 in an unconventional route for replication vesicle trafficking. *Plant Cell* **30**, 2594–2615. https://doi.org/10.1105/tpc.18.00281 (2018).
115. Sasvari, Z., Kovalev, N., Gonzalez, P. A., Xu, K. & Nagy, P. D. Assembly-hub function of ER-localized SNARE proteins in biogenesis of tombusvirus replication compartment. *PLoS Pathog.* **14**, e1007028. https://doi.org/10.1371/journal.ppat.1007028 (2018).
116. Ibrahim, A. *et al.* Plant SNAREs SYP22 and SYP23 interact with Tobacco mosaic virus 126 kDa protein and SYP2s are required for normal local virus accumulation and spread. *Virology* **547**, 57–71. https://doi.org/10.1016/j.virol.2020.04.002 (2020).
117. Heinlein, M. Plant virus replication and movement. *Virology* **479–480**, 657–671. https://doi.org/10.1016/j.virol.2015.01.025 (2015).
118. Harries, P. A. *et al.* Differing requirements for actin and myosin by plant viruses for sustained intercellular movement. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 17594–17599. https://doi.org/10.1073/pnas.0909239106 (2009).
119. Amari, K., Di Donato, M., Dolja, V. V. & Heinlein, M. Myosins VIII and XI play distinct roles in reproduction and transport of tobacco mosaic virus. *PLoS Pathog.* **10**, e1004448. https://doi.org/10.1371/journal.ppat.1004448 (2014).
120. Amari, K., Lerich, A., Schmitt-Keichinger, C., Dolja, V. V. & Ritzenthaler, C. Tubule-guided cell-to-cell movement of a plant virus requires class XI myosin motors. *PLoS Pathog.* **7**, e1002327. https://doi.org/10.1371/journal.ppat.1002327 (2011).
121. Abdelkhalek, A., Ismail, I. A., Dessoky, E. S., El-Hallous, E. I. & Hafez, E. A tomato kinesin-like protein is associated with *Tobacco mosaic* virus infection. *Biotechnol. Biotechnol. Equip.* **33**, 1424–1433. https://doi.org/10.1080/13102818.2019.1673207 (2019).
122. Hofius, D. *et al.* 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 (2007).
123. Lu, L. *et al.* Pc4, a putative movement protein of Rice stripe virus, interacts with a type I DnaJ protein and a small Hsp of rice. *Virus Genes* **38**, 320–327. https://doi.org/10.1007/s11262-008-0324-z (2009).
124. Shimizu, T. *et al.* Identification of a novel tobacco DnaJ-like protein that interacts with the movement protein of tobacco mosaic virus. *Arch. Virol.* **154**, 959–967. https://doi.org/10.1007/s00705-009-0397-6 (2009).
125. Liu, J. Z. & Whitham, S. A. Overexpression of a soybean nuclear localized type-III DnaJ domain-containing HSP40 reveals its roles in cell death and disease resistance. *Plant J.* **74**, 110–121. https://doi.org/10.1111/tpj.12108 (2013).
126. Burbano, R. C. V. *et al.* Screening of *Saccharum* spp. genotypes for sugarcane yellow leaf virus resistance by combining symptom phenotyping and highly precise virus titration. *Crop Prot.* **144**, 105577. https://doi.org/10.1016/j.cropro.2021.105577 (2021).
127. Chinnaraja, C. & Viswanathan, R. Quantification of sugarcane yellow leaf virus in sugarcane following transmission through aphid vector, *Melanaphis sacchari*. *Virusdisease* **26**, 237–242. https://doi.org/10.1007/s13337-015-0267-7 (2015).
128. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2−ΔΔCT method. *Methods* **25**, 402–408. https://doi.org/10.1006/meth.2001.1262 (2001).
129. Peterson, R. A. Finding Optimal Normalizing Transformations via bestNormalize. *The R Journal*. https://doi.org/10.32614/RJ-2021-041 (2021).
130. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2011).
131. Muñoz, F., & Rodriguez, L. S. breedR: Statistical methods for forest genetic resources analysis. In *Trees for the Future: Plant Material in a Changing Climate*. 13. Tulln, Austria (2015).
132. Schloerke, B. *et al.* Extension to ggplot2. *R Package Version 1* (2011).
133. Lê, S., Josse, J. & Husson, F. FactoMineR: An R package for multivariate analysis. *J. Stat. Softw.* **25**, 1–18. https://doi.org/10.18637/jss.v025.i01 (2008).
134. Kassambara, A. & Mundt, F. Factoextra: Extract and visualize the results of multivariate data analyses. *R Package Version* **1**, 2017 (2017).
135. Dinno, A. dunn.test: Dunn's test of multiple comparisons using rank sums. *R package version* **1**(4), 1 (2017).
136. Aljanabi, S. M., Forget, L. & Dookun, A. An improved and rapid protocol for the isolation of polysaccharide- and polyphenol-free sugarcane DNA. *Plant Mol. Biol. Rep.* **17**, 281. https://doi.org/10.1023/A:1007692929505 (1999).
137. Maccheroni, W., Jordão, H., Degaspari, R., & Matsuoka, S. Development of a dependable microsatellite-based fingerprinting system for sugarcane. In *Proceedings of the International Society of Sugar Cane Technologistics*, **27**, 47–52, Durban (2007).
138. Oliveira, K. M. *et al.* Functional integrated genetic linkage map based on EST-markers for a sugarcane (*Saccharum* spp.) commercial cross. *Mol Breed.* **20**, 189–208. https://doi.org/10.1007/s11032-007-9082-1 (2007).

139. Oliveira, K. M. *et al.* Characterization of new polymorphic functional markers for sugarcane. *Genome* **52**, 191–209. https://doi.org/10.1139/g08-105 (2009).
140. Marconi, T. G. *et al.* Functional markers for gene mapping and genetic diversity studies in sugarcane. *BMC Res. Notes* **4**, 264. https://doi.org/10.1186/1756-0500-4-264 (2011).
141. Poland, J. A., Brown, P. J., Sorrells, M. E. & Jannink, J.-L. 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 (2012).
142. Lundin, S., Stranneheim, H., Pettersson, E., Klevebring, D. & Lundeberg, J. Increased throughput by parallelization of library preparation for massive sequencing. *PLoS ONE* **5**, e10029. https://doi.org/10.1371/journal.pone.0010029 (2010).
143. Andrews, S. *FastQC: A Quality Control Tool for High Throughput Sequence Data* (Babraham Bioinformatics, 2010).
144. Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A. & Cresko, W. A. Stacks: An analysis tool set for population genomics. *Mol. Ecol.* **22**, 3124–3140. https://doi.org/10.1111/mec.12354 (2013).
145. Pereira, G. S., Garcia, A. A. F. & Margarido, G. R. A. A fully automated pipeline for quantitative genotype calling from next generation sequencing data in autopolyploids. *BMC Bioinform.* **19**, 398. https://doi.org/10.1186/s12859-018-2433-6 (2018).
146. Glaubitz, J. C. *et al.* TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. *PLoS ONE* **9**, e90346. https://doi.org/10.1371/journal.pone.0090346 (2014).
147. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics (Oxford, England)* **25**, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324 (2009).
148. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359. https://doi.org/10.1038/nmeth.1923 (2012).
149. Paterson, A. H. *et al.* The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**, 551–556. https://doi.org/10.1038/nature07723 (2009).
150. Cardoso-Silva, C. B. *et al.* De novo assembly and transcriptome analysis of contrasting sugarcane varieties. *PLoS ONE* **9**, e88462. https://doi.org/10.1371/journal.pone.0088462 (2014).
151. Melo, A. T. O., Bartaula, R. & Hale, I. GBS-SNP-CROP: A reference-optional pipeline for SNP discovery and plant germplasm characterization using variable length, paired-end genotyping-by-sequencing data. *BMC Bioinform.* **17**, 29. https://doi.org/10.1186/s12859-016-0879-y (2016).
152. Riaño-Pachón, D. M. & Mattiello, L. Draft genome sequencing of the sugarcane hybrid SP80-3280. *F1000Research* **6**, 861. https://doi.org/10.12688/f1000research.11859.2 (2017).
153. Hoang, N. V. *et al.* A survey of the complex transcriptome from the highly polyploid sugarcane genome using full-length isoform sequencing and *de novo* assembly from short read sequencing. *BMC Genom.* **18**, 395. https://doi.org/10.1186/s12864-017-3757-8 (2017).
154. Garsmeur, O. *et al.* 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 (2018).
155. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330 (2011).
156. Knaus, B. J. & Grünwald, N. J. 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 (2017).
157. Gerard, D. Pairwise linkage disequilibrium estimation for polyploids. *Molecular Ecology Resources* **21**, 1230–1242.https://doi.org/10.1111/1755-0998.13349 (2021).
158. Hill, W. G. & Weir, B. S. Variances and covariances of squared linkage disequilibria in finite populations. *Theor. Popul. Biol.* **33**, 54–78. https://doi.org/10.1016/0040-5809(88)90004-4 (1988).
159. Vos, P. G. *et al.* Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. *Theor. Appl. Genet.* **130**, 123–135. https://doi.org/10.1007/s00122-016-2798-8 (2017).
160. Jombart, T. adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403–1405. https://doi.org/10.1093/bioinformatics/btn129 (2008).
161. Stacklies, W., Redestig, H., Scholz, M., Walther, D. & Selbig, J. pcaMethods–a bioconductor package providing PCA methods for incomplete data. *Bioinformatics* **23**, 1164–1167. https://doi.org/10.1093/bioinformatics/btm069 (2007).
162. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
163. Chhatre, V. E. & Emerson, K. J. StrAuto: Automation and parallelization of STRUCTURE analysis. *BMC Bioinform.* **18**, 192. https://doi.org/10.1186/s12859-017-1593-0 (2017).
164. Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software structure: A simulation study. *Mol. Ecol.* **14**, 2611–2620. https://doi.org/10.1111/j.1365-294x.2005.02553.x (2005).
165. Earl, D. A. & vonHoldt, B. M. Structure harvester: A website and program for visualizing structure output and implementing the Evanno method. *Conserv. Genet. Resour.* **4**, 359–361. https://doi.org/10.1007/s12686-011-9548-7 (2012).
166. Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A. & Mayrose, I. Clumpak: A program for identifying clustering modes and packaging population structure inferences across K. *Mol. Ecol. Resour.* **15**, 1179–1191. https://doi.org/10.1111/1755-0998.12387 (2015).
167. Liu, X., Huang, M., Fan, B., Buckler, E. S. & Zhang, Z. Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genet.* **12**, e1005767. https://doi.org/10.1371/journal.pgen.1005767 (2016).
168. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004. https://doi.org/10.1111/j.0006-341x.1999.00997.x (1999).
169. VanRaden, P. M. Efficient methods to compute genomic predictions. *J. Dairy Sci.* **91**, 4414–4423. https://doi.org/10.3168/jds.2007-0980 (2008).
170. Slater, A. T., Cogan, N. O. I., Forster, J. W., Hayes, B. J. & Daetwyler, H. D. Improving genetic gain with genomic selection in autotetraploid potato. *Plant Genome* **9**, 1–15. https://doi.org/10.3835/plantgenome2016.02.0021 (2016).
171. Amadeu, R. R. *et al.* AGHmatrix: R package to construct relationship matrices for autotetraploid and diploid species: A blueberry example. *Plant Genome* **9**, 1–10. https://doi.org/10.3835/plantgenome2016.01.0009 (2016).
172. Freund, Y. & Schapire, R. E. 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 (1997).
173. Quinlan, J. R. Induction of decision trees. *Mach. Learn.* **1**, 81–106. https://doi.org/10.1007/BF00116251 (1986).
174. Friedman, N., Geiger, D. & Goldszmidt, M. Bayesian network classifiers. *Mach. Learn.* **29**, 131–163. https://doi.org/10.1023/A:1007465528199 (1997).
175. Rasmussen, C. E. Gaussian processes in machine learning. In *Summer School on Machine Learning* (eds Bousquet, O. *et al.*) 63–71 (Springer, 2003).
176. Cover, T. & Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **13**, 21–27. https://doi.org/10.1109/TIT.1967.1053964 (1967).
177. Popescu, M. C., Balas, V., Perescu-Popescu, L. & Mastorakis, N. Multilayer perceptron and neural networks. *WSEAS Trans. Circuits Syst.* **8**, 579–588 (2009).
178. Breiman, L. Bagging predictors. *Mach. Learn.* **24**, 123–140. https://doi.org/10.1007/BF00058655 (2001).

179. Cristianini, N. & Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods* (Cambridge University Press, 2000).
180. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
181. Chen, T., & Guestrin, C. Xgboost: A scalable tree boosting system. In *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 785–794 (ACM, New York, 2016).
182. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **63**, 3–42. https://doi.org/10.1007/s10994-006-6226-1 (2006).
183. De Mendiburu, F., & De Mendiburu, M. F. Package 'agricolae'. *R package version*, 1–2 (2020).
184. Voorrips, R. E. MapChart: Software for the graphical presentation of linkage maps and QTLs. *J. Hered* **93**, 77–78. https://doi.org/10.1093/jhered/93.1.77 (2002).
185. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410. https://doi.org/10.1016/s0022-2836(05)80360-2 (1990).

Acknowledgements

We thank Aline C. L. Moraes for assistance in constructing and sequencing the GBS library and Maicon Volpin for assistance in fieldwork. This work was supported by grants from the São Paulo Research Foundation (FAPESP), the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ, Grant 424050/2016-1), the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Computational Biology Program), the Littoral Polytechnic Superior School (ESPOL) and the Secretaría Nacional de Ciencia y Tecnología (SENESYT). RJGP received an MSc fellowship from CAPES (Grant 88887.177386/2018-00) and MSc and PhD fellowships from FAPESP (Grants 2018/18588-8 and 2019/21682-9). AHA received a PhD fellowship from FAPESP (Grant 2019/03232-6). RCVB received a PhD fellowship from PAEDEx-AUIP. CCS received a PD fellowship from FAPESP (Grant 2015/24346-9).

Author contributions

D.P., M.G.A.L., M.C.G., L.R.P. and A.P.S. conceived the project and designed the experiments. R.J.G.P., R.C.V.B., C.C.S., I.A.A. and L.R.P. performed phenotyping. R.J.G.P., R.C.V.B. and A.E.C. performed genotyping. R.J.G.P. and A.H.A. analyzed the data and interpreted the results. R.J.G.P. wrote the manuscript. All authors read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-95116-1.

Correspondence and requests for materials should be addressed to A.P.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.