Genetic variation in a complex polyploid: unveiling the dynamic allelic features of sugarcane

Danilo Augusto Sforça,[a] Sonia Vautrin,[b] Claudio Benicio Cardoso-Silva,[a] Melina Cristina Mancini,[a] María Victoria Romero da Cruz,[a] Guilherme da Silva Pereira,[c] Mônica Conte,[a] Arnaud Bellec,[b] Nair Dahmer,[a] Joelle Fourment,[b] Nathalie Rodde,[b] Marie-Anne Van Sluys,[d] Renato Vicentini,[a] Antônio Augusto Franco Garcia,[c] Eliana Regina Forni-Martins,[a] Monalisa Sampaio,[e] Hermann Hoffmann,[e] Luciana Rossini Pinto,[f] Marcos Guimarães de Andrade Landell,[f] Michel Vincentz,[a] Helene Berges,[b] Anete Pereira Souza[a]

[a]Universidade Estadual de Campinas (UNICAMP), Campinas, SP, Brazil
[b]Centre National de Ressources Genomiques Vegetales (CNRGV), Institut National de la Recherche Agronomique (INRA), Castanet Tolosan, France
[c]Escola Superior de Agricultura Luiz de Queiroz (ESALQ), USP, Piracicaba, SP, Brazil
[d]Universidade de São Paulo, USP, São Paulo, SP, Brazil
[e]Universidade Federal de São Carlos (UFSCAR), Araras, SP, Brazil
[e]Centro de Cana, Instituto Agronômico de Campinas, São Paulo, SP, Brazil

Corresponding author: anete@unicamp.br

Short title: Unveiling the allelic dynamics of sugarcane

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is Anete Pereira de Souza (anete@unicamp.br).

ABSTRACT

Sugarcane (*Saccharum spp*.) is highly polyploid and aneuploid. Modern cultivars are derived from hybridization between *S. officinarum* (x = 10, 2n = 80) and *S. spontaneum* (x = 8, 2n = 40–128). The hypothetical *HP600* and centromere protein C (*CENP-C*) genes from sugarcane were used to elucidate the allelic expression and genomic and genetic behavior of this complex polyploid. The genomically side-by-side genes *HP600* and *CENP-C* were found in two different homeologous chromosome groups with ploidies of eight and ten. The first region (Region01) was a *Sorghum bicolor* ortholog with all haplotypes of *HP600* and *CENP-C* expressed, but *HP600* exhibited an unbalanced haplotype expression. The second region (Region02) was a scrambled sugarcane sequence formed from different noncollinear genes containing duplications of *HP600* and *CENP-C* (paralogs). This duplication occurred before the *Saccharum* genus formation and after the separation of sorghum and sugarcane, resulting in a nonexpressed *HP600* pseudogene and a recombined fusion version of *CENP-C* and orthologous gene Sobic.003G299500 with at least two chimerical gene haplotypes expressed. The genetic map construction supported the difficulty of mapping markers located in duplicated regions of complex polyploid genomes. We thus present an integrated approach to elucidate the homeolog dynamics of polyploid genomes.

INTRODUCTION

The *Saccharum* species are C4 grass and present a high level of ploidy. *S. officinarum* is octaploid (2n = 80), with x = 10 chromosomes, while *S. spontaneum* has x = 8 but presents great variation in the number of chromosomes, with main cytotypes of 2n = 62, 80, 96, 112 or 128. The modern sugarcane cultivars originated from hybridization between these two species and are considered allopolyploid hybrids (Daniels and Roach, 1987; Paterson et al., 2013). The development of these cultivars involved the process of 'nobilization' of the hybrid, with successive backcrosses using *S. officinarum* as the recurrent parent (Bremer, 1961). The resulting hybrids are highly polyploid and aneuploid (D'Hont et al., 1998; Irvine, 1999; Grivet and Arruda, 2002) and have an estimated whole genome size of 10 Gb (D'Hont and Glaszmann, 2001). An in situ hybridization study has shown that the genomes of the commercial hybrids consist of 10-20% chromosomes from *S. spontaneum* and 5-17% recombinant chromosomes between the two species, while the remaining majority of the genome consists of chromosomes from *S. officinarum* (Piperidis and D'Hont, 2001; D'Hont, 2005).

Molecular evidence suggests that polyploid genomes can present dynamic changes in DNA sequence and gene expression, probably in response to genomic shock (genomic remodeling due to the activation of previously deleted heterochromatic elements), and this phenomenon is implicated in epigenetic changes in homologous genes due to intergenomic interactions (McClintock, 1984). The evolutionary success of polyploid species is related to their ability to present greater phenotypic novelty than is observed in their diploid or even absent in parents (Ramsey and Schemske, 2002). Among other factors, this increase in the capacity for phenotypic variation capacity may be caused by regulation of the allelic dosage (Birchler et al., 2005).

The Brazilian sugarcane variety SP80-3280 is derived from a cross between the varieties SP71-1088 × H57-5028 and is resistant to brown rust, caused by *Puccinia melanocephala* (Landel et al., 2005). SP80-3280 was chosen for transcriptome sequencing by SUCEST-FUN (Vettore et al., 2003) and RNAseq (Cardoso-Silva et al., 2014; Nishyiama et al, 2014; Matiello et al., 2015). Biparental

crossing of SP80-3280 has also been used to analyze rust resistance (Balsalobre et al., 2016), quantitative trait loci (QTL) mapping (Costa et al., 2016) and genotyping by sequencing (GBS) (Balsalobre et al., 2017). A Brazilian initiative (Souza et al., 2011) is producing a gene-space genome sequence from SP80-3280, and a draft sugarcane genome based on whole-genome shotgun sequencing was produced (Riaño-Pachón and Mattiello, 2017). In addition, QTL gene synteny from sorghum has been used to map corresponding BACs in SP80-3280 (Mancini et al., 2018).

Three bacterial artificial chromosome (BAC) libraries for different sugarcane varieties have been constructed. The oldest one is for the French variety R570 (Tomkins et al., 1999) and contains 103,296 clones with an average insert size of 130 kb, representing 1.2 total genome equivalents. A mix of four individuals derived from the self-fertilization of the elite cultivar R570 (pseudo F2) was reported by Le Cunff et al. (2008) and contains 110,592 clones with an average insert size of 130 kb, representing 1.4x coverage of the whole genome. In addition, a library of SP80-3280 published by Figueira et al. (2012) contains 36,864 clones with an average insert size of 125 kb, representing 0.4 total genome equivalent coverage.

Sugarcane and sorghum (*Sorghum bicolor*) share a high level of collinearity, gene structure and sequence conservation. de Setta et al. (2014) contributed to understanding the euchromatic regions from R570 and a few repetitive-rich regions, such as centromeric and ribosomal regions, other than defining a basic transposable element dataset. The genomic similarity between sugarcane and sorghum has been frequently used to characterize the sugarcane genome (Jannoo et al., 2007; Garsmeur et al., 2011; Vilela et al., 2017, Mancini et al., 2018), demonstrating the high synteny of sugarcane x sorghum and high gene structure retention among the different sugarcane homeologs. In addition, these works contribute to understanding the genomic and evolutionary relationships among important genes in sugarcane using BAC libraries.

The segregation of chromosomes during cell division is facilitated by the attachment of mitotic spindle microtubules to the kinetochore at the chromosomal

centromere. Only CenH3 (histone H3) and *CENP-C* (centromere protein C) have been shown to bind centromeric DNA (Talbert et al., 2004). The centromere is marked with the histone H3 variant CenH3 (*CENP-A* in human), and *CENP-C* forms part of the inner kinetochore. The *CENP-C* "central domain" makes close contact with the acidic patch of histones H2A/H2B, and the highly conserved "*CENP-C* motif" senses both the acidic patch and recognizes the hydrophobicity of the otherwise nonconserved CenH3 tail, supporting a conserved mechanism of centromere targeting by the kinetochore (Gopalakrishnan et al., 2009; Kato et al., 2013; Sandmann et al., 2017). Sandmann et al. (2017) reported that in *Arabidopsis thaliana,* KNL2, a protein with a *CENPC-k* motif, recognizes centromeric nucleosomes such as the *CENP-C* protein. The CENP-C gene genomic structure in sugarcane has not been detailed.

Genome organization and expression dynamics are poorly understood in complex polyploid organisms, such as sugarcane, mainly because reconstructing large and complex regions of the genome is a challenge. However, an intriguing question is how such a complex genome can function while handling different copy numbers of genes, different allelic dosages and different ploidies of its homo/homeolog groups. For that reason, we examined the genome, transcriptome, evolutionary pattern and genetic interactions/relationships of a *CENP-C*-containing region in a genomic region of the SP80-3280 sugarcane variety (a *Saccharum* hybrid). First, we defined the genome architecture and evolutionary relationships of two physically linked genes, *HP600* (unknown function) and *CENP-C*, in detail. Second, we used the sugarcane SP80-3280 transcriptome to investigate transcription and genomic interactions in each gene (*HP600* and *CENP-C*). Ultimately, we used SNP distribution from these genes to compare the genetic and physical maps.

RESULTS

BAC library construction

The BAC library from the sugarcane variety SP80-3280 resulted in 221.184 clones, arrayed in 576 384-well microtiter plates, with a mean insert size of 110 kb. This BAC library is approximately 2.4 genome equivalents (10 Gb) and 26 monoploid genome equivalents (930 Mb, Figueira et al., 2012). For the sugarcane variety IACSP93-3046, the library construction resulted in 165.888 clones arrayed in 432 384-well microtiter plates, with a mean insert size of 110 kb, which is approximately 1.8 genome equivalents and 19 monoploid genome equivalents.

BAC-end sequencing (BES) results in an overview of the genome and validates the clones obtained through library construction. The SP80-3280 BAC library yielded 650 (84.6%) good BES sequences, of which 319 sequences have repetitive elements, and 92 exhibited similarities with sorghum genes. Excluding hits for more than one gene (probably duplicated genes or family genes), 65 sequences could be mapped to the *S. bicolor* genome (Supplemental Figure 1). The BAC library for IACSP93-3046 yielded 723 (94%) good BES sequences, of which 368 sequences exhibited the presence of repetitive sequences, and 111 exhibited a similarity with some gene. Excluding genes with hits with more than one gene, 74 of the sequences could be mapped to the *S. bicolor* genome (Supplemental Figure 1).

BAC annotation

The gene HP600 was used as target gene and showed strong evidence of being a single-copy gene when the transcripts of HP600 of sorghum, rice and sugarcane was compared. Twenty-two BAC clones from the SP80-3280 library that had the target gene *HP600* (NCBI from MH463467 to MH463488) and a previously sequenced BAC (Mancini et al., 2018; NCBI Accession Number MF737011) were sequenced by Roche 454 sequencing (detailed assembly can be found in Supplemental Table 1). The BACs varied in size from 48 kb (Shy171E23) to 162 kb (Shy432H18), with a mean size of 109 kb. The BACs were compared, and BACs with at least 99% similarity were considered the same haplotype (Figure 1 and Figure 2), resulting in sixteen haplotypes. Indeed, the possibility of one homeolog

being more than 99% similar to another exists, but a real haplotype cannot be distinguished from an assembly mismatch.

Comparisons of the BAC sequences against the sugarcane SP80-3280 genome draft using BLASTN (Riaño-Pachón and Mattiello, 2017) resulted in matches within gene regions, but no genome contig covered a whole BAC, and the BAC transposable elements (TEs) matched with several genome contigs (Supplemental Figure 2). The matches with genes provide further support for our assembly process.

The BACs were first annotated regarding the TEs. The TEs accounted for 21% to 65% of the sequenced bases with a mean of 40% (Supplemental Table 1). Annotation of the TEs in the 22 BACs revealed 618 TEs (220 TEs were grouped in the same type) with sizes ranging from 97 bp to 18,194 bp.

Gene annotation (Supplemental Table 2 and Supplemental Table 3) resulted in three to nine genes per BAC with a mean of five genes per BAC (Supplemental Table 1). The gene Sobic.003G221500, which was used to screen the library, codes for a hypothetical protein called *HP600* in sugarcane that has been found to be expressed in sorghum and rice. A phylogenetic analysis using sorghum, rice and *Arabidopsis thaliana* transcripts revealed that this gene is probably a single-copy gene. The gene Sobic.003G221600 is a *CENP-C* ortholog in sugarcane (*S. officinarum*, haplotypes CENP-C1 and CENP-C2, described by Talbert et al., 2004). The *HP600* and *CENP-C* sugarcane genes, as in *S. bicolor* and *Oryza sativa*, were found to be side by side in the sugarcane haplotypes.

Relationship between Region01 and Region02

Annotation of *HP600* and *CENP-C* in the sixteen BAC haplotypes revealed two groups of BACs. One group had the expected exon/intron organization when compared with *S. bicolor HP600* (five exons in sorghum) and *CENP-C* (fourteen exons in sorghum). This region was further designated as Region01 (Supplemental Table 1 - 10 BACs and 7 haplotypes – Figure 1 - haplotype I to haplotype VII). The other group was found to have fewer exons than expected (when compared with *S. bicolor*) for both *HP600* and *CENP-C*, and it was designated Region02

(Supplemental Table 1 - 13 BACs and 9 haplotypes – Figure 1 - haplotype VIII to haplotype XVI).

A comparison of the BAC haplotypes from Region01 and Region02 revealed an 8-kb shared region. The 8-kb duplication spanned from the last three exons of *HP600* to the last seven exons of *CENP-C*. *HP600* and *CENP-C* were physically linked, but the orientation of the genes was opposite (Supplemental Figure 3, panel B). A phylogenetic tree was constructed to examine the relationships among this 8-kb region (Supplemental Figure 3, panel A). The orthologous region from *S. bicolor* was used as an outgroup, and the separation in the two groups (Region01 and Region02) suggests that the shared 8-kb sequence appeared as the consequence of a sugarcane-specific duplication.

Region01 exhibited high gene collinearity with *S. bicolor*. However, in the BAC haplotype VII, a change in gene order involving the sorghum orthologs Sobic.003G221800 and Sobic.003G221400 was observed (Figure 1, dotted line). We were unable to determine whether this alteration resulted from a duplication or a translocation since we do not have a single haplotype that covers the entire region. Sobic.003G221800 is missing in this position from haplotypes I, II and VI.

Region01 and Region02, except for the genes *HP600* and *CENP-C*, contain different sorghum orthologous genes (Figure 1). Region02 was found to be noncollinear with *S. bicolor* (Figure 1 and Figure 2), which reinforces the notion of a specific duplication in sugarcane. Region02 appeared as a mosaic formed by different sorghum orthologous genes distributed in different chromosomes and arose by duplication after the separation of sorghum and sugarcane.

In Region02, the Sobic.008G134300 orthologous gene was found only in haplotype VIII, and the Sobic.008G134700 ortholog was found in a different position in haplotype IX (Figure 1, dotted line in Region02 and Figure 2). The phylogenetic analysis of Sobic.008G134700 and sugarcane orthologs demonstrated that sugarcane haplotype IX are more closely related to sorghum than to other sugarcane homeologs (Supplemental Figure 4). In addition, the orientation of transcription of the Sobic.008G134700 ortholog in haplotype IX is opposite that of the other sugarcane haplotypes (Figure 1 and Figure 2). This

finding suggests that this gene could be duplicated (paralogs) or translocated (orthologs) in haplotypes X, XIV, XV and XVI. No *S. bicolor* orthologous region that originated from Region02 could be determined, since it contained genes from multiple sorghum chromosomes.

Twenty long terminal repeat (LTR) retrotransposons were located in the two regions, but no LTR retrotransposons were shared among the haplotypes from Region01 and Region02, suggesting that all LTR retrotransposon insertions occurred after the duplication. In addition, ancient LTR retrotransposons could be present, but the sequences among the sugarcane haplotypes are so divergent that they could not be identified. The oldest LTR retrotransposon insertions were dated from 2.3 Mya (from haplotype VIII from Region02, a DNA/MuDR transposon, similar to MUDR1N_SB), which means that there is evidence that this duplication is at least 2.3 Mya old. Four LTR retrotransposons similar to RLG_scAle_1_1-LTR had identical sequences (Region01: Sh083P14_TE0360 – haplotype III and Sh040F02_TE0180 – haplotype XI; Region02: Sh285K15_TE0060 – haplotype XII and Sh452C23_TE0090 – haplotype XIII), which indicates a very recent insertion into the duplication from both regions.

To estimate the genomic diversity in sugarcane haplotypes from both regions (analyzed together and separately), the shared 8-kb region (duplication) was used (Supplemental Table 4), and the SNPs were identified. The diversity in the *HP600* and *CENP-C* genes was analyzed, and one SNP was observed every 43 bases (Region02) and 70 bases (Region01). We searched for SNPs that could distinguish each region (Supplemental Table 5) in the *HP600* and *CENP-C* genes, and one SNP was found for every 56 bases (20 SNPs in total). In addition, small (3-10 bases) and large (30 – 200 bases) insertions were found. These results revealed a high level of diversity in sugarcane, i.e., a high number of SNPs in each region, which could be used to generate molecular markers and to improve genetic maps. In addition, the diversity rate of both regions together could be used as an indicator of a duplicated gene, i.e., a rate < 20 (Supplemental Table 4).

***HP600* and *CENP-C* haplotypes and phylogenetics**

Gene haplotypes, i.e., genes with the same coding sequences (CDSs), from *HP600* and *CENP-C* that have the same coding sequence (i.e., exons) in different BAC haplotypes were considered the same gene haplotype. In Region01, four haplotypes of *HP600* were identified. In sorghum, the size of *HP600* is 187 amino acids (561 base pairs). *HP600* has two different sizes in sugarcane haplotypes of 188 amino acids (564 base pairs (haplotype I/II/VI, haplotype IV/V and haplotype VII) and 120 amino acids (360 base pairs – haplotype III). *HP600* haplotype III has a base deletion at position 77, causing a frameshift that results in a premature stop codon.

In Region02, *HP600* exhibited six haplotypes: haplotype VIII, haplotype IX, haplotype X/XI/XII/XIII/XIV, haplotype XV, and haplotype XVI. *HP600* Haplotype IX carried an insertion of eight bases in the last exon that caused a frameshift.

In *S. bicolor*, *CENP-C* is formed by 14 exons (Talbert et al., 2004) encoding 694 amino acids (2082 base pairs). In sugarcane, the haplotypes from Region01 had 14 exons that give rise to a protein of 708 or 709 amino acids (2124 or 2127 bases). Talbert et al. (2004) described two haplotypes in sugarcane EST clones (Vettore et al., 2003), *CENP-C1* and *CENP-C*2, which correspond to the haplotypes I/II and haplotypes IV/V, respectively. In addition to CENP-C1 and CENP-C2, three other *CENP-C* haplotypes were observed: haplotype III, haplotype VI, and haplotype VIII.

In Region02, the sugarcane duplication of *CENP-C* consisted of the last seven exons (exons eight to fourteen from *CENP-C* in Region01), and six haplotypes were found: haplotype VIII, haplotype IX, haplotypes XI/XII/XIII, haplotype XIV, haplotype XV, and haplotype XVI. The haplotype X BAC sequence finished before the CENP-C gene (Figure 1).

To reconstruct a phylogenetic tree for *HP600* and *CENP-C* from both regions, the orthologs from *O. sativa* and *Zea mays* were searched. The rice *HP600* and *CENP-C* orthologs, LOC_Os01g43060 and LOC_Os01g43050, were recovered, respectively. Maize has gone through tetraploidization since its divergence from sorghum approximately 12 million years ago (Woodhouse et al., 2010). The maize *HP600* ortholog search returned three possible genes with high

similarity: GRMZM2G114380 (chromosome 03), GRMZM2G018417 (chromosome 01) and GRMZM2G056377 (chromosome 01). The *CENP-C* maize ortholog search returned three possible genes with high similarity: GRMZM2G114315 (chromosome 03), GRMZM2G134183 (chromosome 03), and GRMZM2G369014 (chromosome 01).

Given the gene organization among the BACs, sorghum and rice revealed that *HP600* and *CENP-C* were side by side, and the expected orthologs from maize could be GRMZM2G114380 (*HP600*) and GRMZM2G114315 (*CENP-C*) because only these two genes are physically side by side. The other maize orthologs were probably maize paralogs that resulted from specific duplications of the *Z. mays* genome.

Two phylogenetic trees were constructed (Supplemental Figure 5), one for *HP600* (Supplemental Figure 5, panel A) and the other for *CENP-C* (Supplemental Figure 5 panel B), using sugarcane *HP600* and *CENP-C* haplotypes from both regions. The results demonstrated that the haplotypes from Region01 and Region02 are more similar to themselves than they are to those of sorghum or rice. Thus, the results also suggest that Region02 contains paralogous genes from Region01.

The divergence times among sugarcane *HP600* haplotypes and sorghum ranged from 1.5 Mya to 4.5 Mya. For *CENP-C*, the haplotype divergence time rates were 0.3-0.7 Mya, and the comparison with sorghum indicated 4.2-4.5 Mya for the highest values. The estimated sugarcane x sorghum divergence time was 5 Mya (Ming et al., 1998) to 8-9 Mya (Jannoo et al., 2007).

Chromosome number determination and BAC-FISH

The determination of the range of *CENP-C* and *HP600* loci that are present in the sugarcane genome was performed using in situ hybridization. First, the number of chromosomes in sugarcane variety SP80-3280 was defined, but the number of clear and well-spread metaphases for the variety SP80-3280 was less than 10 (Supplemental Table 6). We expanded the analysis to four more varieties of sugarcane (SP81-3250, RB835486, IACSP95-3018 and IACSP93-3046) to

improve the conclusions (Supplemental Figure 6 – Panel A, B, C, D and E – and Supplemental Table 6). The most abundant number of chromosomes was 2n = 112 (range: 2n = 98 to 2n = 118 chromosomes). The chromosome number of the *Saccharum* hybrid cultivar SP80-3280 was found to be 2n = 112 (range: 2n = 108 to 2n = 118 chromosomes - Supplemental Table 6). Vieira et al. (2018) found 2n = 112 chromosomes for IACSP93-3046 variety, corroborating with our data. The 2n = 112 chromosome number should indicate convergence in the number of chromosomes in the *Saccharum* hybrid cultivar.

As second step, we used two varieties with the best chromosome spreads, i.e., IACSP93-3046 and IACSP95-3018, for the CMA/DAPI banding patterns (Supplemental Figure 6 – Panel F, G, H and I). The variety IACSP93-3046 exhibited at least six terminal $CMA^{+}/DAPI^{-}$ bands, one chromosome with $CMA^{+}/DAPI^{o}$ and two chromosomes with adjacent intercalations of CMA^{+} and $DAPI^{+}$ in the same chromosome (Supplemental Figure 6 – Panel F and G). The variety IACSP95-3018 revealed seven terminal $CMA^{+}/DAPI^{-}$ bands, and at least two chromosomes exhibited adjacent CMA^{+} and $DAPI^{+}$; one was in the intercalary position, and the other was in the terminal position (Supplemental Figure 6 – Panel H and I). Additionally, the equal number of chromosomes and the divergent number of bands could indicate different chromosomic arrangements and/or different numbers of homeologs in each variety.

Finally, we performed BAC-FISH in the better metaphases of variety SP80-3280 using Shy064N22 (haplotype VII) from Region01; 64 metaphases with some signal of hybridization were obtained, and for the BAC-FISH of Shy048L15 (haplotype XI) from Region02, 69 were obtained. At least six metaphases for each region were used to determine the number of signals. For BAC Shy064N22 Region01, eight signals could be counted (Figure 3 – Panel A), and for BAC Shy048L15 in Region02, ten signals could be defined (Figure 3 – Panel B). These results detail the numbers of haplotypes in sugarcane for Region01 and Region 02. Moreover, the numbers of BAC haplotypes found in each region are appropriate considering the BAC-FISH results, suggesting a missing haplotype for each region.

The results observed so far suggest differences between the haplotypes, i.e., different TEs, insertions and even gene insertions/translocations. We used an identity of 99% to determinate the same BAC haplotype. The possibility of haplotypes with more than 99% similarity *in vivo* could not be tested with our data, since it is not possible distinguish a mismatch in sequence assembly from a real haplotype.

Expression of *HP600* and *CENP-C* haplotypes

The transcriptomes of the sugarcane variety SP80-3280 from the roots, shoots and stalks were mapped on *HP600* and *CENP-C* (NCBI SRR7274987), and the set of transcripts was used for the transcription analyses. All of the haplotypes of *HP600* from Region01 were covered by the reads, including haplotype III with a premature stop codon. None of the haplotypes of *HP600* from Region02 were found, suggesting *HP600* from Region02 is not expressed (Supplemental Figure 3). For the *CENP-C* gene from Region01, the haplotypes IV/V were found to be expressed. Furthermore, haplotypes I/II, haplotype VI and haplotype VII were fully covered by the reads, except for the first three SNPs, but these SNPs were described in the work of Talbert et al. (2004) under the haplotype CENP-C1, suggesting that the set of reads did not cover this region. For haplotype III, one SNP was not found, but nine exclusive SNPs of this haplotype were represented. Therefore, all haplotypes of *CENP-C* from Region01 were considered to be expressed.

The *CENP-C* haplotypes I/II, III and VI from Region01 have large retrotransposons in the introns (Figure 2 – black rectangles). Additionally, no evidence of substantial influence on expression could be found for this gene, which may indicate the silencing of these LTR retrotransposons, as discussed by Kim and Ziberman (2014).

The mapping of the transcript reads in the *CENP-C* haplotypes from Region02 revealed evidence of a chimerical gene (Figure 1, dotted rectangle and Figure 4). The chimeric gene was formed by the first five exons of the sugarcane orthologous gene of Sobic.003G299500 and the eighth to fourteenth exons of

384 *CENP-C* (Figure 4 – Panel C). RNAseq reads overlapped the region corresponding
385 to the union of the chimerical gene (position 1253 of the *CENP-C* haplotypes from
386 Region02 by 38 reads - Figure 4 – Panel F). This result provided robust evidence
387 for the formation of the chimerical gene and its expression.
388 The sugarcane gene orthologous to Sobic.003G299500 was represented by
389 BAC BAC267H24 (GenBank KF184671) from the sugarcane hybrid R570 as
390 published by de Setta et al. (2014) under the name “SHCRBa_267_H24_F_10”
391 (Figure 4 – Panel D). This finding indicated that the ancestral genes from sorghum
392 (orthologs) were retained in the sugarcane genome (Figure 4 – Panel B and D) and
393 that the chimerical gene was formed by the fusion of a partial duplication of *CENP-*
394 *C* and the sorghum ortholog gene Sobic.003G299500 (Figure 4 – Panel C).
395 Two chimerical *CENP-C* haplotypes from Region02 were fully mapped with
396 transcripts, i.e., haplotypes XI/XII/XIII and haplotype XIV. The chimerical *CENP-C*
397 haplotypes IX and XVI from Region02 were not fully mapped, but exclusive SNPs
398 from these haplotypes were recovered. The *CENP-C* haplotypes VIII and XV from
399 Region02 exhibited no exclusive SNPs in the transcriptome, and evidence for the
400 expression of these two haplotypes remains undefined.
401
402 **How locus number of homeologs influences expression**
403 We searched the SNPs in the BAC sequences and RNAseq reads (i.e., only in the
404 transcriptome of the sugarcane variety SP80-3280 from the roots, shoots and
405 stalks – NCBI SRR7274987) and compared the correspondences to the genes
406 *HP600* and *CENP-C*. For Region01 and Region02, we defined the ploidies as eight
407 and ten, respectively, based on the BAC-FISH data. The numbers of BAC
408 haplotypes recovered for Region01 and Region02 were seven and nine,
409 respectively, which indicated one missing BAC haplotype in each region.
410 The missing BAC haplotypes were determined by searching for SNPs
411 present only in the transcriptome. For the *HP600* haplotypes from Region01 (Table
412 1), six SNPs were found in the transcriptome and not in the BAC haplotypes,
413 including a (GAG)3 -> (GAG)2 deletion. For the *CENP-C* gene (Table 2), eight
414 SNPs were not represented in the genomic haplotypes. The presence of SNPs

only in transcript data corroborates the assumption that (at least) one genomic haplotype was missing in each region.

Using the results obtained from the RNAseq mapping of haplotypes, we also assumed that all haplotypes of the gene *HP600* were expressed in Region01 and that none were expressed in Region02. For *CENP-C*, all haplotypes from Region01 were considered expressed, and it was not possible to identify how many haplotypes were expressed in Region02 (chimerical gene); thus, we used only the nonduplicated portion of *CENP-C* (exons one to seven of the *CENP-C* gene).

We formed three assumptions using the previous results: (I) there is a missing haplotype for each region; (II) all haplotypes of HP600 from Region01 are expressed, and there is no expression of HP600 in Region02; and (III) CENP-C is expressed in both regions, but only in Region01 is it possible to infer that all haplotypes are expressed. Using these premises, we investigated the possibilities of one BAC haplotype being expressed at a higher or lower level than the others. Therefore, if the haplotypes contribute equally to expression, one SNP found in a BAC should have the same ratio (dosage) for the transcriptome data. Since we found evidence for a missing haplotype, two tests were performed: (I) we determined whether the missing BAC haplotype contributed to the dosage of more common SNPs, and (II) we determined whether the missing BAC haplotype contributed to the dosage of the variant SNP.

For the *HP600* haplotypes from Region01 (Table 1), only the SNPs 10 and 1 had significant p-values for hypotheses (I) and (II), respectively. These results suggested that the BAC haplotype ratio does not explain the transcriptome ratio. The transcript frequencies of SNPs 2, 3, and 4 (Table 1) were less than 0.125 (the minimum expected ratio for 1:7). To explain these frequencies, the dosage of the SNPs should be higher than a ploidy of eight (i.e., more than twelve), and our data do not support this possibility. The three variant SNPs came from *HP600* haplotype III. This finding could be evidence of some differential expression of the gene haplotypes, which could suggest that haplotype III is expressed at a lower level than the others for the HP600 gene.

For *CENP-C*, only the nonduplicated portions of the haplotypes from Region01 were used. At least one hypothesis was accepted for 17 (70%) SNPs (Table 2). The mean coverage of the SNPs was 64 reads per SNP, which could be considered a low coverage when an eight-ploidy region (Region01) is being inspected (Table 2). Moreover, the result suggests that the haplotypes from Region01 are equally expressed.

Genetic mapping

For the genetic mapping, 44 SNPs (Supplemental Table 7) were used to develop molecular markers (Figure 5), and they were used to construct a genetic map. The SuperMASSA (Serang et al., 2012) software calculates all possible ploidies for a locus and produces the most probable ploidy. Moreover, it is possible to define a fixed ploidy for a locus. The first option was used to call the dosages, which were ultimately used to construct the genetic map (Figure 6), and this map was compared with the fixed ploidy according to the BAC haplotype results (Figure 5). In fact, when using a Bayesian approach similar to that from the SuperMASSA, providing prior information about the ploidy level might improve the dosage estimates.

The markers from introns and exons were drawn along Region01 (Figure 5, “Location” column), including the duplicated region found in Region02. Among them, seven exhibited no variant presence in genotyping (Figure 5 – “×” marked), but five were detected in the RNAseq reads. Two markers (Figure 5 – “+” marked) were detected only for the “SuperMASSA best ploidy”, which was a ploidy higher than the “SuperMASSA expected ploidy”. In addition, two SNP loci were genotyped two times using different capture primer pairs (SugSNP_sh061/SugSNP_sh084 and SugSNP_sh067/SugSNP_sh092), and, as expected, at higher ploidy levels (> 12), the dosages of the loci diverge. These results could be explained by intrinsic problems in molecular biology that occur during the preparation of the samples, which affects the signal intensity of the Sequenom iPLEX MassARRAY® (Sequenom Inc., San Diego, CA, USA) data.

The SuperMASSA best ploidy was equal to the genomic ploidy for six SNPs (Figure 5), and the allelic dosage confirmed in four of them. When the ploidy for the loci was fixed (8 for Region01 and 18 for Region01 and Region02 SNPs), 24 SNPs had their dosage confirmed by SuperMASSA (Figure 5 – "SuperMASSA expected ploidy" columns). Notably, the estimation of the ploidy could also be a difficult task (Garcia et al., 2013), but when the ploidy used was found in BAC-FISH, the estimated dosage was in agreement with the dosage found in the BACs in 63% (28) of the SNPs (Figure 5).

For the genetic mapping, ten markers were used according to the SuperMASSA best ploidy results. First, attempts were made to add each marker to the existing linkage groups published by Balsalobre et al. (2017), but none of the markers could be linked to the groups. Then, the markers were tested for linkage among themselves. Two linkage groups could be created (Figure 6 – panel A) with 27.4 cM and 32.7 cM, respectively. The two linkage groups were too large; therefore, the markers SugSNP_sh065 and SugSNP_sh099 were excluded, since both markers were in the duplicated region (Figure 6 – panel B).

Then, attempts were made to add the remaining markers to the groups again, and the marker SugSNP_sh005 was inserted into Linkage group 02 (Figure 6 – panel C). The markers that were in the wrong positions according to the physical map (BACs) were also excluded, and the marker SugSNP_sh005 was excluded from Linkage group 01 but remained in Linkage group 02 (Figure 6 – panel C). Then, an attempt was made to form one linkage group with the remaining markers by forcing OneMap to place the markers in a single group. Again, the size of the group was too large (60.3 cM - Figure 6 – panel D). Therefore, the best representation of the region was two linkage groups, Linkage group 01 with 2.1 cM, and Linkage group 02 with 0 cM (Figure 6 – panel E).

DISCUSSION

The genetic, genomic and transcriptome interactions among sugarcane homeologs remain obscure. Several works have attempted to understand these interactions (Jannoo et al., 2007; Wang et al., 2010; Garsmeur et al., 2011; Casu et al., 2012;

Figueira et al., 2012; Garcia et al., 2013; de Setta et al., 2014; Xue et al., 2014; Sun and Joyce, 2017; Vilela et al., 2017; Mancini et al., 2018; and others). The high polyploidy in sugarcane cultivars make the detection of the ploidy of a locus a challenge (Casu et al., 2012; Garcia et al., 2013; Xue et al., 2014; Sun and Joyce, 2017; and other).

The chromosome number of the main Brazilian varieties was determined. The chromosome number determination showed an equal number of chromosomes (2n = 112, range: 2n = 98-118). The aneuploid nature of sugarcane hybrid cultivars (D'Hont, 2005; Piperidis et al., 2010) means that they contain different numbers of homeologous chromosomes. A number of differences in the CMA/DAPI patterns were found among the different varieties analyzed in this study, suggesting differences in chromosome content, i.e., differences in homeologous arrangement. Vieira et al. (2018) analyzed several sugarcane pollen cells showing metaphase chromosomes not lined up at the plate, lagging chromosomes and chromosomal bridges, tetrad cells with micronuclei and dyads with asynchronous behavior. They conclude that the presence of chromatin bridges indicates the indirect occurrence of chromosomal inversions.

For genetic and genomic studies, information about genomic organization is very important. Here, we report the construction of two new BAC libraries for two important Brazilian cultivars, SP80-3280 and SP93-3046, with a larger number of clones and higher sugarcane genome coverage than previously reported (Tomkins et al., 1999; Le Cunff et al., 2008; Figueira et al., 2012). The number of clones in a library is directly related to the number of homeologous regions that can be recovered.

The approach of mapping the BES in the *S. bicolor* genome, already performed for other libraries (Figueira et al., 2012; Kim et al., 2013; Visendi et al., 2016), revealed high synteny with the *S. bicolor* genome and a large number of TEs in the sugarcane genome. Kim et al. (2013) showed BES anchorage of approximately 6.4%, and Figueira et al. (2012) showed anchorage of approximately 22%. Our data showed 10% BES anchorage in the sorghum genome for both libraries constructed. These results are more similar to those of

Kim et al. (2013), since they used only BES >= 300 bp, and we used BES >= 100 bp.

The sugarcane genome has been reported to be composed of approximately 40% TEs (Figueira et al., 2012; Kim et al., 2013; de Setta et al., 2014). We also found that the average percentage of TEs was 40%, but this value has a very large variance among the haplotypes, with a minimum of 21% and a maximum of 65%. Figueira et al. (2012) and De Setta et al. (2014) also revealed an inflation of the sugarcane genome in comparison with the *S. bicolor* genome. De Setta et al. (2014) reported a very significant expansion that mainly occurred in the intergenic and intronic regions and was primarily because of the presence of TE, and we confirmed this report by comparing our data with data on the *S. bicolor* genome. Several studies have reported a very significant sugarcane genome expansion (Jannoo et al., 2007; Wang et al., 2010; Garsmeur et al., 2011; Figueira et al., 2012; Setta et al., 2014; Vilela et al., 2017, Mancini et al., 2018).

A hypothetical gene *HP600* and the *CENP-C* gene were used in this work as a case study. The function of *HP600* is unknown, but the ortholog of this gene is present in the genomes of rice (LOC_Os01g43060), maize (GRMZM2G114380) and sorghum (Sobic.003G221600). Sobic.003G221600 (ortholog of *HP600*) was also found in a QTL for Brix (sugar accumulation) that was mapped by Murray et al. (2008) and based on the sorghum consensus map reported by Mace and Jordan (2011). The *CENP-C* protein is a kinetochore component (Kato et al., 2013 and Sandmann et al., 2017) located next to *HP600*. Here, we have demonstrated the existence of paralogous genes for *HP600* and *CENP-C* that are localized in two different homeologous sugarcane chromosome groups. The BAC haplotypes could be separated into two sugarcane homeologous groups as follows: Region01 contained the collinearity region between sorghum and sugarcane *HP600* and *CENP-C* genes, and Region02 contained their paralogs.

Region01 is a recurrent case of high gene conservation and collinearity among sugarcane homeologs and the *S. bicolor* genome as reported by other authors (Jannoo et al., 2007; Garsmeur et al., 2011; de Setta et al., 2014; Vilela et al., 2017, Mancini et al., 2018). Region02 contains parts of the genes *HP600* and

CENP-C (paralogs). No synteny was found between the sugarcane Region02 and the sorghum genome. In Region02, a third partial gene (ortholog of Sobic.003G299500) was also found next to *CENP-C*, and transcriptome analysis revealed the fusion of the *CENP-C* partial exons with the partial exons of the sugarcane ortholog of Sobic.003G299500 to form a chimerical gene. Region02 is a scrambled sugarcane sequence that was possibly formed from different noncollinear ancestral sequences, but the exonic structure of the genes was retained. The phylogenetic analysis of gene haplotypes from *HP600* and *CENP-C* provided evidence that the multiple genes found in maize are the result of specific duplications in the maize taxa, as expected.

The nature of sugarcane hybrid cultivars, especially the processes of polyploidization (Daniels and Roach, 1987; Paterson et al., 2013) and nobilization (Bremer, 1961), are the main reason for the genomic variability, gene pseudogenization and increases in new genes (McClintock, 1984). It is possible that the structure found in Region02 could be a result of the polyploidization and domestication of sugarcane (Grivet and Arruda, 2002; Cuadrado et al., 2004; D'Hont, 2005; Piperidis et al., 2010). However, the presence of a set of sugarcane homeologs with very similar gene structures leads us to speculate that the occurrence of an ancestral event prior to polyploidization resulted in this structure.

Rearrangement events can also be caused by TEs, but they are normally caused by the formation of a pseudogene (Lai et al., 2004; Ilic et al., 2013). In the case of Region02, multiple events resulted in this region, but the number and types (TE, translocations) of events could not be determined with our data.

BAC-FISH hybridization was used to indicate the ploidy of each region. Eight signals were found for Region01 and 10 for Region02. These results are highly consistent with the BAC haplotype and suggest that at least one BAC haplotype is missing in each region. Casu et al. (2012), Xue et al. (2014) and Sun and Joyce (2017) reported different methods to quantify the copy number of endogenous gene, some of which resulted in odd copy numbers. Sun and Joyce (2017) reported that the low or odd numbers could be explained by the contribution of only the *S. spontaneum* or the *S. officinarum* genome. The presence of genes

599 without collinearity among the sugarcane homeologs could also explain the result
600 as verified for the orthologs Sobic.003G221800 and Sobic.008G134700.
601 The genomic SNP variation in sugarcane coding regions has been
602 estimated to be one SNP every 50 bp (Cordeiro et al. 2006) and one every 86 bp
603 (Cardoso-Silva et al. 2014). For the coding Region01 one SNP was found per 70
604 bases. Feltus et al. (2004) showed that different ratios of SNPs occur across the
605 genome. When we compared Region01 and Region02 one SNP was found per 12
606 bases using only the data for one sugarcane variety SP-803280. In light of the
607 possible existence of at least one more haplotype, this number could be
608 underestimated.
609 Once established, the polyploidy might now fuel evolution by virtue of its
610 polyploid-specific advantages. Vegetative propagation can lead to the retention of
611 genes. Meiosis may or may not play a role in either the origin or maintenance of a
612 polyploid lineage (Freeling, 2017). Vegetative propagation is widely used to
613 propagate sugarcane (even for nondomesticated sugarcanes) and could explain
614 the high variation in sugarcane and the high level of gene retention. However, it is
615 not the only factor, with sugarcane polyploidization and nobilization also
616 contributing to these effects.
617 The homologous gene expression in polyploids can be affected in different
618 ways, i.e., the homologous genes may retain their original function, one or more
619 copies may be silenced, or the genes may diversify in function or expression
620 (Ohno, 1970; Lynch and Force, 2000; Hegarty et al., 2006; Buggs et al., 2011). In
621 complex polyploids, the roles of ploidy and genome composition in possible
622 changes in gene expression are poorly understood (Shi et al., 2015). Even in
623 diploid organisms, this task is difficult, since different interactions can affect the
624 expression of a gene, and not all homologs are guaranteed to contribute to a
625 function (Birchler et al., 2005). The expression of the *HP600* and *CENP-C*
626 haplotypes in Region01 could be confirmed. In Region02, the haplotypes of *HP600*
627 were not found in the transcriptome dataset (Cardoso-Silva et al., 2014; Matiello et
628 al., 2015), but at least two haplotypes of the gene *CENP-C* were expressed.

The gene haplotypes of *HP600* from Region01 exhibited unbalanced expression; i.e., for some reason, some haplotypes were expressed at greater levels than others. These findings could mean that apart from the duplication, *HP600* might be expressed as a single-copy gene wherein only the haplotypes of the *HP600* in Region01 were expressed. In addition, we could not identify the mechanisms contributing to the unbalanced expression. Therefore, the transcripts from different tissues make us speculate that some kind of tissue-specific expression could be occurring.

Numerous molecular mechanisms are involved in the creation of new genes, such as exon shuffling, retrotransposons and gene duplications (reviewed in Long et al., 2003). Gene fusions allow the physical coupling of functions, and their occurrence in the genome increases with the genome size (Snel et al., 2000). Sandmann et al. (2017) describes the function of the protein KNL2, which uses *CENPC-k* motifs to bind DNA sequence independently and interacts with the centromeric transcripts. The *CENPC* motif is characteristic of *CENP-C*. The *CENPC* motif of the rat *CENP-C* protein can bind directly to a chimeric H3/cenH3 nucleosome *in vitro* suggesting that this motif binds to cenH3 nucleosomes *in vivo*. Consequently, it is involved directly in cell division (Kato et al., 2013 and Sandmann et al., 2017). The CENPC motifs described by Sandmann et al. (2017), were compared with those of *CENP-C* genes in *A. thaliana*, *O. sativa*, *Z. mays* and *S. bicolor* (Supplemental Figure 7). The *CENP-C* haplotypes from Region02 (chimerical gene) have the same *CENPC* motif as that in sorghum. The *CENP-C* haplotypes from Region01 have one variation in the second residue of the *CENPC* motif, which is a glycine in sorghum and a valine in *CENP-C* haplotypes from Region01. This result suggests that the chimerical gene retained the ancestral residue at this site, whereas a mutation occurred in *CENP-C* haplotypes from Region01. Therefore, the mutation could have occurred in sorghum and in the haplotypes from Region02, but this is unlikely. This result suggests that the *CENP-C* haplotypes from Region01 and Region02 are able to bind to cenH3 nucleosomes.

659 The presence of the motif in the *CENP-C* haplotypes from the Region02
660 proteins could indicate a chimerical protein with a similar function, specific to
661 sugarcane, that is involved in the organization of centromeric regions. Moreover,
662 the presence of large LTR retrotransposons in the intronic region of the *CENP-C*
663 haplotypes in Region01 does not influence the gene expression. Furthermore, two
664 studies (Saze, et al., 2013 and Wang, et al., 2013) identified the inactivation of the
665 same gene, IBM2/ANTI-SILENCING 1 (ASI1), which causes gene transcripts with
666 methylated intronic transposons that terminate within the elements. The complete
667 mechanisms that control LTR retrotransposon methylation require further
668 clarification (Kim and Ziberman, 2014).
669 These results have several implications for the integration of transcriptome
670 data and genomic data. First, for example, a gene such as *HP600* that
671 demonstrates single-copy behavior in the transcriptome data and the genomic
672 behavior of a duplicated gene can cause bias in genetic mapping. Second, a
673 chimerical gene such as the *CENP-C* haplotypes in Region02 can result in different
674 levels of expression of the duplicated and nonduplicated gene regions in the
675 transcriptome data. Using the *CENP-C* gene as an example, if the gene expression
676 quantification probe recovers the nonduplicated portion of the *CENP-C* gene, it will
677 give an expression level only for the *CENP-C* haplotypes in Region01. In contrast,
678 this probe quantifies the duplicated region of *CENP-C*, it will result in the
679 quantification of *CENP-C* from both Region01 and Region02 and thus overestimate
680 the expression of *CENP-C*. Consequently, analyses of the expression of the gene
681 for functional studies for evaluating the balance of gene expression will be biased.
682 The SNPs were also used to compare the ploidy found in BACs with the
683 results of SuperMASSA software (Garcia et al., 2013). SuperMASSA uses
684 segregation ratios to estimate ploidy, which is not the same as estimating ploidy by
685 chromosome counting because of the differences in estimation and the real ploidy
686 visualized. The SNPs present in a duplication were mapped in a linkage group and
687 demonstrated a high distance between the markers in the linkage map. The size of
688 a genetic map is a function of the recombination fraction, so two factors influence
689 the map size: (I) the number of recombinations found between two markers, and

(II) genotyping errors. In this case, the mapping of duplicated markers is an error and is interpreted by OneMap in a recombination fraction, which inflates the map.

Two markers classified with a ploidy of 10 and one with a ploidy of 8 formed the linkage group 02. The ploidy is not a determinant for the OneMap construction of a linkage group, but the recombination fraction is. In other words, recombination fractions can still be computed between single-dose markers classified in different ploidy levels. In fact, most of nulliplex, simplex and duplex individuals will have the same dosage call using either 8 or 10 as the ploidy level. In addition, the genome data (BACs and BAC-FISH) demonstrated that all markers had the same ploidy of eight and that the physical distances among the markers were too small and thus probably resulted in the lack of recombination. The fact that we obtained two linkage groups can be explained by the possibility that single-dose markers may be linked in repulsion, and insufficient information is available to assemble all of the markers in one group. Trying to calculate the recombination fraction between markers D1 and D2 (according to the nomenclature of Wu et al., 2002) in diploids presents the same obstacle.

For the first time, we observed the relationship between a linkage map and the physical map of a region in sugarcane. Indeed, it is a small region to observe whereas sugarcane has a large genome, and a linkage map is constructed based on the recombination fraction. However, it was possible to observe what happens in the genetic map when a duplicated locus was mapped.

The combination of divergent genomes within a hybrid can lead to immediate, profound and highly varied genome modifications, which could include chromosomal and molecular structural modifications (Shen et al., 2005; Doyle et al., 2008; Soltis and Soltis, 2009; Jiang et al., 2011) as well as epigenetic changes (Chen et al., 2010) and global transcriptomic changes (Hegarty et al., 2006; Buggs et al., 2011). The integration of the genetic, genomic and transcriptomic data was used to explain the interaction of the two regions in sugarcane. *HP600* is a hypothetical gene that is next to the *CENP-C* gene, a kinetochore component responsible for the initiation of nucleosomes. The sugarcane gene haplotypes of *HP600* in Region01 and the *CENP-C* haplotypes in Region01 were duplicated in

721 another group of homeologous chromosomes. The duplication of the *HP600*
722 haplotypes in Region01 resulted in a paralog pseudogene in the *HP600* haplotypes
723 in Region02. The duplication of *CENP-C* in the haplotypes of Region02 resulted in
724 fusion with another gene, which contained the first five exons of the orthologous
725 gene Sobic.003G299500 and exons eight to fourteen of *CENP-C*. The region
726 where this duplication was inserted (Region02) contained at least three more
727 genes that probably arose due to duplication, which indicates that multiple
728 duplication events occurred in this region.
729 The *HP600* and *CENP-C* duplication described in this work occurred
730 sometime after the separation of sugarcane and sorghum and before the
731 polyploidization of the *Saccharum* genus. This result is supported by the following
732 information: (I) the molecular clock time, (II) the genes are present in a
733 homeologous group of chromosomes; and (III) the *CENP-k* motifs of the *CENP-C*
734 haplotypes in Region02 are more similar to sorghum than to its paralog in
735 sugarcane. The formation of a chimeric gene and the scrambled Region02
736 exhibited a specific moment of formation before *Saccharum* polyploidization, which
737 makes us wonder which genomic event could be the result this formation. TEs
738 carrying this region could not be found. It is also possible that TEs were inserted in
739 this region, and the TE sequences were subsequently lost. An event that resulted
740 in some genome instability could also be a reason. Additionally, multiple events
741 could also have occurred.
742 The transcripts from SP80-3280 revealed full expression of the haplotypes
743 of *HP600* in Region01 (in an unbalanced manner) and the lack of expression of the
744 haplotypes *HP600* in Region02. The expression of the *HP600* haplotypes in
745 Region01 can be considered a single-copy gene, despite the presence of the
746 duplication. The *CENP-C* gene can be considered fully expressed, despite the low
747 coverage of the transcriptome data. The *CENP-C* haplotypes in Region02 have
748 four haplotypes that are considered expressed.
749 Currently, only markers with low dosages can be used to construct the
750 genetic map in sugarcane, which is a limitation of the mapping method in
751 polyploids. We attempted to map a duplicated region, which is a difficult task even

for diploid organisms. Again, it is important to observe that we used a sugarcane variety with asexual reproduction and performed the genetic mapping in artificial progeny. We have no idea how the progeny genome responded to the cross, since sugarcane is aneuploid. In addition, the premise that each individual of the progeny did not miss any chromosome in the cross (aneuploidy) and the ploidy of a locus is the same in both the parents and all individuals in the progeny could be biologically untruthful. The genetic mapping demonstrates that there are obstacles that still need to be overcome in the genetic mapping of complex polyploids.

This study sheds light on the influence of the genome arrangement for transcriptome and genetic map analyses in the sugarcane polyploid genome. The integration of genomic sequence arrangements, transcription profiles, cytogenetic organization and the genetic mapping approach might help to elucidate the behavior of gene expression, the genetic structure and successful sequence assembly of the sugarcane genome. Such integrated studies will undoubtedly help to enhance our understanding of complex polyploid genomes including the sugarcane genome.

Particular emphasis should be given to the determination studies of the ploidy level and of the duplication loci with the intention of better understanding complex polyploids. Such studies remain the most original and challenging in terms of understanding the sugarcane genome. From this perspective, this work presents an integrated approach to elucidate the allelic dynamics in polyploid genomes.

METHODS

Plant material

The sugarcane varieties SP80-3280 and SPIAC93-3046 were collected from germplasm at the active site located in the Agronomic Institute of Campinas (IAC) Sugarcane Center in Ribeirão Preto, São Paulo, Brazil. The leaves were collected on dry ice and stored at -80ºC until use.

BAC library construction and BAC-end analyses

The high-molecular-weight (HMW) DNA was prepared from the leaves as described by Peterson et al. (2000) with modifications as described by Gonthier et al. (2010). The HMW DNA was embedded in low melt agarose (Lonza InCert™ Agarose, Lonza Rockland Inc., Rockland, ME, USA) and partially digested with HindIII (New England Biolabs, Ipswich, MA, USA). Next, two size selection steps were performed by pulsed field gel electrophoresis (PFGE) with a Bio-Rad CHEF Mapper system (Bio-Rad Laboratories, Hercules, CA, USA), and the selected DNA was ligated into the pIndigoBAC-5 HindIII-Cloning Ready vector (Epicenter Biotechnologies, Madison, WI, USA) as described by Chalhoub et al. (2004). The insert size was verified by preparing DNA BACs with the NucleoSpin® 96 Plasmid Core Kit (MACHEREY-NAGEL GmbH & Co., Düren, Germany), according to the kit instructions, and the DNA was digested by the NotI (New England Biolabs, Ipswich, MA, USA) restriction enzyme and analyzed by PFGE.

For the BES, 384 random BAC DNAs from each library were prepared with the NucleoSpin® 96 Plasmid Core Kit (MACHEREY-NAGEL GmbH & Co., Düren, Germany), according to the kit instructions. The sequencing reactions were performed according to the manufacturer's instructions for the BigDye Terminator Kit (Applied Biosystems, Foster City, CA, USA). The primers used in the reactions were T7 Forward (5' TAATACGACTCACTATAGG 3') and M13 Reverse (5' AACAGCTATGACCATG 3'). The PCR conditions were 95°C for 1 min followed by 90 cycles of 20 sec at 95°C, 20 sec at 50ºC and 4 min at 60°C. The samples were loaded on a 3730xl DNA Analyzer (Applied Biosystems). Sequence trimming was conducted by processing the traces using the base-calling software PHRED (Ewing and Green, 1998; Ewing et al., 1998), and reads with phred score < 20 were trimmed. The sequences were compared by using BLASTN in the *S. bicolor* genome from Phytozome v10.1 (Goodstein et al., 2012). Only clones with forward and reverse sequence maps in the *S. bicolor* genome, with a maximum distance of 600 kb and with no hits with repetitive elements, were used to anchor the *S. bicolor* genome.

Target gene determination

Transcripts of *S. bicolor*, *Z. mays* and *O. sativa* were obtained from Phytozome v10.1 (Goodstein et al., 2012). Each transcript was queried against itself, and orthologous genes that resulted in redundant sequences were eliminated. From the remaining genes, the gene Sobic.003G221600 (*Sorghum bicolor* v3.1.1 – Phytozome v. 12) was chosen because it was inserted in a QTL for Brix from a study by Murray et al. (2008), which identified the QTL in the SB-03 genome (*S. bicolor* v3.1.1 – Phytozome v. 12). The sequence of the gene Sobic.003G221600 was then used as query in the SUCEST-FUN database (http://sucest-fun.org/ - Vettore et al., 2003) and the transcriptome obtained by Cardoso-Silva et al. (2014) to recover sugarcane transcripts. All the transcripts obtained were aligned (MAFFT; Katoh et al., 2002) to generate phylogenetic trees by the maximum likelihood method (PhylML 3.0; Guindon and Gascuel, 2003).

The sugarcane transcripts were split into exons according to their annotation in *S. bicolor*, *Z. mays* and *O. sativa,* and exon five was used to design the probe to screen both BAC libraries (F: 5' ATCTGCTTCTTGGTGTTGCTG 3', R: 5' GTCAGACACGATAGGTTTGTC 3'). DNA fragments were PCR-amplified from sugarcane SP80-3280 and SPIAC93-3046 genomic DNA with specific primers targeting the gene Sobic.003G221600. The PCR amplification conditions were 95°C for 8 min; 30 cycles of 20 sec denaturation at 95°C, 20 sec of annealing at 60ºC, and a 40 sec extension at 72°C; and a final 10 min extension at 72°C. The probes were sequenced before the screening of the BAC library.

BAC library screening

Both BAC libraries were spotted onto high-density colony filters with the workstation QPix2 XT (Molecular Devices, Sunnyvale, CA, USA). The BAC clones were spotted in duplicate using a 7x7 pattern onto 22 × 22 cm Immobilon-Ny+ filters (Molecular Devices). The whole BAC library from the SP80-3280 sugarcane variety was spotted in four sets of filters, each one with 55 296 clones in duplicate and the whole BAC library from SPIAC93-3046 sugarcane variety was spotted in three sets of filters each with 55,296 clones in duplicate. The filters were processed

as described by Roselli et al. (2017). Probe radiolabeling and filter hybridization were performed as described in Gonthier et al. (2010).

The SP80-3280 BAC library was used to construct a 3D pool. A total of 110,592 clones were pooled into 12 superpools following the protocol used by Paux et al. (2008). The positive BAC clones from the SP80-3280 library were isolated, and one isolated clone was validated by qPCR. The insert size of each BAC was estimated by using an electrophoretic profile of NotI-digested BAC DNA fragments and observed by PFGE (CHEF-DRIII system, Bio-Rad) in a 1% agarose gel in 0.5× TBE buffer under the conditions described in Paiva et al. (2011).

Sequencing and assembly

Twenty-two positive BAC clones were sequenced in pools of 10 clones. One microgram of each BAC clone was used to prepare individual tagged libraries with the GS FLX Titanium Rapid Library Preparation Kit (Roche, Branford, CT, USA). BAC inserts were sequenced by pyrosequencing with a Roche GS FLX Life Sciences instrument (Branford, CT, USA) in CNRGV, Toulouse, France.

The sequences were trimmed with PHRED, vector pIndigoBAC-5 sequences and the *Escherichia coli* str. K12 substr. DH10B complete genome was masked using CROSS_MATCH, and the sequences were assembled with PHRAP (Gordon et al., 1998; Gordon et al., 2001; Gordon, 2003) as described by de Setta et al. (2014). A BLASTN with the draft genome (Riaño-Pachón and Mattiello, 2017) was performed. A search was performed in the NCBI databank to find sugarcane BACs that could possibly have the target gene *HP600*.

Sequence analysis and gene annotation

All the BACs were aligned to verify the presence of redundant sequences of homeologs. BAC clones with more than 99% similarity were considered the same homeolog. BACs that represented the same homeologs were not combined. The BACs were annotated with the gene prediction programs EUGENE (Foissac et al., 2008) and Augustus (Keller et al., 2011). The BAC sequences were also searched for genes with BLASTN and BLASTX against the transcripts of SUCEST-FUN

875 database (http://sucest-fun.org/; Vettore et al., 2003), the CDS of *S. bicolor*, *Z.*
876 *mays* and *O. sativa* from Phytozome v12.0 and the transcripts published by
877 Cardoso-Silva et al. (2014). The BACs were also subjected to BLASTX against
878 Poaceae proteins. The candidate genes were manually annotated using *S. bicolor*,
879 *O. sativa* and *Z. mays* CDS. The sequences with more than 80% similarity and at
880 least 90% coverage were annotated as genes.
881 Repetitive content in the BAC clone sequences was identified with the web
882 program LTR_FINDER (Xu and Wang, 2007). Afterward, the BAC sequences were
883 tested by CENSOR (Kohany et al., 2006) against Poaceae.
884 The phylogenic trees were built by the Neighbor-Joining method (Saitou and
885 Nei, 1987) with nucleic distances calculated with the Jukes-Cantor model (Jukes
886 and Cantor, 1969) in MEGA 7 software (Kumar et al., 2016). The Kimura 2-
887 parameter (Kimura, 1980) was used as the distance mode.
888
889 **Duplication divergence time**
890 The gene contents of *HP600* and *CENP-C* in the duplication regions were
891 compared, and the distance “d” for coding regions was determined by Nei-Gojobori
892 with Jukes-Cantor, available in MEGA 7 software (Kumar et al., 2016). The
893 divergence times of the sequences shared by the duplicated regions in the BACs
894 were estimated by T = d/2r. The duplicated sequences were used to calculate the
895 pairwise distances (d), and “r” was replaced by the mutation rate of 6.5 x 10-9
896 mutations per site per year as proposed by Gaut et al. (1996). For the whole
897 duplication, the distance “d” for noncoding regions was determined with the Kimura
898 2-parameter model and the mutation rate of 1.3 x 10-8 mutations per site per year,
899 as described by Ma and Bennetzen (2004).
900 The insertion ages of the LTR retrotransposons were estimated based on
901 the accumulated number of substitutions between the two LTRs (d) (SanMiguel et
902 al., 1998), using the mutation rate of 1.3 x 10-8 mutations per site per year, as
903 described by Ma and Bennetzen (2004).
904
905 **Gene expression**

The transcriptomes of the sugarcane variety SP80-3280 from the roots, shoots and stalks were mapped on *HP600* and *CENP-C* (NCBI SRR7274987), and the set of transcripts was used for the transcription analyses. The reads from the sugarcane transcriptomes were mapped to the reference genes with the Bowtie2 software 2.2.5 (Langmead and Salzberg, 2012) with default parameters; low-quality reads and unmapped reads were filtered out (samtools -b -F 4); bam files were sorted (samtools sort); and only mapped reads to the genes were extracted from the bam files (samtools fastq) and recorded in a FASTQ format file. A haplotype was considered to be expressed only when the transcript reads were mapped with 100% similarity. SNPs not found in the dataset were searched in the SP80-3280 transcriptomes from Vettore et al. (2003), Talbert et al. (2004) and Cardoso-Silva et al. (2014) to verify the SNP presence in transcripts, but they were not used in the expression analysis.

To test whether the haplotypes had the same proportional ratio in the genome and transcriptome, the transcripts were mapped against one haplotype of the *HP600* haplotypes in Region01 and *CENP-C* with a 90% similarity, and the SNPs found in the transcripts were identified and the coverage and raw variant reads count used to verify the presence of SNPs not found in BACs. An SNP was considered present in the transcripts if it was represented by at least six transcriptome reads (Kim et al., 2016).

We assumed that one haplotype of each region was missing and tested two genomic frequencies for comparison with the transcriptome sequences: (1) the missing haplotype had a higher frequency of the SNP, and (2) the missing haplotype had a lower frequency of the SNP. When the SNP was not found in the genomic data, we assumed that only the missing haplotype contained the variant SNP.

The frequency of the genomic data was used to test the transcriptome data with R Studio (Rstudio team, 2015) and the exact binomial test (*binom.test* - Clopper and Pearson, 1934, Conover, 1971 and Hollander and Wolfe, 1973). A p-value >= 0.05 is equivalent to a 95% confidence interval for considering the genomic ratio equal to the transcriptome ratio.

937

938 **Chromosome number determination and BAC-FISH**

939 The chromosome number determination was performed as described by Guerra
940 (1983) with root tips 0.5–1.5 cm in length, treated with 5 N HCl for 20 min. The
941 slides were stained with Giemsa 2% for 15 min. Chromosome number
942 determination was performed for the varieties SP80-3280, SP81-3250, RB83-5486,
943 RB92-5345, IACSP95-3018 and IACSP93-3046. CMA/DAPI coloration was
944 performed by enzymatic digestion as described by Guerra and Souza (2002). The
945 slides were stained with 10 µg/ml DAPI for 30 min and 10 µg/ml CMA for 1 h.
946 Afterward, the slides were stained with 1:1 glycerol/McIlvaine buffer and visualized.
947 BAC-FISH was performed using the variety SP803280. For mitotic
948 chromosome preparations, root tips 0.5–1.5 cm in length were collected and
949 treated in the dark with p-dichlorobenzene-saturated solution in the dark at room
950 temperature for 2 h, then fixed in a freshly prepared 3:1 mixture (ethanol: glacial
951 acetic acid) at 4°C for 24 h and stored at −20°C until use. After being washed in
952 water, they were digested with the following enzymes: 2% cellulase (w/v) (Serva,
953 Heidelberg, Baden-Wurtemberg State, Germany), 20% pectinase (v/v) (Sigma,
954 Munich, Baviera State, Germany) and 1% Macerozyme (w/v) (Sigma) at 37°C for 1
955 h-2 h (Schwarzacher et al., 1980). The meristems were squashed in a drop of 45%
956 acetic acid and fixed in liquid nitrogen for 15 min. After air-drying, slides with good
957 metaphase chromosome spreads were stored at −20°C.
958 The BACs Shy064N22 and Shy048L15, both from the BAC library for the
959 SP80-3280 variety, were used as probes. The probes were labeled with
960 digoxigenin-11-dUTP (Roche) by nick translation. Bacterial artificial chromosome-
961 fluorescence in situ hybridization (BAC-FISH) was performed as described by
962 Schwarzacher and Heslop-Harrison (2000) with minor modifications. The $C_o t$-100
963 fraction of the sugarcane variety SP80-3280 genomic DNA, which was used to
964 block repetitive sequences, was prepared according to Zwick et al. (1997).
965 Preparations were counterstained and mounted with 2 µg/ml DAPI in Vectashield
966 (Vector, Burlingame, CA, USA).

The sugarcane metaphase chromosomes were observed and photographed, depending on the procedure, with transmitted light or epifluorescence under an Olympus BX61 microscope equipped with the appropriate filter sets (Olympus, Shinjuku-ku, Tokyo, Japan) and a JAI® CV-M4 + CL monochromatic digital camera (JAI, Barrington, N.J., USA). Digital images were imported into Photoshop 7.0 (Adobe, San Jose, Calif., USA) for pseudocoloration and final processing.

Genetic map construction

The BAC haplotypes were used to identify 44 sugarcane SNPs in the *HP600* and *CENP-C* exons. The SNP genotyping method was based on MALDI-TOF analysis performed on a mass spectrometer platform from Sequenom Inc.® as described by Garcia et al. (2013). The mapping population consisted of 151 full siblings derived from a cross between the SP80-3280 (female parent) and RB835486 (male parent) sugarcane cultivars, and the genetic map was constructed as described by Balsalobre et al. (2017).

ACCESSION NUMBERS

Sequence data from this article can be found in the EMBL/GenBank data libraries under the following accession number(s):

BAC sequences: from MH463467 to MH463488.

RNAseq subset data: SRR7274987

SUPPLEMENTAL MATERIAL:

Supplemental Figure 1. BAC-end BLASTN location in Sorghum genome.

Supplemental Figure 2. BAC BLASTN against sugarcane genome contigs.

Supplemental Figure 3. Phylogenetic and physical duplication representation.

Supplemental Figure 4. Evolutionary relationships of the gene Sobic.008G134700.

Supplemental Figure 5. Evolutionary relationships of HP600 and CENP-C.

Supplemental Figure 6. Mitotic metaphases of the sugarcane varieties.

Supplemental Figure 7. CENP-C motifs alignments.

Supplemental Table 1. BACs assembly and annotation.

Supplemental Table 2. Orthologous genes from Region01.

Supplemental Table 3. Orthologous genes from Region02.

Supplemental Table 4. Number of SNPs found in CENP-C and HP600.

Supplemental Table 5. Number of SNPs found in duplicated region.

Supplemental Table 6. Chromosome counts.

Supplemental Table 7. Sequenom iPLEX MassARRAY® primers.

ACKNOWLEDGMENTS

This study was supported by the São Paulo Research Foundation (FAPESP) (2008/52197-4) and Coordination for the Improvement of Higher Education Personnel (CAPES, Computational Biology Program). The first author was supported by a FAPESP Ph.D. fellowship from FAPESP (2010/50119-6). The third and fourth authors were supported by PD fellowships (MM 2014/11482-9 and CC-S 2015/16399-5). AP received a research fellowship from the National Council for Scientific and Technological Development (CNPq).

AUTHOR CONTRIBUTIONS

APS, DAS, ERFM, HB, MV, and AAFG designed the study; AB, DAS, HH, JF, MC, MCM, MVRC, ND, NR and SV performed the research; CBC-S, DAS, GSP, MV, M-AVS and RV contributed new analytical/computational tools; AAFG, AB, APS, CBC-S, DAS, GSP, HB, LRP, MCM, MGAL, MS, MV, and SV analyzed the data; and DAS, MV and APS wrote the paper. All authors critically read the text and approved the manuscript.

REFERENCES

Balsalobre, T.W.A., da Silva Pereira, G., Margarido, G.R.A., Gazaffi, R., Barreto, F.Z., Anoni, C.O., Cardoso-Silva, C.B., Costa, E.A., Mancini, M.C., Hoffmann, H.P., de Souza, A.P., Garcia, A.A., and Carneiro, M.S. (2017). GBS-based single dosage markers for linkage and QTL mapping

allow gene mining for yield-related traits in sugarcane. BMC Genomics **18:** 72.

Balsalobre, T.W.A., Mancini, M.C., Pereira, G.d.S., Anoni, C.O., Barreto, F.Z., Hoffmann, H.P., de Souza, A.P., Garcia, A.A.F., and Carneiro, M.S. (2016). Mixed modeling of yield components and brown rust resistance in sugarcane families. Agron. J. **108**: 1824-1837.

Birchler, J.A., Riddle, N.C., Auger, D.L., and Veitia, R.A. (2005). Dosage balance in gene regulation: biological implications. Trends Genet. **21:** 219-226.

Bremer, G. (1961). Problems in breeding and cytology of sugarcane. Euphytica **10:** 59–78.

Buggs, R.J., Zhang, L., Miles, N., Tate, J.A., Gao, L., Wei, W., Schnable, P.S., Barbazuk, W.B., Soltis, P.S., and Soltis, D.E. (2011). Transcriptomic shock generates evolutionary novelty in a newly formed, natural allopolyploid plant. Curr. Biol. **21:** 551–556.

Cardoso-Silva, C.B., Costa, E.A., Mancini, M.C., Balsalobre, T.W.A., Canesin, L.E.C., Pinto, L.R., Carneiro, M.S., Garcia, A.A.F., de Souza, A.P., and Vicentini, R. (2014). De novo assembly and transcriptome analysis of contrasting sugarcane varieties. PLOS ONE **9:** e88462 doi: 10.1371/journal.pone.0088462.

Casu, R.E., Selivanova, A., and Perroux, J.M. (2012). High-throughput assessment of transgene copy number in sugarcane using real-time quantitative PCR. Plant Cell Rep. **31:** 167–177.

Chalhoub, B., Belcram, H., and Caboche, M. (2004). Efficient cloning of plant genomes into bacterial artificial chromosome (BAC) libraries with larger and more uniform insert size. Plant Biotechnol. J. **2:** 181–188.

Chen, F., He, G., He, H., Chen, W., Zhu, X., Liang, M., Chen, L., and Deng, X.W. (2010). Expression analysis of miRNAs and highly-expressed small RNAs in two rice subspecies and their reciprocal hybrids. J. Integr. Plant Biol. **52:** 971–980.

Clopper, C.J., and Pearson, E.S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika **26:** 404–413.

Conover, J.W. (1971) Practical nonparametric statistics. John Wiley & Sons, New York, Pp 97–104.

Cordeiro, G.M., Eliott, F., McIntyre, C.L., Casu, R.E., and Henry, R.J. (2006). Characterisation of single nucleotide polymorphisms in sugarcane ESTs. Theor. Appl. Genet. **113:** 331-343.

Costa, E.A., Anoni, C.O., Mancini, M.C., Santos, F.R.C., Marconi, T.G., Gazaffi, R., Pastina, M.M., Perecin, D., Mollinari, M., Xavier, M.A., Pinto, L.R., Souza, A.P., and Garcia, A.A.F. (2016). QTL mapping including codominant SNP markers with ploidy level information in a sugarcane progeny. Euphytica **211**: 1–16.

Cuadrado, A., Acevedo, R., Moreno Díaz de la Espina, S., Jouve, N., and De La Torre, C. (2004). Genome remodelling in three modern S. officinarumxS. spontaneum sugarcane cultivars. J. Exp. Bot. **55:** 847-854.

D'Hont, A. (2005). Unraveling the genome structure of polyploids using FISH and GISH; examples of sugarcane and banana. Cytogenet. Genome Res. **109:** 27-33.

D'Hont, A., and Glaszmann, J.C. (2001). Sugarcane genome analysis with molecular markers, a first decade of research Proceedings of the Int Soc. Sugarcane Technol. 24**,** pp 556–559.

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

Daniels, J., and Roach, B.T. (1987). Taxonomy and evolution. In:. Heinz DJ (ed.). Sugarcane improvement through breeding, vol. 1. Elsevier, Amsterdam, pp 7–84.

de Setta, N., Monteiro-Vitorello, C.B., Metcalfe, C.J., Cruz, G.M.Q., Del Bem, L.E., Vicentini, R., Nogueira, F.T.S., Campos, R.A., Nunes, S.L., Turrini, P.C.G., Vieira, A.P., Ochoa Cruz, E.A., Corrêa, T.C., Hotta, C.T., de Mello Varani, A., Vautrin, S., da Trindade, A.S., de Mendonça Vilela, M.,

Lembke, C.G., Sato, P.M., de Andrade, R.F., Nishiyama, M.Y., Cardoso-Silva, C.B., Scortecci, K.C., Garcia, A.A., Carneiro, M.S., Kim, C., Paterson, A.H., Bergès, H., D'Hont, A., de Souza, A.P., Souza, G.M., Vincentz, M., Kitajima, J.P., and Van Sluys, M.A. (2014). Building the Sugarcane genome for biotechnology and identifying evolutionary trends. Genomics **15:** 540.

Doyle, J.J., Flagel, L.E., Paterson, A.H., Rapp, R.A., Soltis, D.E., Soltis, P.S., and Wendel, J.F. (2008). Evolutionary genetics of genome merger and doubling in plants. Annu. Rev. Genet. **42:** 443–461.

Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using Phred. II. Error probabilities. Genome Res. **8:** 186–194.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. (1998). Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. Genome Res. **8:** 175–185.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39:** 783-791.

Feltus, F.A., Wan, J., Schulze, S.R., Estill, J.C., Jiang, N., and Paterson, A.H. (2004). An SNP resource for rice genetics and breeding based on subspecies indica and japonica genome alignments. Genome Res. **14:** 1812–1819.

Figueira, T.R., Okura, V., Rodrigues da Silva, F., Jose da Silva, M., Kudrna, D., Ammiraju, J.S., Talag, J., Wing, R., and Arruda, P. (2012). A BAC library of the SP80-3280 sugarcane variety (Saccharum sp.) and its inferred microsynteny with the sorghum genome. BMC Res. Notes **5:** 185.

Foissac, S., Gouzy, J., Rombauts, S., Mathé, C., Amselem, J., Sterck, L., Vean, de Peer, Y.V., Rouze, P., and Schiex, T. (2008). Genome annotation in plants and fungi: EuGene as a model platform. Curr. Bioinform. **3:** 87–97.

Freeling, M. (2017). Picking up the ball at the K/Pg boundary: the distribution of ancient polyploidies in the plant phylogenetic tree as a spandrel of asexuality with occasional sex. Plant Cell **29**: 202-206.

Freeling, M., Lyons, E., Pedersen, B., Alam, M., Ming, R., and Lisch, D. (2008). Many or most genes in Arabidopsis transposed after the origin of the order Brassicales. Genome Res. **18**: 1924–1937.

Garcia, A.A., Mollinari, M., Marconi, T.G., Serang, O.R., Silva, R.R., Vieira, M.L., Vicentini, R., Costa, E.A., Mancini, M.C., Garcia, M.O., Pastina, M.M., Gazaffi, R., Martins, E.R., Dahmer, N., Sforça, D.A., Silva, C.B., Bundock, P., Henry, R.J., Souza, G.M., van Sluys, M.A., Landell, M.G., Carneiro, M.S., Vincentz, M.A., Pinto, L.R., Vencovsky, R., and Souza, A.P. (2013). SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. Sci. Rep. **3:** 3399.

Garsmeur, O., Charron, C., Bocs, S., Jouffe, V., Samain, S., Couloux, A., Droc, G., Zini, C., Glaszmann, J.C., Van Sluys, M.A., and D'Hont, A. (2011). High homologous gene conservation despite extreme autopolyploid redundancy in sugarcane. New Phytol. **189:** 629-642.

Gaut, B.S., and Doebley, J.F. (1997). DNA sequence evidence for the segmental allotetraploid origin of maize. Proc. Natl Acad. Sci. U.S.A. **94:** 6809–6814.

Gaut, B.S., Morton, B.R., McCaig, B.C., and Clegg, M.T. (1996). Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. Proc. Natl Acad. Sci. U.S.A. **93:** 10274-10279.

Gonthier, L., Bellec, A., Blassiau, C., Prat, E., Helmstetter, N., Rambaud, C., Huss, B., Hendriks, T., Bergès, H., and Quillet, M.C. (2010). Construction and characterization of two BAC libraries representing a deep-coverage of the genome of chicory. (Cichorium intybus L., Asteraceae). BMC Res. Notes **3:** 225.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., and Rokhsar, D.S. (2012). Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. **40:** D1178–D1186.

Gopalakrishnan, S., Sullivan, B.A., Trazzi, S., Della Valle, G., and Robertson, K.D. (2009). DNMT3B interacts with constitutive centromere protein *CENP-*

C to modulate DNA methylation and the histone code at centromeric regions. Hum. Mol. Genet. **18:** 3178–3193.

Gordon, D. (2003), 11.2.1-11.2.43. Viewing and editing assembled sequences using Consed. Curr. Protoc. Bioinformatics. In: Baxevanis AD, Davison DB (eds). John Wiley & Co., New York, pp 11.2.1-11.2.43.

Gordon, D., Abajian, C., and Green, P. (1998). Consed: a graphical tool for sequence finishing. Genome Res. **8:** 195-202.

Gordon, D., Desmarais, C., and Green, P. (2001). Automated finishing with Autofinish. Genome Res. **11:** 614-625.

Grivet, L., and Arruda, P. (2002). Sugarcane genomics: depicting the complex genome of an important tropical crop. Curr. Opin. Plant Biol. **5:** 122-127.

Guerra, M. (1983). O uso de Giemsa em citogenética vegetal – comparação entre a coloração simples e o bandamento. Cienc. Cult. **35:** 190-193.

Guerra, M., and Souza, M.J. (2002) Como observar cromossomos: um guia de técnicas em citogenética vegetal, animal e humana. FUNPEC, Ribeirão Preto, SP, p 131.

Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. **52:** 696-704.

Hegarty, M.J., Barker, G.L., Wilson, I.D., Abbott, R.J., Edwards, K.J., and Hiscock, S.J. (2006). Transcriptome shock after interspecific hybridization in senecio is ameliorated by genome duplication. Curr. Biol. **16:** 1652–1659.

Hollander, M., and Wolfe, D.A. (1973) Nonparametric statistical methods. John Wiley & Sons, New York, pp 15–22.

Ilic, K., SanMiguel, P.J., and Bennetzen, J.L. (2003). A complex history of rearrangement in an orthologous region of the maize, sorghum, and rice genomes. Proc. Natl Acad. Sci. U.S.A. **100**: 12265-12270.

Irvine, J.E. (1999). *Saccharum* species as horticultural classes. TAG Theoretical and Applied Genetics **98**: 186–194.

Jannoo, N., Grivet, L., Chantret, N., Garsmeur, O., Glaszmann, J.C., Arruda, P., and D'Hont, A. (2007). Orthologous comparison in a gene-rich region

among grasses reveals stability in the sugarcane polyploid genome. Plant J. **50:** 574-585.

Jiang, B., Lou, Q., Wu, Z., Zhang, W., Wang, D., Mbira, K.G., Weng, Y., and Chen, J. (2011). Retrotransposon- and microsatellite sequence-associated genomic changes in early generations of a newly synthesized allotetraploid cucumis□×□hytivus Chen & Kirkbride. Plant Mol. Biol. **77:** 225–233.

Jukes, T.H., and Cantor, C.R. (1969) Evolution of protein molecules. Academic Press, New York, pp 21–132.

Kato, H., Jiang, J., Zhou, B.R., Rozendaal, M., Feng, H., Ghirlando, R., Xiao, T.S., Straight, A.F., and Bai, Y. (2013). A conserved mechanism for centromeric nucleosome recognition by centromere protein *CENP-C*. Science **340:** 1110-1113.

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. **30:** 3059-3066.

Keller, O., Kollmar, M., Stanke, M., and Waack, S. (2011). A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics **27:** 757-763.

Kellogg, E.A. (2001). Evolutionary history of the grasses. Plant Physiol. **125:** 1198–1205.

Kim, C., Guo, H., Kong, W., Chandnani, R., Shuang, L.S., and Paterson, A.H. (2016). Application of genotyping by sequencing technology to a variety of crop breeding programs. Plant Sci. **242:** 14-22.

Kim, C., Lee, T.H., Compton, R.O., Robertson, J.S., Pierce, G.J., and Paterson, A.H. (2013). A genome-wide BAC end-sequence survey of sugarcane elucidates genome composition, and identifies BACs covering much of the euchromatin. Plant Mol. Biol. **81:** 139–147.

Kim, M. Y., and Zilberman, D. (2014). DNA methylation as a system of plant genomic immunity. Trends Plant Sci. **19**: 320-326.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16:** 111-120.

Kohany, O., Gentles, A.J., Hankus, L., and Jurka, J. (2006). Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinform. **7:** 474.

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol. Biol. Evol. **33:** 1870-1874.

Lai, J., Ma, J., Swigonová, Z., Ramakrishna, W., Linton, E., Llaca, V., Tanyolac, B., Park, Y.J., Jeong, O.Y., Bennetzen, J.L., and Messing, J. (2004). Gene loss and movement in the maize genome. Genome Res. **14**: 1924-1931.

Landell, M.G.A., Campana, M.P., Figueiredo, P., Vasconcelos ACM, Xavier MA, Bidoia MAP, Prado H, Silva MA, and Miranda LLD (2005) Variedades de cana-de-açúcar pad'ra o centro sul do Brasil. Technical Bulletin IAC **197**: 33.

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

Le Cunff, L., Garsmeur, O., Raboin, L.M., Pauquet, J., Telismart, H., Selvi, A., Grivet, L., Philippe, R., Begum, D., Deu, M., Costet, L., Wing, R., Glaszmann, J.C., and D'Hont, A. (2008). Diploid/polyploid syntenic shuttle mapping and haplotype-specific chromosome walking toward a rust resistance gene (Bru1) in highly polyploid sugarcane (2n approximately 12x approximately 115). Genetics **180:** 649-660.

Long, M., Betrán, E., Thornton, K., and Wang, W. (2003). The origin of new genes: glimpses from the young and old. Nat. Rev. Genet. **4:** 865–875.

Lynch, M., and Force, A.G. (2000). The origin of interspecific genomic incompatibility via gene duplication. Am. Nat. **156:** 590–605.

Ma, J., and Bennetzen, J.L. (2004). Rapid recent growth and divergence of rice nuclear genomes. Proc. Natl Acad. Sci. U.S.A. **101:** 12404–12410.

Mace, E.S., and Jordan, D.R. (2011). Integrating sorghum whole genome sequence information with a compendium of sorghum QTL studies reveals uneven distribution of QTL and of gene-rich regions with significant implications for crop improvement. Theor. Appl. Genet. **123:** 169–191.

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

Mattiello, L., Riaño-Pachón, D.M., Martins, M.C., da Cruz, L.P., Bassi, D., Marchiori, P.E., Ribeiro, R.V., Labate, M.T., Labate, C.A., and Menossi, M. (2015). Physiological and transcriptional analyses of developmental stages along sugarcane leaf. BMC Plant Biol. **15**: 300.

McClintock, B. (1984). The significance of responses of the genome to challenge. Science **226:** 792–801.

Ming, R., Liu, S.C., Lin, Y.R., da Silva, J., Wilson, W., Braga, D., van Deynze, A., Wenslaff, T.F., Wu, K.K., Moore, P.H., Burnquist, W., Sorrells, M.E., Irvine, J.E., and Paterson, A.H. (1998). Detailed alignment of saccharum and sorghum chromosomes: comparative organization of closely related diploid and polyploid genomes. Genetics **150:** 1663–1682.

Moscone, E.A., Matzke, M.A., and Matzke, A.J.M. (1996). The use of combined FISH/GISH in conjunction with DAPI counterstaining to identify chromosomes containing transgene inserts in amphidiploid tobacco. Chromosoma **105:** 231-236.

Murray, S.C., Sharma, A., Rooney, W.L., Klein, P.E., Mullet, J.E., Mitchell, S.E., and Kresovich, S. (2008). Genetic improvement of Sorghum as a biofuel feedstock: I. QTL for stem sugar and grain nonstructural carbohydrates. Crop Sci. **48:** 2165-2179.

Nishiyama Jr, M. Y., Ferreira, S. S., Tang, P. Z., Becker, S., Poertner-Taliana, A., and Souza, G. M. (2014). Full-length enriched cDNA libraries and ORFeome analysis of sugarcane hybrid and ancestor genotypes. PloS one, **9**: e107351.

Ohno, S. (1970) Evolution by gene duplication. Springer-Verlag, New York.

Paiva, J.A.P., Prat, E., Vautrin, S., Santos, M.D., San-Clemente, H., Brommonschenkel, S., Fonseca, P.G., Grattapaglia, D., Song, X., Ammiraju, J.S., Kudrna, D., Wing, R.A., Freitas, A.T., Bergès, H., and Grima-Pettenati, J. (2011). Advancing Eucalyptus genomics: identification and sequencing of lignin biosynthesis genes from deep-coverage BAC libraries. BMC Genomics **12:** 137. doi: 10.1186/1471-2164-12-137.

Paterson, A.H., Moore, P.H., and Tew, T.L. (2013). The gene pool of saccharum species and their improvement. In: Paterson A (ed.). Genomics of the Saccharinae, Springer, Berlin, pp 43–72.

Paux, E., Sourdille, P., Salse, J., Saintenac, C., Choulet, F., Leroy, P., Korol, A., Michalak, M., Kianian, S., Spielmeyer, W., Lagudah, E., Somers, D., Kilian, A., Alaux, M., Vautrin, S., Bergès, H., Eversole, K., Appels, R., Safar, J., Simkova, H., Dolezel, J., Bernard, M., and Feuillet, C. (2008). A physical map of the 1-gigabase bread wheat chromosome 3b. Science **322:** 101–104. doi: 10.1126/science.1161847.

Peterson, D.G., Tomkins, J.P., Frisch, D.A., Wing, R.A., and Paterson, A.H. (2000). Construction of plant bacterial artificial chromosome (BAC) libraries: an illustrated guide. J. Agric. Genomics **5:** 1–100.

Piperidis, G., and D'Hont, A. (2001). Chromosome composition analysis of various Saccharum interspecific hybrids by genomic in situ hybridization (GISH) Proceedings of the Int Soc. Sugcane Technol. 24**,** pp 565–566.

Piperidis, G., Piperidis, N., and D'Hont, A. (2010). Molecular cytogenetic investigation of chromosome composition and transmission in sugarcane. Mol. Genet. Genomics **284:** 65-73.

Ramsey, J., and Schemske, D.W. (2002). Neopolyploidy in flowering plants. Annu. Rev. Ecol. Syst. **33:** 589-639.

Riaño-Pachón, D.M., and Mattiello, L. (2017). Draft genome sequencing of the sugarcane hybrid SP80-3280. F1000Res **6**: 861.

Roselli, S., Olry, A., Vautrin, S., Coriton, O., Ritchie, D., Galati, G., Navrot, N., Krieger, C., Vialart, G., Bergès, H., Bourgaud, F., and Hehn, A. (2017). A

bacterial artificial chromosome (BAC) genomic approach reveals partial clustering of the furanocoumarin pathway genes in parsnip. Plant J. **89:** 1119-1132. doi: 10.1111/tpj.13450.

RStudio Team (2015). RStudio: integrated development for R. RStudio, Inc., Boston, MA. http://www.rstudio.com/.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406-425.

Sandmann, M., Talbert, P., Demidov, D., Kuhlmann, M., Rutten, T., Conrad, U., and Lermontova, I. (2017). Targeting of arabidopsis KNL2 to centromeres depends on the conserved CENPC-k motif in Its C terminus. Plant Cell **29:** 144–155. doi: 10.1105/tpc.16.00720.

SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y., and Bennetzen, J.L. (1998). The paleontology of intergene retrotransposons of maize. Nat. Genet. **20:** 43-45.

Saze, H., Kitayama, J., Takashima, K., Miura, S., Harukawa, Y., Ito, T., and Kakutani, T. (2013). Mechanism for full-length RNA processing of Arabidopsis genes containing intragenic heterochromatin. Nat. Commun. **4**: 2301.

Schwarzacher, T., Ambros, P., and Schweizer, D. (1980). Application of Giemsa banding to orchid karyotype analysis. Plant Syst. Evol. **134:** 293-297.

Schwarzacher, T., and Heslop-Harrison, P. (2000). Practical in situ hybridization. BIOS Scientific.

Serang, O., Mollinari, M., Garcia, A.A.F. (2012) Efficient exact maximum a posteriori computation for bayesian SNP genotyping in polyploids. PLoS One **7**: e30906.

Shen, Y., Lin, X.Y., Shan, X.H., Lin, C.J., Han, F.P., Pang, J.S., and Liu, B. (2005). Genomic rearrangement in endogenous long terminal repeat retrotransposons of rice lines introgressed by wild rice (Zizania latifolia Griseb.). J. Integr. Plant Biol. **47:** 998–1008.

Shi, X., Zhang, C., Ko, D.K., and Chen, Z.J. (2015). Genome-wide dosage-dependent and -independent regulation contributes to gene expression and

1334 evolutionary novelty in plant polyploids. Mol. Biol. Evol. **32:** 2351–2366 doi:
1335 10.1093/molbev/msv116.
1336 **Snel, B., Bork, P., and Huynen, M.** (2000). Genome evolution. Gene fusion
1337 versus gene fission. Trends Genet. **16**: 9–11.
1338 **Soltis, P.S., and Soltis, D.E.** (2009). The role of hybridization in plant speciation.
1339 Annu. Rev. Plant Biol. **60:** 561–588.
1340 **Souza, G.M., Berges, H., Bocs, S., Casu, R., D'Hont, A., Ferreira, J.E., Henry,**
1341 **R., Ming, R., Potier, B., Sluys, M.A. van, Vincentz, M., and Paterson,**
1342 **A.H.** (2011). The sugarcane genome challenge: strategies for sequencing a
1343 highly complex genome. Tropical Plant Biol. **4:** 145–156.
1344 **Sun, Y., and Joyce, P.A.** (2017). Application of droplet digital PCR to determine
1345 copy number of endogenous genes and transgenes in sugarcane. Plant Cell
1346 Rep. **36:** 1775-1783.
1347 **Swigonová, Z., Lai, J., Ma, J., Ramakrishna, W., Llaca, V., Bennetzen, J.L.,**
1348 **and Messing, J.** (2004). Close split of sorghum and maize genome
1349 progenitors. Genome Res. **14:** 1916–1923.
1350 **Talbert, P.B., Bryson, T.D., and Henikoff, S.** (2004). Adaptive evolution of
1351 centromere proteins in plants and animals. J. Biol. **3:** 18.
1352 **Tomkins, J.P., Yu, Y., Miller-Smith, H., Frisch, D.A., Woo, S.S., and Wing, R.A.**
1353 (1999). A bacterial artificial chromosome library for sugarcane. Theor. Appl.
1354 Genet. **99:** 419-424. doi: 10.1007/s001220051252.
1355 **Vettore, A.L., da Silva, F.R., Kemper, E.L., Souza, G.M., da Silva, A.M., Ferro,**
1356 **M.I., Henrique-Silva, F., Giglioti, E.A., Lemos, M.V., Coutinho, L.L.,**
1357 **Nobrega, M.P., Carrer, H., França, S.C., Bacci Júnior, M., Goldman,**
1358 **M.H., Gomes, S.L., Nunes, L.R., Camargo, L.E., Siqueira, W.J., Van**
1359 **Sluys, M.A., Thiemann, O.H., Kuramae, E.E., Santelli, R.V., Marino, C.L.,**
1360 **Targon, M.L., Ferro, J.A., Silveira, H.C., Marini, D.C., Lemos, E.G.,**
1361 **Monteiro-Vitorello, C.B., Tambor, J.H., Carraro, D.M., Roberto, P.G.,**
1362 **Martins, V.G., Goldman, G.H., de Oliveira, R.C., Truffi, D., Colombo,**
1363 **C.A., Rossi, M., de Araujo, P.G., Sculaccio, S.A., Angella, A., Lima,**
1364 **M.M., de Rosa Júnior, V.E., Siviero, F., Coscrato, V.E., Machado, M.A.,**

Grivet, L., Di Mauro, S.M., Nobrega, F.G., Menck, C.F., Braga, M.D., Telles, G.P., Cara, F.A., Pedrosa, G., Meidanis, J., and Arruda, P. (2003). Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. Genome Res. **13:** 2725–2735.

Vieira, M.L.C., Almeida, C.B., Oliveira, C.A., Tacuatiá, L.O., Munhoz, C.F., Cauz-Santos, L.A., Pinto, L.R., Monteiro-Vitorello, C.B., Xavier, M.A., and Forni-Martins, E.R. (2018). Revisiting meiosis in sugarcane: chromosomal irregularities and the prevalence of bivalent configurations. Front. Gen. **9:** 213.

Vilela, M.M., Del Bem, L.E., Van Sluys, M.A., de Setta, N., Kitajima, J.P., Cruz, G.M.Q., Sforça, D.A., de Souza, A.P., Ferreira, P.C.G., Grativol, C., Cardoso-Silva, C.B., Vicentini, R., and Vincentz, M. (2017). Analysis of three sugarcane homo/homeologous regions suggests independent polyploidization events of Saccharum officinarum and Saccharum spontaneum. Genome Biol. Evol. **9:** 266–278.

Visendi, P., Berkman, P.J., Hayashi, S., Golicz, A.A., Bayer, P.E., Ruperao, P., Hurgobin, B., Montenegro, J., Chan, C.K., Staňková, H., Batley, J., Šimková, H., Doležel, J., and Edwards, D. (2016). An efficient approach to BAC based assembly of complex genomes. Plant Methods **12:** 2-2. doi: 10.1186/s13007-016-0107-9.

Wang, J., Roe, B., Macmil, S., Yu, Q., Murray, J.E., Tang, H., Chen, C., Najar, F., Wiley, G., Bowers, J., Van Sluys, M.A., Rokhsar, D.S., Hudson, M.E., Moose, S.P., Paterson, A.H., and Ming, R. (2010). Microcollinearity between autopolyploid sugarcane and diploid sorghum genomes. BMC Genomics **11:** 261.

Wang, X., Duan, C.G., Tang, K., Wang, B., Zhang, H., Lei, M., Lu, K., Mangrauthia, S.K., Wang, P., Zhu, G., Zhao, Y., and Zhu, J.K. (2013). RNA-binding protein regulates plant DNA methylation by controlling mRNA processing at the intronic heterochromatin-containing gene IBM1. Proc. Natl. Acad. Sci. U.S.A. **110**: 15467-15472.

Woodhouse, M.R., Schnable, J.C., Pedersen, B.S., Lyons, E., Lisch, D., Subramaniam, S., and Freeling, M. (2010). Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homologs. PLoS Biol. **8**: e1000409.

Wu, R., Ma, C.X., Painter, I., and Zeng, Z.B. (2002). Simultaneous maximum likelihood estimation of linkage and linkage phases in outcrossing species. Theor. Popul. Biol. **61:** 349–363.

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. **35:** W265–WW268 doi: 10.1093/nar/gkm286.

Xue, B.T., Guo, J.L., Que, Y.X., Fu, Z.W., Wu, L.G., and Xu, L.P. (2014). Selection of suitable endogenous reference genes for relative copy number detection in sugarcane. Int. J. Mol. Sci. **15:** 8846–8862.

Zwick, M.S., Hanson, R.E., Islam-Faridi, M.N., Stelly, D.M., Wing, R.A., Price, H.J., and McKnight, T.D. (1997). A rapid procedure for the isolation of C0t-1 DNA from plants. Genome **40:** 138–142.

FIGURE LEGENDS

Figure 1. Schematic representation of the sugarcane BAC haplotypes from Region01 and Region02. Squares of the same color represent sugarcane genes orthologous to *Sorghum bicolor* genes. Dotted lines connect the homologous genes in sugarcane at different positions. In sugarcane Region02, the *CENP-C* haplotypes in Region02 are represented by two squares (blue and pink), where each square represents a partial gene fusion. The dark gray strip represents the shared region from Region01 and Region02 (duplication). The genes in light gray (from *S. bicolor*) are not found in the sugarcane BACs. The representation is not to scale. The orientation of transcription is indicated by the direction of the arrow at the end of each gene.

Figure 2. Representation of each sugarcane BAC from Region01 and Region02. Arrows and rectangles of the same color represent the homologous genes in sugarcane. Black rectangles represent repeat regions. Yellow lines represent gaps. Similar regions are represented by a gray shadow connecting the BACs. The orientation of transcription is indicated by the direction of the arrow at the end of each gene. Scale representation.

Figure 3. FISH hybridization of the sugarcane BACs. Panel (A): BAC Shy065N22 hybridization in sugarcane variety SP-803280 mitosis showing eight signals for Region01. **Panel (B):** BAC Shy048L15 hybridization in sugarcane variety SP-803280 mitosis showing ten signals for Region02.

Figure 4. Fusion gene formation of *CENP-C* and Sobic003G299500. Panel (A): Sorghum *CENP-C* and Sobic003G299500 genome location. **Panel (B):** Sugarcane genomic *CENP-C* haplotypes in Region01 (all expressed). **Panel (C):** Partially duplicated sugarcane paralogs of *CENP-C* and Sobic003G299500 haplotypes in Region02 (only haplotypes XI/XII/XIII and haplotype XIV have evidence of expression). **Panel (D):** Sugarcane ortholog of Sobic003G299500 found in the sugarcane R570 BAC library. **Panel (E):** Transcripts from sugarcane SP80-3280 mapped against the CDS of sugarcane *CENP-C* haplotypes from Region01. **Panel (F):** Transcripts from sugarcane SP80-3280 mapped against the sugarcane chimerical paralogs of *CENP-C* and Sobic003G299500. As evidence of fusion

gene formation, the transcripts show the fusion point of the paralogs. **Panel (G):** Transcripts from sugarcane SP80-3280 mapped against the CDS of the sugarcane R570 ortholog of Sobic003G299500.

Figure 5. Ploidy and dosage in the sugarcane genomic DNA (BACs) and SuperMASSA estimation. The location of each SNP is shown by one haplotype from Region01 and one haplotype from Region02. “SuperMASSA Best Ploidy” means the SuperMASSA best ploidy with a posteriori probability of >0.8. “SuperMASSA Expected Ploidy” means we fixed the ploidy of the loci in SuperMASSA according to the BAC-FISH and BAC sequencing results. “Genomic Ploidy” means the ploidy of the loci according to the BAC-FISH and BAC sequencing results. “*” means the SNP was found only in the transcriptome.

Figure 6. Linkage map for the duplicated region. Schematic representation of a multiple sugarcane linkage map for sugarcane variety SP80-3280 with information about the sugarcane genome (BACs).

1460 **TABLES**

1461 **Table 1.** Genomic frequencies of the SNPs in the *HP600* haplotypes in Region01 in the genome and transcriptome. The
1462 global expression (in diverse tissues) was used to determine whether the genomic frequency could explain the
1463 transcription frequency (H_0). The binomial test was used to verify H_0. The highlighted p-values reflect the acceptance of
1464 H_0.

									Missing haplotype for more common SNP				Missing haplotype for variant SNP			
SNP	**Name**	**Change**	**Polymorphism Type**	**Position**	**Coverage**	**Variant Coverage**	**Genomic Detected**	**Transcriptome Proportion**	Genomic Variant	Genomic	Genomic Proportion	P-value (binomial test)	Genomic Variant	Genomic	Genomic Proportion	P-value (binomial test)
1	C	G -> C	SNP (transversion)	12	443	101	Yes	0.23	1	7	0.125	2.32E-09	2	6	0.25	2.98E-01*
2	-	-C	Deletion	78	515	28	Yes	0.05	1	7	0.125	1.13E-07	2	6	0.25	4.76E-32
3	T	C -> T	SNP (transition)	133	542	38	Yes	0.07	1	7	0.125	5.16E-05	2	6	0.25	1.62E-27
4	A	G -> A	SNP (transition)	153	577	33	Yes	0.06	1	7	0.125	9.76E-08	2	6	0.25	1.56E-34
5	TT	GG -> TT	Substitution	166	699	137	Yes	0.2	1	7	0.125	1.18E-07	2	6	0.25	8.85E-04
6	T	C -> T	SNP (transition)	263	569	55	No	0.1	1	7	0.125	4.23E-02	1	7	0.125	4.23E-02
7		(GAG)3 -> (GAG)2	Deletion (tandem repeat)	283	654	42	No	0.06	1	7	0.125	4.35E-07	1	7	0.125	4.35E-07
8	C	T -> C	SNP (transition)	429	849	83	No	0.1	1	7	0.125	1.68E-02	1	7	0.125	1.68E-02
9	A	G -> A	SNP (transition)	434	993	69	No	0.07	1	7	0.125	1.68E-08	1	7	0.125	1.68E-08
10	C	G -> C	SNP (transversion)	436	1035	275	Yes	0.27	2	6	0.25	2.51E-01*	3	5	0.375	1.196E-13
11	T	G -> T	SNP (transversion)	463	936	56	No	0.06	1	7	0.125	5.11E-11	1	7	0.125	5.11E-11
12	A	C -> A	SNP (transversion)	519	679	57	No	0.08	1	7	0.125	9.10E-04	1	7	0.125	9.10E-04

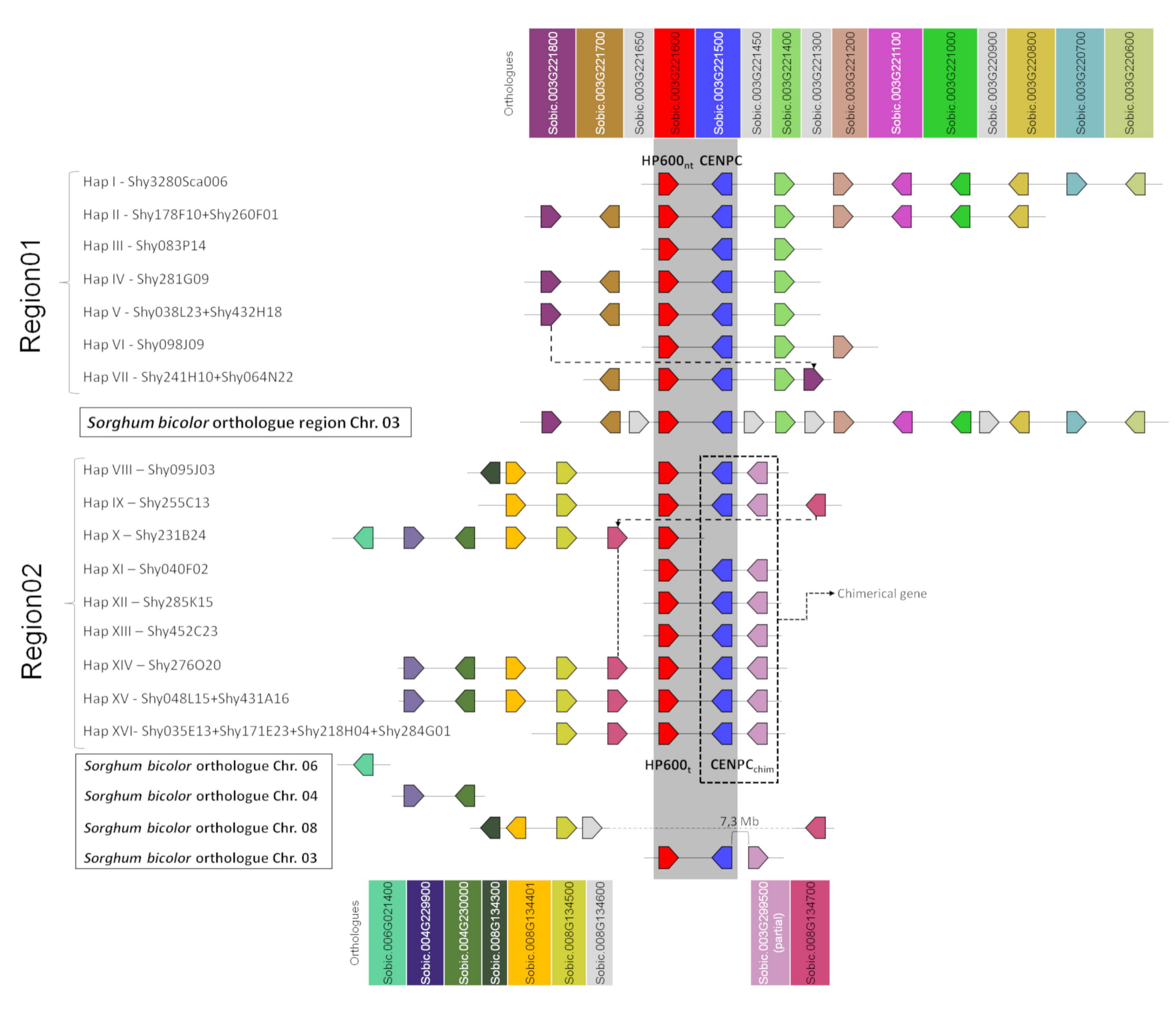
1465

Table 2. Genomic frequencies of the SNPs in the *CENP-C* haplotypes in Region01 and Region02 in the genome and transcriptome. The global expression (in diverse tissues) was used to determine whether the genomic frequency could explain the transcription frequency (H_0). The binomial test was used to verify H_0. The highlighted p-values reflect the acceptance of H_0.

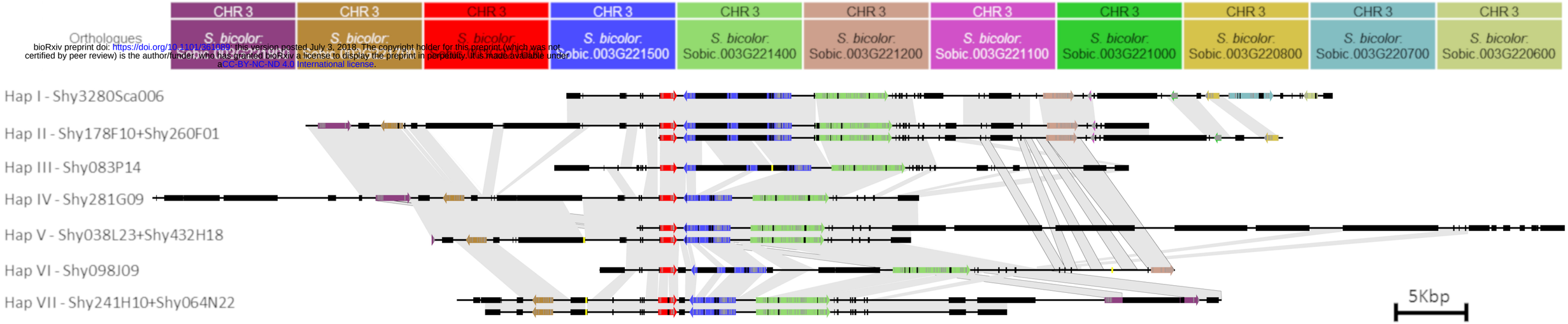
										Missing haplotype for more common SNP				Missing haplotype for variant SNP			
SNP	**Name**	**Change**	**Polymorphism Type**	**Position**	**Coverage**	**Variant Coverage**	**Genomic Detected**	**Transcriptome Proportion**	Genomic Variant	Genomic	Genomic Proportion	P-value (binomial test)	Genomic Variant	Genomic	Genomic Proportion	P-value (binomial test)
1	G	C -> G	SNP (transversion)	106	16	13	Yes	0.81	5	3	0.63	1.95E-01*	4	4	0.5	2.13E-02
2	G	A -> G	SNP (transition)	150	19	8	Yes	0.42	1	7	0.13	1.25E-03	2	6	0.25	1.08E-01*
3	C	G -> C	SNP (transversion)	246	34	7	Yes	0.21	1	7	0.13	1.87E-01*	2	6	0.25	6.93E-01*
4	T	A -> T	SNP (transversion)	369	65	7	Yes	0.11	1	7	0.13	8.51E-01*	2	6	0.25	6.14E-03
5	A	G -> A	SNP (transition)	371	68	19	No	0.28	1	7	0.13	6.21E-04	1	7	0.13	6.21E-04
6	C	T -> C	SNP (transition)	390	64	15	No	0.23	1	7	0.13	1.32E-02	1	7	0.13	1.32E-02
7	G	T -> G	SNP (transversion)	513	46	12	Yes	0.26	3	5	0.38	1.28E-01*	4	4	0.5	1.64E-03
8	A	G -> A	SNP (transition)	518	45	10	Yes	0.22	2	6	0.25	7.34E-01*	3	5	0,375	4.40E-02
9	T	G -> T	SNP (transversion)	731	54	8	Yes	0.15	2	6	0.25	1.14E-01*	3	5	0,375	3.58E-04
10	C	A -> C	SNP (transversion)	1008	56	9	No	0.16	1	7	0.13	4.17E-01*	1	7	0.13	4.17E-01*
11	T	C -> T	SNP (transition)	1061	91	29	Yes	0.32	2	6	0.25	1.46E-01*	3	5	0,375	2.81E-01*
12	T	C -> T	SNP (transition)	1088	77	41	Yes	0.53	4	4	0.50	6.48E-01*	3	5	0,375	6.37E-03
13	T	C -> T	SNP (transition)	1190	76	9	Yes	0.12	2	6	0.25	7.49E-03	3	5	0,375	1.10E-06
14	A	G -> A	SNP (transition)	1209	76	20	No	0.26	1	7	0.13	1.31E-03	1	7	0.13	1.31E-03
15	T	A -> T	SNP (transversion)	1251	62	10	Yes	0.16	2	6	0.25	1.41E-01*	3	5	0,375	3.29E-04
16	G	A -> G	SNP (transition)	1255	62	55	Yes	0.89	6	2	0.75	1.19E-02	5	3	0,625	5.15E-06
17		-ATG	Deletion	1307	75	9	Yes	0.12	1	7	0.13	1.00E+00*	2	6	0.25	7.38E-03
18	G	A -> G	SNP (transition)	1314	90	23	Yes	0.26	1	7	0.13	6.50E-04	2	6	0.25	9.03E-01*
19	G	T -> G	SNP (transversion)	1347	103	13	Yes	0.13	2	6	0.25	2.88E-03	3	5	0,375	3.09E-08
20	A	T -> A	SNP (transversion)	1384	101	37	Yes	0.37	1	7	0.13	5.30E-10	2	6	0.25	1.09E-02
21	G	C -> G	SNP (transversion)	1424	80	9	No	0.11	1	7	0.13	8.66E-01*	1	7	0.13	8.66E-01*
22	A	C -> A	SNP (transversion)	1437	84	10	Yes	0.12	1	7	0.13	1.00E+00*	2	6	0.25	5.12E-03
23	TT	AA -> TT	Substitution	1481	62	7	No	0.11	1	7	0.13	1.00E+00*	1	7	0.13	1.00E+00*
24	G	A -> G	SNP (transition)	1527	106	90	Yes (duplication)	0.85								
25	C	T -> C	SNP (transition)	1540	139	86	Yes (duplication)	0.62								

26	A	T -> A	SNP (transversion)	1584	253	235	Yes (duplication)	0.93								
27	A	G -> A	SNP (transition)	1638	247	39	Yes (duplication)	0.16								
28	C	A -> C	SNP (transversion)	1648	209	106	Yes (duplication)	0.51								
29	A	C -> A	SNP (transversion)	1739	122	16	Yes (duplication)	0.13								
30	T	C -> T	SNP (transition)	1751	132	32	Yes (duplication)	0.24								
31	A	G -> A	SNP (transition)	1753	138	16	Yes (duplication)	0.12								
32	A	C -> A	SNP (transversion)	1762	131	21	No (duplication)	0.16								
33	T	A -> T	SNP (transversion)	1776	125	75	Yes (duplication)	0.6								
34	C	G -> C	SNP (transversion)	1796	88	31	No (duplication)	0.35								
35	G	C -> G	SNP (transversion)	1808	37	25	Yes	0.68	4	3	0.57	0.00E+00*	4	4	0.57	8.90E-01*
36	T	C -> T	SNP (transition)	1808	78	41	Yes (duplication)	0.53								
37	T	C -> T	SNP (transition)	1814	78	27	Yes (duplication)	0.35								
38	T	C -> T	SNP (transition)	1827	68	7	Yes (duplication)	0.1								
39	A	T -> A	SNP (transversion)	1830	65	8	Yes (duplication)	0.12								
40	A	G -> A	SNP (transition)	1839	62	23	Yes (duplication)	0.37								
41	A	G -> A	SNP (transition)	1853	52	6	Yes (duplication)	0.12								
42	C	A -> C	SNP (transversion)	1866	47	30	Yes (duplication)	0.64								
43	A	C -> A	SNP (transversion)	1910	152	34	Yes (duplication)	0.22								
44	A	G -> A	SNP (transition)	1917	158	103	Yes (duplication)	0.65								
45	G	T -> G	SNP (transversion)	1922	165	110	Yes (duplication)	0.67								
46	T	A -> T	SNP (transversion)	1938	170	41	Yes (duplication)	0.24								
47	A	C -> A	SNP (transversion)	2039	196	37	Yes (duplication)	0.19								
48	T	C -> T	SNP (transition)	2043	196	143	Yes (duplication)	0.73								
49	G	T -> G	SNP (transversion)	2080	177	88	Yes (duplication)	0.5								
50	C	A -> C	SNP (transversion)	2123	126	89	Yes (duplication)	0.71								

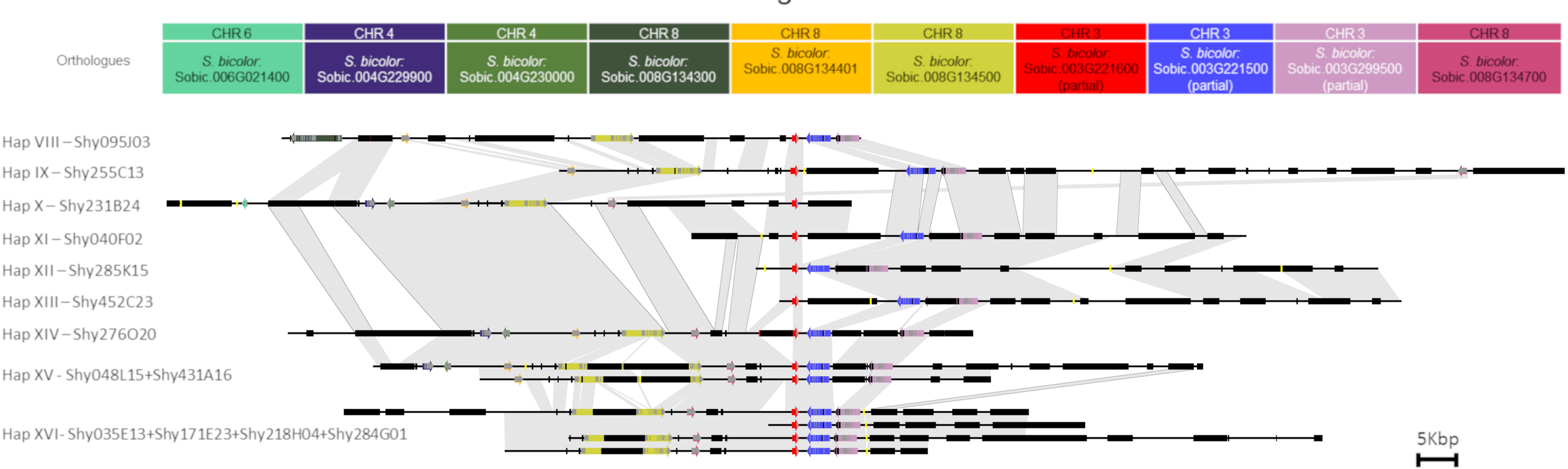
1470

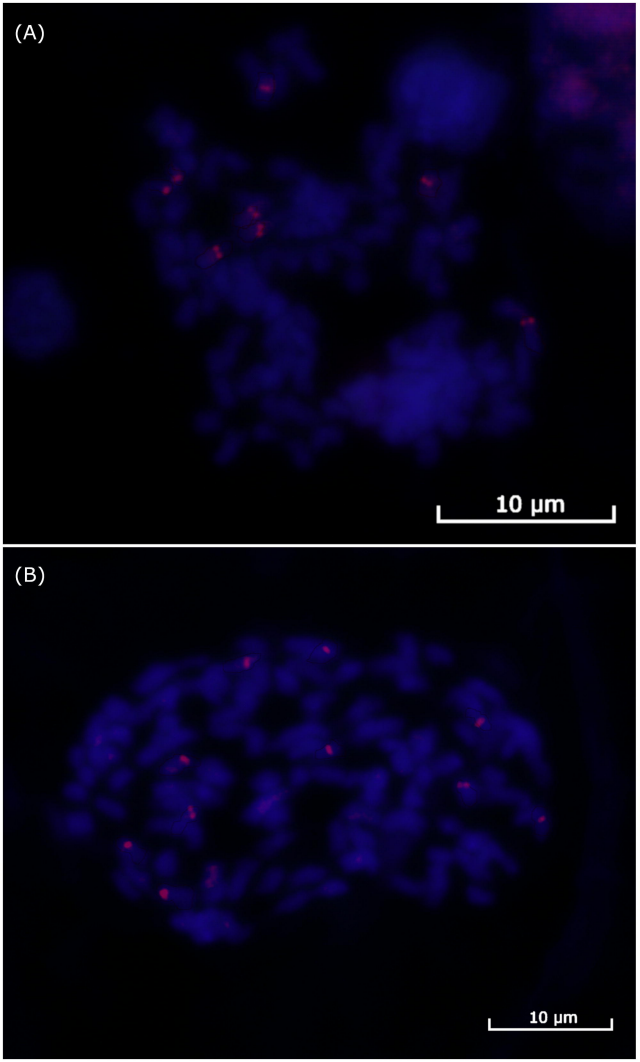

Orthologues
Sobic.003G221800
Sobic.003G221700
Sobic.003G221650
Sobic.003G221600
Sobic.003G221500
Sobic.003G221450
Sobic.003G221400
Sobic.003G221300
Sobic.003G221200
Sobic.003G221100
Sobic.003G221000
Sobic.003G220900
Sobic.003G220800
Sobic.003G220700
Sobic.003G220600
HP600nt CENPC
Region01
Hap I - Shy3280Sca006
Hap II - Shy178F10+Shy260F01
Hap III - Shy083P14
Hap IV - Shy281G09
Hap V - Shy038L23+Shy432H18
Hap VI - Shy098J09
Hap VII - Shy241H10+Shy064N22
Sorghum bicolor orthologue region Chr. 03
Region02
Hap VIII – Shy095J03
Hap IX – Shy255C13
Hap X – Shy231B24
Hap XI – Shy040F02
Hap XII – Shy285K15
Hap XIII – Shy452C23
Hap XIV – Shy276O20
Hap XV - Shy048L15+Shy431A16
Hap XVI- Shy035E13+Shy171E23+Shy218H04+Shy284G01
Chimerical gene
HP600t CENPCchim
Sorghum bicolor orthologue Chr. 06
Sorghum bicolor orthologue Chr. 04
Sorghum bicolor orthologue Chr. 08
Sorghum bicolor orthologue Chr. 03
7,3 Mb
Orthologues
Sobic.006G021400
Sobic.004G229900
Sobic.004G230000
Sobic.008G134300
Sobic.008G134401
Sobic.008G134500
Sobic.008G134600
Sobic.003G299500 (partial)
Sobic.008G134700

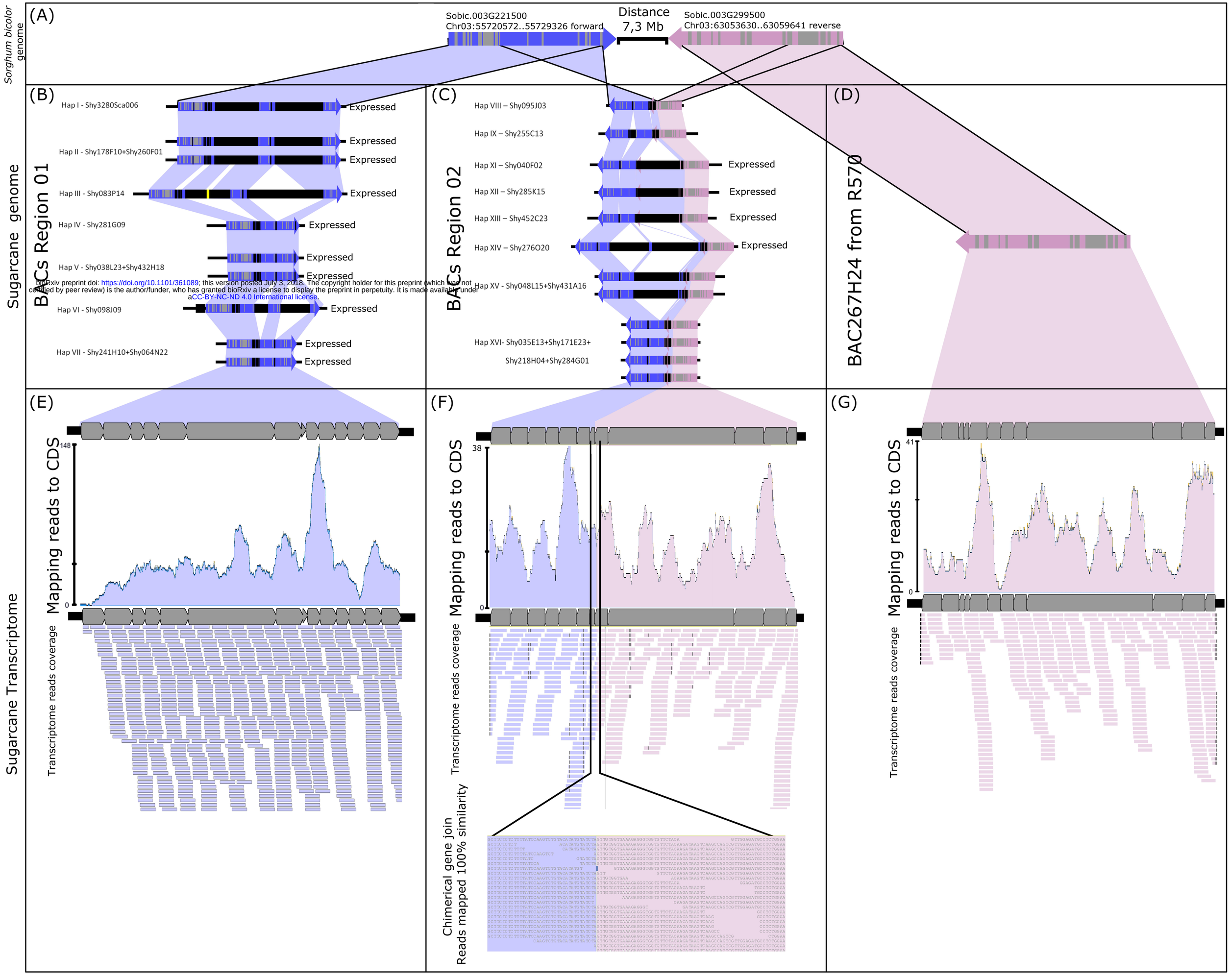
Region01

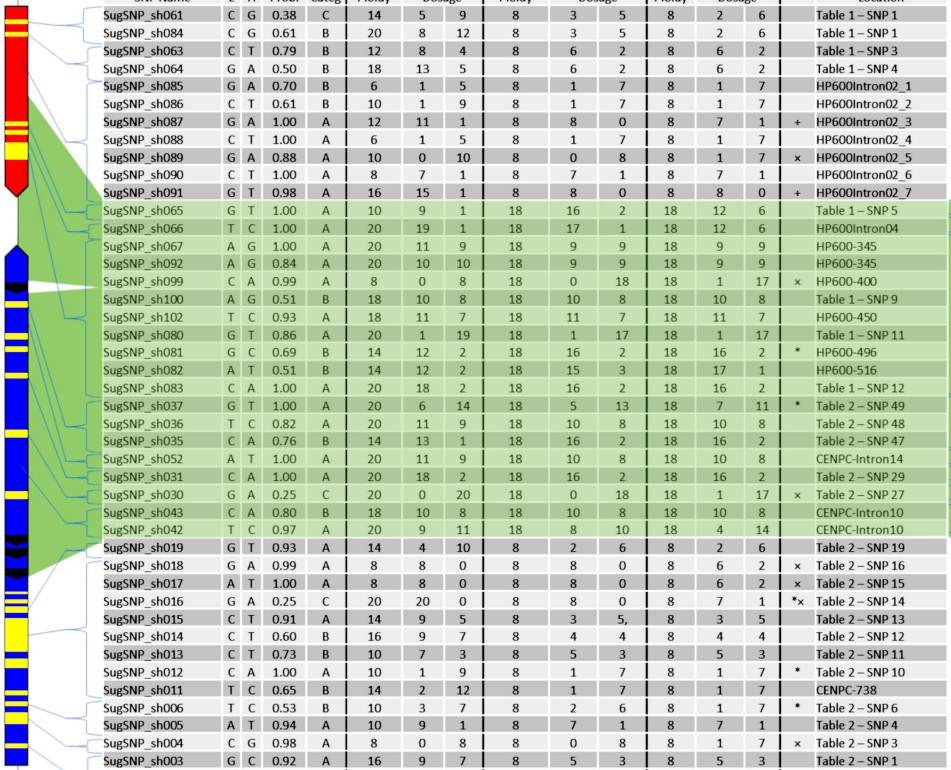


Region02




(A)
10 µm
(B)
10 µm





SNP Name	L	H	Prob.	Categ	SuperMASSA best ploidy			SuperMASSA expected ploidy			Genomic ploidy				Location
					Ploidy	Dosage		Ploidy	Dosage		Ploidy	Dosage			
SugSNP_sh061	C	G	0.38	C	14	5	9	8	3	5	8	2	6		Table 1 – SNP 1
SugSNP_sh084	C	G	0.61	B	20	8	12	8	3	5	8	2	6		Table 1 – SNP 1
SugSNP_sh063	C	T	0.79	B	12	8	4	8	6	2	8	6	2		Table 1 – SNP 3
SugSNP_sh064	G	A	0.50	B	18	13	5	8	6	2	8	6	2		Table 1 – SNP 4
SugSNP_sh085	G	A	0.70	B	6	1	5	8	1	7	8	1	7		HP600Intron02_1
SugSNP_sh086	C	T	0.61	B	10	1	9	8	1	7	8	1	7		HP600Intron02_2
SugSNP_sh087	G	A	1.00	A	12	11	1	8	8	0	8	7	1	+	HP600Intron02_3
SugSNP_sh088	C	T	1.00	A	6	1	5	8	1	7	8	1	7		HP600Intron02_4
SugSNP_sh089	G	A	0.88	A	10	0	10	8	0	8	8	1	7	×	HP600Intron02_5
SugSNP_sh090	C	T	1.00	A	8	7	1	8	7	1	8	7	1		HP600Intron02_6
SugSNP_sh091	G	T	0.98	A	16	15	1	8	8	0	8	8	0	+	HP600Intron02_7
SugSNP_sh065	G	T	1.00	A	10	9	1	18	16	2	18	12	6		Table 1 – SNP 5
SugSNP_sh066	T	C	1.00	A	20	19	1	18	17	1	18	12	6		HP600Intron04
SugSNP_sh067	A	G	1.00	A	20	11	9	18	9	9	18	9	9		HP600-345
SugSNP_sh092	A	G	0.84	A	20	10	10	18	9	9	18	9	9		HP600-345
SugSNP_sh099	C	A	0.99	A	8	0	8	18	0	18	18	1	17	×	HP600-400
SugSNP_sh100	A	G	0.51	B	18	10	8	18	10	8	18	10	8		Table 1 – SNP 9
SugSNP_sh102	T	C	0.93	A	18	11	7	18	11	7	18	11	7		HP600-450
SugSNP_sh080	G	T	0.86	A	20	1	19	18	1	17	18	1	17		Table 1 – SNP 11
SugSNP_sh081	G	C	0.69	B	14	12	2	18	16	2	18	16	2	*	HP600-496
SugSNP_sh082	A	T	0.51	B	14	12	2	18	15	3	18	17	1		HP600-516
SugSNP_sh083	C	A	1.00	A	20	18	2	18	16	2	18	16	2		Table 1 – SNP 12
SugSNP_sh037	G	T	1.00	A	20	6	14	18	5	13	18	7	11	*	Table 2 – SNP 49
SugSNP_sh036	T	C	0.82	A	20	11	9	18	10	8	18	10	8		Table 2 – SNP 48
SugSNP_sh035	C	A	0.76	B	14	13	1	18	16	2	18	16	2		Table 2 – SNP 47
SugSNP_sh052	A	T	1.00	A	20	11	9	18	10	8	18	10	8		CENPC-Intron14
SugSNP_sh031	C	A	1.00	A	20	18	2	18	16	2	18	16	2		Table 2 – SNP 29
SugSNP_sh030	G	A	0.25	C	20	0	20	18	0	18	18	1	17	×	Table 2 – SNP 27
SugSNP_sh043	C	A	0.80	B	18	10	8	18	10	8	18	10	8		CENPC-Intron10
SugSNP_sh042	T	C	0.97	A	20	9	11	18	8	10	18	4	14		CENPC-Intron10
SugSNP_sh019	G	T	0.93	A	14	4	10	8	2	6	8	2	6		Table 2 – SNP 19
SugSNP_sh018	G	A	0.99	A	8	8	0	8	8	0	8	6	2	×	Table 2 – SNP 16
SugSNP_sh017	A	T	1.00	A	8	8	0	8	8	0	8	6	2	×	Table 2 – SNP 15
SugSNP_sh016	G	A	0.25	C	20	20	0	8	8	0	8	7	1	*×	Table 2 – SNP 14
SugSNP_sh015	C	T	0.91	A	14	9	5	8	3	5,	8	3	5		Table 2 – SNP 13
SugSNP_sh014	C	T	0.60	B	16	9	7	8	4	4	8	4	4		Table 2 – SNP 12
SugSNP_sh013	C	T	0.73	B	10	7	3	8	5	3	8	5	3		Table 2 – SNP 11
SugSNP_sh012	C	A	1.00	A	10	1	9	8	1	7	8	1	7	*	Table 2 – SNP 10
SugSNP_sh011	T	C	0.65	B	14	2	12	8	1	7	8	1	7		CENPC-738
SugSNP_sh006	T	C	0.53	B	10	3	7	8	2	6	8	1	7	*	Table 2 – SNP 6
SugSNP_sh005	A	T	0.94	A	10	9	1	8	7	1	8	7	1		Table 2 – SNP 4
SugSNP_sh004	C	G	0.98	A	8	0	8	8	0	8	8	1	7	×	Table 2 – SNP 3
SugSNP_sh003	G	C	0.92	A	16	9	7	8	5	3	8	5	3		Table 2 – SNP 1
SugSNP_sh001	C	A	0.96	A	12	5	7	8	3	5	8	3	5		CENPC-74