Limited allele-specific gene expression in highly polyploid sugarcane

Gabriel Rodrigues Alves Margarido,[1,2] Fernando Henrique Correr,[1,2] Agnelo Furtado,[2] Frederik C. Botha,[2] and Robert James Henry[2]

[1] *Department of Genetics, University of São Paulo, "Luiz de Queiroz" College of Agriculture, Piracicaba 13418-900, Brazil;*
[2] *Queensland Alliance for Agriculture and Food Innovation, University of Queensland, Brisbane 4072, Australia*

Polyploidy is widespread in plants, allowing the different copies of genes to be expressed differently in a tissue-specific or developmentally specific way. This allele-specific expression (ASE) has been widely reported, but the proportion and nature of genes showing this characteristic have not been well defined. We now report an analysis of the frequency and patterns of ASE at the whole-genome level in the highly polyploid sugarcane genome. Very high depth whole-genome sequencing and RNA sequencing revealed strong correlations between allelic proportions in the genome and in expressed sequences. This level of sequencing allowed discrimination of each of the possible allele doses in this 12-ploid genome. Most genes were expressed in direct proportion to the frequency of the allele in the genome with examples of polymorphisms being found with every possible discrete level of dose from 1:11 for single-copy alleles to 12:0 for monomorphic sites. The rarer cases of ASE were more frequent in the expression of defense-response genes, as well as in some processes related to the biosynthesis of cell walls. ASE was more common in genes with variants that resulted in significant disruption of function. The low level of ASE may reflect the recent origin of polyploid hybrid sugarcane. Much of the ASE present can be attributed to strong selection for resistance to diseases in both nature and domestication.

[Supplemental material is available for this article.]

Sugarcane has remarkable potential as a food and bioenergy crop being a C4 grass with high photosynthetic efficiency and biomass yield (Henry 2010), grown mainly for the dual purpose of producing sugar and ethanol. Currently, sugarcane comprises 86% of the sugar crops cultivated worldwide (OECD/FAO 2019). Biofuel obtained from sugarcane may have a key role in addressing climate change concerns (Goldemberg 2007; Souza et al. 2017). Continuing sugarcane breeding efforts are necessary to meet the increasing demands for sugar and ethanol, and a better understanding of molecular processes relevant to carbon partitioning will in turn guide molecular breeding of this crop (Wang et al. 2013).

Genomic analyses in sugarcane (*Saccharum* spp.) are limited by the unique combination of several layers of genomic complexity. All known *Saccharum* accessions are highly polyploid (D'Hont et al. 1998), with evidence of both allo- and autopolyploidy in the evolutionary history of the genus (Kim et al. 2014). Cultivated sugarcane clones are interspecific hybrids between *Saccharum officinarum* and *Saccharum spontaneum*, with varying contributions from each genome, multiple aneuploidy events, and recombination between the parental genomes (D'Hont et al. 1996). Hybrids show a variable number of chromosome copies per homology group (Grivet and Arruda 2002), with different total numbers of chromosomes in different genotypes (Piperidis et al. 2010), and large genomes (D'Hont and Glaszmann 2001). However, there is still a prevalence of $2n = 12x$ among hybrids (Le Cunff et al. 2008; Piperidis and D'Hont 2020). Pompidor et al. (2021) recently proposed the existence of three founding genomes in the genus *Saccharum*, two of them unevenly found in the sugar-rich *S. officinarum*, and the third being observed in the wild *S. spontaneum*.

Genomic resources are being developed for sugarcane (Kandel et al. 2018; Diniz et al. 2019). However, because of the amalgam of obstacles described above, research lags behind that for other major crops (Thirugnanasambandam et al. 2018). One challenge is the assembly of a genomic sequence that fully represents a complete hybrid genome. Despite the recent publication of multiple (partial) sugarcane genome sequences (Garsmeur et al. 2018; Zhang et al. 2018; Souza et al. 2019), sorghum is still a valuable genomic resource and has been extensively used as a genome reference for sugarcane (Grivet and Arruda 2002; Dillon et al. 2007). These two genomes are highly conserved, particularly in the transcribed regions (Wang et al. 2010). Using diploid sorghum as a reference sidesteps issues owing to the multiple alignment of reads to several chromosome copies.

Polyploidy causes a multitude of changes in cell structure and function, both in allo- (Comai 2000; Renny-Byfield and Wendel 2014) and autopolyploids (Yant and Bomblies 2015). It is usually accompanied by a so-called genomic shock and an increase in allelic diversity, which can consequently cause changes to gene expression profiles (Chen and Ni 2006; Feldman and Levy 2009; Baduel et al. 2018). Such alterations include allele-specific expression (ASE), a phenomenon whereby the different alleles of a given gene are unevenly expressed (Gaur et al. 2013). This excessive abundance of one allele stems from differential expression and/or degradation of mRNA molecules originated from different chromosome copies. Allelic imbalance can occur owing to variation in *cis*-acting regulatory regions, nonsense-mediated decay, and

Corresponding authors: gramarga@usp.br, robert.henry@uq.edu.au
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.275904.121.

epigenetic imprinting, among other factors (Castel et al. 2015; Chen et al. 2018).

ASE has been evaluated in diploid and polyploid plants, such as *Arabidopsis* (Zhang and Borevitz 2009), maize (Springer and Stupar 2007), autotetraploid potato (Pham et al. 2017), and allohexaploid wheat (Powell et al. 2017). It has been associated with heterosis in hybrid rice, as a likely cause of dominance and overdominance effects (Shao et al. 2019) with evidence that genes with ASE were under selective pressure through breeding. Hybridization in maize is a cause of ASE, and although not frequent, it can arise from nonsyntenic genes and contribute to hybrid vigour (Baldauf et al. 2020). There is also evidence of conserved ASE between maize and rice orthologs, with accompanying evidence of positive selection (Waters et al. 2013). The joint availability of high-throughput RNA sequencing (RNA-seq) assays and methods for quantitative genotyping of single-nucleotide variants (SNVs) has allowed ASE to be studied in sugarcane (Vilela et al. 2017; Sforça et al. 2019; Cai et al. 2020; Correr et al. 2021). However, these previous investigations focused on a small number of genes, rely on limited genomic sampling, or suffer from potential sources of bias in the genotype calls. There are as yet no genome-wide analyses of ASE in sugarcane that make use of high-depth whole-genome sequencing (WGS) data. Hybridization events in sugarcane include the likely occurrence of allopolyploidy, interspecific hybridization during early breeding endeavors, and crosses between elite genotypes to obtain full-sib progenies, which are all possible causes of ASE. Artificial selection after interspecific hybridization, mainly for high sugar yield and disease resistance, has potentially affected the expression patterns of individual alleles in varying ways. Also, because many genes are involved in the accumulation of sucrose from top to bottom internodes in sugarcane (Whittaker and Botha 1997; Botha and Black 2000), it is conceivable that ASE may play a role in this carbon partitioning process.

We aimed to gauge the genome-wide extent of ASE in culms of elite sugarcane hybrids at different developmental stages and how structural and functional properties of genes affect the rate of ASE in this complex polyploid.

Results

Genome-wide study of polymorphisms in sugarcane hybrids by high sequencing depth

A set of seven sugarcane hybrids were selected from a larger set of 24 hybrids with extensive phenotypic variation (Supplemental Fig. S1) and used to study ASE. According to the phenotypic characterization detailed by Perlo et al. (2020), KQ228 was chosen to represent a high-yield (as measured in tons of cane per hectare), high early-season sugar genotype. In contrast, although SRA5 was also high yielding, it showed low sugar and high fiber content. Q155 and KQB09-20432 (KQB09) represented high-sugar and high-fiber genotypes, respectively. SRA1 showed both low sugar and low fiber content. Finally, MQ239 and Q186 had intermediate behavior with regard to both traits. The seven hybrids are representative of current elite breeding germplasm and are thus all related with multiple recurring common ancestors (Supplemental Fig. S2). We noted a peculiarity in the genetic background of KQB09. The paternal grandfather of KQB09, Hainan92-9, is a *S. spontaneum*. Although KQB09 is a direct descendant of KQ228, this unique background makes it the most genetically dissimilar among the studied hybrids.

The entire genomes of the seven hybrids were sequenced to identify polymorphic sites. KQ228 and SRA5 were sequenced at a depth of coverage of 100× per chromosome copy, or 1200× of the monoploid genome, considering a polyploid genome of 10 Gbp with 12 copies of each chromosome (D'Hont and Glaszmann 2001; Le Cunff et al. 2008). For the other genotypes, 240× of the monoploid genome was obtained (Supplemental Table S1). The high-quality reads were aligned against the sorghum genome reference with between 37.1% and 38.5% of the reads aligning. The specificity of alignment against genic regions was high, with little effect of filtering out low-quality alignments, in agreement with the fact that the genes are low-copy, conserved regions. The effective depth of coverage after alignment was close to the expected values of 1200× and 240×, depending on the genotype (Supplemental Fig. S3A). This approach was especially successful in unambiguously assigning reads to conserved single-copy grass genes (Supplemental Fig. S3B).

A total of 9,321,181 polymorphic sites were identified in the genic regions of the seven genotypes, of which 6,533,916 were biallelic SNVs and 192,157 were multiallelic SNVs. The majority of the remaining sites were insertions and deletions, indicating that the redundancy brought about by the multiple chromosome copies in sugarcane possibly allows for a substantial number of indels to be present. After filtering out low-quality polymorphisms 5,748,251 biallelic SNVs remained. Classifying these sites based on their predicted impact on gene structures resulted in the following distribution of predicted effects: 13,619,337 (79.82%) modifier effects, 1,799,543 (10.55%) SNVs with low impact, 1,608,014 (9.42%) with moderate effects, and 36,392 (0.21%) with high predicted impact. The number of effects is larger than the number of SNVs because each site can be annotated multiple times when in close proximity to multiple genes. For SNVs present within coding regions, 1,567,565 (48.96%) were annotated as silent, or synonymous substitutions; 1,611,587 (50.33%) caused amino acid changes (missense substitutions); and 22,716 (0.71%) were nonsense mutations. A total of 2,140,328 of the 5.75 million high-quality SNVs were polymorphic only in comparison to the sorghum reference but were fixed in these sugarcane genotypes. Because we can only assess ASE for heterozygous sites, these loci were not used for downstream analyses.

Discrete clusters of allele doses revealed by WGS and RNA-seq

The expression data set used corresponded to an RNA-seq assay of sugarcane stalks at five different developmental stages. For each of the genotypes, triplicate libraries were sequenced for internodes of different ages (collection time points C1 and C2) and varying maturity (internodes 5, 8, and 22; the latter also denoted by Ex-5; for details, see Methods). Sequencing generated from 56.0 million to 85.1 million raw reads per library, of which 50.6% to 84.7% remained after preprocessing to remove low-quality reads and contaminant ribosomal RNA, with a median yield of 81.8% (Supplemental Table S2). Hierarchical clustering was used to assess the variation between biological replicates and thus to diagnose potential issues during library prep and sequencing (Supplemental Fig. S4). Samples of lower quality were finally discarded: one biological replicate of KQ228 C2 In5, SRA1 C1 In5, Q186 C2 In5, and Q186 C2 InEx-5.

Alignment rates of the RNA-seq libraries against sorghum were higher than for the WGS data, ranging from 72.4% to 86.0% (Supplemental Table S2). After quantifying allele-specific abundances for each biallelic SNV, poorly covered genomic regions and lowly expressed genes were filtered out, and from

512,827 to 851,295 heterozygous sites per treatment level were analyzed (Table 1).

An initial exploratory data analysis to investigate the relationship between genomic and expressed proportions of the reference allele at each site revealed several features of ASE in sugarcane. The genomic allele proportions showed 11 well-defined clusters, with each cluster centered directly over a fraction of 12 (Fig. 1). This agrees with there being 11 heterozygous classes, with one to 11 doses of the reference allele in a dodecaploid genome. The majority of SNVs showed a single dose of the alternative allele; that is, the major allele was present in 11 copies for most variant sites. Some loci did not cluster tightly with any of the groups, including several with an apparent dose greater than 11 or less than one. In addition to simple random noise, this may indicate events of aneuploidy and/or copy number variants.

Frequency of ASE in sugarcane

A strong overall agreement between genomic and expressed allele ratios was observed. The fraction of reads carrying the reference allele in the RNA-seq data set was directly proportional to the corresponding WGS fraction (Fig. 1). There was very limited bias toward the reference allele, which is a common concern in ASE studies (Castel et al. 2015). Another feature of many of the SNVs was that they showed exclusive expression of one allele, as seen by the masses of points forming the horizontal lines at $Y = 0$ and $Y = 1$. These loci are the likely cause of the departure of the smoothed trend (in pink) from the null expectation (Fig. 1, red diagonal line). These observations were consistent for all genotypes and internodes (Supplemental Fig. S5).

A Bayesian hierarchical beta-binomial model (Correr et al. 2021) confirmed that most SNVs showed no evidence of ASE, with 19,154 to 56,956 heterozygous sites yielding significant tests (Table 1; Supplemental Fig. S6). Similar fractions of SNVs with ASE were found for all genotypes and internodes, except for the four cases in which a lower-quality sample was discarded. Uniformity was also seen when the heterozygous loci were gathered at the gene level. Between 2311 and 5902 genes had at least two SNVs with significant test results and were classified as genes showing allele-specific expression. They correspond to 13.6% to 31.7% of the effectively sampled genes, with an average of 24.1% (Table 1). The full set of genes with ASE, for the 35 combinations of genotypes and internodes, is provided in Supplemental Table S3.

Table 1. Results of the allele-specific expression (ASE) tests

Genotype	Internode	Sites tested	Sites with ASE	Sampled genes	Genes with ASE
KQ228	C1 In5	784,678	50,969	19,254	5652 (29.35%)
	C1 In8	776,315	56,956	18,544	5881 (31.71%)
	C2 In5	670,019	38,142	17,644	4161 (23.58%)
	C2 In8	779,532	54,029	18,370	5577 (30.36%)
	C2 InEx-5	788,762	54,494	18,151	5660 (31.18%)
KQB09	C1 In5	635,981	31,955	18,429	3799 (20.61%)
	C1 In8	682,963	43,452	17,918	4637 (25.88%)
	C2 In5	740,234	34,953	18,698	4060 (21.71%)
	C2 In8	708,982	42,213	17,726	4476 (25.25%)
	C2 InEx-5	769,011	47,203	17,393	4873 (28.02%)
MQ239	C1 In5	591,330	32,415	18,375	3789 (20.62%)
	C1 In8	640,136	36,967	17,448	4025 (23.07%)
	C2 In5	673,876	38,622	17,813	4059 (22.79%)
	C2 In8	708,504	42,413	17,807	4462 (25.06%)
	C2 InEx-5	644,549	30,013	17,125	3377 (19.72%)
Q155	C1 In5	599,518	28,372	18,052	3402 (18.85%)
	C1 In8	670,021	40,431	17,776	4391 (24.70%)
	C2 In5	695,515	38,248	18,057	4167 (23.08%)
	C2 In8	708,493	36,893	17,750	4074 (22.95%)
	C2 InEx-5	714,418	37,491	17,522	4054 (23.14%)
Q186	C1 In5	684,824	37,348	18,729	4224 (22.55%)
	C1 In8	659,836	35,153	17,478	3833 (21.93%)
	C2 In5	577,941	27,231	16,653	2971 (17.84%)
	C2 In8	746,505	41,164	17,836	4342 (24.34%)
	C2 InEx-5	599,414	25,547	16,538	2849 (17.23%)
SRA1	C1 In5	512,827	19,154	16,936	2311 (13.65%)
	C1 In8	648,316	37,685	17,438	3976 (22.80%)
	C2 In5	758,633	41,910	18,393	4482 (24.37%)
	C2 In8	741,704	43,821	18,188	4602 (25.30%)
	C2 InEx-5	655,869	29,526	17,204	3215 (18.69%)
SRA5	C1 In5	628,813	33,587	18,109	3967 (21.91%)
	C1 In8	788,206	52,306	18,292	5467 (29.89%)
	C2 In5	851,295	55,743	18,879	5902 (31.26%)
	C2 In8	815,312	52,494	18,556	5642 (30.41%)
	C2 InEx-5	798,053	49,362	17,898	5264 (29.41%)

For each combination of sugarcane genotype and internode, we show the total number of informative sites, which are heterozygous with at least 50 genomic reads and 10 RNA-seq reads, and those with significant ASE. Sampled genes are those with two or more informative variant sites.

Factors contributing to ASE

When comparing genes showing ASE for the various treatments, a higher similarity was found among samples from the same genotype than from the same internode (Fig. 2). The overlap of genes with ASE for different internodes of the same genotype was 61.6% on average, ranging from 57.8% (SRA1) to 63.9% (KQ228). In some cases, the overlap was as high as 74.6%, as seen for the pair KQ228 C2 In8 and KQ228 C2 InEx-5. On the other hand, the overlap among samples of the same internode ranged from 52.3% (C1 In5) to 57.5% (C2 In8), with an average of 54.8%. Combinations involving different genotypes and different internodes still showed noticeable overlap, ranging from 31.8% to 61.7% (average of 49.5%). Some genes consistently showed ASE for different treatments, whereas there was an added effect of internode and an even greater contribution of the genotype factor.

Clustering the treatments based on the outcomes of ASE tests, that is, the patterns of individual genes with or without ASE, confirmed the closer proximity of samples from the same genotype (Supplemental Fig. S7). The exception was a cluster composed of the more immature internodes from five of the seven genotypes, which were placed to the left of the dendrogram. This indicates that early in stalk development, when all genotypes divert carbon to synthesize cell walls, similar sets of genes may show ASE regardless of the genetic background. Although KQB09 has a unique genomic composition, with recent contribution from *S. spontaneum*, samples from its more mature culms clustered closer to its parent KQ228.

A more detailed comparison of heterozygous SNVs in common for KQ228 and SRA5 showed a strong positive correlation between the genomic allele proportions observed in each genotype ($\rho = 0.92$, P-value $< 10^{-15}$) (Fig. 3A). Rarely was the difference between their most likely doses greater than three or four chromosome copies. A similar trend for the relationship was found between the expressed allele proportions, with $\rho = 0.88$, P-value $< 10^{-15}$ (Fig. 3B). Some loci did show substantial differences between the proportions of the reference allele in the two transcriptomes.

When there was a significant excess of a given allele in both genotypes, the preferential allele expression was often in the same direction (Fig. 3C). In most cases, the bias was toward the reference allele, as seen by the heavier mass of points in the first

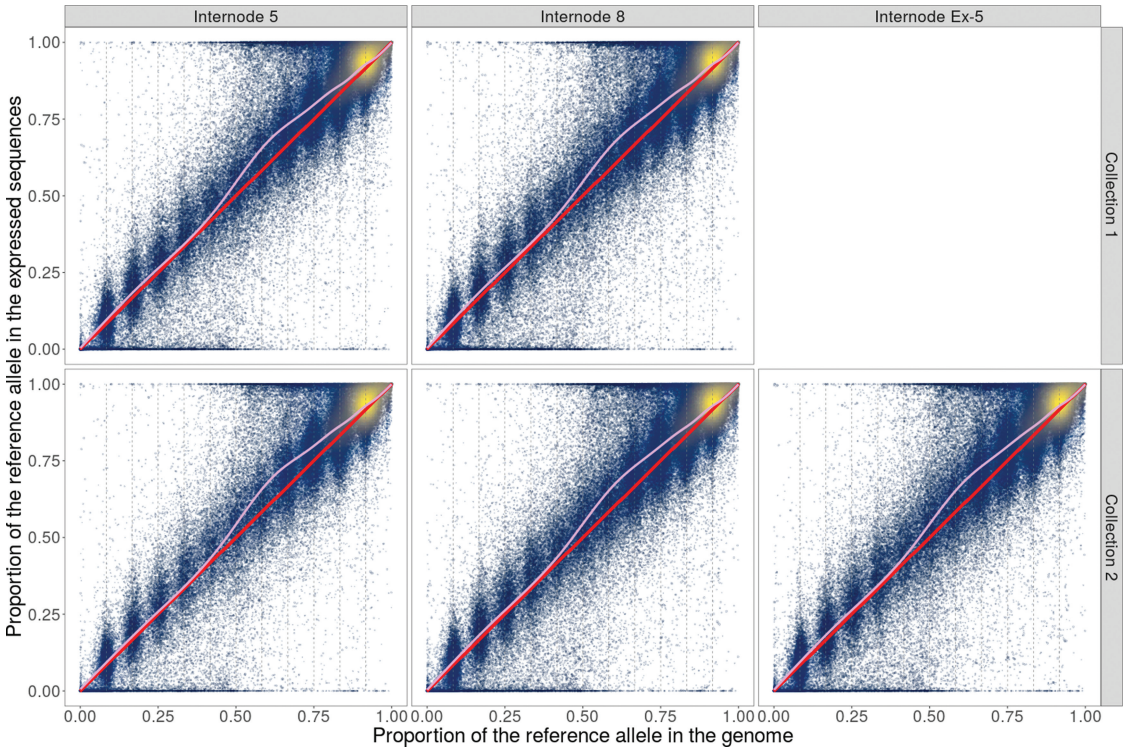

Figure 1. Relationship between expressed and genomic proportions of the reference allele for genotype KQ228. Different internodes are shown in separate panels. The red line indicates the null hypothesis of perfect identity between the expressed proportion of the reference allele (Υ; in the *y*-axis) and the corresponding genomic proportion (denoted by Γ; *x*-axis). Each point represents one single-nucleotide variant (SNV), and lighter colors indicate a higher density of points. The smoothed trend of observed points is shown in pink. Notice the majority of sites with high proportion of the reference allele in the genome. To highlight the discrete nature of clusters, we only show high-depth sites, with 1000 or more genomic reads and 80 or more RNA-seq reads.

quadrant. These observations are likely reflections of their similar genetic background. This may be in part because of the shared domestication history of *Saccharum* genotypes, as well as to strong selective pressures in the recent breeding of sugarcane.

ASE frequency in defense response, cell wall, and stress-response genes

The frequency of occurrence of ASE in genes involved in particular cellular functions was investigated. Gene Set Enrichment Analysis (GSEA) (Mootha et al. 2003; Subramanian et al. 2005) provided evidence of enriched Gene Ontology (GO) terms (The Gene Ontology Consortium et al. 2000). Two functional terms appeared as consistently enriched for most treatments, namely, ADP binding (GO:0043531) and defense response (GO:0006952) (Supplemental Table S4; Supplemental Fig. S8). These terms are in fact closely related, as many resistance proteins have nucleotide-binding domains, and many genes were simultaneously annotated with both terms. Enrichment of UDP-glycosyltransferase activity (GO:0008194) was found for most genotypes and internodes. UDP-glycosyltransferases are a large protein family with multiple roles, including the biosynthesis of many cell wall compounds. The terms sulfotransferase activity (GO:0008146) and sulfation (GO:0051923) were also enriched for many treatments. Additional GO terms were significantly enriched in particular treatments, but no discernible pattern was observable, except for an apparent excess of terms for KQ228 and SRA5, the genotypes sequenced at higher depth. The terms *O*-methyltransferase (GO:0008171) and aromatic compound biosynthetic process (GO:0019438) were enriched in the fiber-rich genotype SRA5, particularly in its upper internodes. Manual inspection revealed that genes annotated with these terms and showing ASE are potentially involved in the synthesis of cell wall components including a gene similar to *ZRP4* (*Zea* Root Preferential), which is involved in the synthesis of suberin and possibly of lignin (Held et al. 1993; Bosch et al. 2011). Strong ASE was observed for a gene similar to herbicide safener binding protein, an *O*-methyltransferase predicted to be involved in the synthesis of lignin precursors (Scott-Craig et al. 1998).

A direct comparison of the contrasting KQ228 and SRA5 genotypes also showed how ASE may contribute to phenotypic variation. A total of 389 genes were found that consistently showed ASE in at least four of the five sampled internodes in KQ228 but in no more than one internode in SRA5. Conversely, 399 genes displayed consistent SRA5-specific ASE. Mapping both gene lists to the KEGG pathways (Kanehisa and Goto 2000) showed that only one of the KQ228-specific genes was assigned to the phenylpropanoid pathway, namely, a β-glucosidase that catalyzes the synthesis of coumarinate (BGLU, EC 3.2.1.21) (Fig. 4). In contrast, the SRA5-specific genes included five additional genes in this pathway, responsible for multiple steps in the biosynthesis and polymerization of monolignols, the precursor molecules of lignin (Vogt 2010). These included *trans*-cinnamate 4-monooxygenase (C4H, EC 1.14.14.91), 4-coumarate-CoA ligase (4CL, EC 6.2.1.12), shikimate O-hydroxycinnamoyltransferase (HCT, EC 2.3.1.133), cinnamoyl-CoA reductase (CCR, EC 1.2.1.44), and a peroxidase (PER, EC 1.11.1.7). It is conceivable that consistent ASE in these genes is important for the fiber-rich phenotype of the SRA5 cultivar. Binning genes into MapMan functional groups (Thimm et al. 2004)

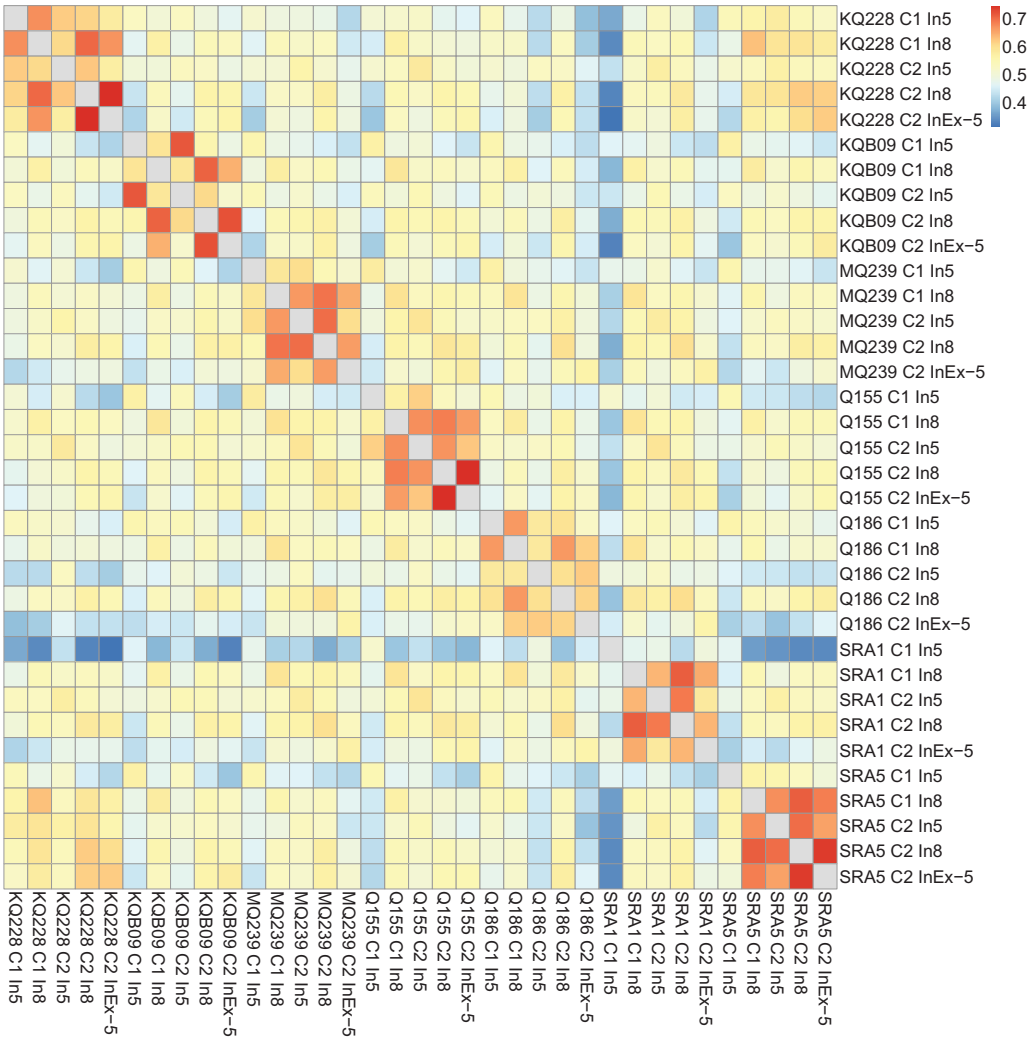

Figure 2. Percentage of genes with allele-specific expression (ASE) detected in common for all genotype × internode combinations. We found higher similarity among different internodes of the same genotype (hot colors in the diagonal blocks), and there was also similarity between the same internode of different genotypes (the light orange cells off the main diagonal).

showed additional SRA5-specific genes related to cell wall organization, such as Cyt-P450 hydroxylase scaffold protein (membrane steroid binding protein [MSBP]), cellulose synthase, endo-1,4-beta-glucanase, and leucine-rich repeat extensin. Eleven genes with KQ228-specific ASE were also associated with cell wall organization but involved in hemicellulose and pectin metabolism. These include extensin beta-1,2-arabinosyltransferase, glucuronosyltransferase, callose synthase, endo-beta-1,4-xylanase, beta-1,3-galactosyltransferase, mannan synthase, beta-galactosidase, and xylosyltransferase. Genes associated with post-Golgi vesicle trafficking, such as SNARE proteins and Rab GTPases, which are important for transporting cell wall components, were found in higher numbers in SRA5 (18 vs. nine in KQ228).

Sugar and starch metabolism is tightly linked to the phenotype of sucrose accumulation. No gene with SRA5-specific ASE was annotated with this term, but we observed six such genes in KQ228: triose phosphate/phosphate translocator, glucose transporter, UDP-D-glucose 4-epimerase, starch branching enzyme, starch synthase, and cytosolic glucanotransferase DPE2. Possibly related solute transporters, including ABCB transporters, hexose transporter, and organic phosphate/glycerol-3-phosphate permease, were found in excess in KQ228 (21 vs. 10). This shows that ASE also affects sucrose accumulation in sugarcane. The genes with consistent ASE in both genotypes also included multiple genes involved in RNA and protein biosynthesis, including large and small ribosomal subunit components, mRNA quality control, and various transcription factors.

Because SRA1 showed low sugar and fiber contents, the genes displaying consistent ASE in this genotype were investigated. Compared with the genes with ASE in SRA5, SRA1 showed 163 specific genes enriched for signaling proteins (10 kinases and three phosphatases), multiprocess regulation (two SnRK1-interacting factors, a pyrophosphatase, and a CBL-dependent protein kinase [CIPK]), redox homeostasis (iron superoxide dismutase and glutathione S-transferase), and the metabolism of terpenoids. These functional groups are associated with stress response, in agreement with the poor phenotype observed for SRA1, and show that ASE is associated with response to environmental and developmental stress. Other stress-related genes with SRA1-specific ASE were a glutaredoxin, two multidrug resistance proteins (MATE), two

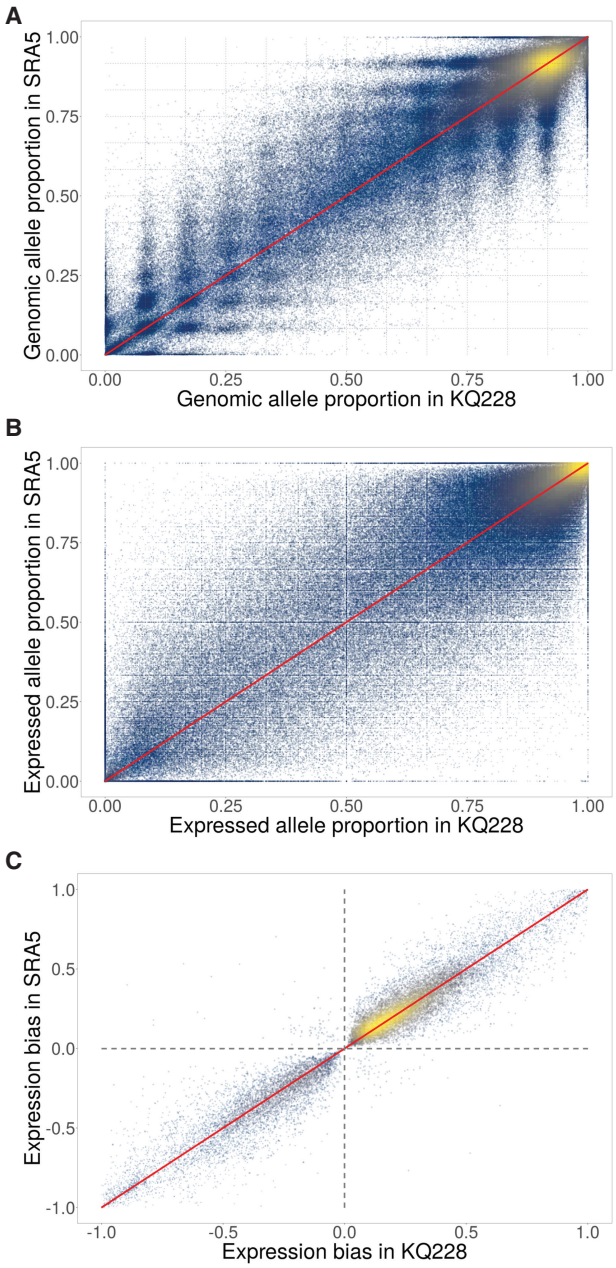

Figure 3. Comparison between KQ228 and SRA5, the two genotypes sequenced at higher depth. (*A*) Relationship between genomic allele ratios (doses) for shared variants, showing overall agreement between the two genotypes. (*B*) Relationship between allele proportions in their expressed sequences. (*C*) Comparison of expression bias for shared SNVs with significant ASE, showing prevalence of higher expression of the reference allele in both genotypes.

interleukin-1 receptor-associated (IRAK) kinases, peroxin, and a polyamine oxidase.

The type of mutation and gene conservation associated with ASE

Many of the SNVs in genes annotated with enriched GO terms showed exclusive expression of the reference allele (Fig. 5A,B). This maximum form of ASE may imply that transcriptional and post-transcriptional regulatory mechanisms have a potential impact on the expression of different alleles for these genes. A comparable tendency was seen for SNVs predicted to have a large impact on protein structure (Fig. 5C). This classification is assigned to SNVs that affect the stop codon, possibly giving rise to truncated or extended proteins in the case of stop gain and stop loss substitutions, respectively. Such gene products are often not fully functional and can be detrimental to cell activity. It is thus not surprising that we observed a single allele to be exclusively present for many of these loci, providing evidence of absence of transcription or post-transcriptional mRNA decay of the other SNV allele (Rivas et al. 2015). At the genome level, the distribution of estimated allele doses per SNV class showed an increase in the frequency of higher doses of the reference allele for variants of higher predicted impact (Supplemental Fig. S9A,B). In addition to the already higher genomic doses, a higher-than-expected abundance was observed for transcripts carrying the reference allele for high-impact SNVs (Supplemental Fig. S9C).

Pham et al. (2017) found that core angiosperm genes responsible for key plant processes were more likely to show ASE in multiple tetraploid potato genotypes. In the present study, genes conserved in a set of 17 monocot species had higher odds (odds ratios > 1.5) of showing ASE than the remaining genes (Fig. 5D). Conversely, we found that genes that are present in single copy in sorghum, rice, and *Brachypodium* were less likely to show preferential expression of alleles. These observations support the notion that conserved genes controlling key biological processes may be enriched among those with ASE, but that single-copy genes may be under some control mechanism limiting ASE in sugarcane. Lower odds ratios were also seen for sorghum-exclusive paralogs, but the number of genes in these groups of paralogs was small and statistical significance was often not attained.

Discussion

Significant ASE in the culms of elite sugarcane hybrids was limited to, on average, one out of four genes, with most heterozygous loci showing agreement between allele ratios in genomic and expressed sequences. Previous studies have shown evidence for ASE in sugarcane (Vilela et al. 2017; Sforça et al. 2019; Cai et al. 2020), but the current genome-wide high-depth analysis shows that ASE is not widespread in this complex polyploid, probably owing to factors regulating gene expression having comparable effects on the different alleles. Because of the high ploidy levels of sugarcane hybrids, it was essential to make use of high sequencing depths to obtain precise estimates of allele doses (Margarido and Heckerman 2015; Gerard et al. 2018). This is particularly relevant if the possibility of bias and overdispersion in allele read counts is considered. To err on the side of being conservative, four RNA-seq libraries were removed from ASE analyses to avoid potential issues in accurately quantifying the expression of different alleles. Nevertheless, these samples were from distinct treatments, such that the effect on statistical power to detect ASE was limited.

Because read alignment was performed against sorghum, the pipeline included various filtering steps to reduce the occurrence of spurious SNVs. The total RNA-seq alignment rates were on par with those seen when using the current sugarcane genomes (Diniz et al. 2019). Alignment of reads from transcribed regions was relatively straightforward, such that the majority of relevant alignments were unambiguous. Ensuring that a large proportion of the reads were effectively and correctly assigned to their corresponding genomic regions is crucial in ASE studies (Castel et al. 2015). The fact that very little (if any) bias toward the reference

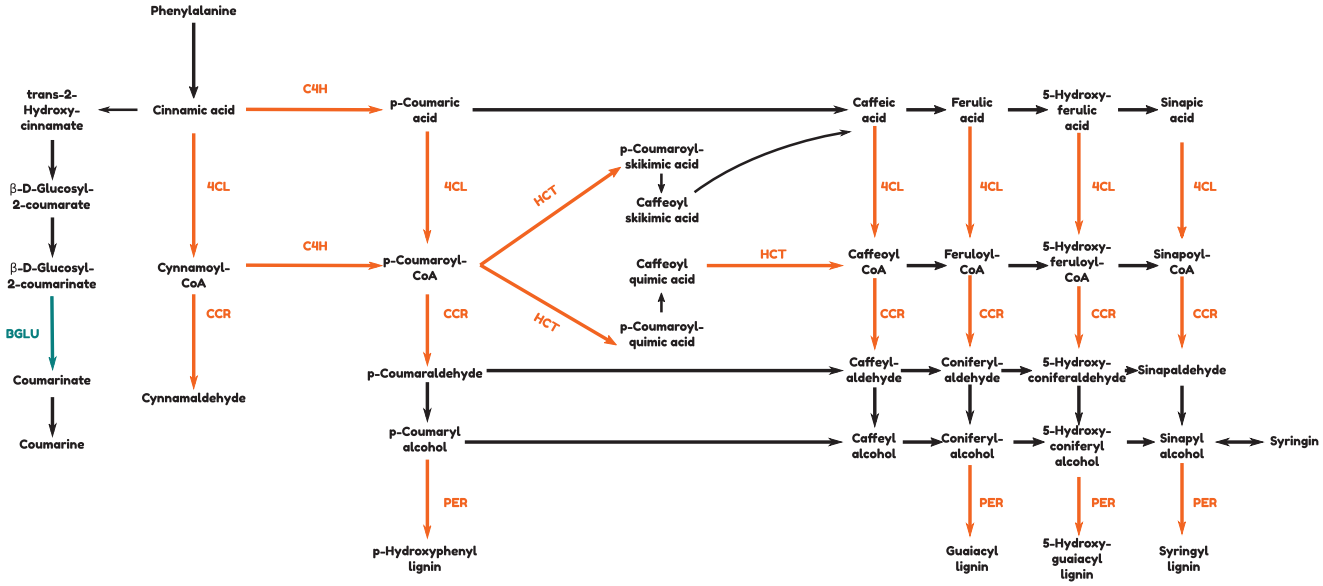

Figure 4. Genes in the phenylpropanoid pathway with consistent ASE for SRA5 and KQ228. Enzymes in orange correspond to genes with consistent ASE in the internodes of SRA5 but not in KQ228. The metabolic step in blue indicates different genes with consistent ASE in either genotype. (BGLU) β-Glucosidase; (C4H) *trans*-cinnamate 4-monooxygenase; (4CL) 4-coumarate-CoA ligase; (HCT) shikimate O-hydroxycinnamoyltransferase; (CCR) cinnamoyl-CoA reductase; (PER) peroxidase.

allele was observed indicates a good overall performance of the pipeline, underlining the reliability of the results.

A beta-binomial model was adapted to investigate ASE in mixed- and high-ploidy organisms (Correr et al. 2021), and this study further modified this method to model uncertainty in estimating allele doses. This approach allowed nonhomogeneous sequencing depth profiles to be naturally handled by the model, with a corresponding adjustment of the null hypothesis for each polymorphism. This statistical analysis strategy likely had an impact on the fraction of genes identified as showing significant ASE. Although there was evidence of ASE for 13.6% to 31.7% of the genes in sugarcane, other studies have reported widely varying values: 13.5% in *Arabidopsis* (Zhang and Borevitz 2009), roughly a third to half of the genes in potato (Pham et al. 2017), ~50% of the genes in maize (Springer and Stupar 2007), and up to 70% among wheat homoeologs (Powell et al. 2017). This variation encompasses differences between diploids, autopolyploids, and allopolyploids, as well as true biological variation in these systems, and includes differences in experimental design and environmental and technical noise. However, a contribution from the distinct data processing and methodologies used in each case cannot be ruled out. Different statistical models and testing strategies, hard thresholds, and other ad hoc criteria can affect the results. In sugarcane, Correr et al. (2021) also observed ASE for ~40% of the sampled genes. However, this study relied on genotyping-by-sequencing of reduced representation libraries, a strategy that is prone to biases and genotyping errors (Scheben et al. 2017). These issues can then strongly influence conclusions about ASE for a set of the polymorphic sites.

In this context, one advantage of the GSEA method used here is that it does not depend on previous testing for ASE but only on the ranking of genes. Genes with large absolute deviations between genomic and expressed allele ratios had higher ranks, and functional terms with many high-ranking genes were tagged as enriched with ASE. A small number of GO terms showed significant enrichment, with the prominent presence of defense-response genes among those with ASE in many genotype and internode combinations. Many resistance proteins have nucleotide-binding domains, and binding to ADP or ATP induces conformational changes that regulate signaling cascades (DeYoung and Innes 2006; Tameling et al. 2006). The higher frequency of ASE among defense-response genes agrees with observations in allohexaploid wheat, in which many of these genes showed homoeologous expression bias in plants challenged with a pathogenic fungus (Powell et al. 2017). Species in the genus *Saccharum* differ broadly with regard to their response to pathogen infection, and sugarcane hybridization and breeding have historically focused on selecting for disease resistance (Cheavegatti-Gianotto et al. 2011; Rott et al. 2013). The seven hybrids have been selected for disease resistance and show different patterns of response against a range of pathogens (Supplemental Table S5). Different alleles may respond differently to particular pathogen strains or races, and breeders may have indirectly selected for higher expression of specific alleles. This ASE probably is also linked to the smaller contribution of the *S. spontaneum* genome to modern hybrids, as a consequence of abnormal chromosome transmission in the initial backcrossing. Selection pressure on disease resistance kept those traits from *S. spontaneum* in the genome, and for most, they are left in low doses.

No homogeneous enrichment was detected for many functional terms associated with carbon partitioning, such as those involved in sucrose accumulation or cell wall biosynthesis, with the exception of UDP-glycosyltransferase activity. Enrichment was observed for assorted transferase terms (including of hexosyl groups), cellulose synthase (UDP-forming) activity (GO:0016760), plant-type primary cell wall biogenesis (GO:0009833), and mannan synthase activity (GO:0051753) for scattered treatments. This indicates that sizeable ASE may occur at least in some genotypes and at some points during the development and maturation of sugarcane stalks. The identification of high-level GO terms associated

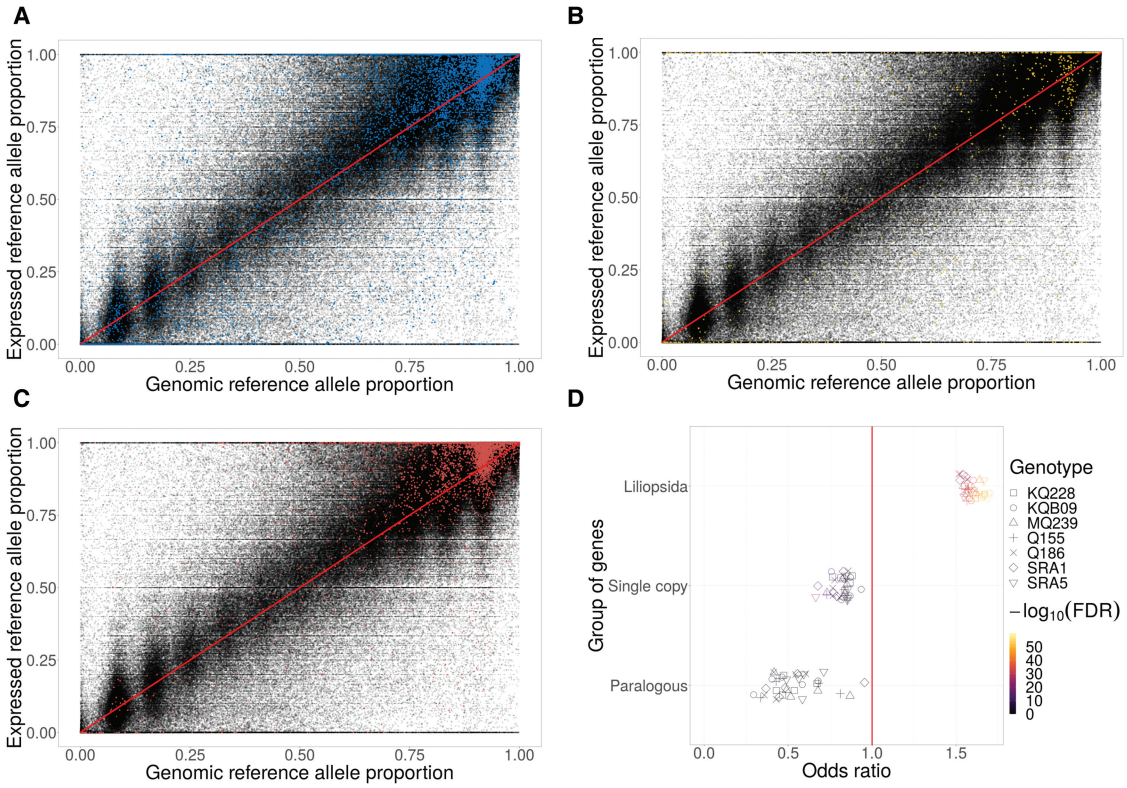

Figure 5. Enrichment results of genes with allele-specific expression. (*A*) SNVs in genes annotated with ontology term defense response (GO:0006952). (*B*) SNVs in genes annotated with ontology term UDP-glycosyltransferase activity (GO:0008194). (*C*) SNVs predicted to have high functional impact. In all cases, note the excess of SNVs with exclusive expression of the reference allele (horizontal line at $\Upsilon = 1$). (*D*) Underrepresentation of single-copy genes among those with ASE (odds ratio < 1); overrepresentation of core genes, conserved in Liliopsida (odds ratio > 1). Multiple points for the same genotype indicate different internodes.

with the synthesis of cell wall compounds in SRA5 opens up the possibility that ASE has direct effects on phenotypes of economic interest. A closer inspection of the phenylpropanoid pathway revealed that many genes involved in the biosynthesis of lignin showed significant ASE for the fiber-rich SRA5 genotype but not for the sugar-rich KQ228.

When analyzing high-impact SNVs, which are predicted to have more drastic effects on protein structure, both higher genomic doses and excessive abundance of the reference allele were discovered. The former implies purifying selection against the alternative allele, whereas the latter hints at its silencing or increased mRNA degradation. This effect of the type of mutation on the incidence of ASE has been reported in humans, similarly with a higher frequency of ASE in nonsense variants (Rivas et al. 2015). These observations can be somewhat noisy because SnpEff treats each polymorphism individually, and multiple adjacent variants may have a joint effect on gene products. However, this is mostly relevant to indels, and the fact that only SNVs were used in this study alleviates the problem. Polymorphisms affecting start and stop codons may be the primary cause of ASE for many genes. If this is in fact the case, the genes bearing these SNVs would be expected to show similar expression patterns across tissues and developmental stages. This agrees with the observation that the transcriptomes of different internodes for the same genotype showed more similar patterns of ASE than other combinations of treatments (Fig. 2). Hence, the control of ASE in sugarcane seems to be largely genetic, which has important implications for plant breeding. The expression of favorable alleles can be directly or indirectly leveraged through selection, at least for a subset of the genes, and these effects are expected to be partly passed onto the progeny of crosses between hybrids. Yet, an effect attributable to internodes and collection time points was also observed. Developmental and environmental cues thus appeared to have some contribution to ASE in these experiments. These effects are also partially accessible to selective breeding, to the extent that they may contribute to genotype × internode interaction. They are also relevant from a biological perspective as they enhance understanding of the biology of sugarcane hybrids.

Clusters of genes conserved in the Liliopsida were overrepresented among those with ASE, whereas single-copy grass genes were underrepresented. Fine-tuning of biological processes that are essential for proper functioning of plant metabolism may include mechanisms favoring the expression of some alleles over the others. In contrast, the lower frequency of ASE among single-copy genes suggests that redundancy may be a relevant factor driving ASE in sugarcane. The extended mutational freedom afforded by gene duplication may allow regulatory mechanisms to tweak the expression of individual allelic copies in these high-ploidy genomes.

The high-depth genomic data analyzed supports the presence of 12 chromosome copies for the majority of polymorphic sites in these seven sugarcane hybrids. Despite the extensive variation in ploidy and chromosome number existent in *Saccharum*, it is compelling to see such consistency among a set of elite hybrids. As is

common in sugarcane breeding programs, these hybrids share a narrow genetic base and are often derived from crosses involving repeated parents. Their common genetic background may have increased the likelihood of producing genomes with similar composition, in spite of their substantial phenotypic variation. However, we could neither infer the genomic origin of each allele nor observe any clear differentiation between putative *S. officinarum*–derived and *S. spontaneum*–derived alleles based on short-range polymorphisms alone. Although the genus *Saccharum* may comprise three distinct subgenomes, these founding genomes are apparently very similar to each other, especially for exonic regions, where the average sequence identity is >99% (Pompidor et al. 2021). Given such high similarity among gene copies, it may not be possible to tell the three subgenomes apart when working at the SNV level in transcribed regions.

The combination of high-depth WGS data and RNA-seq has provided a comprehensive view of the incidence of ASE in sugarcane hybrids. It does not appear to be a pervasive feature of gene expression in this complex crop. Using the sorghum genome as a reference focuses the analysis on genes that are conserved between the two species. An interesting future research opportunity is to investigate sugarcane-exclusive genes and assess whether they are less likely to show ASE. This will be possible when a complete (phased) hybrid sugarcane genome becomes available but will also require improvements in long-read RNA-seq to allow the allelic origin of each read to be unambiguously inferred. In addition, this genomic resource will allow polymorphisms in regulatory regions to be explored and, consequently, to identify *cis*-acting elements that control the occurrence of ASE in sugarcane. This information can then be leveraged into molecular breeding efforts in this important bioenergy crop.

Methods

Biological material and sample collection

Perlo et al. (2020) described a field experiment used to evaluate 24 sugarcane hybrids contrasting in several traits, including fiber content and sugar yield. Later, we performed RNA-seq to assess differential gene expression in internodes collected at five different developmental stages. Briefly, 24 *Saccharum* hybrids were planted in a 6 × 4 Latin square with three replicates, in August 2017 at the Sugar Research Burdekin Station in Burdekin, Queensland, Australia. Each plot was 4 m long with 1.52 m between rows, and environmental conditions were kept uniform via fertilization and furrow irrigation. The genotypes were phenotypically characterized in March, June, and September of 2018 by measuring the soluble solids content, polarity, and fiber percentage. For RNA-seq, three independent biological replicates of stalks were collected 19 wk (the first collection, or C1) and 37 wk (C2) after planting. Internodes 5 and 8 were sampled in both collections, whereas internode Ex-5 was sampled in C2. The latter corresponds to internode 5 from C1, tagged in intact culms to be collected when more mature. In all cases, samples were collected in the morning hours, sliced into smaller fragments, frozen in liquid nitrogen immediately after cutting, and kept at −80°C until processed. We refer the reader to Perlo et al. (2020) for further details about the trial.

Based on these phenotypic traits, we chose seven genotypes to use in our work. KQ228 and SRA5 showed contrasting behavior in terms of sugar and fiber accumulation and were chosen as the main genotypes to explore herein. We also sampled the pair Q155 and KQB09-20432, which showed similar phenotypes to KQ228 and SRA5, respectively. The remaining genotypes, SRA1, MQ239, and Q186, were selected because of their negative or neutral phenotypic traits.

WGS and quality control

Leaves +4 and +5 of each of the seven genotypes were sampled for WGS, where leaf +1 is the first with a visible dewlap. DNA was extracted using the method described by Furtado (2014), modified by adding 5 mL chloroform instead of the phenol:chloroform:isoamyl alcohol (25:24:1) mixture. Following extraction, DNA integrity was checked with a NanoDrop spectrophotometer assay to measure the 260/280 and 230/260 ratios.

High-quality genomic DNA was used for sequencing libraries following the Illumina protocol. The seven libraries were sequenced on one S4 flowcell of a NovaSeq 6000 platform to obtain 2 × 150-bp paired-end reads. The sequencing run was designed to represent the genotypes at different depths of coverage. Libraries for KQ228 and SRA5 were sequenced at five times the volume of the other genotypes, with expected depths of 100× and 20× per chromosome copy, respectively.

Raw whole-genome reads were trimmed using CLC Genomics Workbench v20.0.4 at a quality limit of 0.01 to remove data with low Phred scores (Supplemental Table S1). We also removed reads with more than two ambiguous bases and discarded reads shorter than 50 bases after trimming.

Read alignment and SNV calling

The *Sorghum bicolor* genome 454 v3.0.1 (Paterson et al. 2009; McCormick et al. 2018) was used as a reference to identify SNVs. Briefly, alignment of genomic reads to the sorghum genome was performed with Bowtie 2 v2.4.1 (Langmead and Salzberg 2012); duplicate reads were removed; and only primary alignments were kept. The GATK v4.1.8.1 pipeline (DePristo et al. 2011) was used to identify SNVs and call genotypes. We annotated the predicted impact of each SNV based on their relative genomic position with SnpEff v5.0 (Cingolani et al. 2012). Details about SNV calling are in the Supplemental Methods.

RNA-seq, data preprocessing, and alignment

Total RNA was extracted from three replicates of each genotype × internode combination for a total of 105 samples representing 35 treatments (seven genotypes × five internodes), with each sample comprising a pool of four clonal stalks. Culms from each treatment were first individually pulverized; equal amounts of four stools were combined to form each pool; and pools were again kept at −80°C.

Sequencing libraries were constructed with the standard TruSeq RNA protocol, and the cDNA was sequenced (in pools with additional indexed libraries) on the Illumina NovaSeq 6000. Samples were split into four lanes of one S4 flowcell, with a data volume corresponding to 90 samples per lane. With this multiplexing strategy, the expected yield was of more than 50 million paired-end 100-bp reads for each sample (Supplemental Table S2).

The raw reads were processed to remove Illumina adapters, low-quality bases, and contaminating ribosomal RNA reads. For aligning the RNA-seq reads, we used the same sorghum genome reference and the splice-aware aligner HISAT v2.1.0 (Kim et al. 2015). Details about the processing of RNA-seq reads are in the Supplemental Methods.

Quantification of relative allele proportions

For both the WGS and RNA-seq data sets, we applied the ASEReadCounter tool v4.1.8.1 to estimate the frequency of each allele, in read counts, for each biallelic SNV of each genotype (this tool does not allow for indels or multiallelic sites). At the genomic level, this provided an estimate of the relative allele dose. In this case, we removed PCR/optical duplicates to avoid potential bias in estimating the doses. At the expression level, these counts were used to obtain estimates of the relative expression levels of both alleles. For RNA-seq reads, the duplicate read filter was disabled so as not to bias the estimated allelic proportions for highly expressed genes. In both cases, we disabled the maximum depth limit to ensure that all reads were effectively counted. As a diagnostic metric, we performed hierarchical clustering with R v4.0.5 (R Core Team 2021) to visualize the dissimilarity among samples. To that end, we used SNVs for which the RNA-seq read count was greater than or equal to five for at least half of the samples. Euclidean distances were calculated based on the relative frequency of the reference allele, and the default complete linkage method was used to find clusters.

Assessment of allele-specific expression

Based on the clustering pattern and the amount of rRNA and PCR duplicates, four samples that were of comparatively lower quality than the rest in the RNA-seq data set were excluded. These samples were all from distinct treatments, such that four genotype × internode combinations were represented by two replicates, whereas the remaining 31 treatments had all three replicates. For each of the 35 treatments, we combined the allele counts of the remaining high-quality RNA-seq replicates.

In addition to the raw allele counts, we calculated for each polymorphic site the relative proportion of the reference allele, both for the genomic and expression data sets. Let r_{ik} represent the number of WGS reads carrying the reference allele for SNV i and treatment k, and let g_{ik} indicate the total number of genomic reads for the corresponding locus. We then calculated the observed genomic proportion of the reference allele, denoted by Γ, as $\Gamma_{ik} = \frac{r_{ik}}{g_{ik}}$. Similarly for the transcriptome data, we denote by y_{ik} the number of RNA-seq reads with the reference allele and by n_{ik} the total corresponding read count. The observed expressed proportion of the reference allele is then given by $\Upsilon_{ik} = \frac{y_{ik}}{n_{ik}}$.

To test for ASE, the hierarchical beta-binomial model proposed by Correr et al. (2021) was used with modifications. We first modelled y_{ik} according to a binomial distribution, $y_{ik} \sim \text{Binomial}(n_{ik}, \theta_{ik})$. The a priori distribution of the parameter θ_{ik} was modelled with a beta distribution, $\theta_{ik} \sim \text{Beta}(\alpha_{ik}, \beta_{ik})$, where α_{ik} and β_{ik} indicate the genomic dose of the reference and alternative alleles, respectively. Because these doses are unknown, we obtained estimates by approximating the observed genomic allele proportions, considering a ploidy of 12. In that case, we set $\alpha_{ik} = [12 \times \Gamma_{ik}]$, where the brackets represent the integer closest to $12 \times \Gamma_{ik}$, and $\beta_{ik} = 12 - \alpha_{ik}$.

According to this model, the posterior distribution of θ_{ik} is then $\text{Beta}(y_{ik} + \alpha_{ik},\ n_{ik} - y_{ik} + \beta_{ik})$, but in practice, we used $\text{Beta}(y_{ik} + \alpha_{ik} + 0.5,\ n_{ik} - y_{ik} + \beta_{ik} + 0.5)$ to avoid zero counts. For each site, we obtained the highest density interval (HDI) for this posterior distribution with the R package HDInterval v0.2.2 (https://cran.r-project.org/package=HDInterval). To account for the large number of genotype × SNV combinations, the Bonferroni correction (Bonferroni 1936) was applied to the interval mass, with a global significance level of 0.05 and considering an approximation of 100,000 independent tests. This is because neighboring SNVs in a given gene are not independent, such that the number of tests must be between the number of genes and the number of positions tested.

Finally, to call an SNV as showing significant ASE, we checked whether the HDI included the observed genomic proportion Γ of the reference allele. Because this proportion is an estimate and thus subject to random variation, instead of using a point estimate, we used the 95% confidence interval for the binomial distribution calculated with the Wilson method (Wilson 1927). If there was no overlap between this confidence interval and the posterior HDI, the corresponding SNV was deemed to show significant allele-specific expression. Sites with lower genomic depth of coverage have wider confidence intervals, such that this approach effectively models uncertainty and helps to avoid false positives in less-covered regions.

Functional enrichment tests

Genes were first clustered according to three criteria: (1) genes conserved in a set of 17 monocots, (2) sorghum-exclusive paralogs, and (3) single-copy genes in sorghum, rice, and *Brachypodium*. Poorly covered (fewer than 50 WGS reads) and lowly expressed (fewer than 10 RNA-seq reads) polymorphisms were excluded from the analysis, and a Fisher's exact test was used to test for enrichment of genes with ASE in each cluster. Enrichment of genes annotated with common GO functional terms was performed with GSEAPreranked v4.0.3 (Mootha et al. 2003; Subramanian et al. 2005), with ranking based on the median absolute deviation between expressed and genomic allele ratios. Only genes with 10 or more SNVs and GO terms with five or more genes were considered. See the Supplemental Methods for more details.

Data access

The WGS data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA733812. The RNA-seq data generated in this study have been submitted to the EMBL-EBI European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena/browser/home) under accession number PRJEB44480.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We acknowledge The University of Queensland's Research Computing Centre (RCC) for its support and the Center for Functional Genomics–ESALQ for its computational infrastructure. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior–Brasil (CAPES) grants CAPES-PRINT 88887.466432/2019-00 to G.R.A.M. and CAPES-PRINT 88887.367965/2019-00 to F.H.C. and by grant #2015/22993-7, São Paulo Research Foundation (FAPESP). Australian Research Council and Sugar Research Australia provided funding to R.J.H. and F.C.B.

References

Baduel P, Bray S, Vallejo-Marin M, Kolář F, Yant L. 2018. The "Polyploid Hop": shifting challenges and opportunities over the evolutionary lifespan of genome duplications. *Front Ecol Evol* **6:** 117. doi:10.3389/fevo.2018.00117

Baldauf JA, Vedder L, Schoof H, Hochholdinger F. 2020. Robust non-syntenic gene expression patterns in diverse maize hybrids during root development. *J Exp Bot* **71:** 865–876. doi:10.1093/jxb/erz452

Bonferroni CE. 1936. Teoria statistica delle classi e calcolo delle probabilità. *Pubbl del R Ist Super di Sci Econ e Commer di Firenze* **8:** 3–62.

Bosch M, Mayer CD, Cookson A, Donnison IS. 2011. Identification of genes involved in cell wall biogenesis in grasses by differential gene expression profiling of elongating and non-elongating maize internodes. *J Exp Bot* **62:** 3545–3561. doi:10.1093/jxb/err045

Botha FC, Black KG. 2000. Sucrose phosphate synthase and sucrose synthase activity during maturation of internodal tissue in sugarcane. *Aust J Plant Physiol* **27:** 81–85.

Cai M, Lin J, Li Z, Lin Z, Ma Y, Wang Y, Ming R. 2020. Allele specific expression of Dof genes responding to hormones and abiotic stresses in sugarcane. *PLoS One* **15:** e0227716. doi:10.1371/journal.pone.0227716

Castel SE, Levy-Moonshine A, Mohammadi P, Banks E, Lappalainen T. 2015. Tools and best practices for data processing in allelic expression analysis. *Genome Biol* **16:** 195. doi:10.1186/s13059-015-0762-6

Cheavegatti-Gianotto A, de Abreu HMC, Arruda P, Bespalhok Filho JC, Burnquist WL, Creste S, di Ciero L, Ferro JA, de O Figueira AV, de S Filgueiras T, et al. 2011. Sugarcane (*Saccharum* X *officinarum*): a reference study for the regulation of genetically modified cultivars in Brazil. *Trop Plant Biol* **4:** 62–89. doi:10.1007/s12042-011-9068-3

Chen ZJ, Ni Z. 2006. Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. *Bioessays* **28:** 240–252. doi:10.1002/bies.20374

Chen C, Li T, Zhu S, Liu Z, Shi Z, Zheng X, Chen R, Huang J, Shen Y, Luo S, et al. 2018. Characterization of imprinted genes in rice reveals conservation of regulation and imprinting with other plant species. *Plant Physiol* **177:** 1754–1771. doi:10.1104/pp.17.01621

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Ruden DM, Lu X. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; *iso*-2; *iso*-3. *Fly (Austin)* **6:** 80–92. doi:10.4161/fly.19695

Comai L. 2000. Genetic and epigenetic interactions in allopolyploid plants. *Plant Mol Biol* **43:** 387–399. doi:10.1023/A:1006480722854

Correr FH, Furtado A, Garcia AAF, Henry RJ, Margarido GRA. 2021. Allele expression biases in mixed-ploid species accessions. bioRxiv doi:10.1101/2021.08.26.457296

DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43:** 491–498. doi:10.1038/ng.806

DeYoung BJ, Innes RW. 2006. Plant NBS-LRR proteins in pathogen sensing and host defense. *Nat Immunol* **7:** 1243–1249. doi:10.1038/ni1410

D'Hont A, Glaszmann J-C. 2001. Sugarcane genome analysis with molecular markers, a first decade of research. *Proc Int Soc Sugar Cane Technol* **24:** 556–559.

D'Hont A, Grivet L, Feldmann P, Rao S, Berding N, Glaszmann JC. 1996. Characterisation of the double genome structure of modern sugarcane cultivars (*Saccharum* spp.) by molecular cytogenetics. *Mol Gen Genet* **250:** 405–413. doi:10.1007/BF02174028

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

Dillon SL, Shapter FM, Henry RJ, Cordeiro G, Izquierdo L, Lee LS. 2007. Domestication to crop improvement: genetic resources for *Sorghum* and *Saccharum* (Andropogoneae). *Ann Bot* **100:** 975–989. doi:10.1093/aob/mcm192

Diniz AL, Ferreira SS, Ten-Caten F, Margarido GRA, dos Santos JM, Barbosa GVDS, Carneiro MS, Souza GM. 2019. Genomic resources for energy cane breeding in the post genomics era. *Comput Struct Biotechnol J* **17:** 1404–1414. doi:10.1016/j.csbj.2019.10.006

Feldman M, Levy AA. 2009. Genome evolution in allopolyploid wheat—a revolutionary reprogramming followed by gradual changes. *J Genet Genomics* **36:** 511–518. doi:10.1016/S1673-8527(08)60142-3

Furtado A. 2014. DNA extraction from vegetative tissue for next-generation sequencing. In *Cereal genomics: methods and protocols* (ed. Henry RJ, Furtado A), pp. 1–5. Humana Press, Totowa, NJ.

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

Gaur U, Li K, Mei S, Liu G. 2013. Research progress in allele-specific expression and its regulatory mechanisms. *J Appl Genet* **54:** 271–283. doi:10.1007/s13353-013-0148-y

The Gene Ontology Consortium, Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, et al. 2000. Gene Ontology: tool for the unification of biology. *Nat Genet* **25:** 25–29. doi:10.1038/75556

Gerard D, Ferrão LFV, Garcia AAF, Stephens M. 2018. Genotyping polyploids from messy sequencing data. *Genetics* **210:** 789–807. doi:10.1534/genetics.118.301468

Goldemberg J. 2007. Ethanol for a sustainable energy future. *Science* **315:** 808–810. doi:10.1126/science.1137013

Grivet L, Arruda P. 2002. Sugarcane genomics: depicting the complex genome of an important tropical crop. *Curr Opin Plant Biol* **5:** 122–127. doi:10.1016/S1369-5266(02)00234-0

Held BM, Wang H, John I, Wurtele ES, Colbert JT. 1993. An mRNA putatively coding for an *O*-methyltransferase accumulates preferentially in maize roots and is located predominantly in the region of the endodermis. *Plant Physiol* **102:** 1001–1008. doi:10.1104/pp.102.3.1001

Henry RJ. 2010. Evaluation of plant biomass resources available for replacement of fossil oil. *Plant Biotechnol J* **8:** 288–293. doi:10.1111/j.1467-7652.2009.00482.x

Kandel R, Yang X, Song J, Wang J. 2018. Potentials, challenges, and genetic and genomic resources for sugarcane biomass improvement. *Front Plant Sci* **9:** 151. doi:10.3389/fpls.2018.00151

Kanehisa M, Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **28:** 27–30. doi:10.1093/nar/28.1.27

Kim C, Wang X, Lee T-H, Jakob K, Lee G-J, Paterson AH. 2014. Comparative analysis of *Miscanthus* and *Saccharum* reveals a shared whole-genome duplication but different evolutionary fates. *Plant Cell* **26:** 2420–2429. doi:10.1105/tpc.114.125583

Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12:** 357–360. doi:10.1038/nmeth.3317

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9:** 357–359. doi:10.1038/nmeth.1923

Le Cunff L, Garsmeur O, Raboin LM, Pauquet J, Telismart H, Selvi A, Grivet L, Philippe R, Begum D, Deu M, et al. 2008. Diploid/polyploid syntenic shuttle mapping and haplotype-specific chromosome walking toward a rust resistance gene (*Bru1*) in highly polyploid sugarcane ($2n \sim 12x \sim 115$). *Genetics* **180:** 649–660. doi:10.1534/genetics.108.091355

Margarido GRA, Heckerman D. 2015. ConPADE: genome assembly ploidy estimation from next-generation sequencing data. *PLoS Comput Biol* **11:** e1004229. doi:10.1371/journal.pcbi.1004229

McCormick RF, Truong SK, Sreedasyam A, Jenkins J, Shu S, Sims D, Kennedy M, Amirebrahimi M, Weers BD, McKinley B, et al. 2018. The *Sorghum bicolor* reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *Plant J* **93:** 338–354. doi:10.1111/tpj.13781

Mootha VK, Lindgren CM, Eriksson K-F, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstråle M, Laurila E, et al. 2003. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* **34:** 267–273. doi:10.1038/ng1180

OECD/FAO. 2019. Sugar. In *OECD-FAO agricultural outlook 2019–2028,* pp. 154–165. OECD Publishing, Paris/Food and Agriculture Organization of the United Nations, Rome.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, et al. 2009. The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457:** 551–556. doi:10.1038/nature07723

Perlo V, Botha FC, Furtado A, Hodgson-Kratky K, Henry RJ. 2020. Metabolic changes in the developing sugarcane culm associated with high yield and early high sugar content. *Plant Direct* **4:** e00276. doi:10.1002/pld3.276

Pham GM, Newton L, Wiegert-Rininger K, Vaillancourt B, Douches DS, Buell CR. 2017. Extensive genome heterogeneity leads to preferential allele expression and copy number-dependent expression in cultivated potato. *Plant J* **92:** 624–637. doi:10.1111/tpj.13706

Piperidis N, D'Hont A. 2020. Sugarcane genome architecture decrypted with chromosome-specific oligo probes. *Plant J* **103:** 2039–2051. doi:10.1111/tpj.14881

Piperidis G, Piperidis N, D'Hont A. 2010. Molecular cytogenetic investigation of chromosome composition and transmission in sugarcane. *Mol Genet Genomics* **284:** 65–73. doi:10.1007/s00438-010-0546-3

Pompidor N, Charron C, Hervouet C, Bocs S, Droc G, Rivallan R, Manez A, Mitros T, Swaminathan K, Glaszmann J-C, et al. 2021. Three founding ancestral genomes involved in the origin of sugarcane. *Ann Bot* **127:** 827–840. doi:10.1093/aob/mcab008

Powell JJ, Fitzgerald TL, Stiller J, Berkman PJ, Gardiner DM, Manners JM, Henry RJ, Kazan K. 2017. The defence-associated transcriptome of hexaploid wheat displays homoeolog expression and induction bias. *Plant Biotechnol J* **15:** 533–543. doi:10.1111/pbi.12651

R Core Team. 2021. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Renny-Byfield S, Wendel JF. 2014. Doubling down on genomes: polyploidy and crop plants. *Am J Bot* **101:** 1711–1725. doi:10.3732/ajb.1400119

Rivas MA, Pirinen M, Conrad DF, Lek M, Tsang EK, Karczewski KJ, Maller JB, Kukurba KR, DeLuca DS, Fromer M, et al. 2015. Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science* **348:** 666–669. doi:10.1126/science.1261877

Rott PC, Girard J-C, Comstock JC. 2013. Impact of pathogen genetics on breeding for resistance to sugarcane diseases. In *Proceedings of the ISSCT (International Soc Sugar Cane Technol) Congress* 28, São Paulo, Brazil.

Scheben A, Batley J, Edwards D. 2017. Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. *Plant Biotechnol J* **15:** 149–161. doi:10.1111/pbi.12645

Scott-Craig JS, Casida JE, Poduje L, Walton JD. 1998. Herbicide safener-binding protein of maize: purification, cloning, and expression of an encoding cDNA. *Plant Physiol* **116:** 1083–1089. doi:10.1104/pp.116.3.1083

Sforça DA, Vautrin S, Cardoso-Silva CB, Mancini MC, Romero-da Cruz MV, Pereira GDS, Conte M, Bellec A, Dahmer N, Fourment J, et al. 2019. Gene duplication in the sugarcane genome: a case study of allele interactions and evolutionary patterns in two genic regions. *Front Plant Sci* **10:** 553. doi:10.3389/fpls.2019.00553

Shao L, Xing F, Xu C, Zhang Q, Che J, Wang X, Song J, Li X, Xiao J, Chen L-L, et al. 2019. Patterns of genome-wide allele-specific expression in hybrid rice and the implications on the genetic basis of heterosis. *Proc Natl Acad Sci* **116:** 5653–5658. doi:10.1073/pnas.1820513116

Souza GM, Ballester MVR, de Brito Cruz CH, Chum H, Dale B, Dale VH, Fernandes ECM, Foust T, Karp A, Lynd L, et al. 2017. The role of bioenergy in a climate-changing world. *Environ Dev* **23:** 57–64. doi:10.1016/j.envdev.2017.02.008

Souza GM, Van Sluys MA, Lembke CG, Lee H, Margarido GRA, Hotta CT, Gaiarsa JW, Diniz AL, Oliveira MDM, Ferreira SDS, et al. 2019. Assembly of the 373k gene space of the polyploid sugarcane genome reveals reservoirs of functional diversity in the world's leading biomass crop. *Gigascience* **8:** giz129. doi:10.1093/gigascience/giz129

Springer NM, Stupar RM. 2007. Allele-specific expression patterns reveal biases and embryo-specific parent-of-origin effects in hybrid maize. *Plant Cell* **19:** 2391–2402. doi:10.1105/tpc.107.052258

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci* **102:** 15545–15550. doi:10.1073/pnas.0506580102

Tameling WIL, Vossen JH, Albrecht M, Lengauer T, Berden JA, Haring MA, Cornelissen BJC, Takken FLW. 2006. Mutations in the NB-ARC domain of I-2 that impair ATP hydrolysis cause autoactivation. *Plant Physiol* **140:** 1233–1245. doi:10.1104/pp.105.073510

Thimm O, Bläsing O, Gibon Y, Nagel A, Meyer S, Krüger P, Selbig J, Müller LA, Rhee SY, Stitt M. 2004. MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J* **37:** 914–939. doi:10.1111/j.1365-313X.2004.02016.x

Thirugnanasambandam PP, Hoang NV, Henry RJ. 2018. The challenge of analyzing the sugarcane genome. *Front Plant Sci* **9:** 616. doi:10.3389/fpls.2018.00616

Vilela MDM, Del Bem LE, Van Sluys M-A, De Setta N, Kitajima JP, Cruz GMQ, Sforça DA, de Souza AP, Ferreira PCG, Grativol C, et al. 2017. Analysis of three sugarcane homo/homeologous regions suggests independent polyploidization events of *Saccharum officinarum* and *Saccharum spontaneum*. *Genome Biol Evol* **9:** 266–278. doi:10.1093/gbe/evw293

Vogt T. 2010. Phenylpropanoid biosynthesis. *Mol Plant* **3:** 2–20. doi:10.1093/mp/ssp106

Wang J, Roe B, Macmil S, Yu Q, Murray JE, Tang H, Chen C, Najar F, Wiley G, Bowers J, et al. 2010. Microcollinearity between autopolyploid sugarcane and diploid sorghum genomes. *BMC Genomics* **11:** 261. doi:10.1186/1471-2164-11-261

Wang J, Nayak S, Koch K, Ming R. 2013. Carbon partitioning in sugarcane (*Saccharum* species). *Front Plant Sci* **4:** 201. doi:10.3389/fpls.2013.00201

Waters AJ, Bilinski P, Eichten SR, Vaughn MW, Ross-Ibarra J, Gehring M, Springer NM. 2013. Comprehensive analysis of imprinted genes in maize reveals allelic variation for imprinting and limited conservation with other species. *Proc Natl Acad Sci* **110:** 19639–19644. doi:10.1073/pnas.1309182110

Whittaker A, Botha FC. 1997. Carbon partitioning during sucrose accumulation in sugarcane internodal tissue. *Plant Physiol* **115:** 1651–1659. doi:10.1104/pp.115.4.1651

Wilson EB. 1927. Probable inference, the law of succession, and statistical inference. *J Am Stat Assoc* **22:** 209–212. doi:10.1080/01621459.1927.10502953

Yant L, Bomblies K. 2015. Genome management and mismanagement: cell-level opportunities and challenges of whole-genome duplication. *Genes Dev* **29:** 2405–2419. doi:10.1101/gad.271072.115

Zhang X, Borevitz JO. 2009. Global analysis of allele-specific expression in *Arabidopsis thaliana*. *Genetics* **182:** 943–954. doi:10.1534/genetics.109.103499

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

Received June 17, 2021; accepted in revised form December 19, 2021.