

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

3

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

11

12 ^aUniversidade Estadual de Campinas (UNICAMP), Campinas, SP, Brazil

13 ^bCentre National de Ressources Genomiques Vegetales (CNRGV), Institut
14 National de la Recherche Agronomique (INRA), Castanet Tolosan, France

15 ^cEscola Superior de Agricultura Luiz de Queiroz (ESALQ), USP, Piracicaba, SP,
16 Brazil

17 ^dUniversidade de São Paulo, USP, São Paulo, SP, Brazil

18 ^eUniversidade Federal de São Carlos (UFSCAR), Araras, SP, Brazil

19 ^fCentro de Cana, Instituto Agrônômico de Campinas, São Paulo, SP, Brazil

20

21 Corresponding author: anete@unicamp.br

22

23 Short title: Unveiling the allelic dynamics of sugarcane

24

25

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).

26 **ABSTRACT**

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

45

46 INTRODUCTION

47 The *Saccharum* species are C4 grass and present a high level of ploidy. *S.*
48 *officinarum* is octaploid ($2n = 80$), with $x = 10$ chromosomes, while *S. spontaneum*
49 has $x = 8$ but presents great variation in the number of chromosomes, with main
50 cytotypes of $2n = 62, 80, 96, 112$ or 128 . The modern sugarcane cultivars
51 originated from hybridization between these two species and are considered
52 allopolyploid hybrids (Daniels and Roach, 1987; Paterson et al., 2013). The
53 development of these cultivars involved the process of 'nobilization' of the hybrid,
54 with successive backcrosses using *S. officinarum* as the recurrent parent (Bremer,
55 1961). The resulting hybrids are highly polyploid and aneuploid (D'Hont et al.,
56 1998; Irvine, 1999; Grivet and Arruda, 2002) and have an estimated whole genome
57 size of 10 Gb (D'Hont and Glaszmann, 2001). An in situ hybridization study has
58 shown that the genomes of the commercial hybrids consist of 10-20%
59 chromosomes from *S. spontaneum* and 5-17% recombinant chromosomes
60 between the two species, while the remaining majority of the genome consists of
61 chromosomes from *S. officinarum* (Piperidis and D'Hont, 2001; D'Hont, 2005).

62 Molecular evidence suggests that polyploid genomes can present dynamic
63 changes in DNA sequence and gene expression, probably in response to genomic
64 shock (genomic remodeling due to the activation of previously deleted
65 heterochromatic elements), and this phenomenon is implicated in epigenetic
66 changes in homologous genes due to intergenomic interactions (McClintock,
67 1984). The evolutionary success of polyploid species is related to their ability to
68 present greater phenotypic novelty than is observed in their diploid or even absent
69 in parents (Ramsey and Schemske, 2002). Among other factors, this increase in
70 the capacity for phenotypic variation capacity may be caused by regulation of the
71 allelic dosage (Birchler et al., 2005).

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

77 crossing of SP80-3280 has also been used to analyze rust resistance (Balsalobre
78 et al., 2016), quantitative trait loci (QTL) mapping (Costa et al., 2016) and
79 genotyping by sequencing (GBS) (Balsalobre et al., 2017). A Brazilian initiative
80 (Souza et al., 2011) is producing a gene-space genome sequence from SP80-
81 3280, and a draft sugarcane genome based on whole-genome shotgun
82 sequencing was produced (Riaño-Pachón and Mattiello, 2017). In addition, QTL
83 gene synteny from sorghum has been used to map corresponding BACs in SP80-
84 3280 (Mancini et al., 2018).

85 Three bacterial artificial chromosome (BAC) libraries for different sugarcane
86 varieties have been constructed. The oldest one is for the French variety R570
87 (Tomkins et al., 1999) and contains 103,296 clones with an average insert size of
88 130 kb, representing 1.2 total genome equivalents. A mix of four individuals
89 derived from the self-fertilization of the elite cultivar R570 (pseudo F2) was
90 reported by Le Cunff et al. (2008) and contains 110,592 clones with an average
91 insert size of 130 kb, representing 1.4x coverage of the whole genome. In addition,
92 a library of SP80-3280 published by Figueira et al. (2012) contains 36,864 clones
93 with an average insert size of 125 kb, representing 0.4 total genome equivalent
94 coverage.

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

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

108 centromere. Only CenH3 (histone H3) and *CENP-C* (centromere protein C) have
109 been shown to bind centromeric DNA (Talbert et al., 2004). The centromere is
110 marked with the histone H3 variant CenH3 (*CENP-A* in human), and *CENP-C*
111 forms part of the inner kinetochore. The *CENP-C* "central domain" makes close
112 contact with the acidic patch of histones H2A/H2B, and the highly conserved
113 "*CENP-C* motif" senses both the acidic patch and recognizes the hydrophobicity of
114 the otherwise nonconserved CenH3 tail, supporting a conserved mechanism of
115 centromere targeting by the kinetochore (Gopalakrishnan et al., 2009; Kato et al.,
116 2013; Sandmann et al., 2017). Sandmann et al. (2017) reported that in *Arabidopsis*
117 *thaliana*, KNL2, a protein with a *CENPC-k* motif, recognizes centromeric
118 nucleosomes such as the *CENP-C* protein. The *CENP-C* gene genomic structure
119 in sugarcane has not been detailed.

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

134

135 **RESULTS**

136

137 **BAC library construction**

138 The BAC library from the sugarcane variety SP80-3280 resulted in 221.184 clones,
139 arrayed in 576 384-well microtiter plates, with a mean insert size of 110 kb. This
140 BAC library is approximately 2.4 genome equivalents (10 Gb) and 26 monoploid
141 genome equivalents (930 Mb, Figueira et al., 2012). For the sugarcane variety
142 IACSP93-3046, the library construction resulted in 165.888 clones arrayed in 432
143 384-well microtiter plates, with a mean insert size of 110 kb, which is approximately
144 1.8 genome equivalents and 19 monoploid genome equivalents.

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

156

157 **BAC annotation**

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

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

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

176 The BACs were first annotated regarding the TEs. The TEs accounted for
177 21% to 65% of the sequenced bases with a mean of 40% (Supplemental Table 1).
178 Annotation of the TEs in the 22 BACs revealed 618 TEs (220 TEs were grouped in
179 the same type) with sizes ranging from 97 bp to 18,194 bp.

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

190

191 **Relationship between Region01 and Region02**

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

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

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

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

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

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

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

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

247 To estimate the genomic diversity in sugarcane haplotypes from both
248 regions (analyzed together and separately), the shared 8-kb region (duplication)
249 was used (Supplemental Table 4), and the SNPs were identified. The diversity in
250 the *HP600* and *CENP-C* genes was analyzed, and one SNP was observed every
251 43 bases (Region02) and 70 bases (Region01). We searched for SNPs that could
252 distinguish each region (Supplemental Table 5) in the *HP600* and *CENP-C* genes,
253 and one SNP was found for every 56 bases (20 SNPs in total). In addition, small
254 (3-10 bases) and large (30 – 200 bases) insertions were found. These results
255 revealed a high level of diversity in sugarcane, i.e., a high number of SNPs in each
256 region, which could be used to generate molecular markers and to improve genetic
257 maps. In addition, the diversity rate of both regions together could be used as an
258 indicator of a duplicated gene, i.e., a rate < 20 (Supplemental Table 4).

259

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

261 Gene haplotypes, i.e., genes with the same coding sequences (CDSs), from
262 *HP600* and *CENP-C* that have the same coding sequence (i.e., exons) in different
263 BAC haplotypes were considered the same gene haplotype. In Region01, four
264 haplotypes of *HP600* were identified. In sorghum, the size of *HP600* is 187 amino
265 acids (561 base pairs). *HP600* has two different sizes in sugarcane haplotypes of
266 188 amino acids (564 base pairs (haplotype I/II/VI, haplotype IV/V and haplotype
267 VII) and 120 amino acids (360 base pairs – haplotype III). *HP600* haplotype III has
268 a base deletion at position 77, causing a frameshift that results in a premature stop
269 codon.

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

273 In *S. bicolor*, *CENP-C* is formed by 14 exons (Talbert et al., 2004) encoding
274 694 amino acids (2082 base pairs). In sugarcane, the haplotypes from Region01
275 had 14 exons that give rise to a protein of 708 or 709 amino acids (2124 or 2127
276 bases). Talbert et al. (2004) described two haplotypes in sugarcane EST clones
277 (Vettore et al., 2003), *CENP-C1* and *CENP-C2*, which correspond to the
278 haplotypes I/II and haplotypes IV/V, respectively. In addition to *CENP-C1* and
279 *CENP-C2*, three other *CENP-C* haplotypes were observed: haplotype III, haplotype
280 VI, and haplotype VIII.

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

286 To reconstruct a phylogenetic tree for *HP600* and *CENP-C* from both
287 regions, the orthologs from *O. sativa* and *Zea mays* were searched. The rice
288 *HP600* and *CENP-C* orthologs, LOC_Os01g43060 and LOC_Os01g43050, were
289 recovered, respectively. Maize has gone through tetraploidization since its
290 divergence from sorghum approximately 12 million years ago (Woodhouse et al.,
291 2010). The maize *HP600* ortholog search returned three possible genes with high

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

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

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

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

315

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

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

323 improve the conclusions (Supplemental Figure 6 – Panel A, B, C, D and E – and
324 Supplemental Table 6). The most abundant number of chromosomes was $2n =$
325 112 (range: $2n = 98$ to $2n = 118$ chromosomes). The chromosome number of the
326 *Saccharum* hybrid cultivar SP80-3280 was found to be $2n = 112$ (range: $2n = 108$
327 to $2n = 118$ chromosomes - Supplemental Table 6). Vieira et al. (2018) found $2n =$
328 112 chromosomes for IACSP93-3046 variety, corroborating with our data. The $2n =$
329 112 chromosome number should indicate convergence in the number of
330 chromosomes in the *Saccharum* hybrid cultivar.

331 As second step, we used two varieties with the best chromosome spreads,
332 i.e., IACSP93-3046 and IACSP95-3018, for the CMA/DAPI banding patterns
333 (Supplemental Figure 6 – Panel F, G, H and I). The variety IACSP93-3046
334 exhibited at least six terminal $CMA^+/DAPI^-$ bands, one chromosome with
335 $CMA^+/DAPI^0$ and two chromosomes with adjacent intercalations of CMA^+ and
336 $DAPI^+$ in the same chromosome (Supplemental Figure 6 – Panel F and G). The
337 variety IACSP95-3018 revealed seven terminal $CMA^+/DAPI^-$ bands, and at least
338 two chromosomes exhibited adjacent CMA^+ and $DAPI^+$; one was in the intercalary
339 position, and the other was in the terminal position (Supplemental Figure 6 – Panel
340 H and I). Additionally, the equal number of chromosomes and the divergent
341 number of bands could indicate different chromosomal arrangements and/or
342 different numbers of homeologs in each variety.

343 Finally, we performed BAC-FISH in the better metaphases of variety SP80-
344 3280 using Shy064N22 (haplotype VII) from Region01; 64 metaphases with some
345 signal of hybridization were obtained, and for the BAC-FISH of Shy048L15
346 (haplotype XI) from Region02, 69 were obtained. At least six metaphases for each
347 region were used to determine the number of signals. For BAC Shy064N22
348 Region01, eight signals could be counted (Figure 3 – Panel A), and for BAC
349 Shy048L15 in Region02, ten signals could be defined (Figure 3 – Panel B). These
350 results detail the numbers of haplotypes in sugarcane for Region01 and Region 02.
351 Moreover, the numbers of BAC haplotypes found in each region are appropriate
352 considering the BAC-FISH results, suggesting a missing haplotype for each region.

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

359

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

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

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

380 The mapping of the transcript reads in the *CENP-C* haplotypes from
381 Region02 revealed evidence of a chimerical gene (Figure 1, dotted rectangle and
382 Figure 4). The chimeric gene was formed by the first five exons of the sugarcane
383 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

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

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

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

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

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

451

452 **Genetic mapping**

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

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

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

483 For the genetic mapping, ten markers were used according to the
484 SuperMASSA best ploidy results. First, attempts were made to add each marker to
485 the existing linkage groups published by Balsalobre et al. (2017), but none of the
486 markers could be linked to the groups. Then, the markers were tested for linkage
487 among themselves. Two linkage groups could be created (Figure 6 – panel A) with
488 27.4 cM and 32.7 cM, respectively. The two linkage groups were too large;
489 therefore, the markers SugSNP_sh065 and SugSNP_sh099 were excluded, since
490 both markers were in the duplicated region (Figure 6 – panel B).

491 Then, attempts were made to add the remaining markers to the groups
492 again, and the marker SugSNP_sh005 was inserted into Linkage group 02 (Figure
493 6 – panel C). The markers that were in the wrong positions according to the
494 physical map (BACs) were also excluded, and the marker SugSNP_sh005 was
495 excluded from Linkage group 01 but remained in Linkage group 02 (Figure 6 –
496 panel C). Then, an attempt was made to form one linkage group with the remaining
497 markers by forcing OneMap to place the markers in a single group. Again, the size
498 of the group was too large (60.3 cM - Figure 6 – panel D). Therefore, the best
499 representation of the region was two linkage groups, Linkage group 01 with 2.1 cM,
500 and Linkage group 02 with 0 cM (Figure 6 – panel E).

501

502 **DISCUSSION**

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

506 Figueira et al., 2012; Garcia et al., 2013; de Setta et al., 2014; Xue et al., 2014;
507 Sun and Joyce, 2017; Vilela et al., 2017; Mancini et al., 2018; and others). The
508 high polyploidy in sugarcane cultivars make the detection of the ploidy of a locus a
509 challenge (Casu et al., 2012; Garcia et al., 2013; Xue et al., 2014; Sun and Joyce,
510 2017; and other).

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

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

530 The approach of mapping the BES in the *S. bicolor* genome, already
531 performed for other libraries (Figueira et al., 2012; Kim et al., 2013; Visendi et al.,
532 2016), revealed high synteny with the *S. bicolor* genome and a large number of
533 TEs in the sugarcane genome. Kim et al. (2013) showed BES anchorage of
534 approximately 6.4%, and Figueira et al. (2012) showed anchorage of
535 approximately 22%. Our data showed 10% BES anchorage in the sorghum
536 genome for both libraries constructed. These results are more similar to those of

537 Kim et al. (2013), since they used only BES \geq 300 bp, and we used BES \geq 100
538 bp.

539 The sugarcane genome has been reported to be composed of
540 approximately 40% TEs (Figueira et al., 2012; Kim et al., 2013; de Setta et al.,
541 2014). We also found that the average percentage of TEs was 40%, but this value
542 has a very large variance among the haplotypes, with a minimum of 21% and a
543 maximum of 65%. Figueira et al. (2012) and De Setta et al. (2014) also revealed
544 an inflation of the sugarcane genome in comparison with the *S. bicolor* genome.
545 De Setta et al. (2014) reported a very significant expansion that mainly occurred in
546 the intergenic and intronic regions and was primarily because of the presence of
547 TE, and we confirmed this report by comparing our data with data on the *S. bicolor*
548 genome. Several studies have reported a very significant sugarcane genome
549 expansion (Jannoo et al., 2007; Wang et al., 2010; Garsmeur et al., 2011; Figueira
550 et al., 2012; Setta et al., 2014; Vilela et al., 2017, Mancini et al., 2018).

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

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

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

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

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

591 BAC-FISH hybridization was used to indicate the ploidy of each region.
592 Eight signals were found for Region01 and 10 for Region02. These results are
593 highly consistent with the BAC haplotype and suggest that at least one BAC
594 haplotype is missing in each region. Casu et al. (2012), Xue et al. (2014) and Sun
595 and Joyce (2017) reported different methods to quantify the copy number of
596 endogenous gene, some of which resulted in odd copy numbers. Sun and Joyce
597 (2017) reported that the low or odd numbers could be explained by the contribution
598 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.

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

637 Numerous molecular mechanisms are involved in the creation of new
638 genes, such as exon shuffling, retrotransposons and gene duplications (reviewed
639 in Long et al., 2003). Gene fusions allow the physical coupling of functions, and
640 their occurrence in the genome increases with the genome size (Snel et al., 2000).
641 Sandmann et al. (2017) describes the function of the protein KNL2, which uses
642 *CENPC-k* motifs to bind DNA sequence independently and interacts with the
643 centromeric transcripts. The *CENPC* motif is characteristic of *CENP-C*. The
644 *CENPC* motif of the rat *CENP-C* protein can bind directly to a chimeric H3/cenH3
645 nucleosome *in vitro* suggesting that this motif binds to cenH3 nucleosomes *in vivo*.
646 Consequently, it is involved directly in cell division (Kato et al., 2013 and
647 Sandmann et al., 2017). The *CENPC* motifs described by Sandmann et al. (2017),
648 were compared with those of *CENP-C* genes in *A. thaliana*, *O. sativa*, *Z. mays* and
649 *S. bicolor* (Supplemental Figure 7). The *CENP-C* haplotypes from Region02
650 (chimerical gene) have the same *CENPC* motif as that in sorghum. The *CENP-C*
651 haplotypes from Region01 have one variation in the second residue of the *CENPC*
652 motif, which is a glycine in sorghum and a valine in *CENP-C* haplotypes from
653 Region01. This result suggests that the chimerical gene retained the ancestral
654 residue at this site, whereas a mutation occurred in *CENP-C* haplotypes from
655 Region01. Therefore, the mutation could have occurred in sorghum and in the
656 haplotypes from Region02, but this is unlikely. This result suggests that the *CENP-*
657 *C* haplotypes from Region01 and Region02 are able to bind to cenH3
658 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

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

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

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

711 The combination of divergent genomes within a hybrid can lead to
712 immediate, profound and highly varied genome modifications, which could include
713 chromosomal and molecular structural modifications (Shen et al., 2005; Doyle et
714 al., 2008; Soltis and Soltis, 2009; Jiang et al., 2011) as well as epigenetic changes
715 (Chen et al., 2010) and global transcriptomic changes (Hegarty et al., 2006; Buggs
716 et al., 2011). The integration of the genetic, genomic and transcriptomic data was
717 used to explain the interaction of the two regions in sugarcane. *HP600* is a
718 hypothetical gene that is next to the *CENP-C* gene, a kinetochore component
719 responsible for the initiation of nucleosomes. The sugarcane gene haplotypes of
720 *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

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

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

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

773

774 **METHODS**

775

776 **Plant material**

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

781

782 **BAC library construction and BAC-end analyses**

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

796 For the BES, 384 random BAC DNAs from each library were prepared with
797 the NucleoSpin® 96 Plasmid Core Kit (MACHEREY-NAGEL GmbH & Co., Düren,
798 Germany), according to the kit instructions. The sequencing reactions were
799 performed according to the manufacturer's instructions for the BigDye Terminator
800 Kit (Applied Biosystems, Foster City, CA, USA). The primers used in the reactions
801 were T7 Forward (5' TAATACGACTCACTATAGG 3') and M13 Reverse (5'
802 AACAGCTATGACCATG 3'). The PCR conditions were 95°C for 1 min followed by
803 90 cycles of 20 sec at 95°C, 20 sec at 50°C and 4 min at 60°C. The samples were
804 loaded on a 3730xl DNA Analyzer (Applied Biosystems). Sequence trimming was
805 conducted by processing the traces using the base-calling software PHRED
806 (Ewing and Green, 1998; Ewing et al., 1998), and reads with phred score < 20
807 were trimmed. The sequences were compared by using BLASTN in the *S. bicolor*
808 genome from Phytozome v10.1 (Goodstein et al., 2012). Only clones with forward
809 and reverse sequence maps in the *S. bicolor* genome, with a maximum distance of
810 600 kb and with no hits with repetitive elements, were used to anchor the *S. bicolor*
811 genome.

812

813 **Target gene determination**

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

826 The sugarcane transcripts were split into exons according to their annotation
827 in *S. bicolor*, *Z. mays* and *O. sativa*, and exon five was used to design the probe to
828 screen both BAC libraries (F: 5' ATCTGCTTCTTGGTGTTGCTG 3', R: 5'
829 GTCAGACACGATAGGTTTGTG 3'). DNA fragments were PCR-amplified from
830 sugarcane SP80-3280 and SPIAC93-3046 genomic DNA with specific primers
831 targeting the gene Sobic.003G221600. The PCR amplification conditions were
832 95°C for 8 min; 30 cycles of 20 sec denaturation at 95°C, 20 sec of annealing at
833 60°C, and a 40 sec extension at 72°C; and a final 10 min extension at 72°C. The
834 probes were sequenced before the screening of the BAC library.

835

836 **BAC library screening**

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

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

846 The SP80-3280 BAC library was used to construct a 3D pool. A total of
847 110,592 clones were pooled into 12 superpools following the protocol used by
848 Paux et al. (2008). The positive BAC clones from the SP80-3280 library were
849 isolated, and one isolated clone was validated by qPCR. The insert size of each
850 BAC was estimated by using an electrophoretic profile of NotI-digested BAC DNA
851 fragments and observed by PFGE (CHEF-DRIII system, Bio-Rad) in a 1% agarose
852 gel in 0.5× TBE buffer under the conditions described in Paiva et al. (2011).

853

854 **Sequencing and assembly**

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

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

867

868 **Sequence analysis and gene annotation**

869 All the BACs were aligned to verify the presence of redundant sequences of
870 homeologs. BAC clones with more than 99% similarity were considered the same
871 homeolog. BACs that represented the same homeologs were not combined. The
872 BACs were annotated with the gene prediction programs EUGENE (Foissac et al.,
873 2008) and Augustus (Keller et al., 2011). The BAC sequences were also searched
874 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 phylogenetic 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 6.5×10^{-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 1.3×10^{-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 1.3×10^{-8} mutations per site per year, as
903 described by Ma and Bennetzen (2004).

904

905 **Gene expression**

906 The transcriptomes of the sugarcane variety SP80-3280 from the roots, shoots and
907 stalks were mapped on *HP600* and *CENP-C* (NCBI SRR7274987), and the set of
908 transcripts was used for the transcription analyses. The reads from the sugarcane
909 transcriptomes were mapped to the reference genes with the Bowtie2 software
910 2.2.5 (Langmead and Salzberg, 2012) with default parameters; low-quality reads
911 and unmapped reads were filtered out (samtools -b -F 4); bam files were sorted
912 (samtools sort); and only mapped reads to the genes were extracted from the bam
913 files (samtools fastq) and recorded in a FASTQ format file. A haplotype was
914 considered to be expressed only when the transcript reads were mapped with
915 100% similarity. SNPs not found in the dataset were searched in the SP80-3280
916 transcriptomes from Vettore et al. (2003), Talbert et al. (2004) and Cardoso-Silva
917 et al. (2014) to verify the SNP presence in transcripts, but they were not used in
918 the expression analysis.

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

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

932 The frequency of the genomic data was used to test the transcriptome data
933 with R Studio (Rstudio team, 2015) and the exact binomial test (*binom.test* -
934 Clopper and Pearson, 1934, Conover, 1971 and Hollander and Wolfe, 1973). A p-
935 value ≥ 0.05 is equivalent to a 95% confidence interval for considering the
936 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 *C₀t-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).

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

974

975 **Genetic map construction**

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

983

984 **ACCESSION NUMBERS**

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

987 BAC sequences: from MH463467 to MH463488.

988 RNAseq subset data: SRR7274987

989

990 **SUPPLEMENTAL MATERIAL:**

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

992 Supplemental Figure 2. BAC BLASTN against sugarcane genome contigs.

993 Supplemental Figure 3. Phylogenetic and physical duplication representation.

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

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

996 Supplemental Figure 6. Mitotic metaphases of the sugarcane varieties.

997 Supplemental Figure 7. *CENP-C* motifs alignments.

998 Supplemental Table 1. BACs assembly and annotation.
999 Supplemental Table 2. Orthologous genes from Region01.
1000 Supplemental Table 3. Orthologous genes from Region02.
1001 Supplemental Table 4. Number of SNPs found in CENP-C and HP600.
1002 Supplemental Table 5. Number of SNPs found in duplicated region.
1003 Supplemental Table 6. Chromosome counts.
1004 Supplemental Table 7. Sequenom iPLEX MassARRAY® primers.

1005

1006 **ACKNOWLEDGMENTS**

1007 This study was supported by the São Paulo Research Foundation (FAPESP)
1008 (2008/52197-4) and Coordination for the Improvement of Higher Education
1009 Personnel (CAPES, Computational Biology Program). The first author was
1010 supported by a FAPESP Ph.D. fellowship from FAPESP (2010/50119-6). The third
1011 and fourth authors were supported by PD fellowships (MM 2014/11482-9 and CC-
1012 S 2015/16399-5). AP received a research fellowship from the National Council for
1013 Scientific and Technological Development (CNPq).

1014

1015 **AUTHOR CONTRIBUTIONS**

1016 APS, DAS, ERFM, HB, MV, and AAFG designed the study; AB, DAS, HH, JF, MC,
1017 MCM, MVRC, ND, NR and SV performed the research; CBC-S, DAS, GSP, MV,
1018 M-AVS and RV contributed new analytical/computational tools; AAFG, AB, APS,
1019 CBC-S, DAS, GSP, HB, LRP, MCM, MGAL, MS, MV, and SV analyzed the data;
1020 and DAS, MV and APS wrote the paper. All authors critically read the text and
1021 approved the manuscript.

1022

1023 **REFERENCES**

1024 **Balsalobre, T.W.A., da Silva Pereira, G., Margarido, G.R.A., Gazaffi, R.,**
1025 **Barreto, F.Z., Anoni, C.O., Cardoso-Silva, C.B., Costa, E.A., Mancini,**
1026 **M.C., Hoffmann, H.P., de Souza, A.P., Garcia, A.A., and Carneiro, M.S.**
1027 (2017). GBS-based single dosage markers for linkage and QTL mapping

- 1028 allow gene mining for yield-related traits in sugarcane. *BMC Genomics* **18**:
1029 72.
- 1030 **Balsalobre, T.W.A., Mancini, M.C., Pereira, G.d.S., Anoni, C.O., Barreto, F.Z.,**
1031 **Hoffmann, H.P., de Souza, A.P., Garcia, A.A.F., and Carneiro, M.S.**
1032 (2016). Mixed modeling of yield components and brown rust resistance in
1033 sugarcane families. *Agron. J.* **108**: 1824-1837.
- 1034 **Birchler, J.A., Riddle, N.C., Auger, D.L., and Veitia, R.A.** (2005). Dosage
1035 balance in gene regulation: biological implications. *Trends Genet.* **21**: 219-
1036 226.
- 1037 **Bremer, G.** (1961). Problems in breeding and cytology of sugarcane. *Euphytica*
1038 **10**: 59–78.
- 1039 **Buggs, R.J., Zhang, L., Miles, N., Tate, J.A., Gao, L., Wei, W., Schnable, P.S.,**
1040 **Barbazuk, W.B., Soltis, P.S., and Soltis, D.E.** (2011). Transcriptomic
1041 shock generates evolutionary novelty in a newly formed, natural
1042 allopolyploid plant. *Curr. Biol.* **21**: 551–556.
- 1043 **Cardoso-Silva, C.B., Costa, E.A., Mancini, M.C., Balsalobre, T.W.A., Canesin,**
1044 **L.E.C., Pinto, L.R., Carneiro, M.S., Garcia, A.A.F., de Souza, A.P., and**
1045 **Vicentini, R.** (2014). De novo assembly and transcriptome analysis of
1046 contrasting sugarcane varieties. *PLOS ONE* **9**: e88462 doi:
1047 10.1371/journal.pone.0088462.
- 1048 **Casu, R.E., Selivanova, A., and Perroux, J.M.** (2012). High-throughput
1049 assessment of transgene copy number in sugarcane using real-time
1050 quantitative PCR. *Plant Cell Rep.* **31**: 167–177.
- 1051 **Chalhoub, B., Belcram, H., and Caboche, M.** (2004). Efficient cloning of plant
1052 genomes into bacterial artificial chromosome (BAC) libraries with larger and
1053 more uniform insert size. *Plant Biotechnol. J.* **2**: 181–188.
- 1054 **Chen, F., He, G., He, H., Chen, W., Zhu, X., Liang, M., Chen, L., and Deng, X.W.**
1055 (2010). Expression analysis of miRNAs and highly-expressed small RNAs in
1056 two rice subspecies and their reciprocal hybrids. *J. Integr. Plant Biol.* **52**:
1057 971–980.

- 1058 **Clopper, C.J., and Pearson, E.S.** (1934). The use of confidence or fiducial limits
1059 illustrated in the case of the binomial. *Biometrika* **26**: 404–413.
- 1060 **Conover, J.W.** (1971) Practical nonparametric statistics. John Wiley & Sons, New
1061 York, Pp 97–104.
- 1062 **Cordeiro, G.M., Elliott, F., McIntyre, C.L., Casu, R.E., and Henry, R.J.** (2006).
1063 Characterisation of single nucleotide polymorphisms in sugarcane ESTs.
1064 *Theor. Appl. Genet.* **113**: 331-343.
- 1065 **Costa, E.A., Anoni, C.O., Mancini, M.C., Santos, F.R.C., Marconi, T.G., Gazaffi,**
1066 **R., Pastina, M.M., Perecin, D., Mollinari, M., Xavier, M.A., Pinto, L.R.,**
1067 **Souza, A.P., and Garcia, A.A.F.** (2016). QTL mapping including
1068 codominant SNP markers with ploidy level information in a sugarcane
1069 progeny. *Euphytica* **211**: 1–16.
- 1070 **Cuadrado, A., Acevedo, R., Moreno Díaz de la Espina, S., Jouve, N., and De**
1071 **La Torre, C.** (2004). Genome remodelling in three modern *S. officinarum* × *S.*
1072 *spontaneum* sugarcane cultivars. *J. Exp. Bot.* **55**: 847-854.
- 1073 **D'Hont, A.** (2005). Unraveling the genome structure of polyploids using FISH and
1074 GISH; examples of sugarcane and banana. *Cytogenet. Genome Res.* **109**:
1075 27-33.
- 1076 **D'Hont, A., and Glaszmann, J.C.** (2001). Sugarcane genome analysis with
1077 molecular markers, a first decade of research Proceedings of the Int Soc.
1078 Sugarcane Technol. 24, pp 556–559.
- 1079 **D'Hont, A., Ison, D., Alix, K., Roux, C., and Glaszmann, J.C.** (1998).
1080 Determination of basic chromosome numbers in the genus *Saccharum* by
1081 physical mapping of ribosomal RNA genes. *Genome* **41**: 221–225.
- 1082 **Daniels, J., and Roach, B.T.** (1987). Taxonomy and evolution. In: Heinz DJ (ed.).
1083 Sugarcane improvement through breeding, vol. 1. Elsevier, Amsterdam, pp
1084 7–84.
- 1085 **de Setta, N., Monteiro-Vitorello, C.B., Metcalfe, C.J., Cruz, G.M.Q., Del Bem,**
1086 **L.E., Vicentini, R., Nogueira, F.T.S., Campos, R.A., Nunes, S.L., Turrini,**
1087 **P.C.G., Vieira, A.P., Ochoa Cruz, E.A., Corrêa, T.C., Hotta, C.T., de Mello**
1088 **Varani, A., Vautrin, S., da Trindade, A.S., de Mendonça Vilela, M.,**

- 1089 **Lembke, C.G., Sato, P.M., de Andrade, R.F., Nishiyama, M.Y., Cardoso-**
1090 **Silva, C.B., Scortecci, K.C., Garcia, A.A., Carneiro, M.S., Kim, C.,**
1091 **Paterson, A.H., Bergès, H., D’Hont, A., de Souza, A.P., Souza, G.M.,**
1092 **Vincentz, M., Kitajima, J.P., and Van Sluys, M.A.** (2014). Building the
1093 Sugarcane genome for biotechnology and identifying evolutionary trends.
1094 *Genomics* **15**: 540.
- 1095 **Doyle, J.J., Flagel, L.E., Paterson, A.H., Rapp, R.A., Soltis, D.E., Soltis, P.S.,**
1096 **and Wendel, J.F.** (2008). Evolutionary genetics of genome merger and
1097 doubling in plants. *Annu. Rev. Genet.* **42**: 443–461.
- 1098 **Ewing, B., and Green, P.** (1998). Base-calling of automated sequencer traces
1099 using Phred. II. Error probabilities. *Genome Res.* **8**: 186–194.
- 1100 **Ewing, B., Hillier, L., Wendl, M.C., and Green, P.** (1998). Base-calling of
1101 automated sequencer traces using Phred. I. Accuracy assessment. *Genome*
1102 *Res.* **8**: 175–185.
- 1103 **Felsenstein, J.** (1985). Confidence limits on phylogenies: an approach using the
1104 bootstrap. *Evolution* **39**: 783-791.
- 1105 **Feltus, F.A., Wan, J., Schulze, S.R., Estill, J.C., Jiang, N., and Paterson, A.H.**
1106 (2004). An SNP resource for rice genetics and breeding based on
1107 subspecies indica and japonica genome alignments. *Genome Res.* **14**:
1108 1812–1819.
- 1109 **Figueira, T.R., Okura, V., Rodrigues da Silva, F., Jose da Silva, M., Kudrna, D.,**
1110 **Ammiraju, J.S., Talag, J., Wing, R., and Arruda, P.** (2012). A BAC library
1111 of the SP80-3280 sugarcane variety (*Saccharum* sp.) and its inferred
1112 microsynteny with the sorghum genome. *BMC Res. Notes* **5**: 185.
- 1113 **Foissac, S., Gouzy, J., Rombauts, S., Mathé, C., Amselem, J., Sterck, L.,**
1114 **Veau, de Peer, Y.V., Rouze, P., and Schiex, T.** (2008). Genome
1115 annotation in plants and fungi: EuGene as a model platform. *Curr.*
1116 *Bioinform.* **3**: 87–97.
- 1117 **Freeling, M.** (2017). Picking up the ball at the K/Pg boundary: the distribution of
1118 ancient polyploidies in the plant phylogenetic tree as a spandrel of
1119 asexuality with occasional sex. *Plant Cell* **29**: 202-206.

- 1120 **Freeling, M., Lyons, E., Pedersen, B., Alam, M., Ming, R., and Lisch, D. (2008).**
1121 Many or most genes in Arabidopsis transposed after the origin of the order
1122 Brassicales. *Genome Res.* **18**: 1924–1937.
- 1123 **Garcia, A.A., Mollinari, M., Marconi, T.G., Serang, O.R., Silva, R.R., Vieira,**
1124 **M.L., Vicentini, R., Costa, E.A., Mancini, M.C., Garcia, M.O., Pastina,**
1125 **M.M., Gazaffi, R., Martins, E.R., Dahmer, N., Sforça, D.A., Silva, C.B.,**
1126 **Bundock, P., Henry, R.J., Souza, G.M., van Sluys, M.A., Landell, M.G.,**
1127 **Carneiro, M.S., Vincentz, M.A., Pinto, L.R., Vencovsky, R., and Souza,**
1128 **A.P. (2013).** SNP genotyping allows an in-depth characterisation of the
1129 genome of sugarcane and other complex autopolyploids. *Sci. Rep.* **3**: 3399.
- 1130 **Garsmeur, O., Charron, C., Bocs, S., Jouffe, V., Samain, S., Couloux, A., Droc,**
1131 **G., Zini, C., Glaszmann, J.C., Van Sluys, M.A., and D’Hont, A. (2011).**
1132 High homologous gene conservation despite extreme autopolyploid
1133 redundancy in sugarcane. *New Phytol.* **189**: 629-642.
- 1134 **Gaut, B.S., and Doebley, J.F. (1997).** DNA sequence evidence for the segmental
1135 allotetraploid origin of maize. *Proc. Natl Acad. Sci. U.S.A.* **94**: 6809–6814.
- 1136 **Gaut, B.S., Morton, B.R., McCaig, B.C., and Clegg, M.T. (1996).** Substitution
1137 rate comparisons between grasses and palms: synonymous rate differences
1138 at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcl*.
1139 *Proc. Natl Acad. Sci. U.S.A.* **93**: 10274-10279.
- 1140 **Gonthier, L., Bellec, A., Blassiau, C., Prat, E., Helmstetter, N., Rambaud, C.,**
1141 **Huss, B., Hendriks, T., Bergès, H., and Quillet, M.C. (2010).** Construction
1142 and characterization of two BAC libraries representing a deep-coverage of
1143 the genome of chicory. (*Cichorium intybus* L., Asteraceae). *BMC Res. Notes*
1144 **3**: 225.
- 1145 **Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J.,**
1146 **Mitros, T., Dirks, W., Hellsten, U., Putnam, N., and Rokhsar, D.S. (2012).**
1147 *Phytozome*: a comparative platform for green plant genomics. *Nucleic Acids*
1148 *Res.* **40**: D1178–D1186.
- 1149 **Gopalakrishnan, S., Sullivan, B.A., Trazzi, S., Della Valle, G., and Robertson,**
1150 **K.D. (2009).** DNMT3B interacts with constitutive centromere protein *CENP-*

- 1151 C to modulate DNA methylation and the histone code at centromeric
1152 regions. *Hum. Mol. Genet.* **18**: 3178–3193.
- 1153 **Gordon, D.** (2003), 11.2.1-11.2.43. Viewing and editing assembled sequences
1154 using Consed. *Curr. Protoc. Bioinformatics*. In: Baxevanis AD, Davison DB
1155 (eds). John Wiley & Co., New York, pp 11.2.1-11.2.43.
- 1156 **Gordon, D., Abajian, C., and Green, P.** (1998). Consed: a graphical tool for
1157 sequence finishing. *Genome Res.* **8**: 195-202.
- 1158 **Gordon, D., Desmarais, C., and Green, P.** (2001). Automated finishing with
1159 Autofinish. *Genome Res.* **11**: 614-625.
- 1160 **Grivet, L., and Arruda, P.** (2002). Sugarcane genomics: depicting the complex
1161 genome of an important tropical crop. *Curr. Opin. Plant Biol.* **5**: 122-127.
- 1162 **Guerra, M.** (1983). O uso de Giemsa em citogenética vegetal – comparação entre
1163 a coloração simples e o bandamento. *Cienc. Cult.* **35**: 190-193.
- 1164 **Guerra, M., and Souza, M.J.** (2002) Como observar cromossomos: um guia de
1165 técnicas em citogenética vegetal, animal e humana. FUNPEC, Ribeirão
1166 Preto, SP, p 131.
- 1167 **Guindon, S., and Gascuel, O.** (2003). A simple, fast, and accurate algorithm to
1168 estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**: 696-704.
- 1169 **Hegarty, M.J., Barker, G.L., Wilson, I.D., Abbott, R.J., Edwards, K.J., and**
1170 **Hiscock, S.J.** (2006). Transcriptome shock after interspecific hybridization
1171 in senecio is ameliorated by genome duplication. *Curr. Biol.* **16**: 1652–1659.
- 1172 **Hollander, M., and Wolfe, D.A.** (1973) Nonparametric statistical methods. John
1173 Wiley & Sons, New York, pp 15–22.
- 1174 **Ilic, K., SanMiguel, P.J., and Bennetzen, J.L.** (2003). A complex history of
1175 rearrangement in an orthologous region of the maize, sorghum, and rice
1176 genomes. *Proc. Natl Acad. Sci. U.S.A.* **100**: 12265-12270.
- 1177 **Irvine, J.E.** (1999). *Saccharum* species as horticultural classes. *TAG Theoretical*
1178 *and Applied Genetics* **98**: 186–194.
- 1179 **Jannoo, N., Grivet, L., Chantret, N., Garsmeur, O., Glaszmann, J.C., Arruda,**
1180 **P., and D’Hont, A.** (2007). Orthologous comparison in a gene-rich region

- 1181 among grasses reveals stability in the sugarcane polyploid genome. *Plant J.*
1182 **50**: 574-585.
- 1183 **Jiang, B., Lou, Q., Wu, Z., Zhang, W., Wang, D., Mbira, K.G., Weng, Y., and**
1184 **Chen, J.** (2011). Retrotransposon- and microsatellite sequence-associated
1185 genomic changes in early generations of a newly synthesized allotetraploid
1186 *cucumis* × *hytivus* Chen & Kirkbride. *Plant Mol. Biol.* **77**: 225–233.
- 1187 **Jukes, T.H., and Cantor, C.R.** (1969) Evolution of protein molecules. Academic
1188 Press, New York, pp 21–132.
- 1189 **Kato, H., Jiang, J., Zhou, B.R., Rozendaal, M., Feng, H., Ghirlando, R., Xiao,**
1190 **T.S., Straight, A.F., and Bai, Y.** (2013). A conserved mechanism for
1191 centromeric nucleosome recognition by centromere protein *CENP-C*.
1192 *Science* **340**: 1110-1113.
- 1193 **Katoh, K., Misawa, K., Kuma, K., and Miyata, T.** (2002). MAFFT: a novel method
1194 for rapid multiple sequence alignment based on fast Fourier transform.
1195 *Nucleic Acids Res.* **30**: 3059-3066.
- 1196 **Keller, O., Kollmar, M., Stanke, M., and Waack, S.** (2011). A novel hybrid gene
1197 prediction method employing protein multiple sequence alignments.
1198 *Bioinformatics* **27**: 757-763.
- 1199 **Kellogg, E.A.** (2001). Evolutionary history of the grasses. *Plant Physiol.* **125**:
1200 1198–1205.
- 1201 **Kim, C., Guo, H., Kong, W., Chandnani, R., Shuang, L.S., and Paterson, A.H.**
1202 (2016). Application of genotyping by sequencing technology to a variety of
1203 crop breeding programs. *Plant Sci.* **242**: 14-22.
- 1204 **Kim, C., Lee, T.H., Compton, R.O., Robertson, J.S., Pierce, G.J., and Paterson,**
1205 **A.H.** (2013). A genome-wide BAC end-sequence survey of sugarcane
1206 elucidates genome composition, and identifies BACs covering much of the
1207 euchromatin. *Plant Mol. Biol.* **81**: 139–147.
- 1208 **Kim, M. Y., and Zilberman, D.** (2014). DNA methylation as a system of plant
1209 genomic immunity. *Trends Plant Sci.* **19**: 320-326.

- 1210 **Kimura, M.** (1980). A simple method for estimating evolutionary rates of base
1211 substitutions through comparative studies of nucleotide sequences. *J. Mol.*
1212 *Evol.* **16**: 111-120.
- 1213 **Kohany, O., Gentles, A.J., Hankus, L., and Jurka, J.** (2006). Annotation,
1214 submission and screening of repetitive elements in Repbase:
1215 RepbaseSubmitter and Censor. *BMC Bioinform.* **7**: 474.
- 1216 **Kumar, S., Stecher, G., and Tamura, K.** (2016). MEGA7: Molecular Evolutionary
1217 Genetics Analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33**: 1870-
1218 1874.
- 1219 **Lai, J., Ma, J., Swigonová, Z., Ramakrishna, W., Linton, E., Llaca, V.,**
1220 **Tanyolac, B., Park, Y.J., Jeong, O.Y., Bennetzen, J.L., and Messing, J.**
1221 (2004). Gene loss and movement in the maize genome. *Genome Res.* **14**:
1222 1924-1931.
- 1223 **Landell, M.G.A., Campana, M.P., Figueiredo, P., Vasconcelos ACM, Xavier**
1224 **MA, Bidoia MAP, Prado H, Silva MA, and Miranda LLD** (2005)
1225 Variedades de cana-de-açúcar pad'ra o centro sul do Brasil. Technical
1226 Bulletin IAC **197**: 33.
- 1227 **Langmead, B., and Salzberg, S.L.** (2012). Fast gapped-read alignment with
1228 Bowtie 2. *Nat. Methods* **9**: 357-359.
- 1229 **Le Cunff, L., Garsmeur, O., Raboin, L.M., Pauquet, J., Telismart, H., Selvi, A.,**
1230 **Grivet, L., Philippe, R., Begum, D., Deu, M., Costet, L., Wing, R.,**
1231 **Glaszmann, J.C., and D'Hont, A.** (2008). Diploid/polyploid syntenic shuttle
1232 mapping and haplotype-specific chromosome walking toward a rust
1233 resistance gene (Bru1) in highly polyploid sugarcane (2n approximately 12x
1234 approximately 115). *Genetics* **180**: 649-660.
- 1235 **Long, M., Betrán, E., Thornton, K., and Wang, W.** (2003). The origin of new
1236 genes: glimpses from the young and old. *Nat. Rev. Genet.* **4**: 865–875.
- 1237 **Lynch, M., and Force, A.G.** (2000). The origin of interspecific genomic
1238 incompatibility via gene duplication. *Am. Nat.* **156**: 590–605.
- 1239 **Ma, J., and Bennetzen, J.L.** (2004). Rapid recent growth and divergence of rice
1240 nuclear genomes. *Proc. Natl Acad. Sci. U.S.A.* **101**: 12404–12410.

- 1241 **Mace, E.S., and Jordan, D.R.** (2011). Integrating sorghum whole genome
1242 sequence information with a compendium of sorghum QTL studies reveals
1243 uneven distribution of QTL and of gene-rich regions with significant
1244 implications for crop improvement. *Theor. Appl. Genet.* **123**: 169–191.
- 1245 **Mancini, M.C., Cardoso-Silva, C.B., Sforça, D.A., and Souza, A.P.** (2018).
1246 “Targeted sequencing by gene synteny,” a new strategy for polyploid
1247 species: sequencing and physical structure of a complex sugarcane region.
1248 *Front. Plant Sci.* **9**: 397.
- 1249 **Mattiello, L., Riaño-Pachón, D.M., Martins, M.C., da Cruz, L.P., Bassi, D.,**
1250 **Marchiori, P.E., Ribeiro, R.V., Labate, M.T., Labate, C.A., and Menossi,**
1251 **M.** (2015). Physiological and transcriptional analyses of developmental
1252 stages along sugarcane leaf. *BMC Plant Biol.* **15**: 300.
- 1253 **McClintock, B.** (1984). The significance of responses of the genome to challenge.
1254 *Science* **226**: 792–801.
- 1255 **Ming, R., Liu, S.C., Lin, Y.R., da Silva, J., Wilson, W., Braga, D., van Deynze,**
1256 **A., Wenslaff, T.F., Wu, K.K., Moore, P.H., Burnquist, W., Sorrells, M.E.,**
1257 **Irvine, J.E., and Paterson, A.H.** (1998). Detailed alignment of saccharum
1258 and sorghum chromosomes: comparative organization of closely related
1259 diploid and polyploid genomes. *Genetics* **150**: 1663–1682.
- 1260 **Mosccone, E.A., Matzke, M.A., and Matzke, A.J.M.** (1996). The use of combined
1261 FISH/GISH in conjunction with DAPI counterstaining to identify
1262 chromosomes containing transgene inserts in amphidiploid tobacco.
1263 *Chromosoma* **105**: 231-236.
- 1264 **Murray, S.C., Sharma, A., Rooney, W.L., Klein, P.E., Mullet, J.E., Mitchell, S.E.,**
1265 **and Kresovich, S.** (2008). Genetic improvement of Sorghum as a biofuel
1266 feedstock: I. QTL for stem sugar and grain nonstructural carbohydrates.
1267 *Crop Sci.* **48**: 2165-2179.
- 1268 **Nishiyama Jr, M. Y., Ferreira, S. S., Tang, P. Z., Becker, S., Poertner-Taliana,**
1269 **A., and Souza, G. M.** (2014). Full-length enriched cDNA libraries and
1270 ORFeome analysis of sugarcane hybrid and ancestor genotypes. *PloS one,*
1271 **9**: e107351.

- 1272 **Ohno, S.** (1970) Evolution by gene duplication. Springer-Verlag, New York.
- 1273 **Paiva, J.A.P., Prat, E., Vautrin, S., Santos, M.D., San-Clemente, H.,**
1274 **Brommonschenkel, S., Fonseca, P.G., Grattapaglia, D., Song, X.,**
1275 **Ammiraju, J.S., Kudrna, D., Wing, R.A., Freitas, A.T., Bergès, H., and**
1276 **Grima-Pettenati, J.** (2011). Advancing Eucalyptus genomics: identification
1277 and sequencing of lignin biosynthesis genes from deep-coverage BAC
1278 libraries. *BMC Genomics* **12**: 137. doi: 10.1186/1471-2164-12-137.
- 1279 **Paterson, A.H., Moore, P.H., and Tew, T.L.** (2013). The gene pool of saccharum
1280 species and their improvement. In: Paterson A (ed.). *Genomics of the*
1281 *Saccharinae*, Springer, Berlin, pp 43–72.
- 1282 **Paux, E., Sourdille, P., Salse, J., Saintenac, C., Choulet, F., Leroy, P., Korol,**
1283 **A., Michalak, M., Kianian, S., Spielmeier, W., Lagudah, E., Somers, D.,**
1284 **Kilian, A., Alaux, M., Vautrin, S., Bergès, H., Eversole, K., Appels, R.,**
1285 **Safar, J., Simkova, H., Dolezel, J., Bernard, M., and Feuillet, C.** (2008). A
1286 physical map of the 1-gigabase bread wheat chromosome 3b. *Science* **322**:
1287 101–104. doi: 10.1126/science.1161847.
- 1288 **Peterson, D.G., Tomkins, J.P., Frisch, D.A., Wing, R.A., and Paterson, A.H.**
1289 (2000). Construction of plant bacterial artificial chromosome (BAC) libraries:
1290 an illustrated guide. *J. Agric. Genomics* **5**: 1–100.
- 1291 **Piperidis, G., and D’Hont, A.** (2001). Chromosome composition analysis of
1292 various *Saccharum* interspecific hybrids by genomic in situ hybridization
1293 (GISH) *Proceedings of the Int Soc. Sugcane Technol.* 24, pp 565–566.
- 1294 **Piperidis, G., Piperidis, N., and D’Hont, A.** (2010). Molecular cytogenetic
1295 investigation of chromosome composition and transmission in sugarcane.
1296 *Mol. Genet. Genomics* **284**: 65-73.
- 1297 **Ramsey, J., and Schemske, D.W.** (2002). Neopolyploidy in flowering plants.
1298 *Annu. Rev. Ecol. Syst.* **33**: 589-639.
- 1299 **Riaño-Pachón, D.M., and Mattiello, L.** (2017). Draft genome sequencing of the
1300 sugarcane hybrid SP80-3280. *F1000Res* **6**: 861.
- 1301 **Roselli, S., Olry, A., Vautrin, S., Coriton, O., Ritchie, D., Galati, G., Navrot, N.,**
1302 **Krieger, C., Vialart, G., Bergès, H., Bourgaud, F., and Hehn, A.** (2017). A

- 1303 bacterial artificial chromosome (BAC) genomic approach reveals partial
1304 clustering of the furanocoumarin pathway genes in parsnip. *Plant J.* **89**:
1305 1119-1132. doi: 10.1111/tpj.13450.
- 1306 **RStudio Team** (2015). RStudio: integrated development for R. RStudio, Inc.,
1307 Boston, MA. <http://www.rstudio.com/>.
- 1308 **Saitou, N., and Nei, M.** (1987). The neighbor-joining method: a new method for
1309 reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**: 406-425.
- 1310 **Sandmann, M., Talbert, P., Demidov, D., Kuhlmann, M., Rutten, T., Conrad, U.,
1311 and Lermontova, I.** (2017). Targeting of arabidopsis KNL2 to centromeres
1312 depends on the conserved CENPC-k motif in Its C terminus. *Plant Cell* **29**:
1313 144–155. doi: 10.1105/tpc.16.00720.
- 1314 **SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y., and Bennetzen, J.L.**
1315 (1998). The paleontology of intergene retrotransposons of maize. *Nat.*
1316 *Genet.* **20**: 43-45.
- 1317 **Saze, H., Kitayama, J., Takashima, K., Miura, S., Harukawa, Y., Ito, T., and
1318 Kakutani, T.** (2013). Mechanism for full-length RNA processing of
1319 Arabidopsis genes containing intragenic heterochromatin. *Nat. Commun.* **4**:
1320 2301.
- 1321 **Schwarzacher, T., Ambros, P., and Schweizer, D.** (1980). Application of Giemsa
1322 banding to orchid karyotype analysis. *Plant Syst. Evol.* **134**: 293-297.
- 1323 **Schwarzacher, T., and Heslop-Harrison, P.** (2000). Practical in situ hybridization.
1324 BIOS Scientific.
- 1325 **Serang, O., Mollinari, M., Garcia, A.A.F.** (2012) Efficient exact maximum a
1326 posteriori computation for bayesian SNP genotyping in polyploids. *PLoS*
1327 *One* **7**: e30906.
- 1328 **Shen, Y., Lin, X.Y., Shan, X.H., Lin, C.J., Han, F.P., Pang, J.S., and Liu, B.**
1329 (2005). Genomic rearrangement in endogenous long terminal repeat
1330 retrotransposons of rice lines introgressed by wild rice (*Zizania latifolia*
1331 Griseb.). *J. Integr. Plant Biol.* **47**: 998–1008.
- 1332 **Shi, X., Zhang, C., Ko, D.K., and Chen, Z.J.** (2015). Genome-wide dosage-
1333 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.,**

- 1365 **Grivet, L., Di Mauro, S.M., Nobrega, F.G., Menck, C.F., Braga, M.D.,**
1366 **Telles, G.P., Cara, F.A., Pedrosa, G., Meidanis, J., and Arruda, P. (2003).**
1367 Analysis and functional annotation of an expressed sequence tag collection
1368 for tropical crop sugarcane. *Genome Res.* **13**: 2725–2735.
- 1369 **Vieira, M.L.C., Almeida, C.B., Oliveira, C.A., Tacuatiá, L.O., Munhoz, C.F.,**
1370 **Cauz-Santos, L.A., Pinto, L.R., Monteiro-Vitorello, C.B., Xavier, M.A.,**
1371 **and Forni-Martins, E.R. (2018).** Revisiting meiosis in sugarcane:
1372 chromosomal irregularities and the prevalence of bivalent configurations.
1373 *Front. Gen.* **9**: 213.
- 1374 **Vilela, M.M., Del Bem, L.E., Van Sluys, M.A., de Setta, N., Kitajima, J.P., Cruz,**
1375 **G.M.Q., Sforça, D.A., de Souza, A.P., Ferreira, P.C.G., Grativol, C.,**
1376 **Cardoso-Silva, C.B., Vicentini, R., and Vincentz, M. (2017).** Analysis of
1377 three sugarcane homo/homeologous regions suggests independent
1378 polyploidization events of *Saccharum officinarum* and *Saccharum*
1379 *spontaneum*. *Genome Biol. Evol.* **9**: 266–278.
- 1380 **Visendi, P., Berkman, P.J., Hayashi, S., Golicz, A.A., Bayer, P.E., Ruperao, P.,**
1381 **Hurgobin, B., Montenegro, J., Chan, C.K., Staňková, H., Batley, J.,**
1382 **Šimková, H., Doležel, J., and Edwards, D. (2016).** An efficient approach to
1383 BAC based assembly of complex genomes. *Plant Methods* **12**: 2-2. doi:
1384 10.1186/s13007-016-0107-9.
- 1385 **Wang, J., Roe, B., Macmil, S., Yu, Q., Murray, J.E., Tang, H., Chen, C., Najar,**
1386 **F., Wiley, G., Bowers, J., Van Sluys, M.A., Rokhsar, D.S., Hudson, M.E.,**
1387 **Moose, S.P., Paterson, A.H., and Ming, R. (2010).** Microcollinearity
1388 between autopolyploid sugarcane and diploid sorghum genomes. *BMC*
1389 *Genomics* **11**: 261.
- 1390 **Wang, X., Duan, C.G., Tang, K., Wang, B., Zhang, H., Lei, M., Lu, K.,**
1391 **Mangrauthia, S.K., Wang, P., Zhu, G., Zhao, Y., and Zhu, J.K. (2013).**
1392 RNA-binding protein regulates plant DNA methylation by controlling mRNA
1393 processing at the intronic heterochromatin-containing gene IBM1. *Proc.*
1394 *Natl. Acad. Sci. U.S.A.* **110**: 15467-15472.

- 1395 **Woodhouse, M.R., Schnable, J.C., Pedersen, B.S., Lyons, E., Lisch, D.,**
1396 **Subramaniam, S., and Freeling, M.** (2010). Following tetraploidy in maize,
1397 a short deletion mechanism removed genes preferentially from one of the
1398 two homologs. *PLoS Biol.* **8**: e1000409.
- 1399 **Wu, R., Ma, C.X., Painter, I., and Zeng, Z.B.** (2002). Simultaneous maximum
1400 likelihood estimation of linkage and linkage phases in outcrossing species.
1401 *Theor. Popul. Biol.* **61**: 349–363.
- 1402 **Xu, Z., and Wang, H.** (2007). LTR_FINDER: an efficient tool for the prediction of
1403 full-length LTR retrotransposons. *Nucleic Acids Res.* **35**: W265–W268 doi:
1404 10.1093/nar/gkm286.
- 1405 **Xue, B.T., Guo, J.L., Que, Y.X., Fu, Z.W., Wu, L.G., and Xu, L.P.** (2014).
1406 Selection of suitable endogenous reference genes for relative copy number
1407 detection in sugarcane. *Int. J. Mol. Sci.* **15**: 8846–8862.
- 1408 **Zwick, M.S., Hanson, R.E., Islam-Faridi, M.N., Stelly, D.M., Wing, R.A., Price,**
1409 **H.J., and McKnight, T.D.** (1997). A rapid procedure for the isolation of C0t-
1410 1 DNA from plants. *Genome* **40**: 138–142.
- 1411

1412 **FIGURE LEGENDS**

1413 **Figure 1. Schematic representation of the sugarcane BAC haplotypes from**
1414 **Region01 and Region02.** Squares of the same color represent sugarcane genes
1415 orthologous to *Sorghum bicolor* genes. Dotted lines connect the homologous
1416 genes in sugarcane at different positions. In sugarcane Region02, the *CENP-C*
1417 haplotypes in Region02 are represented by two squares (blue and pink), where
1418 each square represents a partial gene fusion. The dark gray strip represents the
1419 shared region from Region01 and Region02 (duplication). The genes in light gray
1420 (from *S. bicolor*) are not found in the sugarcane BACs. The representation is not to
1421 scale. The orientation of transcription is indicated by the direction of the arrow at
1422 the end of each gene.

1423 **Figure 2. Representation of each sugarcane BAC from Region01 and**
1424 **Region02.** Arrows and rectangles of the same color represent the homologous
1425 genes in sugarcane. Black rectangles represent repeat regions. Yellow lines
1426 represent gaps. Similar regions are represented by a gray shadow connecting the
1427 BACs. The orientation of transcription is indicated by the direction of the arrow at
1428 the end of each gene. Scale representation.

1429 **Figure 3. FISH hybridization of the sugarcane BACs. Panel (A):** BAC
1430 Shy065N22 hybridization in sugarcane variety SP-803280 mitosis showing eight
1431 signals for Region01. **Panel (B):** BAC Shy048L15 hybridization in sugarcane
1432 variety SP-803280 mitosis showing ten signals for Region02.

1433 **Figure 4. Fusion gene formation of *CENP-C* and Sobic003G299500. Panel (A):**
1434 *Sorghum CENP-C* and Sobic003G299500 genome location. **Panel (B):** Sugarcane
1435 genomic *CENP-C* haplotypes in Region01 (all expressed). **Panel (C):** Partially
1436 duplicated sugarcane paralogs of *CENP-C* and Sobic003G299500 haplotypes in
1437 Region02 (only haplotypes XI/XII/XIII and haplotype XIV have evidence of
1438 expression). **Panel (D):** Sugarcane ortholog of Sobic003G299500 found in the
1439 sugarcane R570 BAC library. **Panel (E):** Transcripts from sugarcane SP80-3280
1440 mapped against the CDS of sugarcane *CENP-C* haplotypes from Region01. **Panel**
1441 **(F):** Transcripts from sugarcane SP80-3280 mapped against the sugarcane
1442 chimerical paralogs of *CENP-C* and Sobic003G299500. As evidence of fusion

1443 gene formation, the transcripts show the fusion point of the paralogs. **Panel (G):**
1444 Transcripts from sugarcane SP80-3280 mapped against the CDS of the sugarcane
1445 R570 ortholog of Sobic003G299500.
1446 Figure 5. Ploidy and dosage in the sugarcane genomic DNA (BACs) and
1447 SuperMASSA estimation. The location of each SNP is shown by one haplotype
1448 from Region01 and one haplotype from Region02. “SuperMASSA Best Ploidy”
1449 means the SuperMASSA best ploidy with a posteriori probability of >0.8.
1450 “SuperMASSA Expected Ploidy” means we fixed the ploidy of the loci in
1451 SuperMASSA according to the BAC-FISH and BAC sequencing results. “Genomic
1452 Ploidy” means the ploidy of the loci according to the BAC-FISH and BAC
1453 sequencing results. “*” means the SNP was found only in the transcriptome.
1454 **Figure 6. Linkage map for the duplicated region.** Schematic representation of a
1455 multiple sugarcane linkage map for sugarcane variety SP80-3280 with information
1456 about the sugarcane genome (BACs).
1457
1458
1459

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 .

SNP	Name	Change	Polymorphism Type	Position	Coverage	Variant Coverage	Genomic Detected	Transcriptome Proportion	Missing haplotype for more common SNP				Missing haplotype for variant SNP			
									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

1466 **Table 2.** Genomic frequencies of the SNPs in the *CENP-C* haplotypes in Region01 and Region02 in the genome and
 1467 transcriptome. The global expression (in diverse tissues) was used to determine whether the genomic frequency could
 1468 explain the transcription frequency (H_0). The binomial test was used to verify H_0 . The highlighted p-values reflect the
 1469 acceptance of H_0 .

SNP	Name	Change	Polymorphism Type	Position	Coverage	Variant Coverage	Genomic Detected	Transcriptome Proportion	Missing haplotype for more common SNP				Missing haplotype for variant SNP			
									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

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

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

Sorghum bicolor orthologue Chr. 06

Sorghum bicolor orthologue Chr. 04

Sorghum bicolor orthologue Chr. 08

Sorghum bicolor orthologue Chr. 03

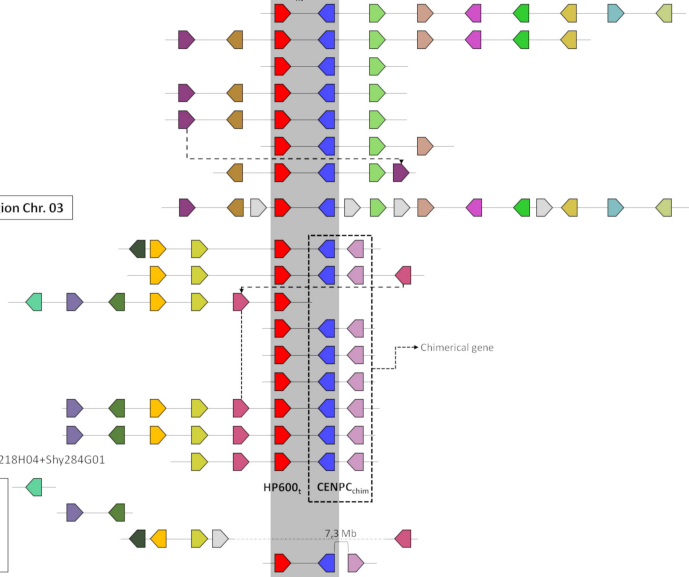
Orthologues



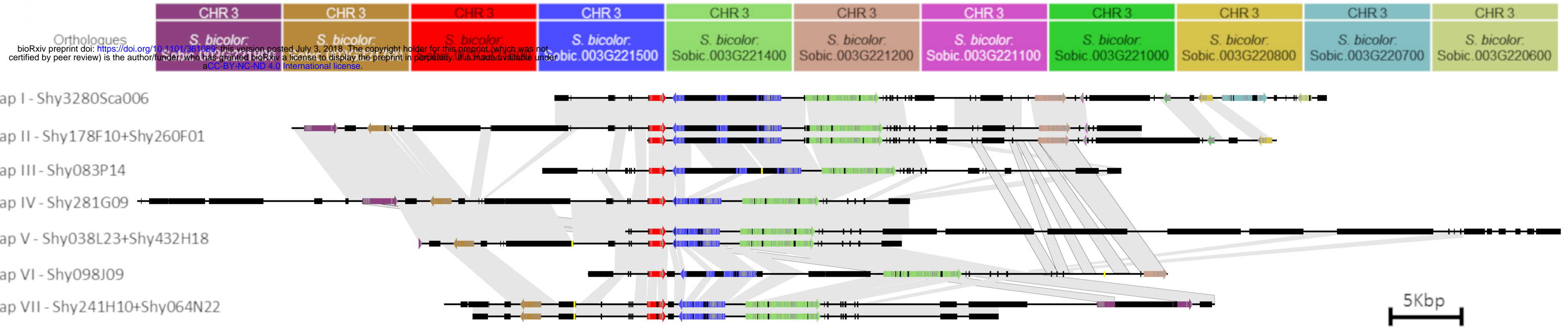
Orthologues



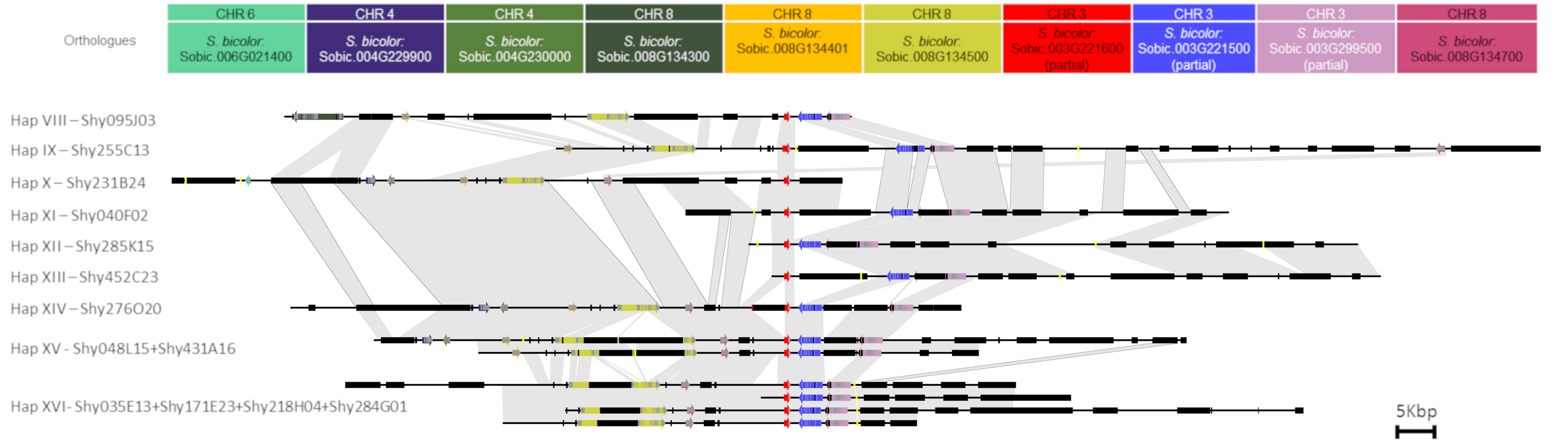
HP600_{nt} CENPC



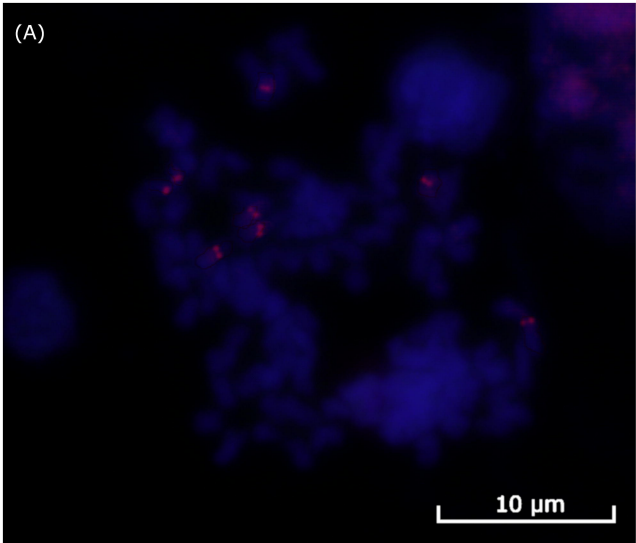
Region01



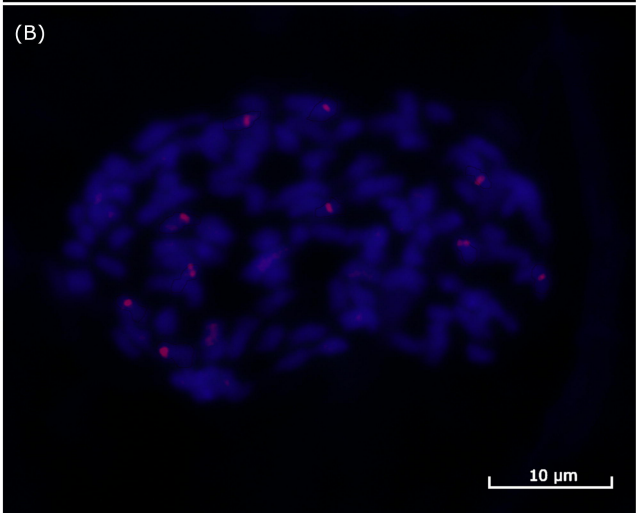
Region02

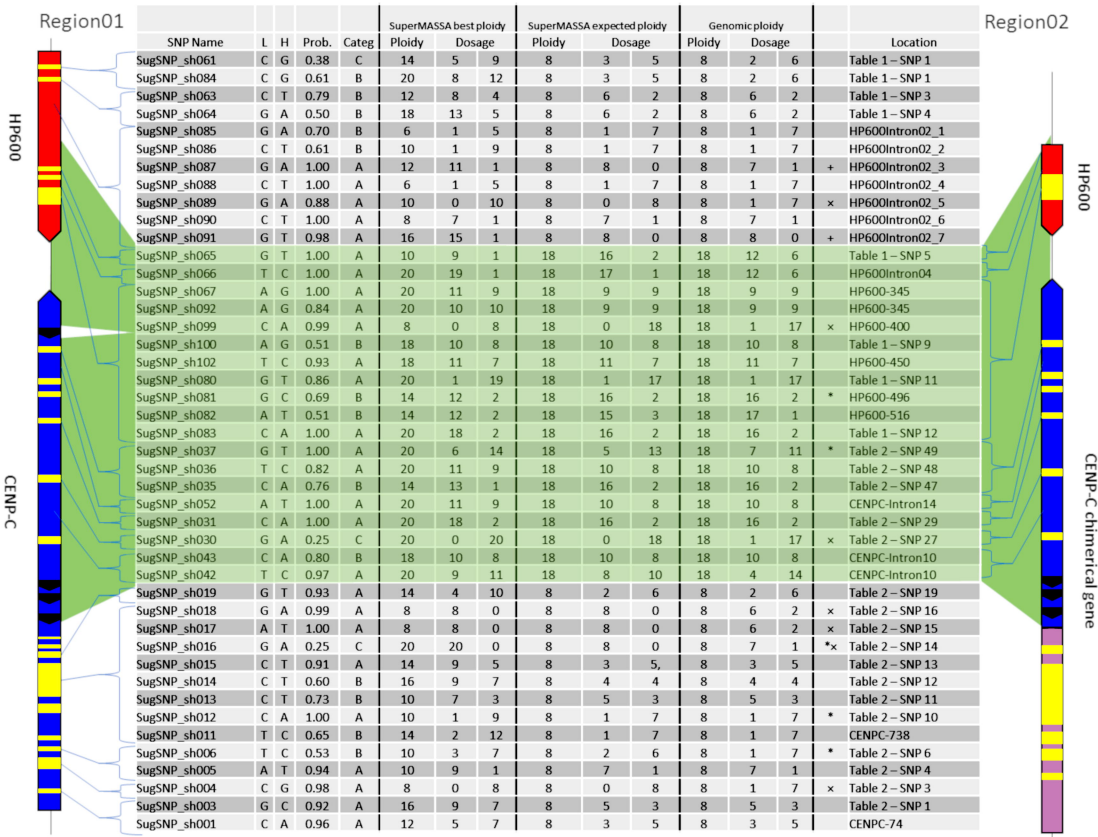


(A)

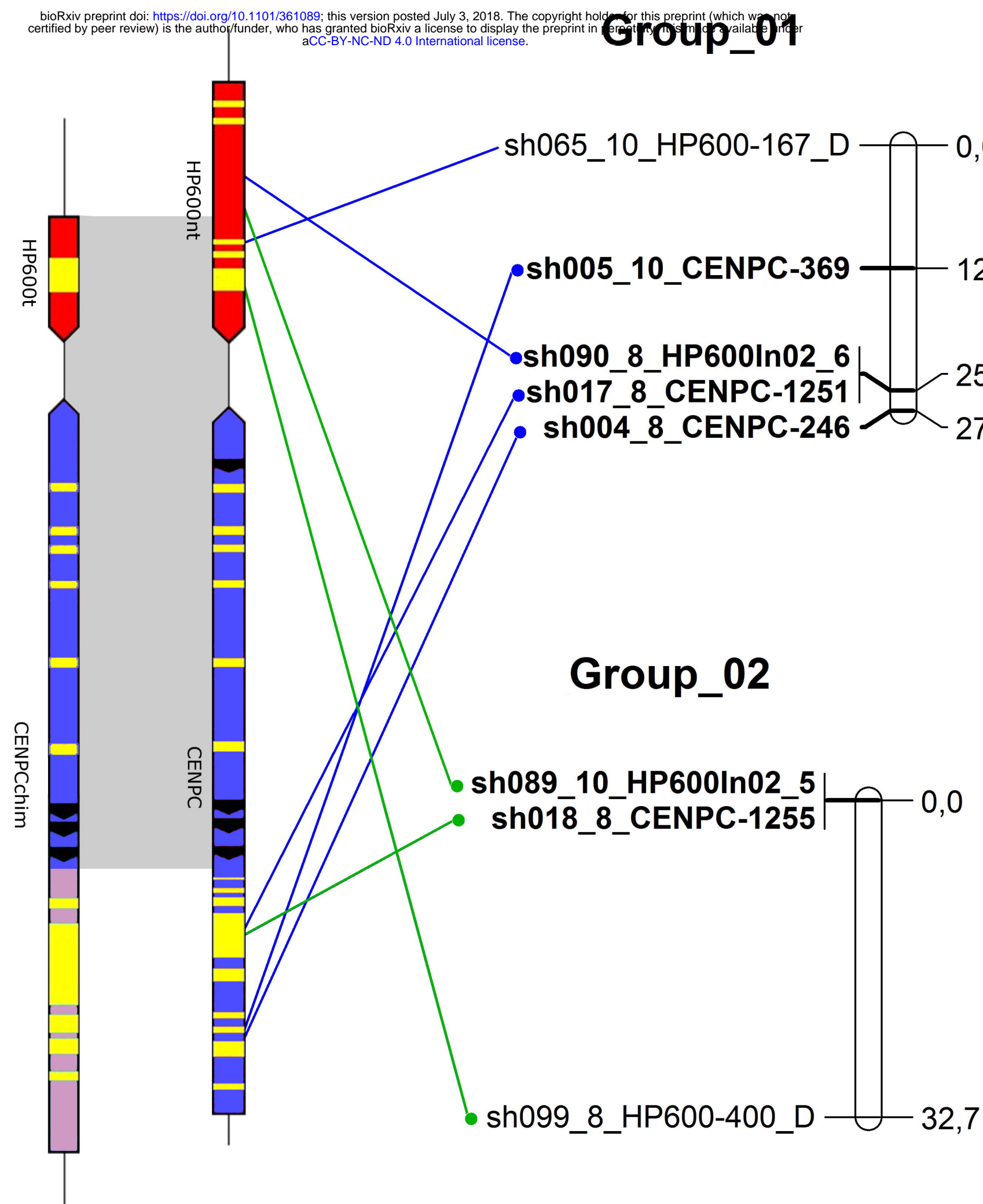


(B)



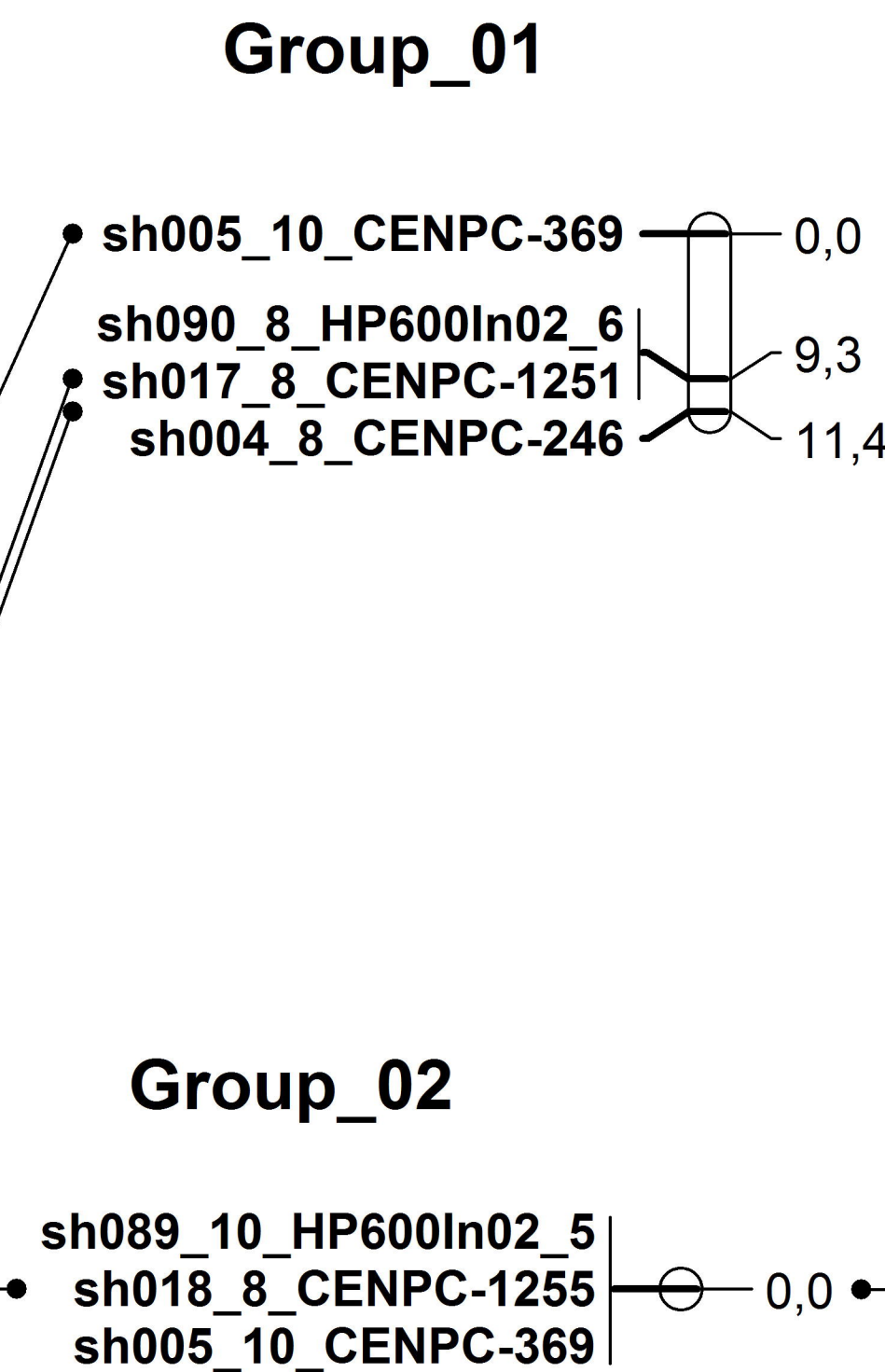


Physical map

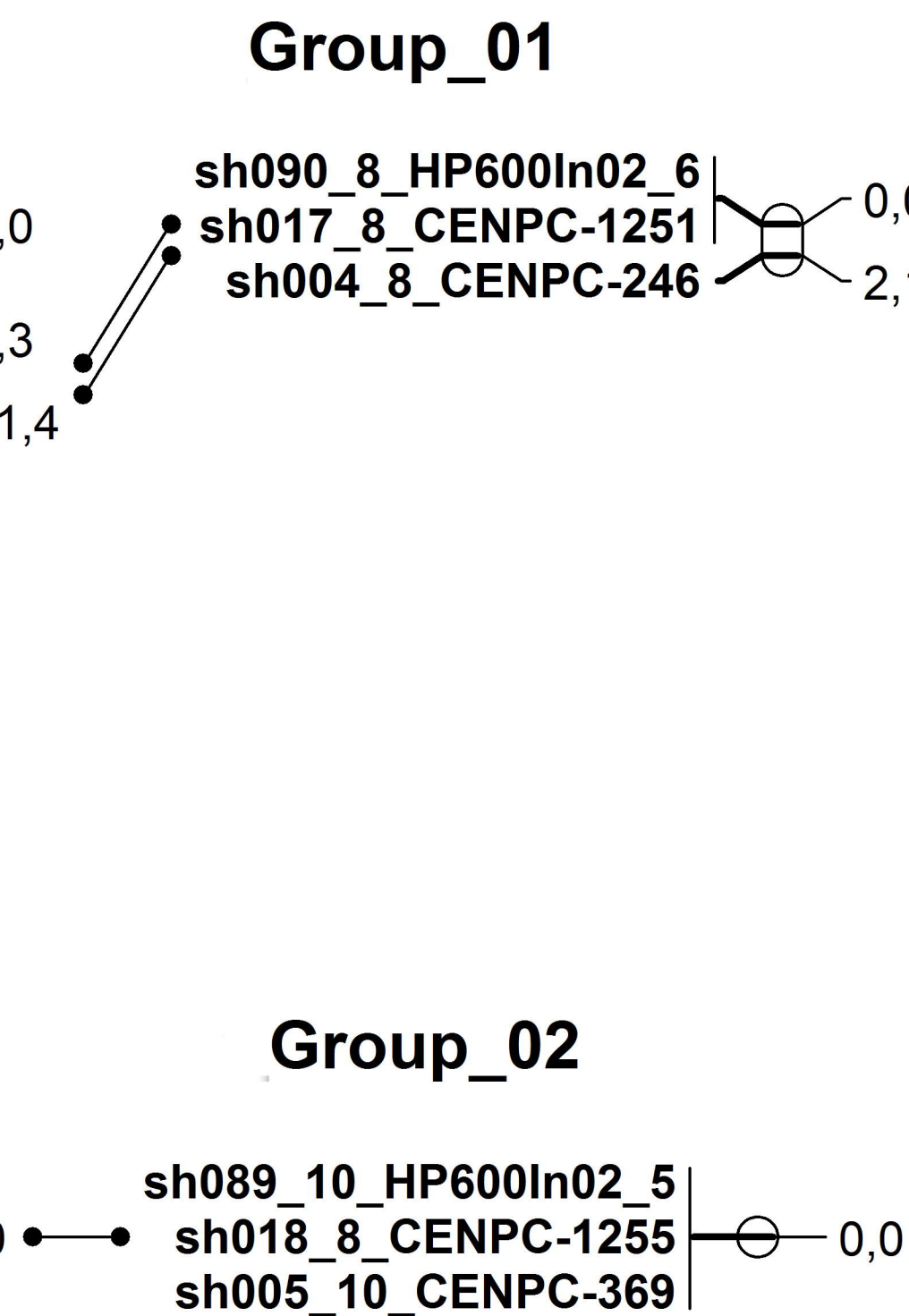


A - Two linkage groups first map

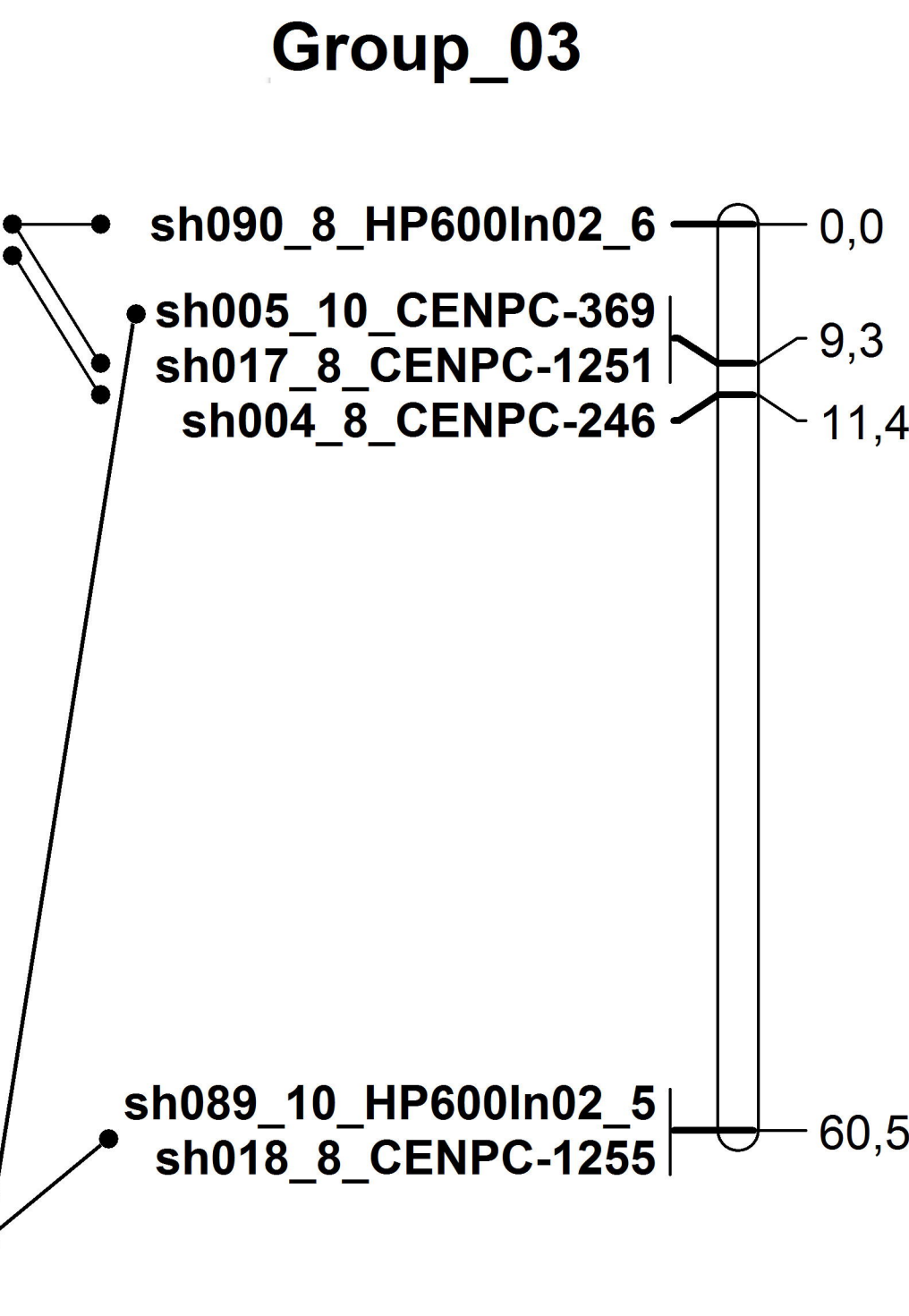
B - Two linkage groups without markers in duplication



C - Two linkage groups without markers in the wrong order according to BACs



D - Trying to create one group with markers in correct order according to BACs



E- Best linkage group using all information about the region

Physical map

