UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE BIOLOGIA

DANILO AUGUSTO SFORÇA

VARIAÇÃO GENÉTICA EM POLIPLOIDES COMPLEXOS: DESVENDANDO A DINÂMICA ALÉLICA EM CANA-DE-AÇÚCAR

GENETIC VARIATION IN COMPLEX POLYPLOIDS: UNVEILING THE DYNAMIC ALLELIC FEATURES OF SUGARCANE

CAMPINAS
2019

DANILO AUGUSTO SFORÇA

VARIAÇÃO GENÉTICA EM POLIPLOIDES COMPLEXOS: DESVENDANDO A DINÂMICA ALÉLICA EM CANA-DE-AÇÚCAR

GENETIC VARIATION IN COMPLEX POLYPLOIDS: UNVEILING THE DYNAMIC ALLELIC FEATURES OF SUGARCANE

Tese apresentada ao Instituto de Biologia da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do Título de Doutor em Genética e Biologia Molecular na área de Genética Vegetal e Melhoramento.

Thesis presented to the Institute of Biology of the University of Campinas in partial fulfillment of the requirements for the degree of [Doctor in Genetics and Molecular Biology in the area of Plant Genetics and Genetic Breeding.

ESTE ARQUIVO DIGITAL CORRESPONDE À VERSÃO FINAL DA TESE DEFENDIDA PELO ALUNO DANILO AUGUSTO SFORÇA E ORIENTADO PELA ANETE PEREIRA DE SOUZA

Orientadora: ANETE PEREIRA DE SOUZA

CAMPINAS
2019

Agência(s) de fomento e nº(s) de processo(s): FAPESP, 2010/50119-6; CNPq, 142950/2010-0; CAPES

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Biologia
Mara Janaina de Oliveira - CRB 8/6972

Sf57v

Sforça, Danilo Augusto, 1985-
Variação genética em poliploides complexos : desvendando a dinâmica alélica em cana-de-açúcar / Danilo Augusto Sforça. – Campinas, SP : [s.n.], 2019.

Orientador: Anete Pereira de Souza.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Biologia.

1. Cana-de-açúcar. 2. Poliploide. 3. Mapas físicos. 4. Homologia (Biologia). 5. Mapeamento cromossômico. I. Souza, Anete Pereira de, 1962-. II. Universidade Estadual de Campinas. Instituto de Biologia. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Genetic variation in complex polyploids : unveiling the dynamic allelic features of sugarcane
Palavras-chave em inglês:
Sugarcane
Polyploidy
Physical maps
Homology (Biology)
Chromosome mapping
Área de concentração: Genética Vegetal e Melhoramento
Titulação: Doutor em Genética e Biologia Molecular
Banca examinadora:
Anete Pereira de Souza [Orientador]
Gabriel Rodrigues Alves Margarido
Maria Lúcia Carneiro Vieira
Américo José Carvalho Viana
Rodrigo Gazaffi
Data de defesa: 05-02-2019
Programa de Pós-Graduação: Genética e Biologia Molecular

COMISSÃO EXAMINADORA

Profa. Dra. Anete Pereira de Souza

Dr. Gabriel Rodrigues Alves Margarido

Dra. Maria Lúcia Carneiro Vieira

Dr. Américo José Carvalho Viana

Dr. Rodrigo Gazaffi

Os membros da Comissão Examinadora acima assinaram a Ata de Defesa, que se encontra no processo de vida acadêmica do aluno.

DEDICATÓRIA

À minha mãe, por todo amor, apoio,

incentivo, por acreditar em mim,

por me inspirar todos os dias...

Dedico.

AGRADECIMENTOS

Especialmente à Profa. Anete Pereira de Souza, por toda confiança, apoio, paciência, orientação e oportunidade de realizar este trabalho, por me ensinar a fazer pesquisa.

À Dra. Hélène Bergès, do "*Centre National de Ressources Génomiques Végétales*" (CNRGV/INRA/Toulouse), por ter me recebido tão bem em seu grupo de pesquisa, pela valiosa contribuição na construção da biblioteca genômica em BACs e pela valiosa amizade.

Ao Prof. Michel Vincentz por todo o conhecimento compartilhado com muita paciência e dedicação, pelo apoio e incentivo no desenvolvimento deste trabalho, pela disponibilidade em discutir e ensinar e pela sabedoria transmitida.

À Dra. Tatiana de Campos e Dra. Adna Cristina Barbosa por todo o ensinamento, por despertar o cientista em mim, pela paciência e pelo companheirismo.

Aos grandes amigos do "*Centre National de Ressources Genomiques Vegetales*" (CNRGV/INRA): Arnauld, Sonia, Elisa, Gensèric, Laetitia, Joelle, David, Nadine e Sthéphane, pela disponibilidade em compartilhar seus ensinamentos sobre bibliotecas em BACs e pela amizade, confiança e apoio.

À minha família, em especial à minha mãe e meu irmão por todo o companheirismo e esforço durante a trajetória.

À Profa. Eliana Regina Forni-Martins e Dra. María Victoria Romero da Cruz, pelos ensinamentos e grande disponibilidade durante as análises citogenéticas.

À Melina Mancini, pela ajuda e momentos maravilhosos durante a jornada, pelo tempo dedicado e pelos conselhos.

À Aline Moraes e Patricia Zambon, pela ajuda, confiança e por compartilhar o tempo, pelas conversas e conselhos.

À Dra. Prianda Laborda, pelos conselhos, pelas conversas enigmáticas sobre para onde vamos e de onde viemos, pela paciência e pelo repeito.

Aos amigos Cláudio Benicio e Guilherme Toledo pelos momentos de descontração e de trabalho duro.

Aos colegas do laboratório de Análises Genética e Molecular / CBMEG e Barracão, pelos conselhos e ajuda nos momentos de dúvidas, pelos momentos de descontração, pelo carinho durante esses anos de convivência.

À Universidade Estadual de Campinas e ao Programa de Pós-Graduação em Genética e Biologia Molecular na área de Genética Vegetal e Melhoramento, pela qualidade do ensino e estrutura oferecida e oportunidade de realizar o doutorado.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pela concessão de bolsa de estudos (2010/50119-6) e suporte financeiro ao projeto (2008/52197-4).

A todos aqueles que de forma direta ou indireta tenham contribuído para a realização deste trabalho.

RESUMO

A cana-de-açúcar é utilizada principalmente na produção de açúcar e etanol. As cultivares atuais são originárias de uma série de hibridizações naturais e artificiais e retrocruzamentos. Esse processo de domesticação e melhoramento resultou em um genoma altamente complexo, com elevados níveis de ploidia e aneuploidia, além de um genoma de aproximadamente 10 Gb. Sorgo, além de possuir alta sintenia com cana-de-açúcar, é o ancestral mais próximo com genoma completamente sequenciado e anotado. A presente tese teve como objetivo estudar as diferenças genéticas, a arquitetura gênica e a expressão gênica de uma região do genoma de cana-de-açúcar, assim como gerar ferramentas moleculares para este e outros estudos envolvendo o genoma desta espécie.

Genes de cópia única foram buscados em transcritos de arroz, milho e sorgo e o resultado comparado com sequências expressas de cana-de-açúcar. Treze genes candidatos foram encontrados, dos quais o gene *HP600* (em sorgo Sobic.003G221600) encontrava-se, além de em provável cópia única, em um QTL localizado em sorgo para *Brix*. Foi construída uma biblioteca de BAC para a variedade SP80-3280 e outra para a IACSP93-3046, com o objetivo de acessar o genoma de cana-de-açúcar. Foram desenvolvidos dois métodos de seleção de clones: macroarranjos (desenvolvido para ambas as bibliotecas) e mistura ordenada de clones (*Pool3D* – desenvolvido para metade da biblioteca SP80-3280).

As bibliotecas de BACs das variedades SP80-3280 e IACSP93-3046 resultaram em 221.184 e 165.888 clones, ambas com um tamanho médio de 110 kb, representando aproximadamente 2,4 e 1,8 vezes o genoma de cana-de-açúcar, respectivamente. Os treze genes foram utilizados para selecionar BACs nas duas bibliotecas utilizando macroarranjos e o Pool3D na biblioteca SP80-3280. Os BACs resultantes dos dois métodos de seleção da biblioteca da SP80-3280 para o gene *HP600* foram utilizados para estudar a arquitetura gênica e transcritos do gene para estudar a expressão gênica em cana-de-açúcar.

O gene da Proteína Centromérica C (*CENP-C*) foi encontrado ao lado do gene *HP600* em cana-de-açúcar e sorgo, e ambos foram utilizados para exemplificar a expressão alélica e o comportamento genômico e genético. Os genes *HP600* e *CENP-C* foram encontrados em dois grupos cromossômicos homeólogos. O primeiro grupo (*Region01*), com ploidia oito, representa a região sintênica de *Sorghum bicolor*, com todos os haplótipos dos dois genes expressos, porém os haplótipos do

HP600 exibiram expressão diferencial. O segundo grupo de homeologos (*Region02*), com ploidia dez, é uma região formada a partir de diferentes genes não-colineares com sorgo contendo duplicações dos genes *HP600* e *CENP-C* (parálogos). Essa duplicação ocorreu após a separação de sorgo e cana-de-açúcar, resultando em um pseudogene HP600 não expresso e uma versão fusionada e recombinada do *CENP-C* com um terceiro gene (ortólogo do Sobic.003G299500) com pelo menos dois haplótipos quiméricos expressos.

O mapa genético evidenciou que marcadores em regiões duplicadas são incorporados ao mapa com distância genética enviesadas, afetando diretamente o mapeamento genético. Esta tese apresenta a complexidade envolvida na genética, genômica e expressão gênica de cana-de-açúcar e na dinâmica genômica e alélica, o que pode ser útil para a compreensão de outros genomas poliploides.

ABSTRACT

Sugarcane is mainly used in the production of sugar and ethanol. The contemporaneous cultivars originate from a series of natural and artificial hybridizations and backcrossing. The process of domestication and breeding resulted in a complex genome with high levels of ploidy, aneuploidy and approximately 10Gb genome. In addition, Sorghum has high sintenia with sugarcane and it is the closest ancestor with genome completely sequenced and annotated. The aim of this thesis was to study the genetic architecture, genomic and gene expression of a region in sugarcane, as well as generate molecular tools involving sugarcane genome.

Single copy genes were searched in rice, maize and sorghum transcripts and the result compared to expressed sequences of sugarcane. Thirteen candidates' genes were found, which the HP600 gene (in sorghum Sobic.003G221600) was in silgle copy and also located in a QTL in sorghum for Brix. A BAC library was constructed for the sugarcane SP80-3280 variety and another for IACSP93-3046 with the objective of accessing the sugarcane genome. Two methods of clone selection were developed: macroarrays (developed for both libraries) and orderly clone mixing (Pool3D - developed for half of the SP80-3280 library).

The BAC libraries construction of the varieties SP80-3280 and IACSP93-3046 resulted in 221,184 and 165,888 clones, both with an average size of 110 kb, representing approximately 2.4 x and 1.8 x the sugarcane genome, respectively. The thirteen genes were used to select BACs from both libraries using macroarrays and the Pool3D selection of the SP80-3280 library. The BACs of the SP80-3280 library resulting from the two selections for the *HP600* gene were used to study the genomic architecture and transcripts of the gene to study the genic expression in sugarcane.

The gene of Centromeric Protein C (*CENP-C*) was found side-by-side the HP600 gene in sugarcane and sorghum, and both were used to exemplify allelic expression and genomic and genetic behavior. The *HP600* and *CENP-C* genes were found in two homeologue chromosomic groups. The first group (Region01), with ploidy eight, represents the synthenic region of *Sorghum bicolor*, with all the haplotypes of the two genes expressed, but the *HP600* haplotypes exhibited differential expression. The second group of homoelogues (Region02), with ploidy ten, is a region formed from different non-collinear genes with sorghum containing duplications of the *HP600* and *CENP-C* genes (paralogues). This duplication occurred after sorghum and sugarcane separation, resulting in an *HP600* pseudogene and a fused and recombined version

of *CENP-C* with a third gene (ortholog of Sobic.003G299500), with at least two expressed chimeric haplotypes.

The genetic map evidenced that markers in duplicate regions are mapped in linkage groups with bias in the genetic distance, affecting the genetic mapping. This thesis presents the complexity involved in genetics, genomics and gene expression of sugarcane and in genomic and allelic dynamics, which may be useful for understanding other polyploid genomes.

SUMÁRIO

INTRODUÇÃO **15**

REVISÃO BIBLIOGRÁFICA **18**

Classificação taxonômica 18
Importância Econômica 19
Origem, domesticação e melhoramento da cana-de-açúcar 21
Aspectos genômicos da cana-de-açúcar 23
Bibliotecas genômicas de grandes insertos 25
BAC-FISH 27
Mapeamento genético em cana-de-açúcar 28
Mapas Físicos em Cana-de-Açúcar 31

OBJETIVOS **33**

Objetivo Geral 33
Objetivos Específicos 33

CAPÍTULO I **34**

Desenvolvimento de ferramentas genômicas para estudos genéticos em cana-de-açúcar 34
Introdução 35
Material e Métodos 47
Construção das Bibliotecas de BAC 47
Isolamento do DNA nuclear de alto peso molecular e preparação dos plugs *47*
Teste de digestão parcial do DNA de alto peso molecular (HMW) *49*
Seleções de tamanho *49*
Isolamento do DNA selecionado da agarose *50*
Ligação e transformação *50*
Estimativa do tamanho do inserto *50*
Sequenciamento das pontas de BACs 51
Busca por genes de cópia única 51
Seleção de clones 52
Screening por Macroarranjos *52*
Screening por Pool 3D *56*
Sequenciamento completo dos clones e montagem 58
Resultados 58
Construção das Bibliotecas de BAC 58
Genes em cópia única 68
Seleção de clones baseada em hibridação: Macroarranjos 69
Seleção de clones baseada em PCR: Pool3D 73
Discussão 77

Conclusão 78

CAPÍTULO II 79

Abstract 80
Introduction 80
Material and Methods 83
Plant Material 83
BAC Library Construction and BAC-End Analyses 83
BAC Library Screening 85
Sequencing and Assembly 85
Sequence Analysis and Gene Annotation 86
Duplication Divergence Time 87
Gene Expression 87
Chromosome Number Determination and BAC-FISH 88
Genetic Map Construction 89
Results 89
BAC Library Construction 89
BAC Annotation 92
Relationship between Region01 and Region02 95
HP600 and CENP-C Haplotypes and Phylogenetics 97
Chromosome Number Determination and BAC-FISH 98
Expression of HP600 and CENP-C Haplotypes 103
Comparison with Others Saccharum Genome 106
How the Locus Number of Homeologs Influences Expression 107
Genetic Mapping 112
Discussion 116
References 124
Supplementary Figures 134
Supplementary Tables 149

RESULTADOS COMPLEMENTARES 157

RESUMO DOS RESULTADOS 160

Capítulo I 160
Capítulo II 160

CONCLUSÕES GERAIS 163

PERSPECTIVAS 166

REFERÊNCIAS BIBLIOGRÁFICAS 168

ANEXOS 190

Anexo I 190

"TARGETED SEQUENCING BY GENE SYNTENY" A NEW STRATEGY FOR POLYPLOID SPECIES: SEQUENCING AND PHYSICAL STRUCTURE OF A COMPLEX SUGARCANE REGION 190
ANEXO II 199
DECLARAÇÃO BIOÉTICA E/OU BIOSSEGURANÇA 199
ANEXO III 201
DECLARAÇÃO DIREITOS AUTORAIS 201

INTRODUÇÃO

A cana-de-açúcar é cultivada em uma área estimada em 26 milhões de hectares no mundo (FAO, 2018), sendo que no Brasil, a produção estimada para a safra 2018/19 é de 625,96 milhões de toneladas. O Brasil é o maior produtor mundial de cana-de-açúcar, seguido pela Índia e China (FAO, 2018). Os principais derivados dessa cultura são o açúcar e o etanol, na safra 2018/19 deverá atingir 35,48 milhões de toneladas de açúcar e 28,16 bilhões de litros a de etanol. Atualmente, a cana-de-açúcar é considerada uma das grandes alternativas para o setor de biocombustíveis na produção de etanol e aos respectivos subprodutos. Além da produção de etanol e açúcar, as unidades de produção têm como objetivo operar com maior eficiência. Este fato vem sendo comprovado pela geração de energia elétrica utilizando os subprodutos derivados da cadeia de produção de açúcar e/ou etanol, auxiliando na redução dos custos e contribuindo para a sustentabilidade da atividade (CONAB, 2018).

Segundo Daniels e Roach (1987), a cana-de-açúcar é uma espécie alógama, da família Poaceae (Gramineae) e do gênero *Saccharum*. Seis espécies de cana-de-açúcar constituem este gênero: *Saccharum officinarum* L., *Saccharum spontaneum* L., *Saccharum robustum* E.W. Brandes & Jeswiet ex Grassl, *Saccharum barberi* Jesw., *Saccharum sinense* Jeswiet. e *Saccharum edule* Hassk. A espécie *S. officinarum* L. destaca-se por suas boas características agronômicas e altos teores de sacarose, sendo denominada como 'cana nobre' (Matsuoka et al., 1999; Landell e Bressiani, 2008), porém suscetível às principais doenças que acometem a cultura. A *S. spontaneum* por sua vez, apresenta menor teor de açúcar, porém com maior robustez e resistência às doenças. A estratégia foi reunir alto teor de açúcar e robustez vinda de duas espécies diferentes em uma única planta, através de cruzamentos, principalmente entre *S. officinarum* com *S. spontaneum*, seguido do retrocruzamento com aquela que detém a característica mais importante para a cultura, teor de açúcar. Assim, a cana-de-açúcar cultivada atualmente é um híbrido interespecífico, apresentando genoma extremamente complexo, grande, poliploide e aneuploide. A cana-de-açúcar cultivada também é fértil.

O genoma da cana-de-açúcar cultivada é o resultado do cruzamento majoritário entre *S. officinarum* com *S. spontaneum*, seguido do retrocruzamento com *S. officinarum*, porém *S. barberi* e *S. sinense* também foram utilizadas nesses primeiros cruzamentos artificiais. Acredita-se que o genoma haploide de cana-de-açúcar

cultivada seja próximo ao de *Sorghum bicolor* L., em torno de 800–900 Mb, com um tamanho total em torno de 10 Gb, e o número básico de cromossomos é potencialmente presente em oito a doze exemplares homeólogos (D'Hont e Glaszmann, 2001). O número básico de cromossomos de *S. officinarum* é 10, enquanto que o de *S. spontaneum* é oito (D'Hont et al.,1998). Garsmeur e colaboradores (2018) propõem que em *S. spontaneum*, fusões cromossômicas aconteceram, reduzindo o número básico de cromossomos 10 para oito, mas mantendo o conteúdo genômico. D'Hont e colaboradores (2005) propõem que o genoma da cana-de-açúcar cultivada seja composto por 70 a 80% dos cromossomos de *S. officinarum*, 10 a 20% de *S. spontaneum* e 10% são recombinante entre as duas.

Apesar de pouco elucidada, existe certo consenso na literatura que o pareamento cromossômico ocorre preferencialmente pela configuração bivalente e que as irregularidades no decorrer da meiose são relativas à segregação irregular dos cromossomos, formando gametas aneuplóides.

Sendo assim, este trabalho objetivou entender a dinâmica por trás da expressão de dois genes e a arquitetura gênica que controla as múltiplas cópias desses genes. O gene *HP600* (em sorgo Sobic.003G221600) foi considerado em cópia única quando buscado em arroz, arabdopsis e sorgo, e foi localizado em um QTL relacionado a acúmulo de açúcar em sorgo (Murray et al., 2008). O gene da Proteína Centromérica C (*CENP-C*) está localizado ao lado do gene *HP600* em cana-de-açúcar e sorgo. Ambos os genes foram utilizados para exemplificar o comportamento genômico e genético em cana-de-açúcar. Suas sequências foram separadas em haplótipos de acordo com a variação de bases (SNPs). Cada haplótipo formado foi comparado a um conjunto de transcritos e foi verificado a expressão alélica de cada um.

Os genes *HP600* e *CENP-C* foram encontrados em dois grupos cromossômicos homeólogos diferentes com ploidias oito e dez. A primeira região (*Region01*), ploidia oito, representa a região ortóloga a *Sorghum bicolor*. Regiões ortólogas de sorgo foram descritas em diversos trabalhos, utilizando diferentes genes e regiões genômicas (Garsmeur et al., 2011; Kim et al, 2014; De Setta et al., 2014; Vilela et al., 2017; Zhang et al., 2018; Mancini et al., 2018). Especificamente sobre a expressão dos genes *HP600* e *CENP-C*, todos os haplótipos da *Region01* se mostraram expressos, porém os haplótipos do gene *HP600* exibiram uma expressão

desbalanceada, ou seja, a proporção de dos alelos no genoma não é explicada pela proporção encontrada no transcriptoma. De alguma maneira, haplótipos do gene HP600 estariam mais ou menos expressos que outros.

A segunda região (*Region02*), ploidia dez, é formada a partir de diferentes genes não-colineares contendo duplicações dos genes *HP600* e *CENP*-C (parálogos). Essa duplicação ocorreu antes da formação do gênero *Saccharum* e após a separação de sorgo e cana-de-açúcar. O resultado foi uma pseudogenização do gene *HP600*, que não possui evidência de expressão e uma versão fusionada e recombinada do *CENP-C* em um terceiro gene, ortólogo a um gene em sorgo (Sobic.003G299500) com pelo menos dois haplótipos quiméricos com evidências de ser expresso. A construção do mapa genético sugere que marcadores em regiões duplicadas são traduzidos para o mapa como frequência de recombinação e, consequentemente como distância genética, distorcendo o mapa genético em genomas de poliploides complexos.

Todos esses resultados descrevem uma região de baixa sintenia entre cana-de-açúcar e *S. bicolor*, formada por duplicação(ões) ocorrido(s) em um ancestral do gênero *Saccharum* (*Region02*). Foram descritos evidencia de expressão gênica em uma duplicação com formação de um gene quimérico (Region02), além do comportamento de marcadores genéticos em uma região duplicada. A construção do mapa genético reforçou a dificuldade de mapear locos em regiões duplicadas em genomas de poliploides complexos. Foi mostrada a complexidade envolvida na genética da cana-de-açúcar e na dinâmica genômica e alélica. Os resultados desse trabalho serão de grande importância para possibilitar a utilização de marcadores moleculares no estudo genético de organismos poliploides. Dessa forma, será possível refinar o entendimento da genética de cana-de-açúcar e acelerar o melhoramento genético dessa cultura, além de expandir e aplicar esse conhecimento a outros organismos de origem poliploide.

REVISÃO BIBLIOGRÁFICA

Classificação taxonômica

O "*complexo Saccharum*" é um grupo informal composto por cinco gêneros: *Erianthus*, *Miscanthus*, *Narenga*, *Sclerostachya*, e *Saccharum*. A cana-de-açúcar pertence à família Poaceae, tribo Andropogoneae e gênero *Saccharum* (Cronquist, 1981). A cana-de-açúcar cultivada é derivada de uma série de hibridizações de espécies do gênero *Saccharum*. É uma planta alógama, herbácea, cultivada em regiões tropicais e subtropicais. O gênero *Saccharum* é caracterizado pelo alto nível de poliploidia e aneuploidia (Naidu e Sreenivasan, 1987; Roach e Daniels, 1987). Seis espécies de cana-de-açúcar constituem este gênero: *S. officinarum* L., *S. spontaneum* L., *S. robustum* E.W. Brandes & Jeswiet ex Grassl, *S. barberi* Jesw., *S. sinense* Jeswiet. e *S. edule* Hassk.

S. officinarum é conhecida como a espécie produtora de açúcar ou "cana nobre", por apresentar qualidades agronômicas e industriais correspondentes aos mais importantes critérios de seleção: colmos grossos, alto teor de sacarose e baixo conteúdo de fibra e amido (Bremer, 1961; Roach, 1986). *S. officinarum* é uma espécie euploide, octaploide, com 2n = 80, e número básico de cromossomos x = 10, na sua grande maioria (Bremer, 1930; Li e Price, 1967; Price e Daniels, 1968). Acredita-se que alguns poucos clones que não apresentam este número cromossômico se originaram de hibridação com outras espécies (Bremer, 1924).

S. spontaneum é uma espécie rústica, nível muito elevado de ploidia, com vasta expansão geográfica, que vai do Japão ao leste da África, passando pelo sudeste da Ásia, pelo continente indiano, pelo Oriente Médio e a bacia mediterrânea (Brandes et al., 1939). Os clones desta espécie contêm baixo teor de sacarose, entretanto possuem resistência a pragas e doenças, capacidade de rebrota de soqueira, alto vigor, presença de rizomas e grande adaptabilidade (Naidu e Sreenivasan, 1987). O número de cromossomos varia de 2n = 40 a 128, sendo os citótipos mais frequentes os múltiplos de 8, sugerindo que o número básico desta espécie é x = 8 (Panje e Babu, 1960; Sreenivasan et al., 1987; Burner, 1987). De fato, D'Hont e colaboradores (1998) confirmaram este número por hibridização in situ de rDNAs, além de indicar que o nível de ploidia varia entre 8 e 12.

S. robustum distingue-se de *S. spontaneum* pela ausência de rizomas, inflorescência grande, haste mais espessa e maior altura (Stevenson, 1965). Foram

encontrados em *S. robustum* dois citótipos euplóides, com 2n = 60 ou 2n = 80, e citótipos aneuplóides, variando de 2n = 63 a 205 (Price, 1957; Price, 1965).

Acredita-se que *S. barberi* e *S. sinense* derivem de hibridações naturais entre as espécies *S. officinarum* e *S. spontaneum* (Price, 1968; D'Hont et al., 1998). Distinguem-se dos clones de *S. officinarum* devido suas características florais, alto teor de fibras e a sua grande rusticidade devido a maior tolerância aos estresses ambientais (Daniels e Roach, 1987). Seus números cromossômicos variam entre 2n = 81 a 124 em *S. sinense* e 2n = 111 a 120 em *S. barberi* (Sreenivasan et al., 1987).

S. edule é um grupo menor de cana estéril. Os clones desta espécie são provavelmente originados da espécie *S. robustum*, bem como de origem interespecífica com *S. robustum* como doadora do gameta feminino (Grivet et al., 2004). Os perfis moleculares mitocondriais e cloroplásticos dos clones de *S. edule* são associados ao perfil mais frequente observados na espécie *S. robustum* (D'Hont et al., 1993; Sobral et al., 1994). Os clones constituem uma série de poliploides com 2n = 60, 70 ou 80 cromossomos (Roach, 1972).

Importância Econômica

A cana-de-açúcar, juntamente com a beterraba, é a base da indústria açucareira mundial. Estima-se que cerca de 70% do açúcar produzido no mundo é proveniente da cana-de-açúcar e 30% da beterraba (FAO, 2018). Esse número reflete a importância atingida pelo açúcar da cana-de-açúcar na alimentação humana. Ela é cultivada em mais de 115 países, essencialmente em países tropicais devido à sua baixa tolerância ao frio (FAO, 2018).

A cana-de-açúcar é cultivada em uma área estimada em 26 milhões de hectares, com uma produção mundial aproximada de 1,890 bilhões de toneladas, apresentando melhor rendimento em climas tropicais (dados de 2016, FAO, 2018). No Brasil, a produção de cana-de-açúcar, estimada para a safra 2018/19, é de 625,96 milhões de toneladas. A área colhida está estimada em 8,61 milhões de hectares, queda de 1,3% se comparada com a safra 2017/18. O Brasil é o maior produtor mundial de cana-de-açúcar, seguido pela Índia e China, respectivamente (FAO, 2018). O país tem duas regiões produtoras, Centro-Sul e no Norte-Nordeste, com safras alternadas, podendo manter sua presença no mercado mundial ao longo

de todo o ano. Tem o menor custo de produção do mundo e ainda possui potencial de expansão de área plantada e de produtividade.

Quase todos os estados brasileiros produzem cana, mas o maior produtor é São Paulo, com aproximadamente 55% da produção nacional na safra 2018/19, sendo produzidos 337 milhões de toneladas. O estado de São Paulo será responsável por uma área de produção de cana-de-açúcar estimada em 4,4 milhões de hectares para a safra de 2018/19, 52% da área total plantada nacionalmente (CONAB, 2018).

Os principais derivados da cana-de-açúcar são o açúcar e o etanol. A produção de açúcar na safra 2018/19 deverá atingir 35,48 milhões de toneladas, retração de 6,3% ao produzido na safra 2017/18. A redução é reflexo da maior produção mundial de açúcar, fazendo com que a produção de açúcar seja mudada para a produção de etanol. A produção de etanol deverá atingir 28,16 bilhões de litros na produção de 2018/19. Já a produção de etanol anidro (utilizado na mistura com gasolina), deverá ter aumento de 7% alcançando 11,86 bilhões de litros, influenciada pelo consumo de gasolina nos últimos anos (FAO, 2018).

O Brasil é o maior produtor e exportador de açúcar do mundo, responsável por 45% da exportação de açúcar mundial em 2016 (FAO, 2018). Também em 2016, o Brasil exportou 28 milhões de toneladas de açúcar, sendo que a Índia importou 2,3 milhões de toneladas e a China, 2,1 milhões de toneladas. Já em relação ao etanol, em 2016, o Brasil exportou cerca de 1,3 milhões de litros e os Estados Unidos importaram 566 mil toneladas e a Coreia do Sul, 510 mil toneladas (www.novacana.com).

De modo geral, há uma maior conscientização em relação aos efeitos indesejáveis da utilização de combustíveis fósseis no balanço de carbono na atmosfera e seus efeitos desastrosos no aquecimento global. Nesse contexto, a agroindústria sucroalcooleira se mostra muito favorável, uma vez que o etanol é obtido de fonte renovável. Atualmente, a cana-de-açúcar é considerada uma das grandes alternativas para o setor de biocombustíveis na produção de etanol e aos respectivos subprodutos. Além da produção de etanol e açúcar, as unidades de produção têm busca do operar com maior eficiência, inclusive com geração de energia elétrica, auxiliando na redução dos custos e contribuindo para a sustentabilidade da atividade (CONAB, 2018).

Origem, domesticação e melhoramento da cana-de-açúcar

A cana-de-açúcar foi domesticada há cerca de 8.000 anos por sucessivos eventos de hibridizações naturais e artificiais. Acredita-se que a domesticação da *S. officinarum* ocorreu a partir da *S. robustum* no sul da Ásia, região da Nova Guiné em 6.000 a.C., que era cultivada em jardins apenas para serem mascadas. *S. officinarum* difundiu-se para as ilhas do Sul do Pacífico, Índia e China, através de expedições australianas por volta de 1500 a 1000 a.C.. Nesta época, *S. barberi* e *S. sinense* apareceram respectivamente na Índia e China. Em 500 d.C., a cana-de-açúcar aparece na Pérsia, no norte da África e ilhas do mediterrâneo (Brandes, 1956; Daniels e Roach 1987; Grivet et al., 2004).

No século XV, os Portugueses e Espanhóis propagaram a cultura nas ilhas do Atlântico. Na ocasião da sua segunda viagem, Cristóvão Colombo trouxe a cana-de-açúcar para as Américas e foi no Haiti que a cana-de-açúcar foi cultivada primeiramente. Durante os séculos XVI e XVII a extensão da cultura da cana-de-açúcar na América, principalmente no Brasil e no Caribe, estava estreitamente ligada às colonizações europeias (Machado, 2003).

Oficialmente, foi Martim Afonso de Souza que, em 1532, trouxe a primeira muda de cana-de-açúcar ao Brasil e iniciou seu cultivo na Capitania de São Vicente. A partir das Capitanias de Pernambuco e da Bahia os engenhos de açúcar se multiplicaram, dando início a uma indústria que encontrou no Brasil seu campo fértil para uma rápida expansão e perpetuação. Após um início repleto de dificuldades, a produção de açúcar prosperou e passados menos de 50 anos o Brasil já detinha o monopólio mundial da produção (Machado, 2003).

Até meados do século XVIII, o desenvolvimento de plantações de cana-de-açúcar realizou-se a partir de um único clone, ou de um número pequeno de clones, denominado *Creoula*. Tratava-se de um clone de *S. barberi* ou de um híbrido desta espécie com *S. officinarum*. Porém era pouco rústico e suscetível a doenças, com cultivo limitado a terras com alta fertilidade. Até meados do século XIX, este clone foi substituído pelo clone *Bourbon*, que devido a sua susceptibilidade, foi substituído pelo clone *Cheribon* e *Tanna* (Stevenson, 1965).

No inicio século XIX, os clones de *S. officinarum* eram a única fonte de cultivares (canas “nobres”), derivado de coletas nas ilhas do Pacífico, eram suscetíveis a diversas doenças, como podridão da raiz, mosaico, gomose e mal de “*Sereh*”. Neste cenário surgiram os programas de melhoramento de cana-de-açúcar.

As primeiras hibridizações artificiais ocorreram em Java (1858) e Barbados (1859), onde, independentemente, foi observado que o híbrido produzia sementes viáveis (Stevenson, 1965).

Ming e colaboradores (2006) definem a história do programa de melhoramento (hibridizações artificiais) de cana-de-açúcar em cinco períodos: (i) Cruzamento entre as canas "nobres" (*S. officinarum*) para a produção de cultivares "nobres"; (ii) Hibridação entre as cultivares "nobres" e outras espécies do gênero *Saccharum*, principalmente *S. spontaneum* – processo chamado de nobilização; (iii) Cruzamento entre os híbridos nobilizados; (iv) Cruzamento entre híbridos com estágio de seleção avançada; (v) Aumento da base genética.

(i) **Cruzamento entre as canas "nobres" (*S. officinarum*) para a produção de cultivares "nobres".** No começo dos anos 1900, seleção de cultivares derivadas de progênies de cruzamentos de polinização aberta entre as canas "nobres" resultaram nas cultivares "nobres". Porém essas cultivares eram susceptíveis a doenças e insetos e limitados a ambientes tropicais. Dessa maneira surgiu a necessidade de se aumentar a base genética de cana-de-açúcar para aumentar a resistência a doenças e insetos (Stevenson, 1965).

(ii) **Hibridação entre as cultivares "nobres" e outras espécies do gênero *Saccharum*, principalmente *S. spontaneum* – processo chamado de nobilização.** A nobilização é quando as canas "nobres" (*S. officinarum*) são polinizadas com pólen de outras espécies de *Saccharum*, como *S. spontaneum*, seguidos de sucessivos retrocruzamentos com as canas "nobres", gerando descendentes férteis (Bremer, 1961). O evento chave desse processo ocorre com o advento da cultivar 'POJ2878', feito em Java, no ano de 1921 (Jeswiet 1929). Na Índia, a nobilização utilizando três espécies (*S. officinarum* com *S. barberi* e *S. spontaneum*) gerou as cultivares precedidas pela sigla "co". Depois dos anos de 1930, a nobilização foi raramente utilizada nos programas de melhoramento (Stevenson, 1965; Simmonds, 1976; Ethirajan, 1987).

(iii) **Cruzamento entre os híbridos nobilizados.** A nobilização gerou importantes cultivares híbridas (e férteis) utilizados na produção de

açúcar depois dos anos de 1930, além de serem utilizados para cruzamento e seleção de outras importantes variedades. Dentre eles, destacam-se “POJ3016”, “POJ3067”, “Co312”, “POJ2978”, “H32-8560”, “Co419”, “B37161”, “B37172” e “NCo310”. Este último ocupou grandes áreas até o fim da década de 1980 em diversos países (Stevenson, 1965; Anonymous, 1945; Nuss e Brett, 1995; Tew, 1987).

(iv) **Cruzamento entre híbridos com estágio de seleção avançada.** A partir dos anos de 1950, as cultivares de cana-de-açúcar eram selecionadas de estágios avançados de melhoramento, mas sempre partindo das mesmas cultivares nobilizadas nos anos de 1920 (Tew, 2003).

(v) **Aumento da base genética.** As cultivares modernas são derivadas de não mais que 15–20 genótipos de cultivares nobilizadas, e podem ter suas bases genéticas rastreadas até as cultivares nobilizadas em Java e Índia (Roach 1989). A base genética das cultivares modernas é mais restrita do que os dos clones originalmente nobilizados (Walker, 1987). Por isso, os programas de melhoramentos atuais utilizam clones diferentes dos utilizados na nobilização para aumentar a base genética, o que tem sido feito desde 1965 (Kennedy e Rao, 2000).

Os clones de cana-de-açúcar cultivados atualmente no mundo são fruto de sucessivas hibridações entre espécies do gênero *Saccharum* naturais e artificiais, conforme descrito anteriormente.

Aspectos genômicos da cana-de-açúcar

A cana-de-açúcar cultivada atualmente é um hibrido interespecífico entre espécies do gênero *Saccharum,* principalmente *S. officinarum* e *S. spontaneum*, resultando em genoma complexo, grande, poliploide, aneuploide (com diferença do número de cromossomos homeólogos) e o hibrido é fértil. O genoma “monoplóide” – termo que se refere à quantidade de pares de bases do número básico de cromossomos – da *S. officinarum* compreende 930Mpb e da *S. spontaneum* 750Mpb. O tamanho do genoma monoploide de *S. officinarum*, *S. spontaneum* e das

cultivares modernas é comparável ao do sorgo (750 Mb - *Sorghum bicolor* L. Moench, Paterson et al., 2009) e duas vezes maior que o do arroz (430 Mb - *Oryza sativa* L. – Goff et al, 2012). Nas cultivares modernas o genoma "monoploide" é estimado em 800–900 Mb, com um tamanho total em torno de 10 Gb, e o número básico de cromossomos é potencialmente presente em oito a doze exemplares homeólogos (D'Hont e Glaszmann, 2001). Por conseguinte, o tamanho total do genoma das cultivares modernas de cana-de-açúcar é maior que o do milho (*Zea mays* L.) 5500Mpb (2n = 20), sorgo, 1600Mpb (2n = 20), ou arroz, 860Mpb (2n = 24), refletindo a alta poliploidia das cultivares de cana-de-açúcar (D'Hont e Glaszmann 2001).

O número básico de cromossomos no gênero *Saccharum* possui uma série de divergências, porém os números mais prováveis são 8 ou 10 (Nishiyama 1956; Bremer 1961, Sreenivasan et al. 1987). Hibridização fluorescente *in situ* (FISH-*Fluorescence in Situ Hybridization*) e genes rDNA indicaram que os 80 cromossomos de *S. officinarum* se encontram organizados em 8 cópias homólogas, de um conjunto básico de 10 cromossomos diferentes (2n = 10x = 80), enquanto que os 40 a 128 cromossomos de *S. spontaneum* estão organizados em 5 a 12 cópias homólogas, com conjunto básico de 8 cromossomos diferentes (2n = 8x = 40-128) (D'Hont et al 1998; Ha et al 1999).

Acreditava-se que pelo menos um evento de poliploidização era compartilhado entre ambas as espécies *S. officinarum* e *S. spontaneum* (Kim et al, 2014), porém *S. officinarum* e *S. spontaneum* surgiram de eventos de poliploidização distintos (Vilela et al, 2017; Garsmeur et al, 2018). A diferença no número básico de cromossomos implica que diferenças estruturais separam os dois genomas. Como consequência, duas organizações cromossômicas distintas coexistem nas cultivares modernas de cana-de-açúcar (Garsmeur et al, 2018).

FISH em cultivares modernas revelaram uma variação no número cromossômico de 100–130 cromossomos, demonstrando uma variação no conteúdo de cromossomos dependendo da cultivar. Estima-se que 70-80% dos cromossomos são derivados de *S. officinarum*, 10–20% de *S. spontaneum*, e aproximadamente 10% seja uma recombinação entre os cromossomos das duas espécies (D'Hont et al., 1996; Piperidis e D'Hont 2001; Piperidis et al., 2010).

Bibliotecas genômicas de grandes insertos

As bibliotecas contendo grandes insertos de DNA, além do mapeamento físico do genoma, têm permitido a clonagem de genes e a análise da estrutura gênica de vários organismos (Peterson et al., 2000). O princípio da técnica é a obtenção de DNA de alto peso molecular (*High Molecular Weight DNA*; HMW-DNA). Com o sequenciamento de *long reads*, o DNA de alto peso molecular tem ganhado uma importância ainda maior em pesquisas genéticas e genômicas, uma vez que essas bibliotecas requerem HMW-DNA em sua construção (Schadt et al., 2010; Goodwin et al., 2016). O DNA com tamanho de poucos milhares de bases é relativamente estável (Mulcahy et al., 2016), porém os fragmentos grandes tem sido negligenciados, não possuindo informações sobre sua estabilidade (Anchordoquy e Molina, 2007).

A construção de bibliotecas genômicas utilizando o DNA de alto peso molecular como DNA recombinante gera diferentes tipos de tecnologias, de acordo com o hospedeiro e o vetor usado. Dentre essas tecnologias, destacam-se cromossomos artificiais de levedura (*Yeast Artificial Chromosome* - YAC, Burke et al, 1987), cromossomos artificiais de bactérias (*Bacterial Artificial Chromosomes* - BAC, Shizuya et al., 1992), cromossomo artificial de bacteriófago P1 (*Bacteriophage P1-Derived Artificial Chromosome* – PAC, Loannou et al., 1994), na transformação de plantas competentes em BIBAC (ou BAC binário, Hamilton et al., 1996), clonagem de grandes insertos baseada em plasmídeos (*Large-Insert Plasmid-Based Clone* – PBC, Tao e Zhang, 1998) e transformação de cromossomos artificiais competentes (*Transformation-Competent Artificial Chromosomes* – TAC, Liu et al., 1999).

Dentre os vetores que podem carregar grandes fragmentos, o YAC, que possui capacidade de clonagem de até 1.000 kb, parece ser mais vantajoso que os outros tipos (Burke et al, 1987). No entanto, os YACs possuem várias desvantagens, como o elevado número de clones quiméricos, a instabilidade e a dificuldade na purificação do inserto (Peterson et al., 2000). Embora todos os tipos de bibliotecas de grande fragmentos tenham fornecido ferramentas para clonagem posicional e pesquisa genômica avançada (Burke et al, 1987; Shizuya et al, 1992; Loannou et al, 1994; Hamilton et al, 1996, 1999; Tao e Zhang, 1998; Liu et al, 1999, Chang et al, 2001), BAC emergiu como o sistema de clonagem mais usados para pesquisa genômica devido a sua baixa frequência de clones quiméricos, manutenção estável de fragmentos grandes (100 - 300 kb) e facilidade de purificação e manipulação de

DNA clonado (Shizuya et al. 1992; Tao e Zhang, 1998; Chang et al. 2003; Song et al. 2003; Wu et al. 2004; Ren et al. 2005). Nas bibliotecas de BACs, cada clone é armazenado individualmente. Essa característica, combinada ao desenvolvimento de métodos de *fingerprinting* de DNA e sequenciamento, permite conhecer a estrutura genômica de espécies geneticamente complexas (Peterson et al., 2000, Garsmeur et al., 2011; Kim et al, 2014; De Setta et al., 2014; Vilela et al., 2017; Zhang et al., 2018; Mancini et al., 2018, Garsmeur et al., 2018).

Os BACs não são vetores criados a partir de cromossomos artificiais *per se* (ao contrário do que o seu nome sugere), mas são fatores bacterianos do tipo F modificados. A replicação do fator F em *Escherichia coli* é estritamente controlada e os plasmídeos que contém esse fator são mantidos em baixo número de cópias nas células, reduzindo assim o potencial para recombinação entre fragmentos de DNA carregado (Shizuya et al.,1992). Apesar de serem capazes de carregar insertos de até 500 kb, a média dos tamanhos de fragmentos clonados são de 80 a 200 kb (Peterson et al., 2000). Os vetores BAC contêm as características de seleção comuns à maior parte dos vetores, como resistência a antibióticos e um sítio de clonagem múltipla associado a um gene repórter. A presença do fator F impede que mais de um BAC coexista simultaneamente em uma mesma célula bacteriana (Yüksel e Paterson, 2005).

Bibliotecas genômicas de BACs foram construídas para várias espécies vegetais, como: soja (*Glycine max* (L.) Merr. – Marek e Shoemaker, 1997), tomate (*Solanum lycopersicum* L. - Budiman et al., 2000), café (*Coffea arabica* L. - Noir et al., 2004), arroz (*Oryza sativa* L. - Ammiraju et al., 2006), girassol (*Helianthus annuus* L. - Bouzidi et al., 2006), ervilha (*Pisum sativum* L. – Coyne et al, 2007), milho (Wei et al., 2009) e algodão (*Gossypium arboreum* L. - Hu et al., 2010).

As bibliotecas em BACs podem auxiliar estudos genômicos avançados, como a clonagem posicional de genes de herança simples ou quantitativa (*Quantitative Trait Loci* – QTL, Zhang et al., 2007), mapeamento físico genômico por *fingerprinting* (Tao et al., 2001; Ren et al., 2003; Wu et al., 2005; Zhang et al., 2006, 2011), mapeamento físico cromossômico (Pedrosa et al., 2002; Tang et al., 2009; Wai et al., 2010), isolamento de genes (Coyne et al., 2007; Paiva et al., 2011), estudos de sintenia (Ma et al., 2010), análises do genoma funcional em larga escala (Chang et al., 2011; Johnson e Wade-Martins, 2011) e sequenciamento genômico (Venter et al, 1996; Zang e Wu, 2001; Sato et al., 2011).

A facilidade de sequenciamento dos clones de BAC possibilitou o desenvolvimento da estratégia de BAC-end Sequencing (Venter et al., 1998). O sequenciamento de BAC-ends (sequenciamento das extremidades dos insertos) é feita utilizando *primers* que se anelam no vetor, algumas bases antes de onde o inserto foi inserido, resultando no sequenciamento das pontas dos insertos. Os BAC-ends são úteis para a busca de marcadores microssatélites e SNPs, e podem ser utilizados no desenvolvimento e a saturação de mapas de ligação (Han et al., 2011; Rabbi et al., 2012). Esta sobreposição também pode ser realizada em mapas cromossômicos, resultando na integração de ambos os mapas (Tang et al., 2009; Fônseca et al., 2010; Febrer et al., 2010).

A primeira biblioteca de BACs para cana-de-açúcar foi construída em 1999 usando o DNA da variedade francesa R570 e representa em torno de 1,3 vezes o seu genoma total (10 Gb - Tomkins et al. 1999). Figueira e colaboradores (2012) construíram uma biblioteca de BACs de domínio privado para a variedade SP80-3280, representando 0,46 vezes de cobertura do seu genoma.

BAC-FISH

A FISH permite posicionar sequências em relação à eucromatina, heterocromatina, centrômeros e telômeros (Jiang e Gill, 1996). Utilizando como sondas os BACs contendo segmentos de interesse, é possível realizar o mapeamento cromossômico de sequências de cópias únicas, como genes ou marcadores geneticamente mapeados. Quando os BACs utilizados são selecionados com base nestes marcadores, é possível a integração dos mapas de ligação e cromossômicos, e o mapa integrado gerado pode ser usado para estimar as frequências de recombinação em diferentes regiões do genoma, além de esclarecer distorções dos mapas de ligação (Pedrosa et al., 2002).

Utilizar os BACs como sondas resultou em diversos trabalhos onde mapas genéticos e cromossômicos puderam ser integrados e, consequentemente, a identificação do numero cromossômico relacionado diretamente aos grupos de ligação. Dentre os trabalhos, destacam-se os feitos com batata (Dong et al., 2000), tomate (2n = 24) e *Gossypium arboreum* L. (2n = 26) (Wang et al., 2008) e feijão comum (*Phaseolus vulgaris* L.) (2n = 22) (Pedrosa-Harand et al., 2009, Fonseca et al., 2010).

BACs cromossomos específicos foram usados para o mapeamento comparativo entre espécies próximas do mesmo gênero ou tribo, graças à conservação de sequências de nucleotídeos observada entre táxons próximos (Pedrosa et al., 2002; Lysak et al., 2005). Como exemplo de estudos comparativos tem-se o caso de *Brassicaceae*, onde um conjunto de BACs cromossomos-específicos foi utilizado como sonda no mapeamento das seguintes espécies: *Arabidopsis thaliana* (n = 5), *A. lyrata* (n = 8), *Capsella rubella* (n = 8), *Neslia paniculata* (n = 7), *Turritis glabra* (n = 7) e *Hornungia alpina* (n = 6). A partir do pressuposto número cromossômico ancestral, n = 8, foi possível elucidar os mecanismos de evolução, formadores dos cariótipos de *A. thaliana* e das espécies relacionadas (Lysak et al., 2006).

Zhang e colaboradores (2017) isolaram e caracterizaram sequências repetitivas de centrômeros utilizando FISH em *S. spontaneum*, *S. officinarum* e *S. robustum*. Esses resultados mostraram as diferentes composições do genoma dessas três espécies, relativas à presença de satélites e transposons. Vieira e colaboradores (2018) utilizaram FISH para verificar o comportamento meiótico na variedade SP93-3046. Seu trabalho demonstrou a prevalência de pareamento bivalente, porém diversas anomalias cromossômicas foram detectadas.

Mapeamento genético em cana-de-açúcar

Marcadores moleculares são definidos como fragmentos de DNA que permitem a distinção de variações alélicas dentro do genoma de indivíduos da mesma espécie e entre espécies (Borém e Santos 2004). Os marcadores moleculares são ferramentas valiosas no estudo de genomas complexos como o da cana-de-açúcar (Daugrois et al., 1996), contribuindo para maior conhecimento da sua complexidade genética e genômica. Os marcadores moleculares auxiliam na geração de informações sobre a diversidade entre as cultivares (Lu et al., 1994; Jannoo et al., 1999; Lima et al., 2002), identificação de genitores (Glaszmann et al., 1990; D'Hont et al., 1993; Lu et al., 1994; Burnquist et al., 1992; Jannoo et al., 1999; Nair et al., 1999), e identificação de variedades de gêneros do complexo *Saccharum* (AL-Janabi et al., 1994; Besse et al., 1996; Alix et al., 1998, 1999). Ainda têm sido utilizados como uma ferramenta na identificação de variedades, no controle da progênie e no monitoramento de introgressão (D'Hont et al., 1995; Harvey et al., 1998; Cordeiro et al., 2000; Piperidis e D'Hont, 2001).

Single nucleotide polymorphisms (SNPs) são marcadores moleculares que ocorrem quando um único nucleotídeo na sequência do genoma é alterado. Esses polimorfismos, juntamente com as deleções e inserções, são responsáveis pela maior parte da variação genética nos organismos (Cho et al., 1999; Rafalsky e Tingey, 2008) e são amplamente distribuídos pelo genoma, sendo mais abundantes em regiões não transcritas (Mogg et al., 2002; Bundock e Henry, 2004; Giancola et al., 2006; Masouleh et al., 2009). A disponibilidade de marcadores abundantes no genoma facilita a construção de mapas de alta resolução, bem como o mapeamento associativo baseado em desequilíbrio de ligação (Rafalski, 2002). Com o avanço das tecnologias de sequenciamento e a maior disponibilidade de bancos de dados de sequências expressas, a identificação e o uso de SNPs vem crescendo em plantas.

Uma das estratégias para descoberta de SNPs é a utilização de bases de dados de ESTs (*Expressed Sequence Tags*). As bases de dados de ESTs que foram utilizadas na identificação de SNPs para cana-de-açúcar até o momento foram o SUCEST (Vettore et al., 2003; Grivet et al., 2004; McIntyre et al., 2006; Garcia et al., 2013; Costa et al., 2016) e o Plantdb (Cordeiro et al., 2006).

Os recentes avanços nas tecnologias de sequenciamento permitiram a produção de grandes quantidades de dados e redução do custo por base. Bundock e colaboradores (2009) mostraram que o uso de NGS é mais rentável que as técnicas anteriormente empregadas, e por isso ela vem sendo utilizada para o sequenciamento e re-sequenciamento de genomas inteiros. Algumas das aplicações são descobrir grande número de polimorfismos SNPs através do sequenciamento do genoma de vários indivíduos (Elshire et al., 2011), explorar a diversidade genética (Heslot et al., 2013; Lu et al., 2013; Romay et al., 2013), realizar o mapeamento de QTLs (Poland et al., 2012; Spindel et al., 2013; Liu et al., 2014), e mapeamento associativo (Genome-Wide Association Study, GWAS - Byrne et al., 2013; Crossa et al., 2013; Donato et al., 2013; Mascher et al., 2013; Sonah et al., 2013; Uitdewilligen et al., 2013; Chen & Lipka et al., 2016), além de possibilitar a caracterização de germoplasma.

Uma das abordagens para genotipagem utilizando NGS é o GBS e tem se mostrado um grande aliado para identificação em larga escala de polimorfismos genéticos (Elshire et al., 2011; Lu et al., 2013) e pode possibilitar significativo avanço no entendimento da genética da cana-de-açúcar (Elshire et al., 2011; Poland et al.,

2012; Beissinger et al., 2013; Spindel et al., 2013; Glaubitz et al., 2014; Heffelfinger et al., 2014; Liu et al., 2014; Jiang et al., 2016).

Marcadores moleculares podem ser usados para a construção de mapas genéticos. Os mapas genéticos, ou mapas de ligação, identificam a posição de genes ou marcadores moleculares correspondentes à sua ordem linear nos cromossomos. Os primeiros mapas genéticos foram baseados em marcadores morfológicos e citológicos, seguido por isoenzimas e por fim, baseados em marcadores de DNA (Carneiro e Vieira 2002). Novos marcadores moleculares foram surgindo ao longo dos anos e o uso de diferentes marcadores moleculares para a construção de mapas genéticos apresenta como resultado final mapas com maior acurácia e resolução (Ball et al. 2010).

O passo inicial para a construção de um mapa genético é a escolha da população de mapeamento, que deve ser originada de genitores com maior distância genética entre si, objetivando explorar ao máximo o polimorfismo a ser revelado na população segregante (Paterson et al. 1991), além do desequilíbrio de ligação entre os locos. Tradicionalmente, as linhagens endogâmicas, oriundas de retrocruzamentos e populações F2, são utilizadas na construção dos mapas genéticos (Tanksley, 1993) e são bem estabelecidas em espécies diploides. Contudo, cerca de 75% das espécies vegetais são poliploides (Henry, 2008) restringindo a aplicação de técnicas genético-estatísticas na construção de seus mapas genéticos (Pastina et al. 2010, Gazaffi et al. 2010).

Em especial, para cana-de-açúcar, a dificuldade na construção dos mapas genéticos aumenta devido: (i) ao alto nível de ploidia associado a aneuploidia, resultando em um complexo padrão de segregação cromossômica durante a meiose (Heinz e Tew 1987), (ii) a população de mapeamento é derivada de cruzamento entre genitores altamente heterozigotos, com números diferentes de alelos por loco, resultando em diversas proporções de segregações dos marcadores na progênie (Wu et al. 2002, Lin et al. 2003) e (iii) as fases de ligação entre os marcadores são desconhecidas (Pastina et al. 2012). Contornando esta situação, Wu et al. (2002) propuseram a utilização de marcadores SDRF (*single-dose restriction fragment*) para a construção dos mapas genéticos, independente do nível de ploidia da planta. Estes marcadores estão presentes em cópia única em um dos genitores, segregando na progênie na proporção mendeliana de 1:1, ou em uma única cópia em ambos genitores, segregando na proporção de 3:1.

A maioria dos mapas genéticos publicados para cana-de-açúcar baseia-se na estratégia de pseudo-testcross (Grattapaglia e Sederoff, 1994), que resulta na construção de dois mapas individuais, um para cada genitor (Daugrois et al. 1996, Ming et al. 2001, 2002, AlJanabi et al. 2007), explorando apenas a heterozigose em um dos genitores (segregação 1:1). Refinando o mapeamento genético em poliploides, Garcia e colaboradores (2006) propuseram um mapa integrado baseando-se na metodologia proposta por Wu et al. (2002), incorporando os marcadores em heterozigose em ambos os genitores (segregação 3:1), os quais atuam como pontes entre os genomas, identificando regiões de homologia (Maliepaard et al. 1997). Esta estratégia de mapa integrado foi utilizado por Oliveira et al. (2007). Garcia et al. (2013) constatou que não existem razões biológicas consistentes em assumir que os locos em dose única se encontram em maiores proporções no genoma. Este fato discorda de muitos estudos em cana-de-açúcar que afirmam que marcadores em dose única estão presentes em maiores proporções (Aitken et al. 2005).

Diversos trabalhos de mapeamento de QTLs foram desenvolvidos para cana-de-açúcar através de análises de marcas individuais (revisados em Pastina et al., 2012). Algumas exceções existem, visto que alguns trabalhos incluíram marcadores com segregação 3:1 (revisados em Pastina et al., 2012; Singh et al., 2013; Margarido et al., 2015) e segregação 1:2:1 (Costa et al., 2016). O uso de GBS também já foi incorporado ao mapeamento de cana-de-açúcar, resultando em um mapa genético e identificação de QTLs (Balsalobre et al., 2017).

Mapas Físicos em Cana-de-Açúcar

O mapa físico é a representação do genoma de uma espécie. Nos últimos 20 anos, os geneticistas moleculares alcançaram avanços significativos no desenvolvimento de recursos moleculares e no aprimoramento da compreensão geral do genoma da cana-de-açúcar (Garsmeur et al, 2018). Devido à complexidade do genoma das variedades cultivadas, conforme exposto até aqui, a montagem do genoma físico é praticamente impossível com as tecnologias de sequenciamento e de montagem disponíveis até o momento.

A complexidade do genoma de cana-de-açúcar reside (i) na poliploidia das variedades (2n = 100-130), que resultam em uma variação do número de cópias

para determinado locos; (ii) na aneuploidia, que causa a variação do número cromossômico entre variedades; (iii) no tamanho do genoma (10Gb); (iv) no fato das variedades comerciais serem um híbrido interespecífico entre *S. officinarum* e *S. spontaneum*.

Estudar o genoma de *S. officinarum* e *S. spontaneum*, separadamente, é um recurso empregado para estudos genômicos (D'Hont et al., 1996; Piperidis e D'Hont 2001; Piperidis et al., 2010; Kim et al, 2014; Zhang et al., 2018). A utilização de BACs para estudos genômicos tem se tornado a melhor maneira de se obter informações genômicas confiáveis. O sorgo se tornou um genoma referência para cana-de-açúcar, devido à conservação microssintenica entre as duas espécies (Jannoo et al., 2007; Le Cunff et al., 2008; Paterson et al., 2009; Garsmeur et al., 2011; De Setta et al., 2014; Vilela et al., 2017; Mancini et al., 2018; Garsmeur et al, 2018).

Riaño-Pachón e Mattiello (2017) desenvolveram um draft do genoma da variedade SP80-3280, com aproximadamente 200.000 *contigs* utilizando sequenciamento Illumina. Mancini e colaboradores (2018) obtiveram um novo método de recuperar regiões de interesse baseados em QTLs. Os autores utilizaram QTLs preditos em sorgo para buscar regiões sintênicas em cana-de açúcar. Garsmeur e colaboradores (2018) propuseram um mapa físico de regiões ricas em gene, o qual chamaram de genoma "mosaico monoplóide". Os autores fizeram um sequenciamento massivo de mais de 5000 BACs e os alinharam seguindo a sintenia cana- sorgo.

Apesar de laboriosas, até o momento, a utilização de BACs se mostram mais promissoras para recuperar regiões do genoma de variedades de cana-de-açúcar cultivadas. O genoma de cana-de-açúcar, com certeza, é um dos genomas mais interessantes e complexos. O comportamento genético, genômico, expressão gênica e interação entre os alelos em cana-de-açúcar é algo que estamos muito longe de entender completamente.

OBJETIVOS

Objetivo Geral

Criar ferramentas para o estudo genético, genômico e evolutivo de cana-de-açúcar utilizando duas variedades brasileiras. Identificar alelos de um gene supostamente em dose única, estudar sua arquitetura gênica e comportamento genético, genômico e de expressão gênica em cana-de-açúcar.

Objetivos Específicos

- Identificar genes de provável cópia única que possam ser utilizados para estudos genéticos e genômicos em cana-de-açúcar.
- Construir duas Bibliotecas de BACs:
  - SP80-3280.
  - IACSP93-3046.
- Construir uma ferramenta de seleção de clones (*Pool 3D*) para a biblioteca da SP80-3280.
- Caracterizar as duas bibliotecas de BACs construídas utilizando o sequenciamento das extremidades (BAC-end).
- Construir Macroarranjos para as Bibliotecas das variedades de SP80-3280 e IACSP93-3046.
- Selecionar clones por Macroarranjos e *Pool 3D* de genes de cópia única para a biblioteca da variedade de cana-de-açúcar SP80-3280.
- Escolher um provável gene de cópia única para ter sua região genômica estudada.
- Isolar e analisar por *fingerprint* os clones selecionados contendo o gene de interesse.
- Sequenciar e anotar os BACs identificados para o gene de possível cópia única escolhido.
- Analisar filogeneticamente os genes dos BACs selecionados.
- Estudar a microssintenia entre os BACs.
- Distinguir os haplótipos dos genes possivelmente expressos.
- Localizar fisicamente os BACs em cromossomos hom(e)ólogos por BAC-FISH.

CAPÍTULO I

Desenvolvimento de ferramentas genômicas para estudos genéticos em cana-de-açúcar

Introdução

Determinar o genoma da cana-de-açúcar tem se demonstrado um desafio devido ao seu alto grau de complexidade (Garsmeur et al., 2018). Mesmo após o advento das poderosas técnicas de sequenciamento de segunda ("*Next Generation Sequencing*" - NGS) e terceira geração ("*Long reads*"), este desafio ainda persiste. Dentre os trabalhos que encararam esse desafio, destacam-se o de Riaño-Pachón e Mattiello (2017), o de Garsmeur e colaboradores (2018) e Zhang e colaboradores (2018). Riaño-Pachón e Mattiello (2017) desenvolveram um *draft* do genoma da variedade SP80-3280, com aproximadamente 200.000 *contigs* utilizando sequenciamento NGS (*Short reads* – Illumina). Já Garsmeur e colaboradores (2018) chegaram a um genoma mais assertivo para a variedade R570, utilizando BACs, mapas genéticos e a sintenia entre sorgo e cana. Porém, este último ainda não representa um genoma das variedades, mas sim um genoma "mosaico monoplóide" das regiões ricas em genes. Por outro lado, Zhang e colaboradores (2018) apresentaram o genoma em nível de alelos para *S. spontaneum*, uma das espécies envolvida no processo de domesticação das cultivares modernas de cana-de-açúcar.

O Brasil ocupa o posto de maior produtor mundial de cana-de-açúcar (FAO, 2018), e produzir ferramentas que possam auxiliar no melhoramento de variedades brasileiras têm sido o objetivo de vários grupos brasileiros. Dentre as variedades brasileiras mais pesquisadas, destacam-se as variedades SP80-3280 e a SPIAC93-3046. Apesar da base genética estreita das cultivares de cana-de-açúcar, utilizar variedades que são plantadas atualmente no Brasil e/ou são parentais em cruzamentos brasileiros, pode ser a maneira mais eficiente de se transformar dados genéticos e genômicos em ferramentas que auxiliem diretamente o melhoramento de cana-de-açúcar brasileiro.

Apesar das variedades SP803280 e IACSP93-3046 não se destacarem como as mais plantadas atualmente no Brasil, ambas são muito utilizadas para estudos genéticos e genômicos (Vettore et al., 2003; Landel et al., 2005; Souza et al., 2011; Cardoso-Silva et al., 2014; Garcia et al., 2013; Nishyiama et al, 2014; Matiello et al., 2015; Balsalobre et al., 2016; Costa et al., 2016; Balsalobre et al., 2017; Riaño-Pachón e Mattiello, 2017; Mancini et al., 2018; Manechini, et al, 2018; Vieira et al., 2018; dentre outros).

A variedade SP80-3280 é parental da população de mapeamento originada do cruzamento SP80-3280 x RB835486, utilizada pelo Programa de Melhoramento Genético da Cana-de-açúcar da UFSCar/RIDESA. Destaca-se pelo alto teor de sacarose e produtividade em soqueira; perfilhamento intermediário e com bom fechamento das entrelinhas, devido ao crescimento inicial vigoroso; alto teor de fibra, tombamento regular e média exigência em fertilidade do solo; boa brotação de soqueira; sensibilidade média a herbicidas e resistência ao carvão, mosaico e ferrugem; tolerante à escaldadura; não tem mostrado sintomas da síndrome do amarelecimento; apresenta suscetibilidade à broca (Socicana, 2018).

A variedade IACSP93-3046 é parental da população de mapeamento originada do cruzamento IACSP93-3046 x IACSP95-3018 utilizada pelo Programa de Melhoramento da Cana do Instituto Agronômico de Campinas, Centro de Cana. Possui alta produção agroindustrial com características de uniformidade de diâmetro de colmo, o que proporciona maior eficiência na colheita mecânica e manual. Tem alto teor de sacarose no meio e no fim de safra, proporcionando ganhos qualitativos para esses períodos. O período de utilização industrial é longo, adequado para a colheita de junho a outubro. Pode ser cultivada em solos de menor fertilidade e responde significativamente quando plantada em ambientes de maior potencial. Resistente ao carvão, à escaldadura e à ferrugem, com excelente brotação de soqueiras e é adaptada para o Centro-Sul do Brasil (Rural Centro, 2018).

Apesar dessas diferentes características fenotípicas, a base genética entre todas as cultivares é muito estreita. Dez cruzamentos iniciais (Figura 01) entre *S. officinarum*, *S. spontaneum*, *S. Barbieri* e *S. sinense* (nobilização), definem o pedigree das variedades SP80-3280 (Figura 02), IACSP93-3046 (Figura 03) e R570 (Figura 04).

A variedade SP80-3280 possui informações mais completas de pedigree, e foi possível recuperar os parentais femininos utilizados nos policruzamentos. Foi possível recuperar sete gerações de cruzamentos entre os dois genitores da SP80-3280 até se chegar às espécies fundadoras. Os cruzamentos são oriundos de diversas variedades desenvolvidas no início do século XX, com números cromossômicos diferentes, variando entre 2n = 64 a 2n = 148, como as variedades POJ213, POJ2878, POJ2364 e Co281.

A variedade IACSP93-3046 provém de policruzamento diretamente em seu genitor masculino. Apesar disso, possui sete gerações de cruzamentos, igualmente

com variações em seus números cromossômicos de 2n = 64 a 2n = 148. Possui em seu pedigree variedades muito utilizadas, como POJ213, POJ2878 e POJ100.

A variedade R570 é uma variedade francesa, não plantada no Brasil, porém é alvo de diversos estudos genéticos e genômicos de grupos internacionais (Garsmeur et al., 2011; Vilela et al., 2017; Garsmeur et al., 2018; dentre outros). Ao contrário das variedades brasileiras, a R570 não provém de policruzamentos, sendo possível encontrar seu pedigree quase que completo, exceto pela variedade RF72, que não foi possível encontrar os parentais. Possui cinco gerações de cruzamentos, com números cromossômicos entre as variedades variando de 2n = 80 a 2n = 146. Outro detalhe interessante é que o avô materno e a avó paterna são a mesma cultivar POJ2878, sendo fruto de um cruzamento entre meios-irmãos. Em consequência, as bases genéticas da R570 são mais estreitas quando comparada com a SP80-3280 e a IACSP93-3046.

Observando o pedigree de apenas três cultivares, as três compartilham a variedade POJ2878. E se for considerado somente os programas de melhoramento brasileiros, constata-se uma relação mais estreita, compartilhando pelo menos duas cultivares POJ213 e POJ2878.

O grande desafio no estudo genético e molecular da cana-de-açúcar é saber como aplicar as leis de genética clássica a um organismo com alto nível de ploidia e com dosagem alélica podendo variar em diferentes locos. Assim, os locos em cópia única podem ser os mais simpleglos de se iniciar estudos genéticos e genômicos de plantas com o genoma complexo. Na literatura encontra-se uma divergência de quantidade atribuída aos locos únicos. Há trabalhos que relatam que são representados por mais da metade do genoma (Aitken et al., 2007) enquanto outros atribuem pouco mais de 20% do genoma (Garcia et al., 2013). Sendo assim, genes de cópia única são aqueles que se apresentam em um único loco em determinado genoma. Encontrar esses genes em organismos com o genoma sequenciado é relativamente simples. Porém, em organismos em que não se tenha o genoma sequenciado é quase impossível ter certeza se o gene é realmente de cópia única. A estratégia mais utilizada é realizar comparações utilizando genomas próximos como modelo.

No caso de cana-de-açúcar, um organismo sem o genoma sequenciado, a utilização de genomas do sorgo, milho, arroz e arabidopsis (*Arabidopsis thaliana* (L.) Heynh) como referência é a melhor ferramenta para análises genômicas

comparativas, por serem parentes próximos de cana e apresentarem organização estrutural genômica similares, ou por serem considerados como planta modelo em estudos, no caso de arabidopsis.

Além da utilização de organismos filogeneticamente próximos, outra ferramenta utilizada para acessar o genoma de cana-de-açúcar são os BACs, que tem se mostrado uma ótima e eficiente estratégia para recuperar informações genômicas. Além disso, associar tais informações a transcriptomas, QTLs, mapeamento genético e sintenia com outras espécies pode ser a melhor maneira de se encontrar respostas sobre a genética e genômica de cana-de-açúcar. Assim, o desenvolvimento de uma ferramenta robusta para a integração de todos esses dados se torna necessário e pode trazer benefícios acadêmicos e econômicos para o Brasil, nos colocando em posição de destaque perante o mundo.

Entendendo esse cenário foi preciso criar ferramentas biológicas de alta qualidade para acessar de maneira eficaz o genoma da cana-de-açúcar. A solução encontrada foi construir duas bibliotecas de BACs para as variedades brasileiras SP80-3280 e SPIAC93-3046 e toda a metodologia de validações dos clones, plataformas de seleções e sequenciamentos de genes de interesse.

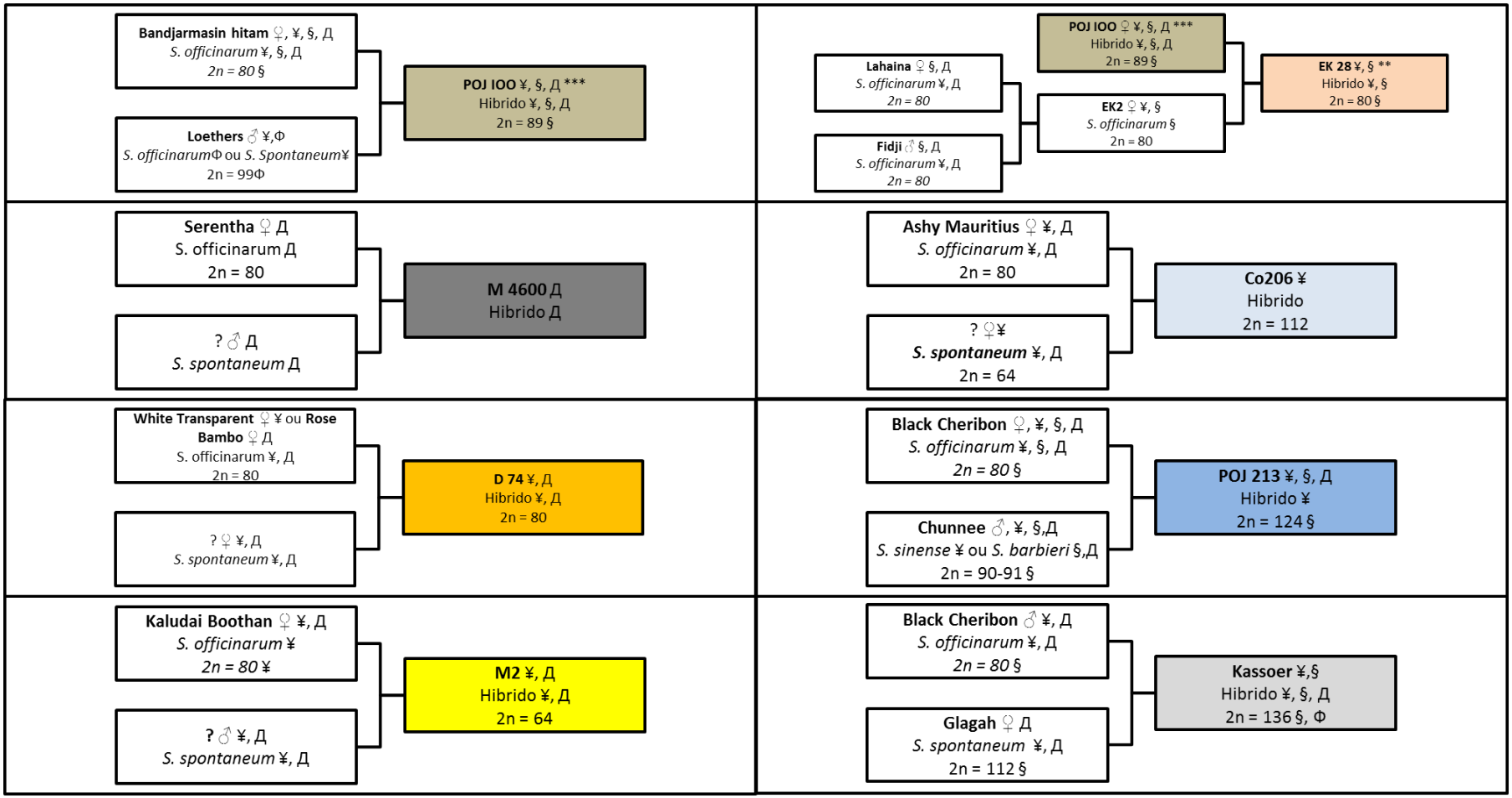


Figura 01: Acessos fundadores dos primeiros cruzamentos de cana-de-açúcar, que deram origem as variedades SP80-3280, SPIAC933046 e R570. Os acessos *Bendjermasin Hitam*, *Loethers*, *Fidji*, *Black Cheribon*, *Chunee* e *Glagah* são cana-de-açúcar de Java e os acessos *Saretha*, *White Transparent*, *Kaludai Boothan*, *Ashy Mauritius* e *Black Cheribon* da Índia. Os cruzamentos

iniciais feitos pelo *Proefstation Oost Java* são nomeados como cultivares POJ e aqueles conduzidos em Coimbatore, Índia, são nomeadas como cultivares Co.

¥: Stokes e Tysdal, 1962; §: Gupta e Tsuchiya,1991; Д: TROPGENE, CIRAD; Φ: Bremer, 1925;

*** POJ IOO foi originado de uma panícula de cruzamento aberto de Bandjermasin Hitam. Loethers é normalmente indicado como o parental masculino, mas é apenas uma suposição (Stokes e Tysdal, 1962).

** Origem questionável.

?: Desconhecido.

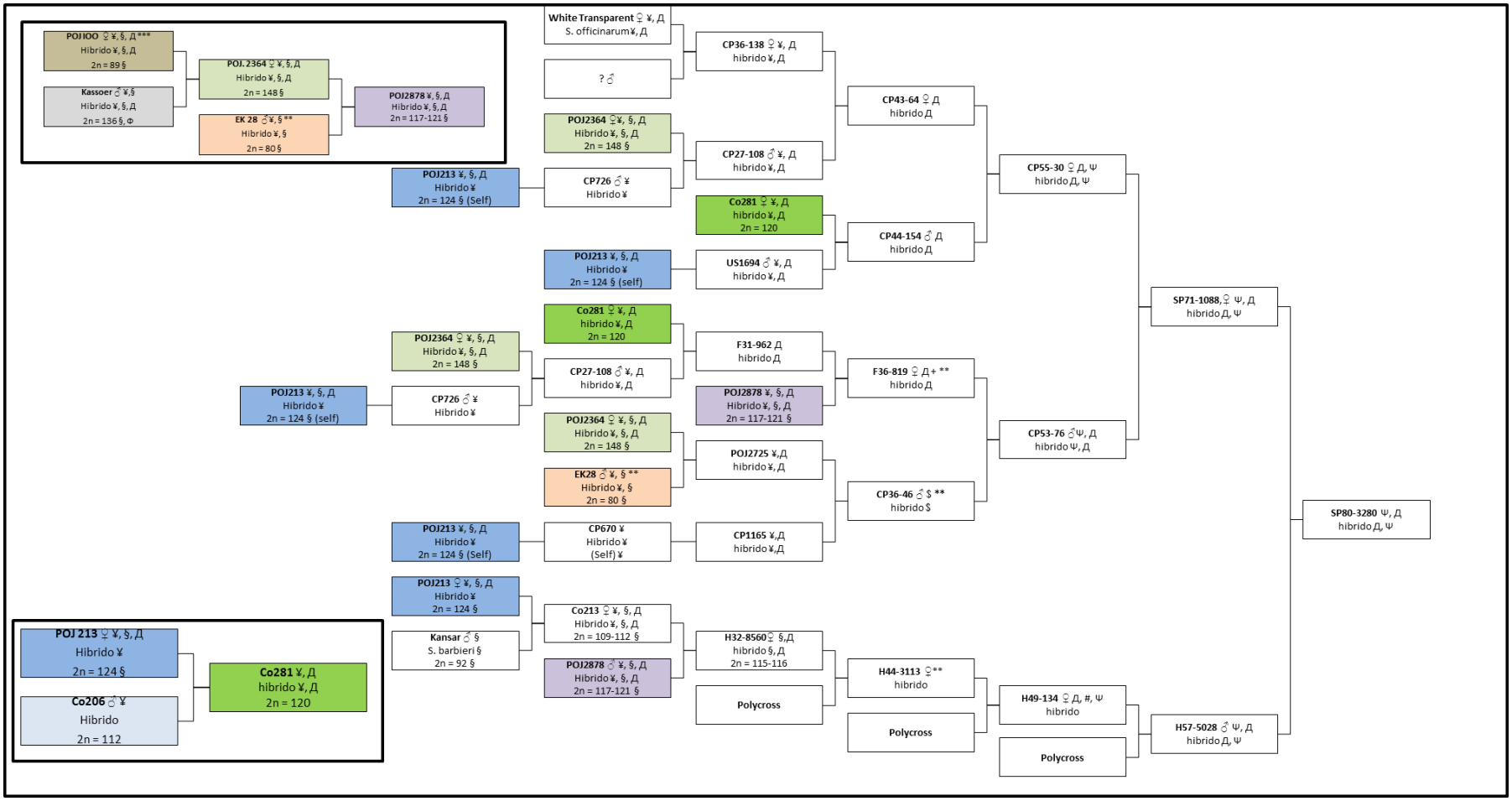


Figura 02: Pedigree da variedade SP80-3280. A partir de dados de literatura, foi possível determinar o pedigree completo da variedade SP80-3280. Os cruzamentos iniciais feitos pelo *Proefstation Oost Java* são nomeados como cultivares *POJ* e aqueles conduzidos em Coimbatore, Índia, são nomeados como cultivares *Co*. Os cultivares POJ213 e Co281 domina os cruzamentos iniciais da variedade SP80-3280. POJ213 destaca-se por introduzir a variabilidade de *S. barberi* (assim como cruzamentos

derivados do acesso *Kansar*) através do acesso *Chunee*, porém é susceptível ao mosaico, podridão vermelha e carvão. Co281 é fruto do cruzamento entre POJ213 e Co206, tolerante ao mosaico, mal de *Sereh* e a podridão vermelha. POJ2878 pode ser encontrado no pedigree de quase todas as variedades dominantes cultivadas em todo o mundo. *SP* e *IAC* variedades criadas no Brasil. *CP*, *US* e *H* dos EUA.

A SP80-3280 tem o parental H57-5028, um hibrido provindo de policruzamento. Policruzamento (*Polycross*) é quando um grande número de genótipos é utilizado como parental masculino, através da mistura do pólen, o que impede a identificação da fonte de pólen. O policruzamento é vantajoso por que resulta em um grande número de sementes obtidas quando comparado aos cruzamentos bi-parentais, bem como a maior variabilidade dentro da população F1 gerada.

¥: Stokes e Tysdal, 1962; §: Gupta e Tsuchiya,1991; Д: TROPGENE, CIRAD; Φ: Bremer, 1925; Ψ: Marconi et al, 2011; $: Catálogo Nacional de variedades “RB” de cana de açúcar – Março de 2010; # Price, 1969; &: Vieira et al, 2018; @: Srivastava et al, 1994;

*** POJ IOO foi originado de uma panícula de cruzamento aberto de Bandjermasin Hitam. Loethers é normalmente indicado como o parental masculino, mas é apenas uma suposição (Stokes e Tysdal, 1962).

** Origem questionável

+: TROPGENE (CIRAD) afirma que F36-819 é uma autofecundação de POJ2878 e Barbosa e colaboradores (2001) afirma ser um cruzamento entre F31-962 x POJ2878

?: Desconhecido.

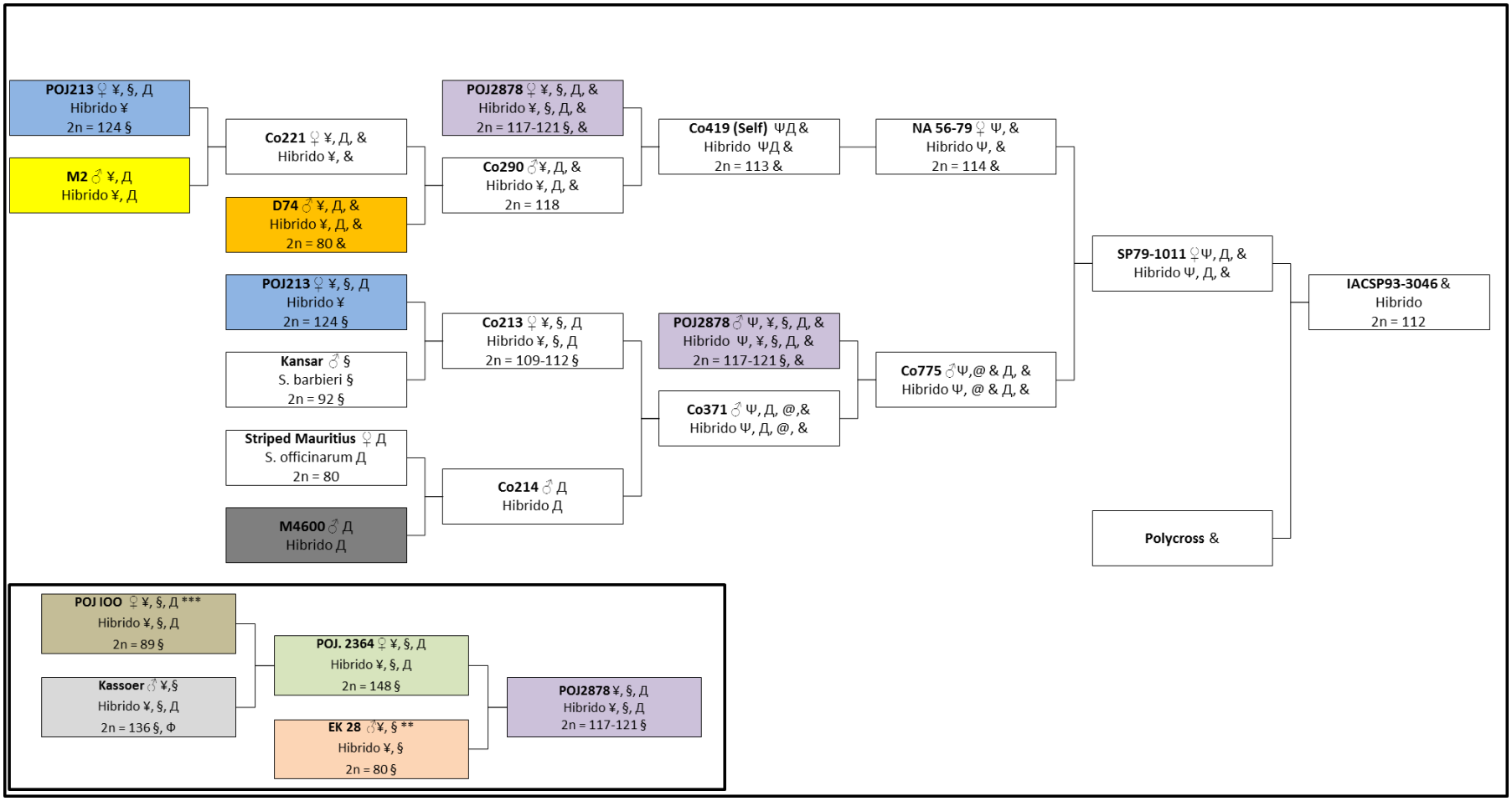


Figura 03: Pedigree da variedade IACSP93-3046. A partir de dados de literatura, foi possível determinar o pedigree completo da variedade IACSP93-3046. Os cruzamentos iniciais feitos pelo *Proefstation Oost Java* são nomeados como cultivares *POJ* e aqueles conduzidos em Coimbatore, Índia, são nomeados como cultivares *Co*. A variedade POJ2878 pode ser encontrado no pedigree de quase todas as variedades dominantes cultivadas em todo o mundo. A POJ2878 apresentou uma inovação para

época, com características muito importantes: crescimento rápido, bom perfilhamento, caules retos, internódios longos, alta produção, forte sistema de raízes, folhas não muito largas, imunidade doença de Sereh, mosaico e podridão radicular, alto teor de açúcar, alta porcentagem de germinação. *SP* e *IAC* variedades criadas no Brasil. *NA* na Argentina. *CP*, *US* e *H* nos EUA. A IACSP93-3046 possui o parental masculino vindo de policruzamento.

¥: Stokes e Tysdal, 1962; §: Gupta e Tsuchiya,1991; Д: TROPGENE, CIRAD; Φ: Bremer, 1925; Ψ: Marconi et al, 2011; $: Catálogo Nacional de variedades “RB” de cana de açúcar – Março de 2010; # Price, 1969; &: Vieira et al, 2018; @: Srivastava et al, 1994.

*** POJ IOO foi originado de uma panícula de cruzamento aberto de Bandjermasin Hitam. Loethers é normalmente indicado como o parental masculino, mas é apenas uma suposição (Stokes e Tysdal, 1962).

** Origem questionável

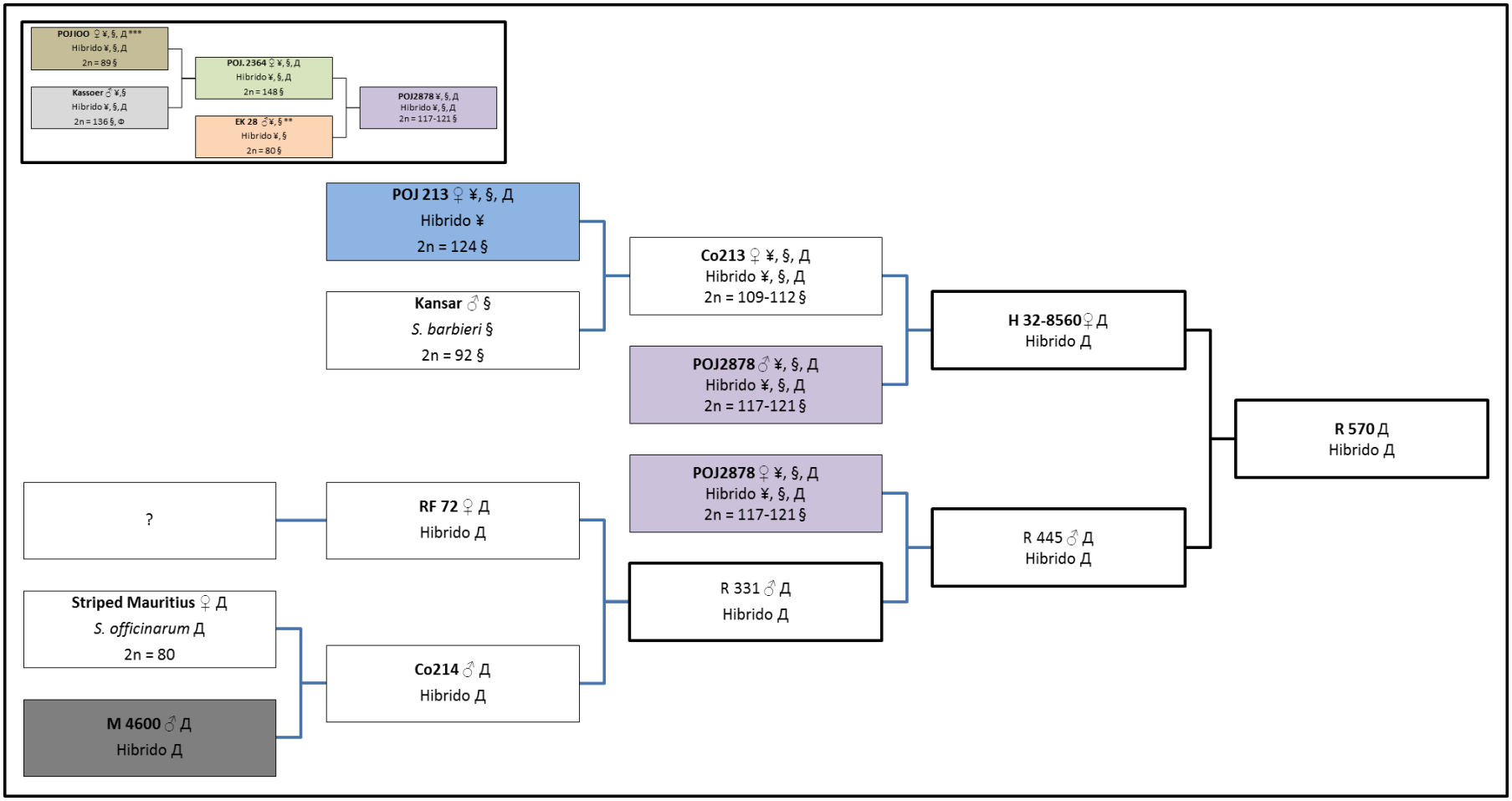


Figura 04: Pedigree da variedade R570. A partir de dados de literatura, foi possível determinar o pedigree da variedade R570. Os cruzamentos iniciais feitos pelo *Proefstation Oost Java* são nomeados como cultivares *POJ* e aqueles conduzidos em Coimbatore, Índia, são nomeados como cultivares *Co*. A variedade POJ2878 pode ser encontrado no pedigree de quase todas as variedades dominantes cultivadas em todo o mundo e na R570 contribui como parental das duas variedades que originam a R570. Isso

diminui a variabilidade, tanto geneticamente como agronomicamente, quando comparada com as variedades brasileiras SP80-3280 e IACSP93-3046. *SP* e *IAC* variedades criadas no Brasil. *H* nos EUA. A IACSP93-3046 possui o parental masculino vindo de policruzamento.

¥: Stokes e Tysdal, 1962; §: Gupta e Tsuchiya,1991; Д: TROPGENE, CIRAD; Φ: Bremer, 1925; Ψ: Marconi et al, 2011; $: Catálogo Nacional de variedades “RB” de cana de açúcar – Março de 2010; # Price, 1969; &: Vieira et al, 2018; @: Srivastava et al, 1994.

*** POJ IOO foi originado de uma panícula de cruzamento aberto de Bandjermasin Hitam. Loethers é normalmente indicado como o parental masculino, mas é apenas uma suposição (Stokes e Tysdal, 1962).

** Origem questionável.

+: TROPGENE (CIRAD) afirma ser uma autofecundação de POJ2878 e Barbosa et al (2001) afirma ser um cruzamento entre F31-962 x POJ2878.

?: Desconhecido.

Material e Métodos

Construção das Bibliotecas de BAC

Duas bibliotecas de BAC para as variedades brasileiras SP80-3280 e SPIAC93-3046 foram construídas em colaboração com o CNRGV/INRA (*Centre National de Ressources Genomiques Vegetales / Institut National de la Recherche Agronomique*) na França. O material vegetal utilizado na extração do DNA nuclear de alto peso molecular (*High Molecular Weigth – HMW*) foi obtido de palmitos de cana-de-açúcar (meristema apical caulinar protegido por folhas novas - SP80-3280) e/ou da folha jovem (SP80-3280 e IAC SP93-3046), ambos coletados no campo, transportado em gelo seco e armazenado em ultrafreezer. A figura 05 representa o processo completo para a obtenção das bibliotecas de BAC, desde a coleta do material vegetal, passando pelas etapas de seleção de tamanho até o armazenamento correto das placas.

O protocolo para a construção de Bibliotecas de BACs utilizado foi descrito por Peterson e colaboradores (2000) com modificações descritas por Gonthier e colaboradores (2010). De maneira resumida, os tópicos "*Isolamento do DNA nuclear de alto peso molecular e preparação dos plugs*", "*Teste de digestão parcial do DNA de alto peso molecular (HMW)*", "*Seleções de tamanho*", "*Isolamento do DNA selecionado da agarose*", "*Ligação e transformação*" e "*Estimativa do tamanho do inserto*" abordam as principais etapas.

Isolamento do DNA nuclear de alto peso molecular e preparação dos *plugs*

Cerca de 10 mg tecidos vegetais foram triturados em nitrogênio líquido, os núcleos isolados primeiramente por lise celular, seguida por tampão contendo β–mercaptoetanol e centrifugações. Os núcleos foram incorporados a *plugs* de agarose de baixo ponto de fusão (Agarose *low-melting point*, LMP) por incubação com leve agitação em tampão de lise. A construção do banco de BAC requer a geração de grandes fragmentos de DNA, variando de 100 a 350kb.

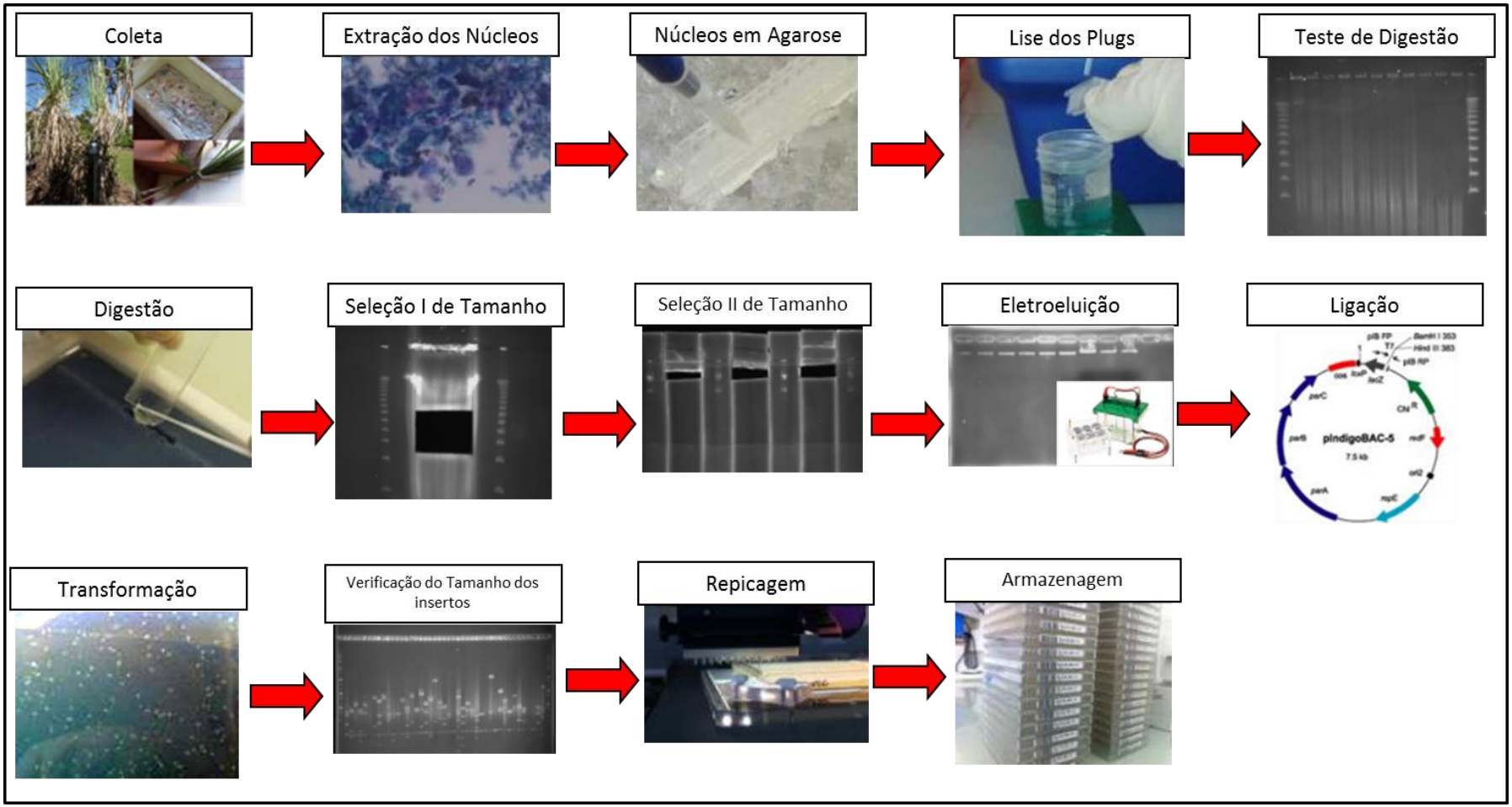


Figura 05: Processo utilizados para a construção da biblioteca de BAC.

Teste de digestão parcial do DNA de alto peso molecular (HMW)

Os *plugs* contendo DNA HMW foram parcialmente digeridos com a enzima de restrição HindIII. Uma série de digestões parciais em diferentes concentrações da enzima foram necessárias para obtenção de maior quantidade de fragmentos nos tamanhos desejados.

Foram testados valores de concentração da enzima de restrição HindIII, variando entre 0 (onde é possível verificar se há ou não degradação do DNA) até 100 U/ml (Tabela 01). Foram preparadas duas diluições da enzima: diluição 1- 2 U/µl e diluição 2- 0,2 U/µl. Para o teste foram usados ¼ de *plug* para cada uma das concentrações e a resolução foi obtida em gel de agarose para *Pulsed Field* 1% TBE 0.25X.

Tabela 01: Variação das diluições da enzima de restrição HindIII para testar a melhor concentração, usada para digestão parcial dos plugs.

Identificação	Quantidade da solução da Enzima adicionada ao tubo	Quantidade de Enzima (Unidades)	Concentração da enzima final em cada tubo (unidades/ml)
A	0	0	0
B	0,625µl da diluição 2	0,125	0,5
C	1,25µl da diluição 2	0,25	1
D	1,875µl da diluição 2	0,375	1,5
E	2,5µl da diluição 2	0,5	2
F	3,75µl da diluição 2	0,75	3
G	5µl da diluição 2	1	4
H	7,5µl da diluição 2	1,5	6
I	1,25µl da diluição 1	2,5	10
J	2,5µl da diluição 1	5	20
K	2,5µl da enzima pura	25	100

Seleções de tamanho

Uma vez que as condições ótimas para a produção dos fragmentos entre 100 e 350kb foram determinadas, utilizou-se de 8 a 15 *plugs* (dependendo da concentração dos mesmos) para as seleções de tamanho utilizando eletroforese de campo pulsado (Pulsed Field / PFGE). A primeira seleção removeu o DNA menor que 100kb, estes fragmentos poderão competir com a extremidade do vetor e

resultar em uma biblioteca com fragmentos muito pequenos. A segunda seleção removeu fragmentos menores que tenham se misturados aos maiores e condensa os fragmentos alvos (entre 100kb e 300kb) para que possam ser incisados do gel em um menor fragmento de agarose e posteriormente eletroeluídos.

Isolamento do DNA selecionado da agarose

Os fragmentos parcialmente digeridos, recuperado da segunda seleção de tamanho, foram eletroeluídos da agarose para serem usados nas reações de ligação. Normalmente são incisados três blocos de agarose da segunda seleção de tamanho. Desta maneira, procede-se com os três individualizados, ou seja, uma ligação para cada fração.

Ligação e transformação

Cada fração de DNA eletroeluído foi ligado ao vetor de clonagem pIndigoBAC5 e posteriormente transformados em *E. coli* DH10B, pelo método de eletroporação. As ligações foram feitas utilizando uma proporção aproximada de 3 ng de fragmentos para 1 ng de vetor pIndigoBAC-5 HindIII Cloning Ready (Epicentre). As transformações foram realizadas adicionando 13 µL da ligação em um volume de 100 µL de células competentes ElectroMAX DH10B T1 Phage Resistent Cells (Invitrogem). Esse volume foi então separado em cinco cuvetas de eletrotransformação e eletroporada a 3000V. Para a verificação de clones positivos, 50 µL da transformação foram plaqueadas em placas de Petri e avaliadas quanto à eficiência de transformação.

Estimativa do tamanho do inserto

Para cada fração foram escolhidos randomicamente até 43 clones de BAC e crescidos em meio líquido LB com cloranfenicol a 37° C por 18 h. Os plasmídeos foram isolados através de lise alcalina, a estimativa do tamanho dos insertos para cada fração foi realizada através da digestão dos clones BAC com a enzima de restrição NotI e a resolução através de PFGE em 1% de agarose em 0,5 X TBE.

Cada fração foi analisada individualmente quanto ao tamanho do inserto, se a fração possuísse um tamanho médio inferior a 100kb ou vetores vazios (falsos

positivos) ela era descartada. Caso contrário, as colônias brancas (que possuem o inserto) foram repicadas usando o robô Qbot (Genetix) e armazenadas em placas de 384 poços contendo meio LB glicerolado. Foram feitas e armazenadas separadamente em ultrafreezer -80ºC três cópias da biblioteca (uma de estoque, uma para desenvolver os trabalhos posteriores e uma para ser enviada para o Brasil).

Sequenciamento das pontas de BACs

O sequenciamento de pontas de BACs é um método usado para se conseguir sequências aleatórias do genoma de maneira rápida. O DNA de BAC foi extraído e a extremidade pareada foi sequenciada usando o sequenciador Sanger 3500xL (Applied Biosystems). O programa Phred foi usado para chamadas de base e o Cross-match foi usado para aparar sequências de vetores. As sequências foram analisadas utilizando o RepeatMasker (Smith et al., 2013) para localizar possíveis elementos repetitivos. Os programas Gramene *SSRIT Tool* (Temnykh et al., 2001) e Blast2go (Götz et al., 2008) foram usados para encontrar microssatélites e pesquisar genes homólogos no banco de proteínas NCBI não redundantes (Nr), respectivamente.

Busca por genes de cópia única

Sequências expressas disponíveis na base de dados do SUCEST (sucest-fun.org) foram comparadas com as sequências expressas de sorgo que codificavam genes de cópias únicas. Posteriormente essas sequências pré-selecionadas foram comparadas com sequências de arroz e de Arabidopsis, também comparada com genes de cópia única nessas espécies. Buscas no OrthoDB (Kriventseva, et al, 2018) também foram realizadas para confirmação das análises utilizando genes de gramíneas.

Em todas as sequências selecionadas foi utilizado o programa Mega7 (Sudhir et al, 2017) para agrupar sequências de acordo com sua identidade e checar se os genes selecionados são dose única. Para cada gene comprovado como dose única foram desenvolvidos *primers* para serem utilizados na seleção de clones positivos na biblioteca de BACs da SP80-3280 e da SP93-3046.

Os pares de *primers* foram testados para um padrão claro de amplificação e bandas únicas. Para tal, o DNA total das variedades SP80-3280 e SPIAC93-3046 foram usados como *template* para a amplificação dos fragmentos especificados por cada par de *primer*. As reações continham cerca de 20 ng de DNA genômico, 1X tampão de PCR; 1,5 mM de MgCl2; 1 U de enzima Taq DNA polimerase; 0,2 mM de cada dNTP; 0,3 µM de cada *primer*, e água ultra-pura para um volume final de 16 µl. As amplificações foram realizadas em termociclador (Eppendorf) utilizando o seguinte programa: desnaturação inicial a 94°C por 5 min., seguida de 30 ciclos de 94°C por 40 segundos, 54°C por 40 segundos e 72°C por 1 min., e extensão final a 72 °C por 8 min. As amostras foram purificadas utilizando os procedimentos do kit QIAquick 96 PCR Purification Kit (Qiagen). Para verificar as amplificações, alíquotas de 5 µl de cada reação foram submetidas à eletroforese em gel de agarose 1% (p/v) e visualizadas sob luz UV, utilizando-se como marcador de massa molecular o ladder 100 pb.

Seleção de clones

A seleção dos clones foi baseada em duas estratégias: em hibridação por Macroarranjos e em PCR por construção de Pool 3D (Kim et al., 1996). Ambas as técnicas são apresentadas nos tópicos "*Screening por Macroarranjos*" e "*Screening por Pool 3D*"

Screening por Macroarranjos

Todos os clones obtidos pela construção da biblioteca foram espotados em pontos duplos usando uma matriz de 6 x 6 em membranas de nylon (Millipore), de 22.2 x 22.2 cm, carregadas positivamente, usando o robô Qpix2 (Genetix). Cada conjunto de duas membranas continha todos os clones da biblioteca. As membranas de nylon duplamente *espotadas* foram transferidas para placas do tipo Q-Tray contendo LB-agar+12,5 µg/ml de cloranfenicol. Os clones bacterianos cresceram por 17 h a 37°C, e, em seguida, foram transferidos para 4°C até as colônias ficarem confluentes.

As membranas foram mantidas durante 4 min em papel Whatmann 3 MM, saturado com tampão de desnaturação (0,5 M de NaOH; 1,5 M NaCl), tratadas durante 10 min. a 100°C com o mesmo tampão, e neutralizadas durante 10 min. em papel Whatmann 3MM saturado com tampão de neutralização (1,5 M Tris-HCl; pH

7,4; 1,5 M NaCl). Imediatamente as membranas foram incubadas a 37°C durante 45 min. com 250 mg/l de proteinase K em 100 mM de Tris-HCl pH 8,0; EDTA 50 mM; NaCl 0,5 M. Finalmente, as membranas foram secas durante 45 min. a 80°C e fixadas sob luz violeta (UV-crosslinked) durante 50 seg (120,000 µJ.cm-2). Em seguida, as membranas foram guardadas a temperatura ambiente até o uso. Os produtos de amplificação dos pares de primers desenvolvidos para cada gene de copia única (sondas) foram utilizados para o *screening* da biblioteca. Os filtros de alta densidade foram preparados usando o robô QBot (Genetix). Os clones foram posicionados em pontos duplos usando um arranjo 7x7 em filtro de nitrocelulose (Hybond NT). Esse padrão de agrupamento permite que 55.296 clones sejam representados por filtro (em duplicata). Foram necessários quatro filtros para representar a biblioteca da SP80-3280 e três filtros para a biblioteca da IACSP93-3046. Os clones foram crescidos por 18h, os filtros processados e, finalmente, o DNA para cada clone foi permanentemente fixado numa posição conhecida do filtro.

As sondas foram marcadas radioativamente e hibridizadas nas membranas, utilizando de um a quatro sondas em pool para cada membrana. Para a marcação das sondas foram utilizados, aproximadamente, 100 ng de DNA por membrana e o kit “Ready-To-Go DNA Labelling Beads (-dCTP)” (GE-Healthcare). O DNA foi desnaturado durante 10 min. a 100°C, resfriado imediatamente no gelo por 2 min. e, em seguida, centrifugado (brevemente). Neste tubo que contêm o DNA foi adicionada a esfera do kit que contém a enzima Klenow, os hexanucleotideos dATP, dTTP, dGTP e, cuidadosamente (com os aparatos necessários para proteção radiológica), 4,5 µl de [α^{-33} P] dCTP 50 µCi. Essa solução foi vortexada suavemente e deixada por 1 h e 30 min. a 37°C. Após este tempo, a sonda foi purificada utilizando o kit "Ilustra Probe Quant G-50 Micro Columms" (GE), que retêm os fragmentos de DNA menores que 50 pb e os dNTPs não incorporados. Primeiramente, as colunas foram preparadas para receber o DNA e, em seguida, os 50 µl de sonda marcada foram adicionados à coluna e centrifugados durante 2 min., a 14000 rpm. Após recuperar a sonda purificada, foi feita sua contagem utilizando 1 µl da sonda purificada e 4 ml de líquido de cintilação em tubo específico para o contador de cintilação líquida Triathler LSC (Hidex).

Antes de iniciar a hibridização, as membranas foram incubadas durante 15 min. em 6X SSC a 50°C. Em seguida, as membranas foram pré-hibridizadas durante 3 h e 30 min. em 50 ml de tampão de hibridização (6X SSC; 5X Denhardt; 0,5% SDS;

100 µg.ml-1 DNA de esperma de salmão desnaturado) a 68°C. As hibridizações foram feitas com altas condições de estringência a 68°C, overnight, usando 50 ml de tampão de hibridização fresco suplementado com a sonda desnaturada. As membranas foram lavadas por 15 min., a 50°C, em tampão 2X SSC - 0,1% SDS, seguindo-se uma segunda lavagem a 50°C, por 30 min., em tampão 0,5X SSC-0,1% SDS. Finalmente, elas foram envoltas em um filme plástico, expostas ao General Purpose PhosphorImager Screen (Amersham Biosciences) por três dias e, finalmente escaneadas usando o Storm System (Amersham Bioscience), com uma resolução de 50 µm. As análises para a identificação dos clones positivos foram feitas usando o software HDFR (Incogen). Os clones positivos foram identificados e rearranjados. Uma vez que a hibridização não é específica, podendo selecionar clones falsos positivos, uma nova reação de PCR com os primers que originaram as sondas foi utilizada para a validação dos clones positivos. A Figura 06 representa um esquema resumido da construção dos Macroarranjos.

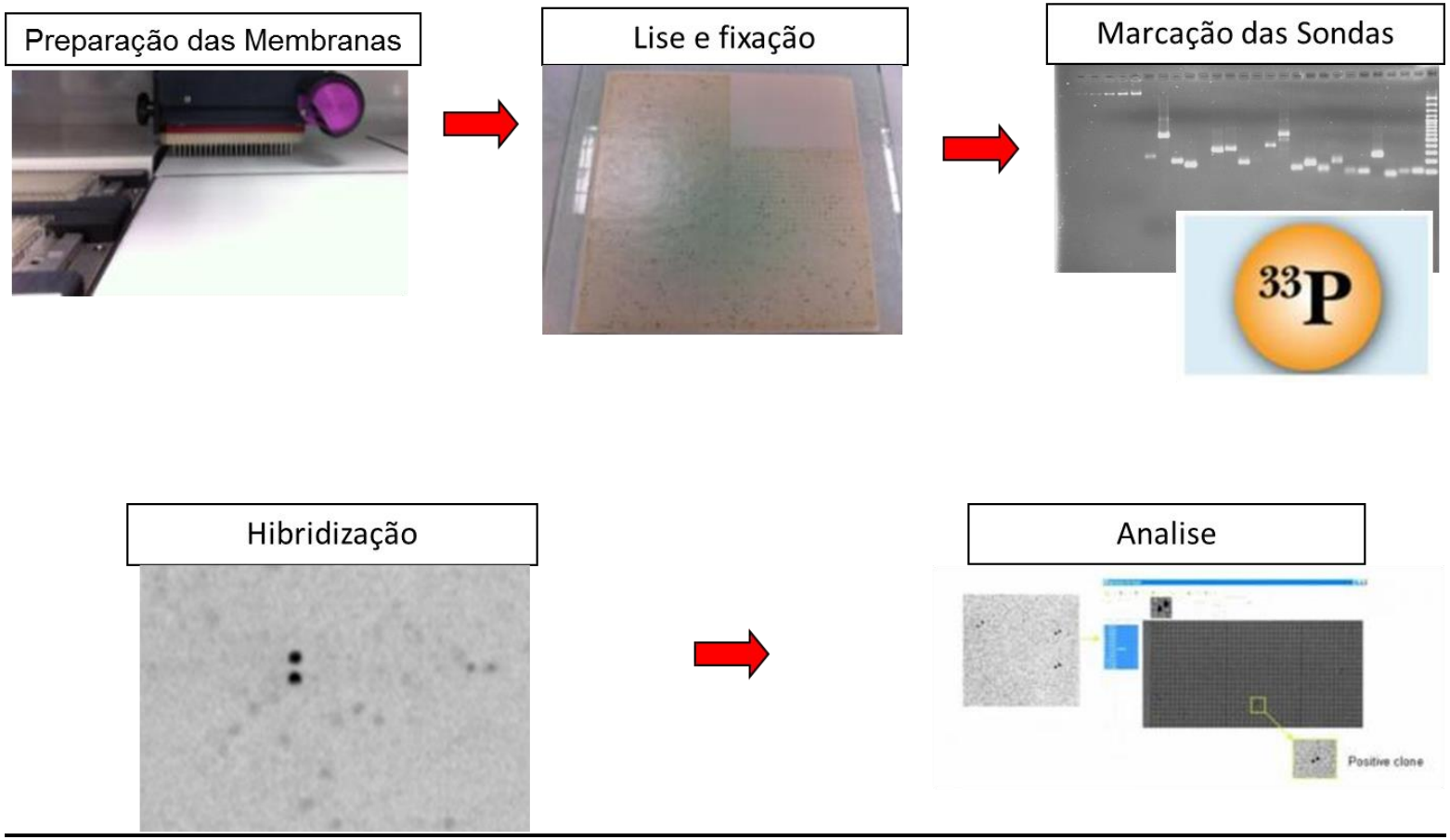


Figura 06: Esquema mostrando a construção dos Macroarranjos.

Screening por *Pool 3D*

A técnica é descrita por Kim e colaboradores (1996). É uma técnica baseada em misturas ordenadas das colônias e é possível fazer o *screening* da biblioteca utilizando-se poucas reações de PCR. As desvantagens são tempo para que o *pool* seja construído e o custo elevado. Porém, os resultados são mais específicos, uma vez que é baseado em PCR, e depois de construído, a seleção por Pool 3D acaba sendo mais rápida e não demanda radioatividade. Assim como para a hibridização, é necessário o desenho de *primers* específicos. A Figura 07 é a representação esquemática da construção e utilização do *Pool3D*.

Primeiramente as placas de 384 poços das bibliotecas foram divididas em *superpools*, depois em *pools* de placas, *pools* de linhas e *pools* de coluna. Os *superpools* são a mistura dos clones de 24 placas de 384. Com esta ferramenta é possível descobrir em qual *pool* de placas seu clone alvo se encontra. Os *pools* de placas são a mistura de cada placa individual e dessa maneira é possível identificar em qual placa, dentro de cada *superpool*, seu clone alvo se encontra. O *pool* de linha mistura todos os clones da linha A de cada *superpool*, depois todos os clones da linha B de cada *superpool* e assim por diante, resultando em 16 *pools* de linha para cada *superpool*. Por fim, para os *pools* de coluna, misturam-se todos os clones da coluna um de cada *superpool*, depois a coluna dois e assim por diante até se obter 24 *pools* de coluna para cada *superpool*. Com isso, um PCR (*superpool*) é suficiente para verificar a presença de um clone dentro do (Superpool) e mais 64 PCRs (24 placas, 16 linhas e 24 colunas) para identificar a coordenada do clone.

Cada mix de clones (*pool*) foi amplificado com Phi29 DNA polymerase, que aumenta a quantidade de DNA disponível. Assim, usou-se cada mix como *template* para as reações de PCR e cada amplificação positiva gera uma coordenada identificando o *superpool*, o *pool* de placa, linha e coluna e, consequentemente, o(s) clone(s) alvo (Figura 06). Os clones selecionados foram, ainda, verificados por RT-PCR, eliminando falsos positivos.

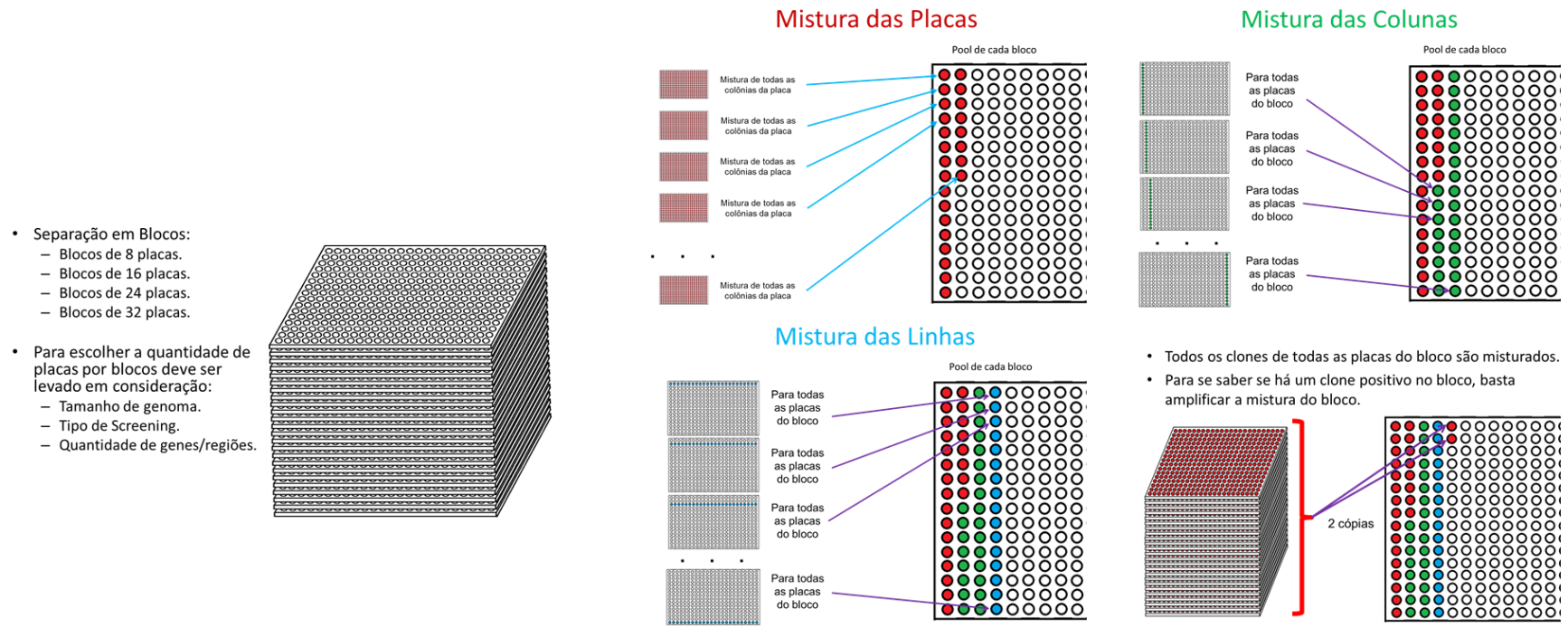


Figura 07: Esquema mostrando a construção e a utilização do Pool 3D. O esquema representa o pool de placas, 24 colunas e 16 linhas.

Sequenciamento completo dos clones e montagem

O sequenciamento foi realizado no INRA-CNRGV – Toulouse – França através do sequenciador 454 Titanium (GE). Os clones sequenciados foram montados utilizando o programa phred/phrap/consed (Gordon *et al*, 1998; Gordon *et al*, 2001; Gordon, 2004). Os *reads* que continham sequências do vetor e *E. coli* foram mascarados utilizando o programa *cross match* (Gordon *et al*, 1998; Gordon *et al*, 2001; Gordon, 2004), posteriormente unidos utilizando o programa phrap e visualizados no consed.

Resultados

Construção das Bibliotecas de BAC

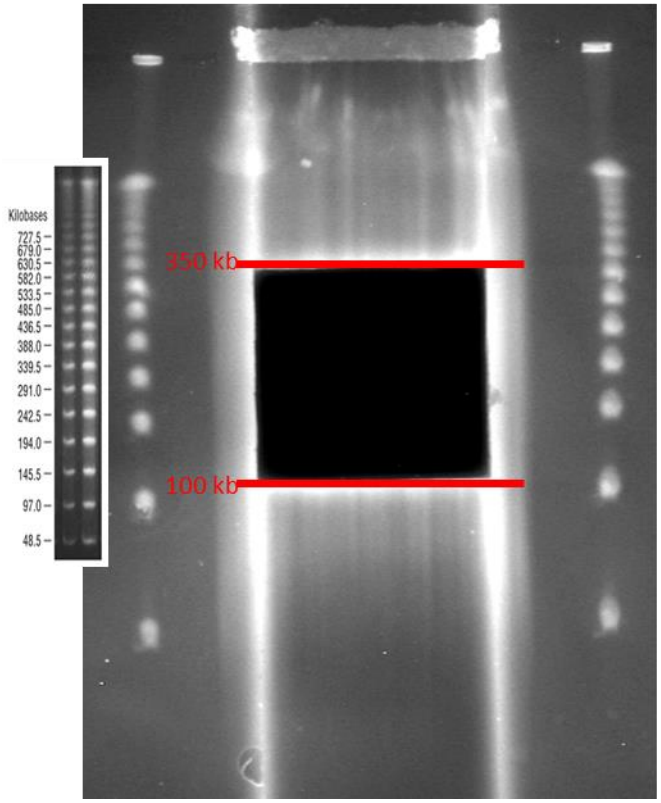
Para a construção da biblioteca da variedade SP80-3280 foram feitas três extrações de núcleos, uma a partir de folhas em estágio primordial e duas a partir de folhas jovens. A extração com folhas resultou em 38 *plugs* uniformes. Os *plugs* foram digeridos em tampão de lise e tratados com inibidor de protease Phenylmethylsulfonyl Fluoride (PMSF). O teste de digestão mostrou melhores resultados com concentrações de enzima de 0,7 U/ml, 1.0 U/ml e 1.2 U/ml e todas foram usadas para a construção da biblioteca. Foram feitas sete seleções de tamanhos, sendo duas a partir do palmito e cinco a partir da folha jovem. Um total de 21 frações foram isoladas e visualizadas em gel de agarose 1% TAE 1X e foi possível obter transformantes positivos para todas as eletroeluições de todas as frações.

Para a construção da biblioteca da variedade SPIAC93-3046 foram feitas duas extrações de núcleos, uma a partir de folhas em estágio primordial e outra a partir da folha jovem. A extração a partir do palmito não demonstrou DNA presente nos *plugs*, apenas a extração a partir da folha jovem foi utilizada e o teste de digestão mostrou melhores resultados com concentrações de enzima de 1 U/ml e 1.5 U/ml. Foram feitas três seleções de tamanhos, resultando em nove frações e todas renderam transformantes positivos.

A primeira seleção (Figura 08a) remove o DNA menor que 100kb, pois estes fragmentos poderão competir com a extremidade do vetor e resultar em uma biblioteca com fragmentos muito pequenos. Os fragmentos parcialmente digeridos,

recuperados da segunda seleção de tamanho, foram eletroeluídos da agarose para serem usados nas reações de ligação. A segunda seleção (Figura 08b) “condensa” os fragmentos alvos (entre 100kb e 300kb) para que possam ser incisados do gel em um menor fragmento de agarose e posteriormente eletroeluídos. Foram incisados três blocos de agarose da segunda seleção de tamanho (Figura 08b). Desta maneira, procede-se com os três individualizados – ou seja, uma ligação para cada “fração”. A ligação entre o vetor de clonagem e grandes fragmentos de DNA é um passo crítico, uma vez que o vetor se liga facilmente a fragmentos menores e qualquer quantidade de fragmentos pequenos são preferencialmente ligados.

(a) Seleção de tamanho I :



(b) Seleção de tamanho II:

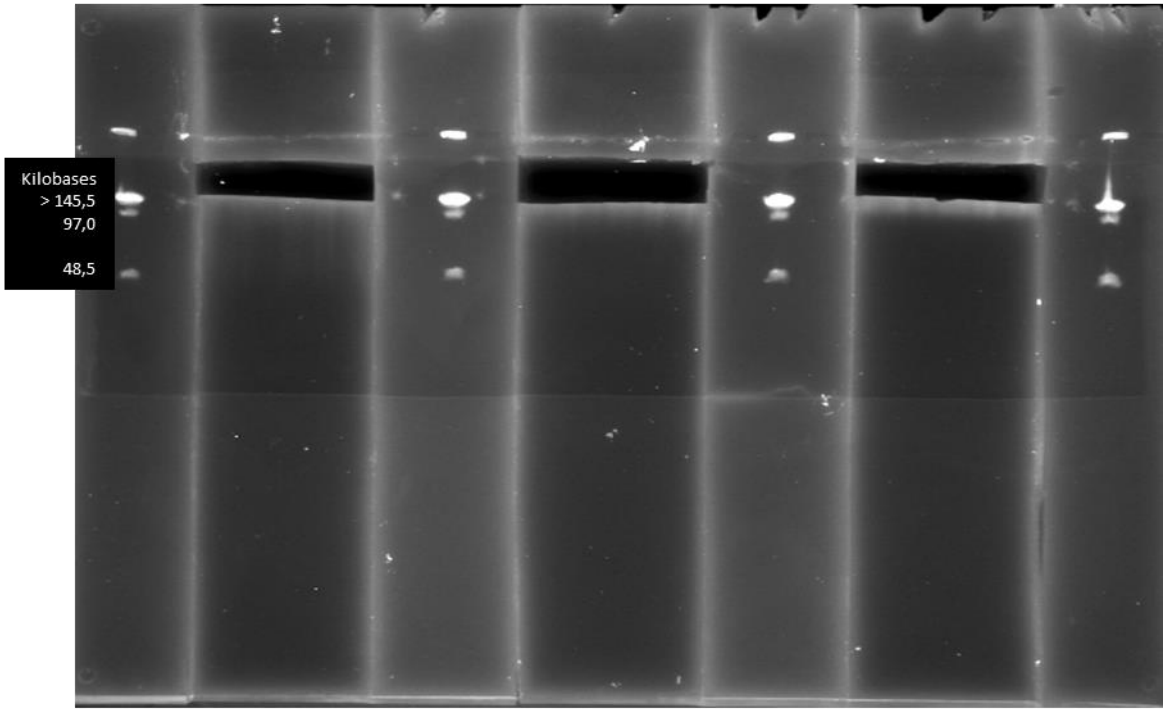


Figura 08: Géis de agarose 1% TAE 1X PFGE. (a) Primeira seleção de tamanho: Recuperação dos fragmentos entre 100-300 kb. (b) Segunda seleção de tamanho, onde os fragmentos são condensados (observar marcador) e os fragmentos maiores que 100 kb são recuperados.

Cada fração foi analisada individualmente quanto ao tamanho do inserto (Figura 09). Nesta etapa, foi possível verificar (1) se as colônias possuem insertos (falsos positivo); (2) o tamanho aproximado de cada inserto; (3) falhas na miniprep (não é possível visualizar nem vetor, nem inserto).

Estes resultados foram usados para verificar se a fração era adequada para ser usada na construção da biblioteca. Quando a fração possuía altas quantidades de vetor sem inserto e insertos menores que 95 kb, a fração foi descartada. As frações ainda foram descartadas quando a eficiência de transformantes foi muito baixa, uma vez que é necessária uma série de transformações para se conseguir poucos clones.

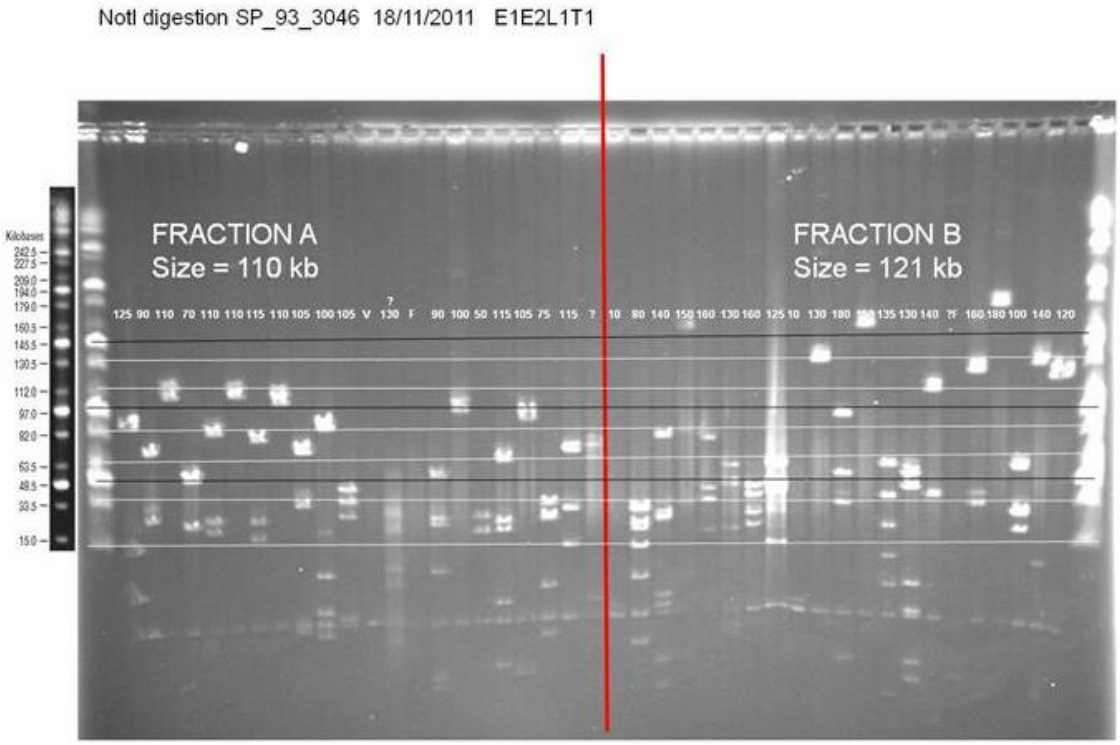


Figura 09: Estimativa do tamanho dos insertos. Os vetores foram digeridos utilizando a enzima NotI (NEB biolabs) e submetidos a corrida em gel de agarose 1% TAE 1X PFGE. Duas bandas (além da banda do vetor, comum a todos) na mesma pista mostram que o inserto também foi digerido, então se soma os tamanhos das bandas. Biblioteca SPIAC 93-3046 para as frações A (tamanho médio de 110Kb) e B (tamanho médio de 121Kb).

Após esta validação, as frações transformadas com tamanho ideal de insertos, sem vetores vazios e com boa eficiência foram plaqueadas em Q-Trays (placa de

Petri específicas para o Robo Q-Pix) e repicadas em meio líquido e armazenadas em ultrafreezer -80ºC.

Para a variedade SP 80-3280 foi possível obter 221.184 clones distribuídos em 576 placas de 384 poços, com um tamanho médio de 106kb, cobrindo 2,34 vezes o genoma da cana-de-açúcar (10 Gb). Para a variedade IACSP93-3046 foi possível obter 165.888 clones distribuídos em 432 placas de 384 poços, com um tamanho médio de 108kb, cobrindo 1,8 vezes o genoma da cana-de-açúcar (Tabela 02 e Figura 10).

Tabela 02: Cobertura das bibliotecas SP80-3280 e IACSP93-3046 estimada utilizando o genoma da cana-de-açúcar de aproximadamente 10Gb, de acordo com cada transformação, bem como o tamanho médio em kb e número de microplacas. Os códigos das transformações indicam o número da transformação e a letra a que fração de eletroeluição ela pertence (A, B ou C).

Variedade	Transformação	Tamanho médio (Kb)	Placas de 384	Cobertura
SP 80-3280	1A	98	30	0,11
	1B	98	112	0,42
	2A	120	56	0,26
	3A	115	56	0,25
	4A	120	91	0,42
	5A	100	96	0,37
	5B	100	135	0,52
Totais		106,07	576	2,35
SP IAC 93-3046	1A	82	19	0,06
	2A	81	107	0,33
	3A	110	40	0,17
	4A	121	40	0,19
	4B	121	52	0,24
	4C	121	37	0,17
	4D	121	60	0,28
	4E	121	39	0,18
	4F	121	38	0,18
Totais		108,36	432	1,80

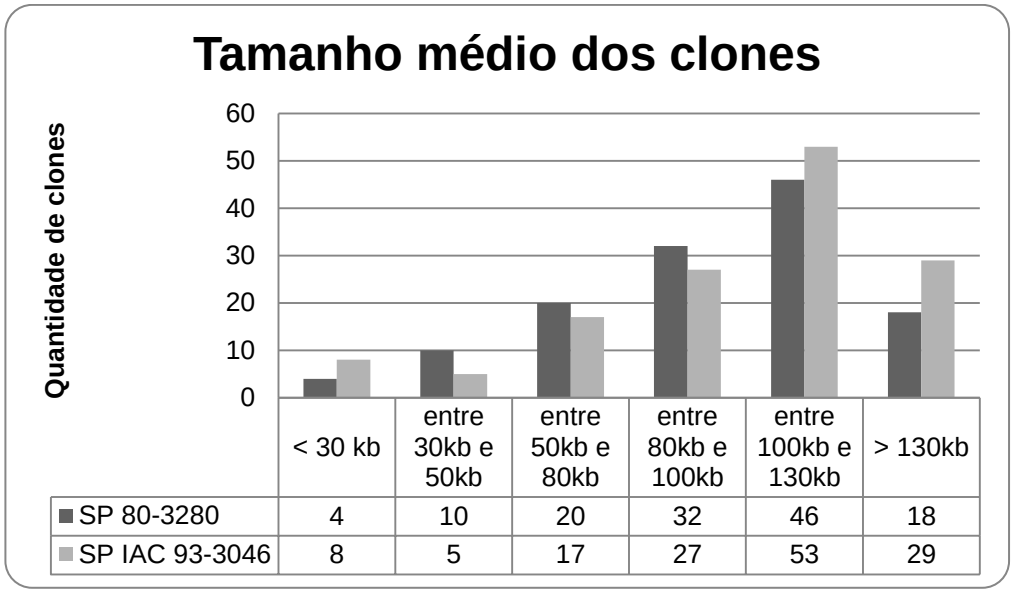


Figura 10. Quantidade de clones por intervalo de tamanho. As barras em verde escuro representam os clones da biblioteca SP80-3280, os em verde claro os clones da IACSP93-3046. Foram utilizados aproximadamente 100 insertos aleatórios para cada uma das bibliotecas. Destaque para poucos insertos inferiores a 80 kb e maior concentração acima de 100kb.

Sequenciamento das pontas de BACs

Inicialmente, o protocolo de sequenciamento de pontas não gerava sequências de qualidade, acreditando-se devido a baixa quantidade de DNA e pouco número de ciclos de PCR. Esse protocolo era o descrito em literatura (Paterson et al., 2000) e precisou ser modificado para se adequar à rotina do laboratório. O protocolo padrão previa a utilização de 2ug de DNA BAC e 25 ciclos de amplificação na reação de sequenciamento. Após testes e variando a quantidade de DNA BAC usado e a quantidade de ciclos, chegou-se a um protocolo otimizado onde se usou 5ug de DNA BAC e 99 ciclos de amplificação na reação de sequenciamento.

Para os BAC-end sequenciados da biblioteca SP80-3280 e IACSP93-3046, 650 (84,6%) e 723 (94,1%) sequências apresentaram mais de 400 bases com alta qualidade, respectivamente. Para SP80-3280, 103 sequências (13,4%) foram excluídas por tamanho (menores que 99 bases), e para IACSP93-3046, 45 sequências (5,9%). Foi possível encontrar 319 sequências de elementos repetitivos de diversas classes (Tabela 03) na biblioteca SP80-3280 (30% de todas as bases analisadas) e 368 na biblioteca IACSP93-3046 (30,75%).

Para a biblioteca SP80-3280 foram encontrados microssatélites em 27 sequências, enquanto 17 deles apresentaram motivos dinucleotídeos (63%) e os 10 restantes apresentaram motivos trinucleotídeos (37%). Para a biblioteca IACSP93-3046, foram encontrados microssatélites em 158 sequências, sendo que 123 deles apresentaram motivos dinucleotídeos (77,8%), 31 apresentaram motivos trinucleotídeos (19,6%) e os demais quatro apresentaram motivos tretranucleotídeos (2,5 %).

Para biblioteca da variedade SP80-3280 foram 26 sequências com alguma homologia a genes e 46 tiveram semelhança com os genes com função conhecida (Figura 11). Para o IACSP93-3046 resultou em 35 sequências com alguma homologia a genes e 52 com similaridade com genes anotados com função conhecida (Figura 12). Dentre os genes recuperados com anotação conhecida (Figura 11 e 12), quando observado os tipos de processos biológicos mais representados por esses genes, destacam-se genes relacionados a processos celulares e metabólicos. Quando funções moleculares são observadas, destacam-se genes relacionados a reconhecimento de sítios e atividade catalítica.

Tabela 03: Elementos repetitivos encontrados nas sequencias de BAC-END.

	SP80-3280			IACSP93-3046		
	Número de Elementos	Tamanho em bases	Porcentagem das sequencias	Número de Elementos	Tamanho em bases	Porcentagem das sequencias
Bases Masked	-	-	30,0%	-	-	30,75%
RETROELEMENTS	278	109302 bp	26.17 %	298	121689 bp	25.66 %
SINEs	-	-	-	1	167 bp	0.04 %
LINEs	6	3065 bp	0.73 %	15	5488	1.16 %
RTE/Bov-B	-	-	-	4	2084 bp	0.44 %
L1/CIN4	6	3065 bp	0.73%	11	3404 bp	0.72 %
LTR Elements	272	106237 bp	25.43 %	282	116034 bp	24.47 %
Ty1/Copia	153	71260 bp	17.06 %	153	72169 bp	15.22 %
Gypsy/DIRS1	119	34977 bp	8.37 %	129	43865 bp	9.25 %
DNA TRANSPOSONS	18	3912 bp	0.94 %	39	8485 bp	1.79 %
hobo-Activator	6	1669 bp	0.40 %	7	1382 bp	0.29 %
Tc1-IS630-Pogo	4	789 bp	0.19 %	6	1044 bp	0.22 %
En-Spm	-	-	-	1	783 bp	0.17 %
Tourist/Harbinger	3	301	0.07 %	12	1571 bp	0.33 %
Unclassified	23	8794 bp	2.10 %	31	11851 bp	2.50 %
Total interspersed repeats	-	114808 bp	27.49 %	-	130425 bp	27.51 %
Small RNA	1	127 bp	0,03 %	1	167 bp	0.04 %
Simple Repeats	61	2527 bp	0.61 %	78	3289 bp	0.69 %
Low Complexity	13	660 bp	0.16 %	9	435 bp	0.09 %

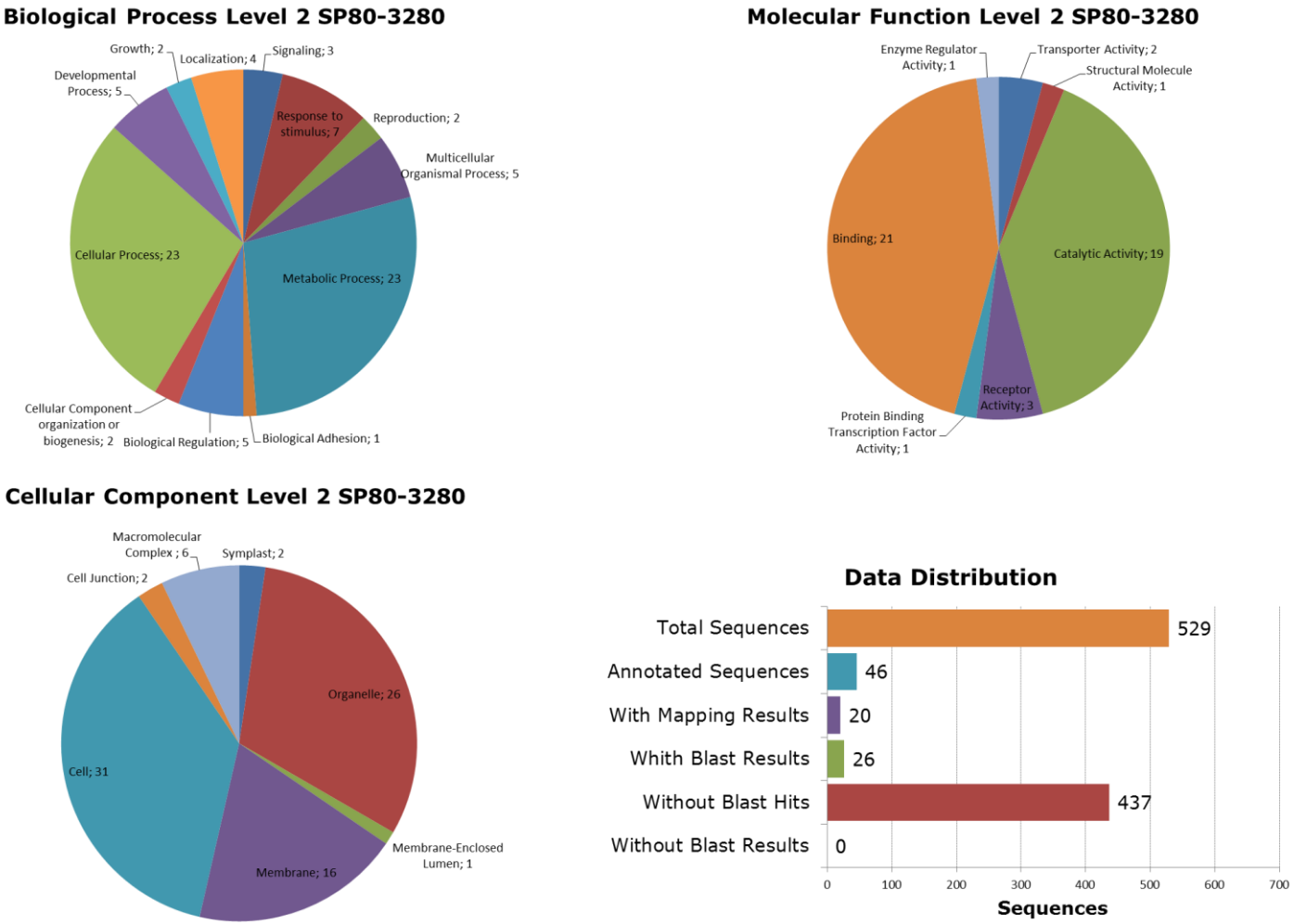


Figure 11: Resultado do Blast2Go para os BAC-End da biblioteca SP80-3280.

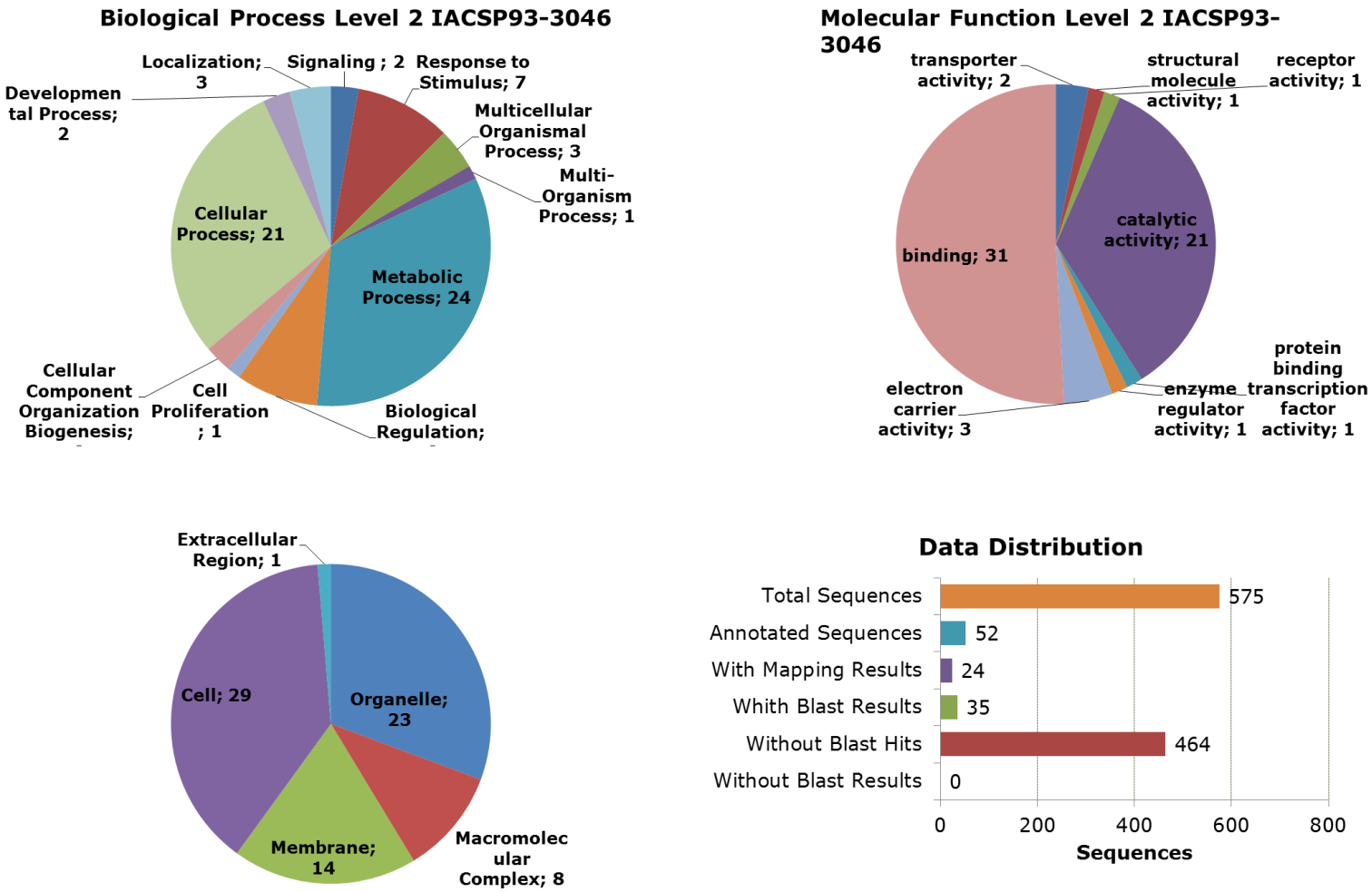


Figure 12: Resultado do Blast2Go para os BAC-End da biblioteca IACSP93-3046

Genes em cópia única

Foram encontrados 12 genes em possível cópia única. As análises foram feitas pelo grupo do Professor Michel Vincentz do CBMEG/UNICAMP (Centro de Biologia Molecular e Engenharia Genética / Universidade Estadual de Campinas). Os transcritos localizados em provável cópia única em *S. bicolor* tiveram seus ortólogos buscados no genoma de *O. sativa*, *A. thaliana*. Foi gerada uma arvore filogenética para cada um dos 12 genes utilizando o programa Mega7 (Sudhir et al, 2017). Para os 12 genes e seus ortólogos (de *O. sativa*, *A. thaliana* e *S. bicolor)* o agrupamento de cada gene mostrou a existência de apenas um gene filogeneticamente próximo, descartando a possibilidade de múltiplas cópias (Figura 13).

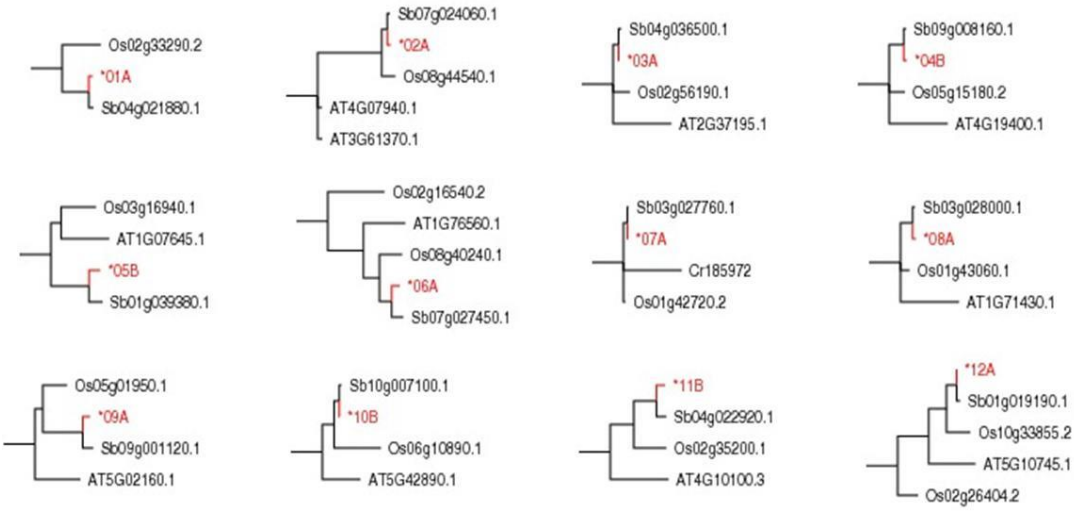


Figura 13. Comparação entre as sequências de genes cópia única de *Oryza sativa (*Os), *Arabidopsis thaliana* (AT) e *Sorghum bicolor* (Sb). Em vermelho o código de cada gene, segundo a Tabela 04.

De dois a quatro pares de *primers* foram desenvolvidos em todos os 12 genes cópia única. Os padrões de amplificação foram checados utilizando DNA genômico para a variedade SP80-3280. Três transcritos apresentaram padrão de amplificação inespecífico e um não amplificou, mesmo com o desenvolvimento de mais de um par de primer para cada gene. Os oito genes restantes que possuíram amplificação satisfatória foram utilizados na seleção de clones positivos (Tabela 04).

Os pares de primers foram utilizados para a seleção via Pool3D e os *amplicons* resultantes dos primers foram utilizados como sondas para a seleção em Macroarranjos.

Seleção de clones baseada em hibridação: Macroarranjos

A biblioteca SP80-3280 foi separada em quatro blocos de 144 placas para a construção dos macroarranjos. Já a biblioteca SPIAC93-3046 em três blocos de 144 placas. Utilizando o robô Qpix (Genetix) cada bloco foi impresso em duplicata em seis membranas de nylon de alta densidade em uma distribuição 7x7. As membranas contendo os clones foram colocadas em meio sólido contendo cloranfenicol, crescidas por 18h, tratadas, e o DNA dos clones permanentemente fixados a membrana.

Os produtos de amplificação dos oito genes (sondas) foram gerados e quantificados através da comparação por lambda DNA em diferentes concentrações. As sondas foram então purificadas, marcadas radioativamente e hibridizadas nas membranas. Por renderem poucos clones positivos, em torno de quatro a oito clones por membrana por gene, decidiu-se agrupar até quatro sondas por membrana, dessa forma, menos hibridizações foram feitas.

Após a hibridização, as membranas foram impressas em um Cassete com Screen 35x43cm, o scanner Storm 820 (GE) foi utilizado para leitura do screen. O programa *High Density Filter Reader* (versão 3) foi usado para analisar as membranas de macroarranjos 7x7 (Figura 14). Os clones positivos foram rearranjados e validados por PCR para cada um dos oito genes de interesse.

As hibridizações das membranas foram feitas em parceria com o CNRGV, sob supervisão da Dra. Hélène Berges. Para as duas bibliotecas, as hibridizações resultaram em 658 clones, sendo 302 e 356 clones para a biblioteca da SP80-3280 e IACSP93-3046, respectivamente. Os clones foram validados para cada um dos genes (Tabela 04), resultando em 388 (59%) clones validados e considerados positivos para pelo menos um gene, sendo 178 e 210 clones para a biblioteca da SP80-3280 e IACSP93-3046, respectivamente. A baixa eficiência dos resultados é inerente à técnica de hibridização, onde a baixa similaridade da sonda x BAC clone já é suficiente para que a hibridização aconteça, gerando muitos falsos positivos.

As hibridizações foram feitas utilizando até quatro sondas na mesma membrana ao mesmo tempo e em consequência disso, sondas que falharam na hibridização só

puderam ser detectadas na hora da validação. A intensidade do clone na exposição da membrana também pode causar a presença de falso positivo. A intensidade mínima para que o clone seja considerado positivo é definida no software e com a mistura de sondas de diversos genes, uma sonda menor pode mostrar um sinal mais fraco e com isso ser considerado negativo.

Tabela 04: Resultados obtidos na seleção dos clones e considerados positivos para cada gene na biblioteca das variedades SP80-3280 e SPIAC93-3046.

	Código do Gene	Origem Sorgo	SUCEST	Amplificação	Gene	Número de clones positivos da Biblioteca SP 80-3280	Número de clones positivos da Biblioteca SPIAC93-3046
1	**SC-01A**	Sb03g027760	SCRULB1057C11.g	OK	Proteína desconhecida	7	11
2	**SC-02B**	Sb07g024060	SCBGLR1098G02.g	OK	Proteína com função desconhecida (DUF3245)	21	19
3	**SC-03A**	Sb04g036500	SCBGLB2073E09.g	OK	Proteína hipotética	38	44
4	**SC-04B**	Sb09g001120	SCEZAD1081D01.g	OK	Proteína hipotética	26	31
5	**SC-05B**	Sb10g007100	SCAGLB1071C03.g	Inespecífico	Glyoxalase/Bleomycin Dioxygenase superfamilia	--------	----------
6	**SC-06A**	Sb06g027340	SCUTFL1063B08.g	Inespecífico	CP12 (Função desconhecida)	--------	-------
7	**SC-07A**	Sb07g027450	SCRFLR1055C03.g	OK	Proteína desconhecida	25	35
8	**SC-08A/HP600**	Sb03g028000	SCVPRZ2038E10.g	OK	Proteína Hipotética	25	28
9	**SC-09A**	Sb01g019190	SCEPRT2046C10.g / Sb5028971	OK	Chaperona DnaJ	16	23
10	**SC-10B**	Sb01g039380	SCMCCL6050C07.g	Inespecífico	Família SCP-2 sterol transfer	--------	---------
11	**SC-11B**	Sb09g008160	SCJFRZ2013A09.g	Não amplificou	Família MoaD	--------	---------
12	**SC-12A**	Sb04g021880	SCBGFL5079D09.g	OK	Proteína desconhecida	20	19

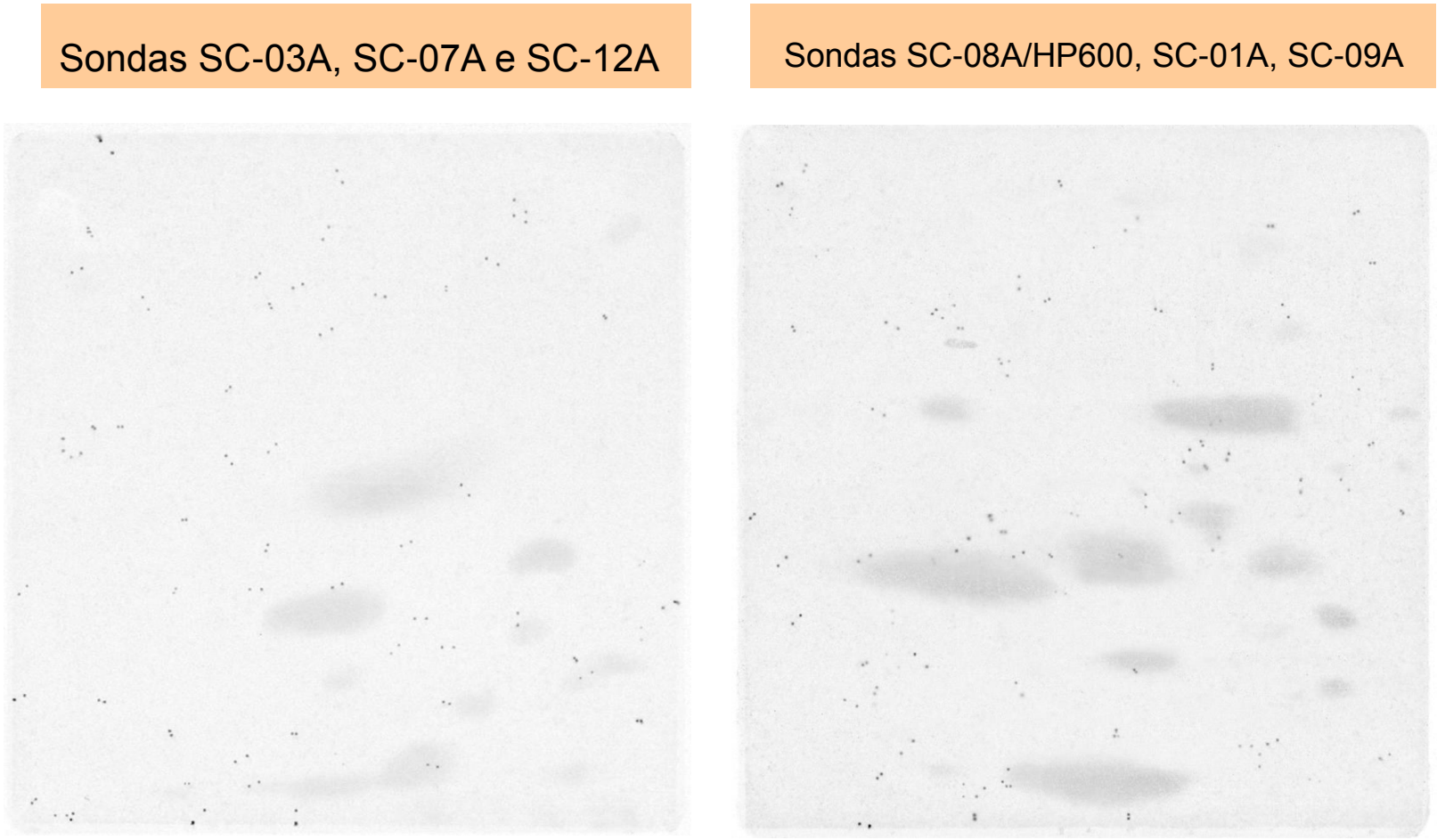


Figura 14: Hibridização demonstrando duas membranas da biblioteca da variedade IACSP933046 com 55.296 clones em duplicata. Pontos escuros mostram os clones positivos.

Seleção de clones baseada em PCR: Pool3D

Com o objetivo de diminuir o tempo utilizado para a construção de membranas e a não utilização de sondas radioativas, a técnica de criação de *pools* é utilizada. Foram construídos 12 *Superpools* (Tabela 05) para a biblioteca da SP80-3280, o que equivale à metade da biblioteca ou 110.592 clones. Cada *superpool* representa 24 placas da biblioteca, esse número de placas foi escolhido uma vez que representa, em teoria, um genoma "monoploide", ou seja, em cada *superpool*, existiria um clone positivo para dado gene único. Para cada *superpool* positivo, os pools de linha, coluna e placa foram amplificados para recuperar a coordenada dos clones positivos.

Tabela 05: Placas utilizadas para a montagem do Pool3D para a biblioteca SP80-3280.

Superpool	Placas	Clones	Tamanho Médio Inserto
Superpool 01	1-24	9216	98
Superpool 02	25-48	9216	98
Superpool 03	49-72	9216	98
Superpool 04	73-96	9216	120
Superpool 05	193-216	9216	115
Superpool 06	217-240	9216	100
Superpool 07	241-264	9216	100
Superpool 08	265-288	9216	100
Superpool 09	289-312	9216	100
Superpool 10	313-336	9216	120
Superpool 11	361-384	9216	100
Superpool 12	481-504	9216	100
Total	288	110592	104

O Gene SC-08A/HP600 e o SC-07A, foram utilizados para testar o Pool3D e os clones positivos recuperados com os obtidos pela técnica de Macroarranjos. O Pool3D recuperou os mesmos clones que já tinham sido recuperados pelos Macroarranjos, o que demonstra que ambas as técnicas são eficazes para a recuperação de clones que possuem gene/sequencias alvo.

Sequenciamento completo dos clones

O gene SC-08A/HP600 foi escolhido para o sequenciamento completo dos BACs e teve 22 dos 25 BACs para a variedade SP80-3280. O gene *HP600* (em sorgo Sobic.003G221600) provavelmente encontra-se em cópia única em cana e foi localizado em um QTL relacionado a acúmulo de açúcar em sorgo (Murray et al., 2008). O gene da Proteína Centromérica C (*CENP-C*) está localizado ao lado do gene *HP600* em cana-de-açúcar e sorgo. Ambos os genes foram utilizados para exemplificar o comportamento genômico e genético em cana-de-açúcar.

O tamanho esperado para cada clone foi estimado extraindo seus plasmídeos, digerindo o DNA com enzima NotI (New England Biolabs) e avaliando em gel de campo pulsado e na presença de um marcador conhecido (Mid Range Marker, New England Biolabs). Para determinar quais seriam sequenciados, iniciou-se com o sequenciamento das pontas de todos os clones para checar se algum deles possuía o gene nas pontas. Nenhum clone apresentou tal resultado. Os 25 clones foram digeridos com a enzima de restrição HindIII e submetidos a uma análise de *fingerprint* (Figura 15) para verificar se os clones eram iguais entre si. Apesar de os clones Shy035E13 e Shy218H04 terem sido considerados iguais, optou-se por sequenciá-los como uma maneira de verificar a qualidade da montagem, uma vez que ambos deveriam representar o mesmo haplótipo.

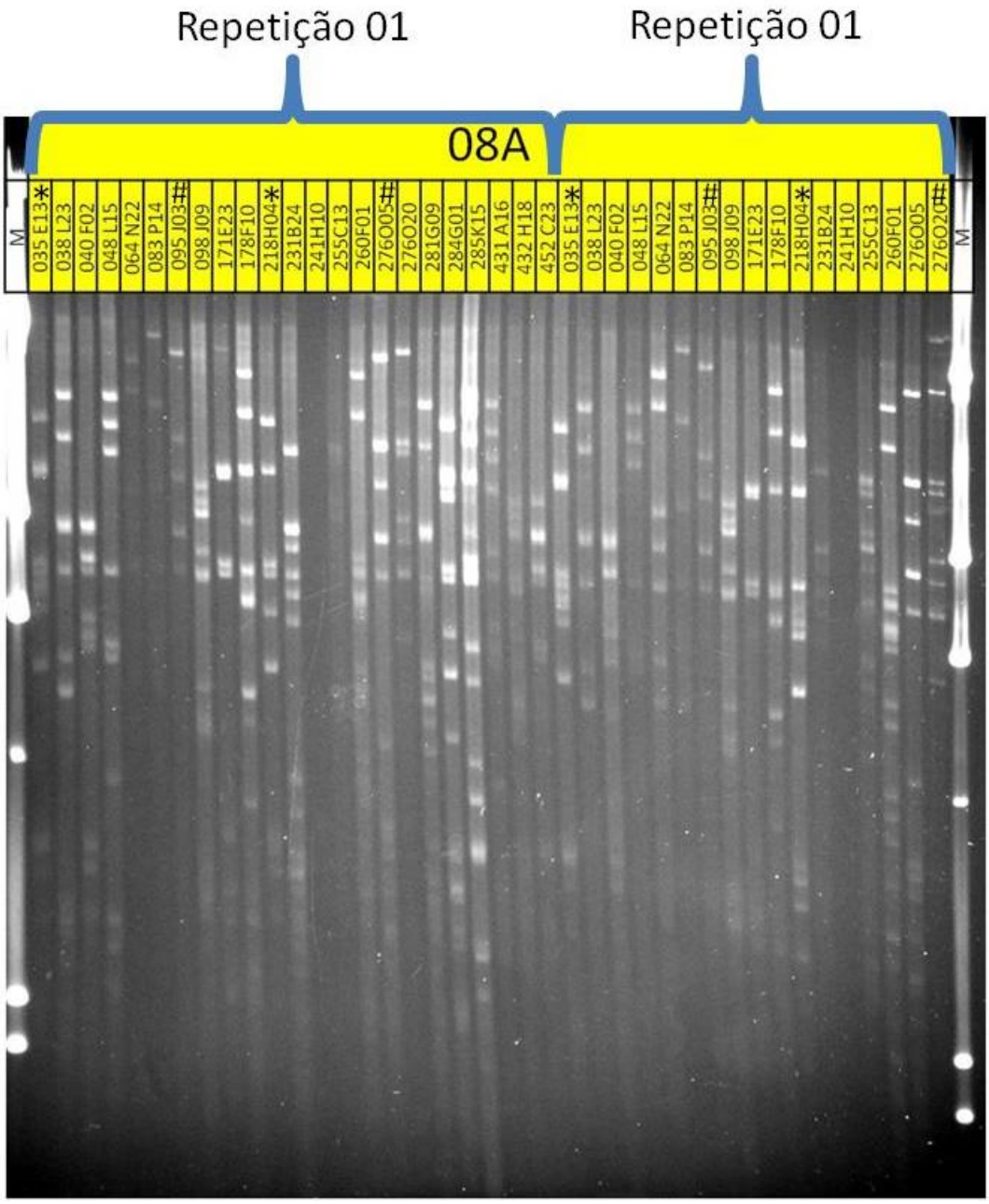


Figura 15. Gel de agarose 0,7% TAE 1X. Digestão com enzima HindIII (New England Biolabs) de 23 clones em duplicata positivos para o gene 08A. Clones indicados com "*" (035E13 e 218H04) e "#" (095J03 e 276O05) foram considerados iguais.

O sequenciamento resultou em uma cobertura média de 48 vezes cada BAC, sendo a cobertura mais baixa de 13x e a maior de 71x. Uma cobertura boa é considerada acima de 30x, sendo que 16 BACs obtiveram essa cobertura. O

tamanho médio esperado das sequências foi de 109 Kb, sendo o menor BAC com tamanho de 48Kb e o maior 162Kb (Tabela 06).

Tabela 06: *Resumo da montagem dos clones BACs.*

BAC	Reads	Tamanho médio das Reads (Bases)	Total de bases sequenciada	Tamanho (Bases)	Cobertura
Shy038L23	17,577	406	7,131,876	84,182	85
Shy064N22	10,412	413	4,297,624	91,701	47
Shy083P14	4,877	394	1,920,945	99,905	19
Shy098J09	6,157	402	2,474,363	98,874	25
Shy178F10	15,961	445	7,104,720	111,364	64
Shy241H10	31,589	290	9,146,219	134,894	68
Shy260F01	4,346	450	1,954,785	148,093	13
Shy281G09	34,579	298	10,298,703	130,914	79
Shy432H18	15,86	424	6,727,971	162,512	41
Shy035E13	6,188	398	2,462,177	105,606	23
Shy040F02	7,402	399	2,955,331	89,075	33
Shy048L15	4,909	384	1,884,792	83,194	23
Shy095J03	5,315	412	2,190,264	90,786	24
Shy171E23	6,349	448	2,845,515	48,796	58
Shy218H04	20,668	299	6,172,319	68,037	91
Shy231B24	22,282	296	6,588,557	107,057	62
Shy255C13	15,045	440	6,615,510	151,415	44
Shy276O20	16,919	443	7,490,042	105,869	71
Shy284G01	12,607	447	5,630,527	122,961	46
Shy285K15	14,393	436	6,273,629	99,815	63
Shy431A16	10,993	392	4,308,085	132,49	33
Shy452C23	16,361	283	4,637,892	100,514	46
Média	15,723	429	8,201,178	108,831	71

Os clones que resultaram em dois *contigs*, foram unidos verificando as sequências finais de cada *contig*: as pontas que terminaram com sequências do vetor, eram a extremidade e as sem essas sequências eram as internas. As internas foram unidas com uma sequencia de 100 "N", indicando um gap.

Para os clones que tiveram três *contigs*, os das pontas foram identificados através da sequência que continha sequência do vetor. Aquele que não tinha essa sequência foi identificado como o do centro. Os *contigs* só foram unidos quando foi

possível identificar grandes sequências de repetições (por exemplo, $(AT)_n$) nas pontas dos *contigs*. Quando não foram identificadas tais sequências, o clone não foi montado.

Discussão

Bibliotecas de BACs são interessantes para estudos genômicos, uma vez que cada clone BAC carrega apenas um fragmento de alelo/haplótipo. O acesso direto ao genoma haploide permite o estudo individualizado de cada alelo de determinado loco. O fato de cada fragmento do DNA clonado em BACs estar individualizado e armazenado separadamente, ainda permite a acessar rapidamente alelos/haplótipos de regiões de interesse. Essas duas características (além das discutidas na Introdução) fizeram com que a técnica de Bibliotecas de BACs fosse escolhida para o estudo genômico de cana-de-açúcar.

Porém, a alta poliploidia de cana-de-açúcar requer uma grande quantidade de clones BACs para cobrir o genoma e estudos baseados em diferenciação de alelos/haplótipos possam ser desenvolvidos. Dessa maneira, as bibliotecas de BACs de cana-de-açúcar desenvolvidas tiveram que conter uma grande quantidade de clones, sendo as duas maiores bibliotecas de BACs para cana-de-açúcar. Para verificar a qualidade das Bibliotecas, a técnica de BAC-end foi utilizada, como uma maneira de caracterizar as bibliotecas, demonstrando a clonagem de fragmentos aleatórios, quantidade de elementos repetitivos e a possibilidade de se ancorar alguns desses fragmentos ao genoma de *S. bicolor*.

A grande quantidade de clones demanda a criação de ferramentas que selecione clones contendo locos de interesse dentro das bibliotecas de maneira eficiente. A melhor maneira de se selecionar um clone contendo um loco de interesse é desenvolver pares de *primers* que flanqueiem o loco de interesse e amplifica-los em cada um dos clones. Porém, em um conjunto de aproximadamente 200.000 clones, este trabalho seria impraticável. Para contornar isso, ferramentas de seleção de clones foram construídas: macroarranjos e *Pool3D*. Cada uma das ferramentas possui limitações e ambas acabam por atingir o objetivo final: selecionar clones que contenham um loco de interesse.

Ambas as técnicas (macroarranjos e *Pool3D*) recuperam, inicialmente, clones que não contém o loco de interesse (falsos positivos). Macroarranjos seleciona

clones através de hibridização e é inerente a técnica que ocorra hibridizações inespecíficas. O Pool3D apresenta falsos positivos quando se tem mais de um clone positivo dentro de um *Superpool*. Isso acontece porque é retornada uma coordenada de placa, linha e coluna, quando um único clone é positivo, é retornada a coordenada do clone. Quando mais de um clone é positivo, são retornados duas coordenadas de placa, duas de linha e duas de coluna. Para dois clones positivos, devem-se testar todos os clones formados pela combinação entre essas coordenadas (oito clones), formando os falsos positivos. As duas técnicas requerem uma amplificação final (validação) de cada possível clone recuperado para a distinção dos falsos positivos.

Conclusão

Os resultados apresentados mostraram a boa qualidade das duas bibliotecas de BAC construídas e podem fornecer recursos para obtenção de um perfil superficial do genoma da cana-de-açúcar, bem como uma base para o sequenciamento BAC-by-BAC de grande parte do conjunto de genes básicos da cana-de-açúcar. Os métodos de seleção, Pool3D e Macroarranjos, foram eficientes na identificação de clones que possuem gene ou sequencias alvo. A escolha de qual utilizar depende dos recursos e equipamentos disponíveis em cada laboratório.

A seleção para os genes únicos contido nesse trabalho foi feita em massa, utilizando grandes quantidades de dados, contrastando em gigantescos bancos de dados. Apesar do gene escolhido para este trabalho (SC-08A) ter sido considerado único e aparecer como único em sorgo (*S. bicolor*), arroz (*O. sativa*).

A possibilidade desse gene se encontrar duplicado também em cana-de-açúcar apareceu quando a filogenia das sequências dos BACs sugeriu dois grandes grupos, o que é discutido no capitulo II.

CAPÍTULO II

Gene Duplication in Sugarcane Genome: Allele Interactions and Evolutionary Patterns

Danilo Augusto Sforça,[1] Sonia Vautrin,[2] Claudio Benicio Cardoso-Silva,[1] Melina Cristina Mancini,[1] María Victoria Romero da Cruz,[1] Guilherme da Silva Pereira,[3] Mônica Conte,[1] Arnaud Bellec,[2] Nair Dahmer,[1] Joelle Fourment,[2] Nathalie Rodde,[2] Marie-Anne Van Sluys,[4] Renato Vicentini,[1] Antônio Augusto Franco Garcia,[3] Eliana Regina Forni-Martins,[1] Monalisa Sampaio Carneiro,[5] Hermann Paulo Hoffmann,[5] Luciana Rossini Pinto,[6] Marcos Guimarães de Andrade Landell,[6] Michel Vincentz,[1] Helene Berges,[2] Anete Pereira Souza[1]

[1]Universidade Estadual de Campinas (UNICAMP), Campinas, SP, Brazil

[2]Centre National de Ressources Genomiques Vegetales (CNRGV), Institut National de la Recherche Agronomique (INRA), Castanet Tolosan, France

[3]Escola Superior de Agricultura Luiz de Queiroz (ESALQ), USP, Piracicaba, SP, Brazil

[4]Universidade de São Paulo, USP, São Paulo, SP, Brazil

[5]Universidade Federal de São Carlos (UFSCAR), Araras, SP, Brazil

[6]Centro de Cana, Instituto Agronômico de Campinas (IAC), Ribeirão Preto, SP, Brazil

Artigo submetido à revista *Frontiers in Plant Science* em 16 de Janeiro de 2019

Manuscrito disponível no bioRxiv desde 03 de Julho de 2018

doi: https://doi.org/10.1101/361089

Abstract

Sugarcane (*Saccharum spp*.) is highly polyploid and aneuploid. Modern cultivars are derived from hybridization between *S. officinarum* and *S. spontaneum*. This combination results in a genome exhibiting variable ploidy among different loci, a huge genome size (approximately 10 Gb) and a high content of repetitive regions. An approach using genomic, transcriptomic and genetic mapping can improve our knowledge of the behavior of genetics in sugarcane. The hypothetical *HP600* and Centromere Protein C (*CENP-C*) genes from sugarcane were used to elucidate the allelic expression and genomic and genetic behaviors of this complex polyploid. The physically linked side-by-side genes *HP600* and *CENP-C* were found in two different homeologous chromosome groups with ploidies of eight and ten. The first region (Region01) was a *Sorghum bicolor* ortholog region with all haplotypes of *HP600* and *CENP-C* expressed, but *HP600* exhibited an unbalanced haplotype expression. The second region (Region02) was a scrambled sugarcane sequence formed from different noncollinear genes containing partial duplications of *HP600* and *CENP-C* (paralogs). This duplication resulted in a non-expressed *HP600* pseudogene and a recombined fusion version of *CENP-C* and the orthologous gene Sobic.003G299500 with at least two chimeric gene haplotypes expressed. It was also determined that it occurred before *Saccharum* genus formation and after the separation of sorghum and sugarcane. A linkage map was constructed using markers from nonduplicated Region01 and for the duplication (Region01 and Region02). We compare the physical and linkage maps, demonstrating the possibility of mapping markers located in duplicated regions with markers in nonduplicated region. Our results contribute directly to the improvement of linkage mapping in complex polyploids and improve the integration of physical and genetic data for sugarcane breeding programs. Thus, we describe the complexity involved in sugarcane genetics and genomics and allelic dynamics, which can be useful for understanding complex polyploid genomes.

Introduction

The *Saccharum* species is a C4 grass and presents a high level of ploidy. *S. officinarum* L. is an octaploid (2n = 80) with x = 10 chromosomes, while *S. spontaneum* L. has x = 8 but presents great variations in the number of

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

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

The Brazilian sugarcane variety SP80-3280 is derived from a cross between the varieties SP71-1088 × H57-5028 and is resistant to brown rust caused by *Puccinia melanocephala* (Landell et al., 2005). SP80-3280, which is one of the main Brazilian cultivars (Manechini et al., 2018), was chosen for transcriptome sequencing by SUCEST-FUN (Vettore et al., 2003) and RNAseq (Cardoso-Silva et al., 2014; Nishiyama et al., 2014; Mattiello et al., 2015). Biparental crossing of SP80-3280 has also been used to analyze rust resistance (Balsalobre et al., 2016), quantitative trait loci (QTL) mapping (Costa et al., 2016) and genotyping by sequencing (GBS) (Balsalobre et al., 2017). A Brazilian initiative (Souza et al., 2011) is producing a gene-space genome sequence from SP80-3280, and a draft sugarcane genome based on whole-genome shotgun sequencing was produced (Riaño-Pachón and

Mattiello, 2017). Additionally, QTL gene synteny from sorghum has been used to map corresponding bacterial artificial chromosomes (BACs) in SP80-3280 (Mancini et al., 2018).

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

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

Genome organization and expression dynamics are poorly understood in complex polyploid organisms, such as sugarcane, mainly because reconstructing large and complex regions of the genome is a challenge. However, an intriguing question is how such a complex genome can function while handling different copy numbers of genes, different allelic dosages and different ploidies of its homo/homeolog groups. Attempting to elucidate such question, we choose two physically linked genes: an unknown function gene *HP600*, in single copy in diploid grass group (OrthoDB, Kriventseva et al., 2018), and the gene *CENP-C* (Centromere Protein C, Gopalakrishnan et al., 2009; Kato et al., 2013; Sandmann et al., 2017; Talbert et al., 2004), involved in cell division, locallyzed next to HP600. We examined

the genome, transcriptome, evolutionary patterns and genetic interactions/relationships of *HP600* and *CENP-C* in a genomic region from the SP80-3280 sugarcane variety (a *Saccharum* hybrid). First, we defined the genome architecture and evolutionary relationships of *HP600* and *CENP-C*, in detail. Second, we used the sugarcane SP80-3280 transcriptome to investigate transcription interactions in each gene (*HP600* and *CENP-C*). Ultimately, we used molecular markers developed from these genes to genotype a segregating population and construct a linkage map and compare it with the physical map.

Material and Methods

Plant Material

The sugarcane varieties were collected from germplasms from the Sugarcane Plant Breeding Program at the active site located in the Agronomic Institute of Campinas (IAC) Sugarcane Center in Ribeirão Preto, São Paulo, Brazil. SP80-3280 and SPIAC93-3046 youngest leaves from one plant of each variety were collected from adult plants and immediately storage on dry ice for transportation and final stored at -80ºC until use. These leaves are used to BAC library construction. For the cytogenetic experiments, IACSP95-3018, IACSP93-3046, RB835486, SP80-3280 and SP81-3250 internodes were collected from adult plants. The internodes were placed in cotton soaked in water and was left for rooting. Roots were collected when it reached 5–15mm.

BAC Library Construction and BAC-End Analyses

The high-molecular-weight (HMW) DNA was prepared from the leaves as described by Peterson et al. (2000) with modifications as described by Gonthier et al. (2010). The HMW DNA was embedded in low-melt agarose (Lonza InCert™ Agarose, Lonza Rockland Inc., Rockland, ME, USA) and partially digested with HindIII (New England Biolabs, Ipswich, MA, USA). Next, two size selection steps were performed by pulsed field gel electrophoresis (PFGE) with a Bio-Rad CHEF Mapper system (Bio-Rad Laboratories, Hercules, CA, USA), and the selected DNA was ligated into the pIndigoBAC-5 HindIII-Cloning Ready vector (Epicenter Biotechnologies, Madison, WI, USA) as described by Chalhoub et al. (2004). The

insert size was verified by preparing DNA BACs with the NucleoSpin® 96 Plasmid Core Kit (MACHEREY-NAGEL GmbH & Co., Düren, Germany) according to the kit instructions, and the DNA was digested by the NotI (New England Biolabs, Ipswich, MA, USA) restriction enzyme and analyzed by PFGE.

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

Target Gene Determination

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

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

BAC Library Screening

Both BAC libraries were spotted onto high-density colony filters with the QPix2 XT workstation (Molecular Devices, Sunnyvale, CA, USA). The BAC clones were spotted in duplicate using a 7x7 pattern onto 22 × 22 cm Immobilon-Ny+ filters (Molecular Devices). The whole BAC library from the SP80-3280 sugarcane variety was spotted on four sets of filters, each one with 55,296 clones in duplicate, and the whole BAC library from SPIAC93-3046 sugarcane variety was spotted on three sets of filters each with 55,296 clones in duplicate. The filters were processed as described by Roselli et al. (2017). Probe radiolabeling and filter hybridization were performed as described in Gonthier et al. (2010).

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

Sequencing and Assembly

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

inserts were sequenced by pyrosequencing with a Roche GS FLX Life Sciences instrument (Branford, CT, USA) in CNRGV, Toulouse, France.

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

Sequence Analysis and Gene Annotation

All the BACs were aligned to verify the presence of redundant homeolog sequences. BAC clones with more than 99% similarity were considered the same homeolog. BACs that represented the same homeologs were not combined. The BACs were annotated with the gene prediction programs EUGENE (Sylvain et al., 2008) and Augustus (Keller et al., 2011). The BAC sequences were also searched for genes with BLASTN and BLASTX against the transcripts from the SUCEST-FUN database (http://sucest-fun.org/; (Vettore et al., 2003)), the CDS of *S. bicolor*, *Z. mays* and *O. sativa* from Phytozome v12.0 and the transcripts published by Cardoso-Silva et al. (2014). The BACs were also subjected to BLASTX against Poaceae proteins. The candidate genes were manually annotated using *S. bicolor*, *O. sativa* and *Z. mays* CDS. The sequences with more than 80% similarity and at least 90% coverage were annotated as genes.

Repetitive content in the BAC clone sequences was identified with the web program LTR_FINDER (Xu and Wang, 2007). Afterward, the BAC sequences were tested by CENSOR (Kohany et al., 2006) against Poaceae.

The phylogenic trees were built by the Neighbor-Joining method (Saitou and Nei, 1987) with nucleic distances calculated with the Jukes-Cantor model (Jukes and Cantor, 1969) in the MEGA 7 software (Kumar et al., 2016). The Kimura 2-parameter (Kimura, 1980) was used as the distance mode.

Duplication Divergence Time

The gene contents of *HP600* and *CENP-C* in the duplication regions were compared, and the distance “d” for coding regions was determined by Nei-Gojobori with Jukes-Cantor, which is available in the MEGA 7 software (Kumar et al., 2016). The divergence times of the sequences shared by the duplicated regions in the BACs were estimated by T = d/2r. The duplicated sequences were used to calculate the pairwise distances (d), and “r” was replaced by the mutation rate of 6.5 x 10-9 mutations per site per year as proposed by Gaut et al. (1996). For the whole duplication, the distance “d” for noncoding regions was determined with the Kimura 2-parameter model and a mutation rate of 1.3 x 10-8 mutations per site per year as described by Ma and Bennetzen (2004).

The insertion ages of the long terminal repeat (LTR) retrotransposons were estimated based on the accumulated number of substitutions between the two LTRs (d) (SanMiguel et al., 1998) using a mutation rate of 1.3 x 10-8 mutations per site per year as described by Ma and Bennetzen (2004).

Gene Expression

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

To test whether the haplotypes had the same proportional ratio in the genome and transcriptome, the transcripts were mapped against one haplotype of the *HP600* haplotypes in Region01 and *CENP-C* with a 90% similarity. The SNPs found in the

transcripts were identified and the coverage and raw variant reads count was used to verify the presence of SNPs not found in BACs. An SNP was considered present in the transcripts if it was represented by at least six transcriptome reads (Kim et al., 2016).

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

The frequency of the genomic data was used to test the transcriptome data with RStudio Team (2015) and the exact binomial test (*binom.test* - (Clopper and Pearson, 1934; Conover, 1971; Hollander et al., 1973)). A p-value ≥ 0.05 is equivalent to a 95% confidence interval for considering the genomic ratio equal to the transcriptome ratio.

Chromosome Number Determination and BAC-FISH

The chromosome number determination was performed as described by Guerra (1983) with root tips that were 0.5–1.5 cm in length and treated with 5 N HCl for 20 min. The slides were stained with 2% Giemsa for 15 min. Chromosome number determination was performed for the SP80-3280, SP81-3250, RB83-5486, RB92-5345, IACSP95-3018 and IACSP93-3046 varieties. CMA/DAPI coloration was performed by enzymatic digestion as described by Guerra and Souza (2002). The slides were stained with 10 µg/ml DAPI for 30 min and 10 µg/ml CMA for 1 h. Afterwards, the slides were stained with 1:1 glycerol/McIlvaine buffer and visualized.

BAC-FISH was performed using the SP803280 variety. For the mitotic chromosome preparations, root tips that were 0.5–1.5 cm in length were collected and treated in the dark with p-dichlorobenzene-saturated solution at room temperature for 2 h, fixed in a freshly prepared 3:1 mixture (ethanol:glacial acetic acid) at 4°C for 24 h and stored at −20°C until use. After being washed in water, the root tips were digested with the following enzymes: 2% cellulase (w/v) (Serva, Heidelberg, Baden-Wurtemberg State, Germany), 20% pectinase (v/v) (Sigma, Munich, Baviera State, Germany) and 1% Macerozyme (w/v) (Sigma) at 37°C for 1-2

h (Schwarzacher et al., 1980). The meristems were squashed in a drop of 45% acetic acid and fixed in liquid nitrogen for 15 min. After air-drying, slides with good metaphase chromosome spreads were stored at −20°C.

The Shy064N22 and Shy048L15 BACs, both from the BAC library for the SP80-3280 variety, were used as probes. The probes were labeled with digoxigenin-11-dUTP (Roche) by nick translation. Bacterial artificial chromosome-fluorescence in situ hybridization (BAC-FISH) was performed as described by Schwarzacher and Heslop-Harrison (2000) with minor modifications. The C_ot-100 fraction of the SP80-3280 sugarcane variety genomic DNA, which was used to block repetitive sequences, was prepared according to Zwick et al. (1997). Preparations were counterstained and mounted with 2 µg/ml DAPI in Vectashield (Vector, Burlingame, CA, USA).

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

Genetic Map Construction

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

Results

BAC Library Construction

The BAC library from the SP80-3280 sugarcane variety resulted in 221.184 clones arrayed in 576 384-well microtiter plates with a mean insert size of 110 kb. This BAC library is approximately 2.4 genome equivalents (10 Gb) and 26 monoploid

genome equivalents (930 Mb, (Figueira et al., 2012)). For the IACSP93-3046 sugarcane variety, the library construction resulted in 165.888 clones arrayed in 432 384-well microtiter plates with a mean insert size of 110 kb, which is approximately 1.8 genome equivalents and 19 monoploid genome equivalents.

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

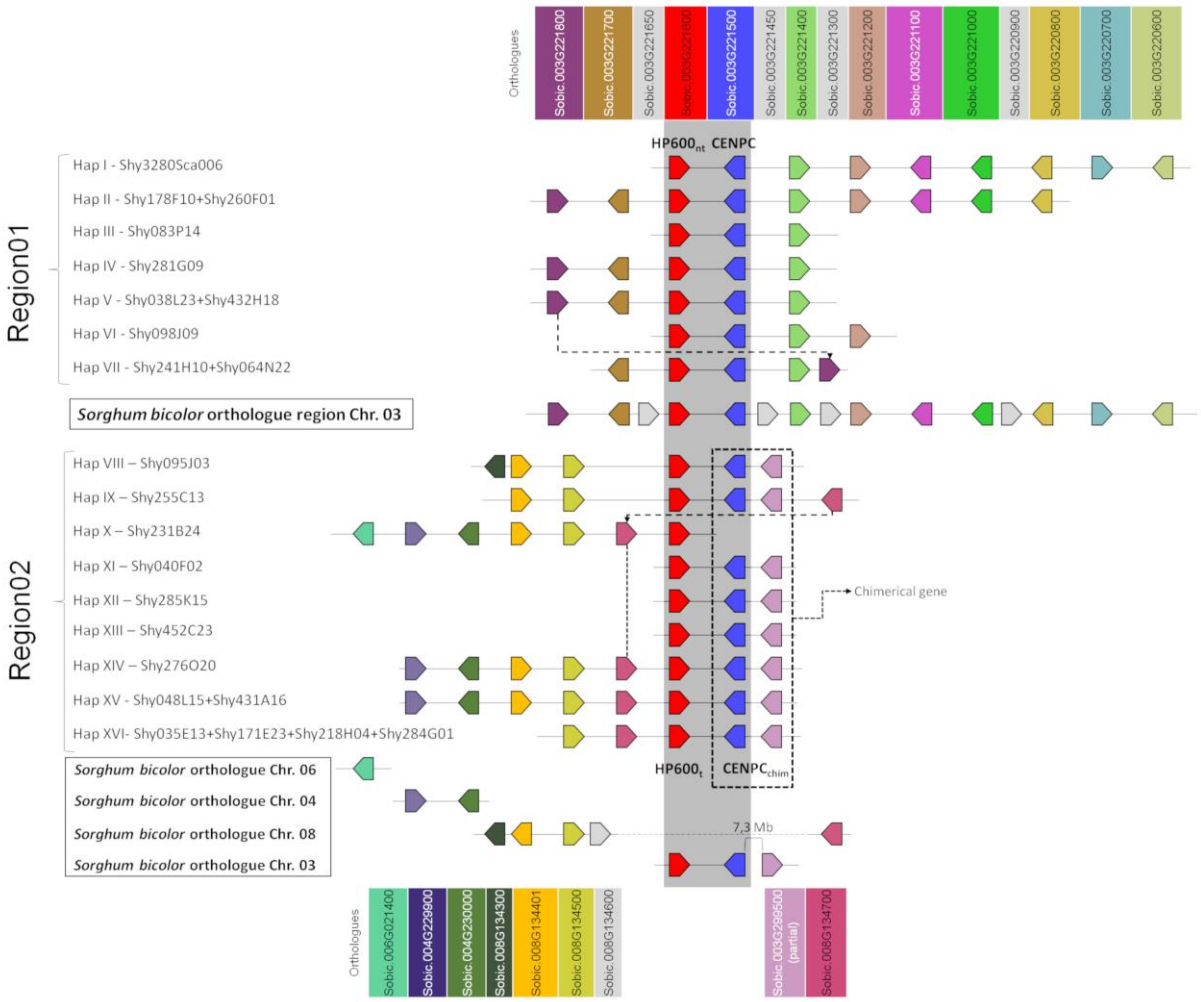


Figure 1. Schematic representation of the sugarcane BAC haplotypes from Region01 and Region02. Squares of the same color represent sugarcane genes orthologous to *Sorghum bicolor* genes. Dotted lines connect the homologous genes in sugarcane at different positions. In sugarcane Region02, the *CENP-C*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

BAC Annotation

The *HP600* gene was used as a target gene and showed strong evidence of being a single-copy gene when the *HP600* transcripts from sorghum, rice and sugarcane were compared. Twenty-two BAC clones from the SP80-3280 library that had the *HP600* target gene (NCBI from MH463467 to MH463488) and a previously sequenced BAC (Mancini et al. (2018); NCBI Accession Number MF737011) were sequenced by Roche 454 sequencing (see Supplementary Table 1, Supplementary Material). The BACs varied in size from 48 kb (Shy171E23) to 162 kb (Shy432H18), with a mean size of 109 kb. The BACs were compared, and BACs with at least 99% similarity were considered the same haplotype (Figures 1 and 2), resulting in sixteen haplotypes. Indeed, the possibility of one homeolog being more than 99% similar to another exists, but a real haplotype cannot be distinguished from an assembly mismatch.

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

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

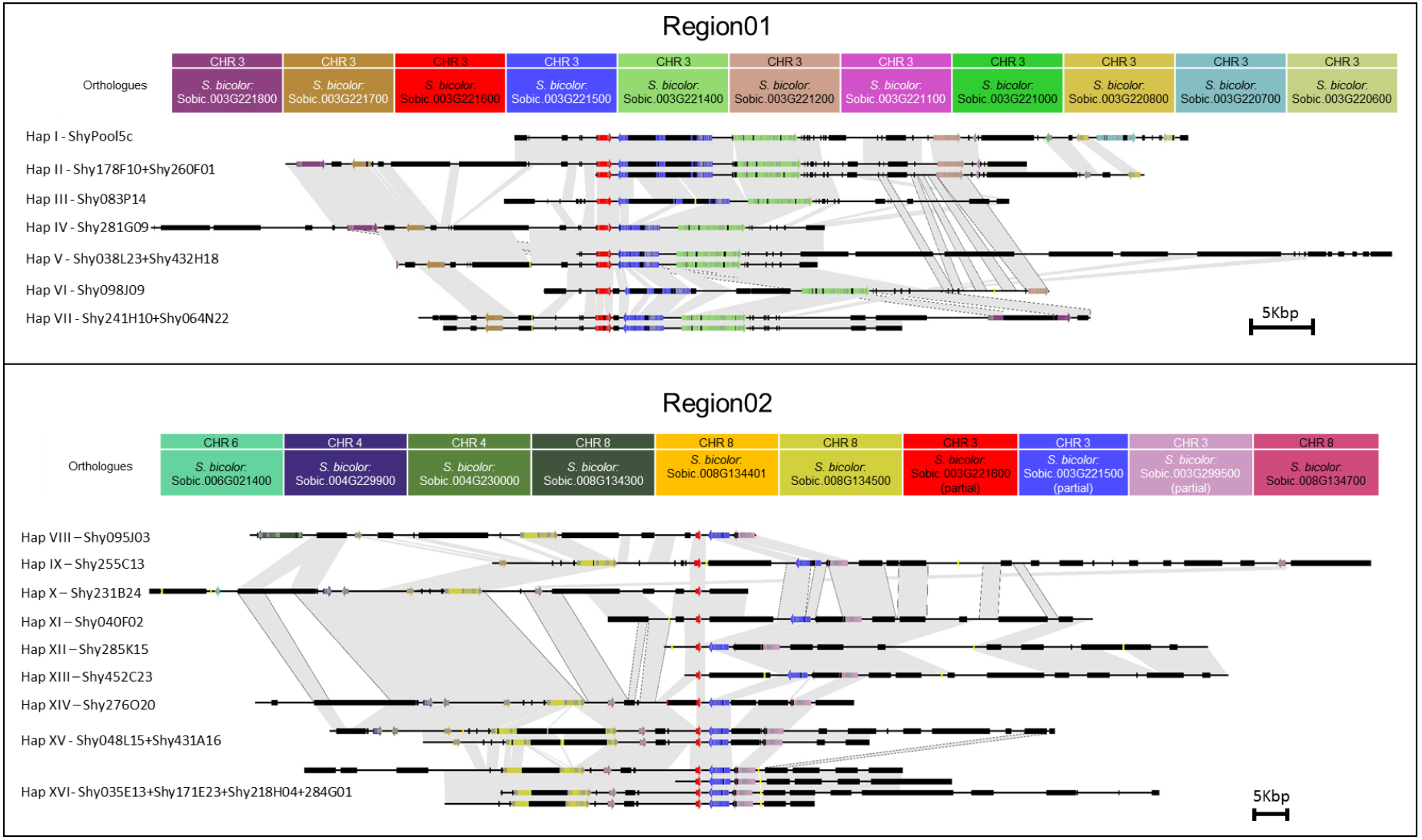

Region01
Orthologues
CHR 3 S. bicolor: Sobic.003G221800
CHR 3 S. bicolor: Sobic.003G221700
CHR 3 S. bicolor: Sobic.003G221600
CHR 3 S. bicolor: Sobic.003G221500
CHR 3 S. bicolor: Sobic.003G221400
CHR 3 S. bicolor: Sobic.003G221200
CHR 3 S. bicolor: Sobic.003G221100
CHR 3 S. bicolor: Sobic.003G221000
CHR 3 S. bicolor: Sobic.003G220800
CHR 3 S. bicolor: Sobic.003G220700
CHR 3 S. bicolor: Sobic.003G220600
Hap I - ShyPool5c
Hap II - Shy178F10+Shy260F01
Hap III - Shy083P14
Hap IV - Shy281G09
Hap V - Shy038L23+Shy432H18
Hap VI - Shy098J09
Hap VII - Shy241H10+Shy064N22
5Kbp
Region02
Orthologues
CHR 6 S. bicolor: Sobic.006G021400
CHR 4 S. bicolor: Sobic.004G229900
CHR 4 S. bicolor: Sobic.004G230000
CHR 8 S. bicolor: Sobic.008G134300
CHR 8 S. bicolor: Sobic.008G134401
CHR 8 S. bicolor: Sobic.008G134500
CHR 3 S. bicolor: Sobic.003G221600 (partial)
CHR 3 S. bicolor: Sobic.003G221500 (partial)
CHR 3 S. bicolor: Sobic.003G299500 (partial)
CHR 8 S. bicolor: Sobic.008G134700
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+284G01
5Kbp

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

Relationship between Region01 and Region02

Annotation of *HP600* and *CENP-C* in the sixteen BAC haplotypes revealed two groups of BACs. One group had the expected exon/intron organization compared with *S. bicolor HP600* (five exons in sorghum) and *CENP-C* (fourteen exons in sorghum). This region was further designated as Region01 (see Supplementary Table 1, Supplementary Material - 10 BACs and 7 haplotypes – Figure 1 - haplotype I to haplotype VII). The other group was found to have fewer exons than expected (compared with *S. bicolor*) for both *HP600* and *CENP-C* and was designated Region02 (see Supplementary Table 1, Supplementary Material - 13 BACs and 9 haplotypes – Figure 1 - haplotype VIII to haplotype XVI).

A comparison of the BAC haplotypes from Region01 and Region02 revealed an 8-kb shared region. The 8-kb duplication spanned from the last three exons of *HP600* to the last seven exons of *CENP-C*. *HP600* and *CENP-C* were physically linked, but the orientation of the genes was opposite (see Supplementary Figure 2, panel B, Supplementary Material). A phylogenetic tree was constructed to examine the relationships among this 8-kb region (see Supplementary Figure 2, Panel A, Supplementary Material). The orthologous region from *S. bicolor* was used as an outgroup, and the separation in the two groups (Region01 and Region02) suggests that the shared 8-kb sequence appeared as a consequence of a sugarcane-specific duplication.

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

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

In Region02, the Sobic.008G134300 orthologous gene was found only in haplotype VIII, and the Sobic.008G134700 ortholog was found in a different position in haplotype IX (Figure 1, dotted line in Region02 and Figure 2). The phylogenetic analysis of Sobic.008G134700 and sugarcane orthologs demonstrated that sugarcane haplotype IX is more closely related to sorghum than to other sugarcane homeologs (see Supplementary Figure 3, Supplementary Material). Additionally, the orientation of transcription of the Sobic.008G134700 ortholog in haplotype IX is opposite that of the other sugarcane haplotypes (Figures 1 and 2). This finding suggests that this gene could be duplicated (paralogs) or translocated (orthologs) in haplotypes X, XIV, XV and XVI. No *S. bicolor* orthologous region that originated from Region02 could be determined, as it contained genes from multiple sorghum chromosomes.

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

To estimate the genomic diversity in sugarcane haplotypes from both regions (analyzed together and separately), the shared 8-kb region (duplication) was used (see Supplementary Table 4, Supplementary Material), and the SNPs were identified. The diversity in the *HP600* and *CENP-C* genes was analyzed, and one SNP was observed every 43 bases (Region02) and 70 bases (Region01). We searched for SNPs that could distinguish each region (see Supplementary Table 5, Supplementary Material) in the *HP600* and *CENP-C* genes, and one SNP was found for every 56 bases (20 SNPs in total). Additionally, small (3-10 bases) and large (30 – 200 bases) insertions were found. These results revealed a high level of diversity in sugarcane,

i.e., a high number of SNPs in each region, which could be used to generate molecular markers and to improve genetic maps. Moreover, the diversity rate of both regions together could be used as an indicator of a duplicated gene, i.e., a rate < 20 (see Supplementary Table 4, Supplementary Material).

HP600 and *CENP-C* Haplotypes and Phylogenetics

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

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

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

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

To reconstruct a phylogenetic tree for *HP600* and *CENP-C* from both regions, the orthologs from *O. sativa* and *Zea mays* L. were searched. The rice *HP600* and

CENP-C orthologs, LOC_Os01g43060 and LOC_Os01g43050, respectively, were recovered. Maize has gone through tetraploidization since its divergence from sorghum approximately 12 million years ago (Woodhouse et al., 2010). The maize *HP600* ortholog search returned the following three possible genes with high similarity: GRMZM2G114380 (chromosome 03), GRMZM2G018417 (chromosome 01) and GRMZM2G056377 (chromosome 01). The *CENP-C* maize ortholog search returned the following three possible genes with high similarity: GRMZM2G114315 (chromosome 03), GRMZM2G134183 (chromosome 03), and GRMZM2G369014 (chromosome 01).

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

Two phylogenetic trees were constructed (see Supplementary Figure 4, Supplementary Material), one for *HP600* (see Supplementary Figure 4, Panel A, Supplementary Material) and the other for *CENP-C* (see Supplementary Figure 4, Panel B, Supplementary Material), using sugarcane *HP600* and *CENP-C* haplotypes from both regions. The results demonstrated that the haplotypes from Region01 and Region02 are more similar to themselves than they are to those of sorghum or rice. Thus, the results also suggest that Region02 contains paralogous genes from Region01.

The divergence times among sugarcane *HP600* haplotypes and sorghum ranged from 1.5 Mya to 4.5 Mya. For *CENP-C*, the haplotype divergence time rates were 0.3-0.7 Mya, and the comparison with sorghum indicated 4.2-4.5 Mya for the highest values. The estimated sugarcane x sorghum divergence time was 5 Mya (Ming et al., 1998) to 8-9 Mya (Jannoo et al., 2007; Zhang et al., 2018a).

Chromosome Number Determination and BAC-FISH

The determination of the range of *CENP-C* and *HP600* loci that are present in the sugarcane genome was performed using in situ hybridization. First, the number of chromosomes in the SP80-3280 sugarcane variety was defined, but the number of

clear and well-spread metaphases for the SP80-3280 variety was less than 10 (see Supplementary Table 6, Supplementary Material). We expanded the analysis to four more sugarcane varieties (SP81-3250, RB835486, IACSP95-3018 and IACSP93-3046) to improve the conclusions (see Supplementary Figure 5 – Panels A-E – and Supplementary Table 6, Supplementary Material). The most abundant number of chromosomes was 2n = 112 (range: 2n = 98 to 2n = 118 chromosomes). The chromosome number for the *Saccharum* hybrid cultivar SP80-3280 was found to be 2n = 112 (range: 2n = 108 to 2n = 118 chromosomes - see Supplementary Table 6, Supplementary Material). Vieira et al. (2018) also identified 2n = 112 chromosomes in the IACSP93-3046 variety.

As a second step, we used two varieties with the best chromosome spreads, i.e., IACSP93-3046 and IACSP95-3018, for the CMA/DAPI banding patterns (see Supplementary Figure 5 – Panels F-I, Supplementary Material). The IACSP93-3046 variety exhibited at least six terminal $CMA^{+}/DAPI^{-}$ bands, one chromosome with $CMA^{+}/DAPI^{o}$ and two chromosomes with adjacent intercalations of CMA^{+} and $DAPI^{+}$ in the same chromosome (see Supplementary Figure 5 – Panels F and G, Supplementary Material). The IACSP95-3018 variety revealed seven terminal $CMA^{+}/DAPI^{-}$ bands, and at least two chromosomes exhibited adjacent CMA^{+} and $DAPI^{+}$, one of which was in the intercalary position and the other was in the terminal position (see Supplementary Figure 5 – Panels H and I, Supplementary Material). Additionally, an equal number of chromosomes and the divergent number of bands could indicate different chromosomic arrangements and/or different numbers of homeologs in each variety.

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

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

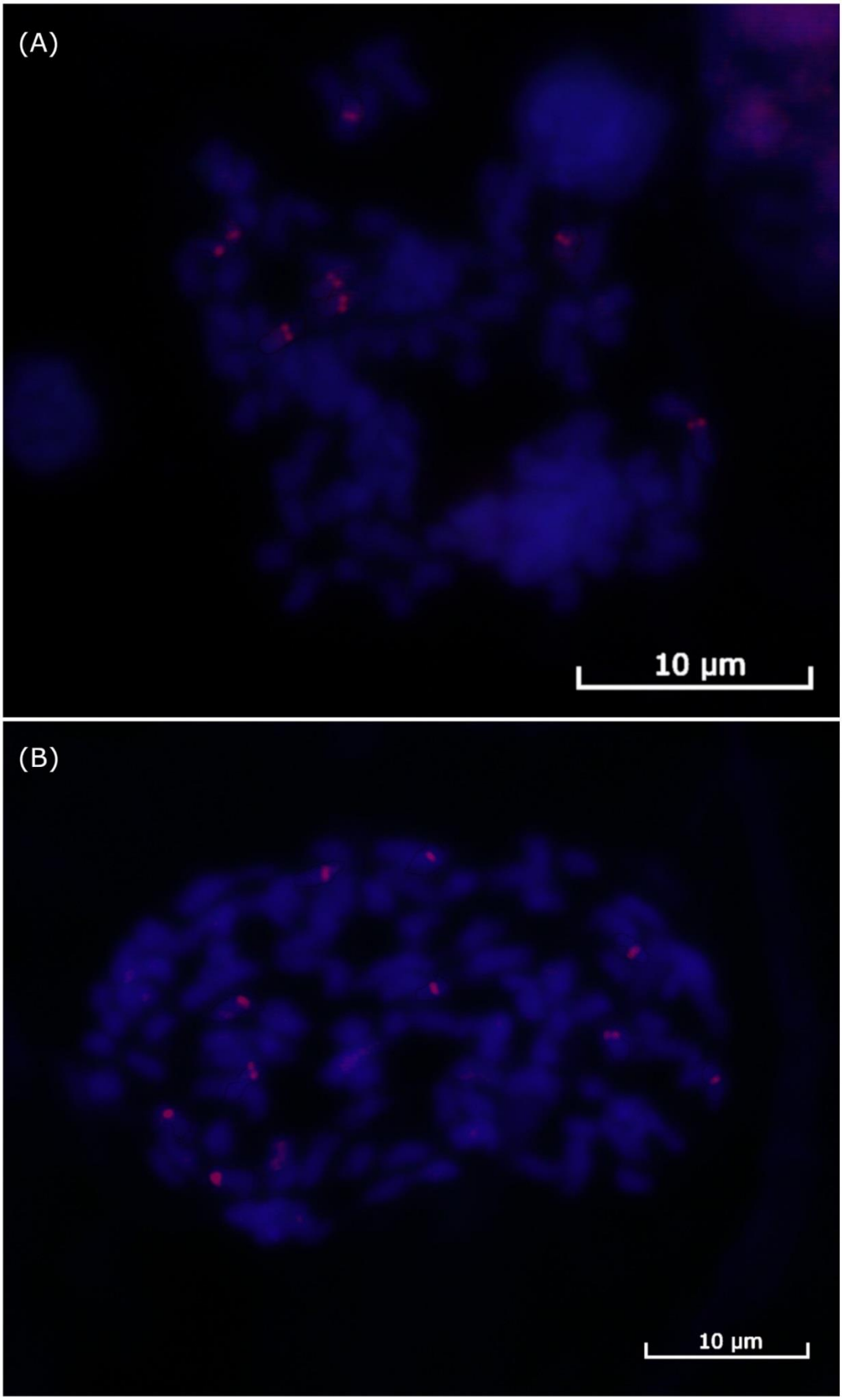

(A)
10 µm
(B)
10 µm

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

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

The transcriptomes of the SP80-3280 sugarcane variety from the roots, shoots and stalks were mapped on *HP600* and *CENP-C* (NCBI SRR7274987), and the set of transcripts was used for the transcription analyses. All of the *HP600* haplotypes from Region01 were covered by the reads, including haplotype III with a premature stop codon. None of the *HP600* haplotypes from Region02 were found, suggesting that *HP600* is not expressed from Region02 (see Supplementary Figure 2, Supplementary Material). For the *CENP-C* gene from Region01, haplotypes IV/V were found to be expressed. Furthermore, haplotypes I/II, haplotype VI and haplotype VII were fully covered by the reads, except for the first three SNPs, but these SNPs were described in the work of Talbert et al. (2004) under the *CENP-C*1 haplotype, suggesting that the set of reads did not cover this region. For haplotype III, one SNP was not found, but nine exclusive SNPs from this haplotype were represented. Therefore, all *CENP-C* haplotypes from Region01 were considered to be expressed.

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

The mapping of the transcript reads in the *CENP-C* haplotypes from Region02 revealed evidence of a chimeric gene (Figure 1, dotted rectangle and Figure 4). The chimeric gene was formed by the first five exons from the Sobic.003G299500 sugarcane orthologous gene and the eighth to fourteenth exons from *CENP-C* (Figure 4 – Panel C). RNAseq reads overlapped the region corresponding to the union of the chimeric gene (position 1253 from the *CENP-C* haplotypes from Region02 by 38 reads - Figure 4 – Panel F). This result provided robust evidence for the formation of the chimeric gene and its expression.

The sugarcane gene orthologous to Sobic.003G299500 was represented by BAC BAC267H24 (GenBank KF184671) from the sugarcane hybrid R570 as published by De Setta et al. (2014) under the name “SHCRBa_267_H24_F_10” (Figure 4 – Panel D). This finding indicated that the ancestral genes from sorghum (orthologs) were retained in the sugarcane genome (Figure 4 – Panels B and D) and

that the chimeric gene was formed by the fusion of a partial duplication of *CENP-C* and the sorghum ortholog gene Sobic.003G299500 (Figure 4 – Panel C).

Two chimeric *CENP-C* haplotypes from Region02 were fully mapped with transcripts, i.e., haplotypes XI/XII/XIII and haplotype XIV. The chimeric *CENP-C* haplotypes IX and XVI from Region02 were not fully mapped, but exclusive SNPs from these haplotypes were recovered. The *CENP-C* haplotypes VIII and XV from Region02 exhibited no exclusive SNPs in the transcriptome, and evidence for the expression of these two haplotypes remains undefined.

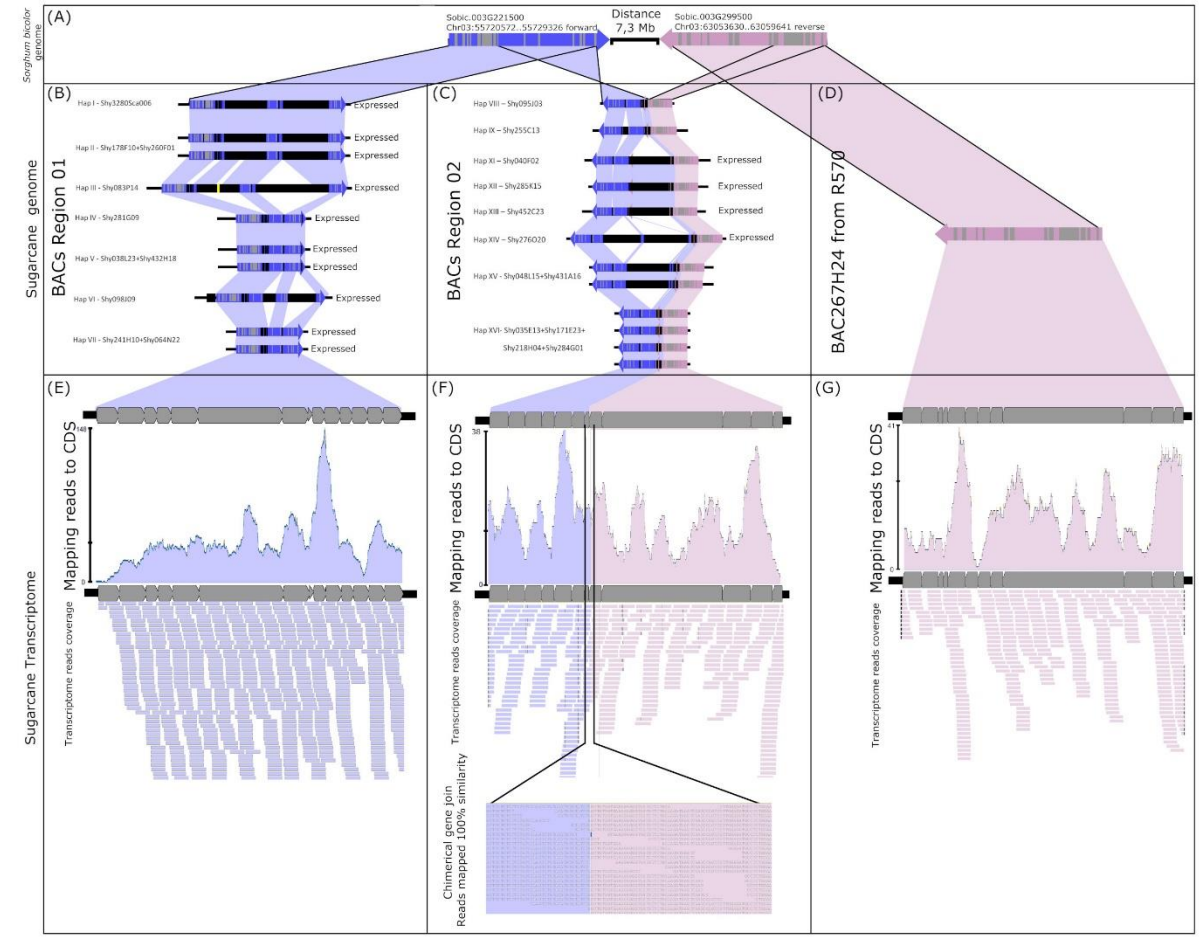


Figure 4. Fusion gene formation of *CENP-C* and Sobic003G299500. Panel (**A**): Sorghum *CENP-C* and Sobic003G299500 genome location. Panel (**B**): Sugarcane genomic *CENP-C* haplotypes in Region01 (all expressed). Panel (**C**): Partially duplicated sugarcane paralogs of *CENP-C* and Sobic003G299500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

Comparison with Others *Saccharum* Genome

A search for the *HP600* and *CENP-C* genes against the sugarcane R570 mosaic monoploid genome (Garsmeur et al., 2018) returns no hits, indicating that both genes ware not represented in the R570 BACs. Comparisons of the BAC sequences against the sugarcane SP80-3280 genome draft using BLASTN (Riaño-Pachón and Mattiello, 2017) resulted in matches within gene regions, but no genome contig covered a whole BAC, and the BAC TEs matched with several genome contigs (see Supplementary Figure 6, Supplementary Material). The matches with gene regions provide further support for our assembly process.

A BLAST search of the all genes recovered from Region01 and Region02 against the *S. spontaneum* genome (Zhang et al., 2018b) resulted in the recovery of the chromosome of each gene in *S. spontaneum* (see Supplementary Table 02 and Supplementary Table 03). *HP600* was found in chromosomes Chr2D, Chr3B, Chr3C and Chr3D from *S. spontaneum*. In chromosome Chr2D, *HP600* was found as in Region02 BACs, and in chromosomes Chr3B and Chr3C, as in Region02 BACs. In chromosome Chr3D, *HP600* was duplicated at positions 14833330 and 35428849, both of which had the same architecture as in Region01 (five exons).

Both the *CENP-C* and chimeric *CENP-C* sequences were used to search for the *CENP-C* gene. The *CENP-C* gene was found in *S. spontaneum* chromosomes Chr3B and Chr3C. Chromosome Chr3D had a duplication of *CENP-C* at position 14835786 (complete gene) and position 35431299 (partial, last six exons – not found in our data). In chromosome Chr7B, the 9 first exons were found, but this architecture was not found in our data. The chimeric *CENP-C* gene was found in chromosomes Chr2A and Chr2D.

Regarding these results, Region01 is present in Chr03B, Chr3C and Chr3D (only in position 14835786), with *HP600* and *CENP-C* physically side by side (see Supplementary Figure 7, Supplementary Material). Region02 is only represented in chromosome Chr02D with a duplication composed of *HP600* and the *CENP-C* chimera physically side by side. Another copy of the *CENP-C* chimera was found in chromosome Chr2A, but without the presence of *HP600*. Additionally, the Sobic.003G299500 ortholog gene, which was fused with *CENP-C*, was also found with its complete sequence (as demonstrated in Figure 4D) in chromosome Chr3A at

position 16992405 and duplicated in two positions, 32628152 and 60347125, in chromosome Chr3C.

How the Locus Number of Homeologs Influences Expression

We searched the SNPs in the BAC sequences and RNAseq reads (i.e., only in the transcriptome of the SP80-3280 sugarcane variety from the roots, shoots and stalks – NCBI SRR7274987) and compared the correspondences to the *HP600* and *CENP-C* genes. For Region01 and Region02, we defined the ploidies as eight and ten, respectively, based on the BAC-FISH data. The numbers of BAC haplotypes recovered for Region01 and Region02 were seven and nine, respectively, which indicated one missing BAC haplotype in each region.

The missing BAC haplotypes were determined by searching for SNPs present only in the transcriptome. For the *HP600* haplotypes from Region01 (Table 1), six SNPs were found in the transcriptome and not in the BAC haplotypes, including a (GAG)3 -> (GAG)2 deletion. For the *CENP-C* gene (Table 2), eight SNPs were not represented in the genomic haplotypes. The presence of SNPs only in the transcript data corroborates the assumption that (at least) one genomic haplotype was missing in each region.

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

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

determined whether the missing BAC haplotype contributed to the dosage of more common SNPs and (II) we determined whether the missing BAC haplotype contributed to the dosage of the variant SNP.

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

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

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

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

									Missing haplotype for more common SNP				Missing haplotype for variant SNP			
SNP	**Name**	**Change**	**Polymorphism Type**	**Position**	**Coverage**	**Variant Coverage**	**Genomic Detected**	**Transcriptome Proportion**	Genomic Variant	Genomic	Genomic Proportion	P-value (binomial test)	Genomic Variant	Genomic	Genomic Proportion	P-value (binomial test)
1	G	C -> G	SNP (transversion)	106	16	13	Yes	0.81	5	3	0.63	1.95E-01*	4	4	0.5	2.13E-02
2	G	A -> G	SNP (transition)	150	19	8	Yes	0.42	1	7	0.13	1.25E-03	2	6	0.25	1.08E-01*
3	C	G -> C	SNP (transversion)	246	34	7	Yes	0.21	1	7	0.13	1.87E-01*	2	6	0.25	6.93E-01*
4	T	A -> T	SNP (transversion)	369	65	7	Yes	0.11	1	7	0.13	8.51E-01*	2	6	0.25	6.14E-03
5	A	G -> A	SNP (transition)	371	68	19	No	0.28	1	7	0.13	6.21E-04	1	7	0.13	6.21E-04
6	C	T -> C	SNP (transition)	390	64	15	No	0.23	1	7	0.13	1.32E-02	1	7	0.13	1.32E-02
7	G	T -> G	SNP (transversion)	513	46	12	Yes	0.26	3	5	0.38	1.28E-01*	4	4	0.5	1.64E-03
8	A	G -> A	SNP (transition)	518	45	10	Yes	0.22	2	6	0.25	7.34E-01*	3	5	0,375	4.40E-02
9	T	G -> T	SNP (transversion)	731	54	8	Yes	0.15	2	6	0.25	1.14E-01*	3	5	0,375	3.58E-04
10	C	A -> C	SNP (transversion)	1008	56	9	No	0.16	1	7	0.13	4.17E-01*	1	7	0.13	4.17E-01*
11	T	C -> T	SNP (transition)	1061	91	29	Yes	0.32	2	6	0.25	1.46E-01*	3	5	0,375	2.81E-01*
12	T	C -> T	SNP (transition)	1088	77	41	Yes	0.53	4	4	0.50	6.48E-01*	3	5	0,375	6.37E-03
13	T	C -> T	SNP (transition)	1190	76	9	Yes	0.12	2	6	0.25	7.49E-03	3	5	0,375	1.10E-06
14	A	G -> A	SNP (transition)	1209	76	20	No	0.26	1	7	0.13	1.31E-03	1	7	0.13	1.31E-03
15	T	A -> T	SNP (transversion)	1251	62	10	Yes	0.16	2	6	0.25	1.41E-01*	3	5	0,375	3.29E-04
16	G	A -> G	SNP (transition)	1255	62	55	Yes	0.89	6	2	0.75	1.19E-02	5	3	0,625	5.15E-06
17		-ATG	Deletion	1307	75	9	Yes	0.12	1	7	0.13	1.00E+00*	2	6	0.25	7.38E-03
18	G	A -> G	SNP (transition)	1314	90	23	Yes	0.26	1	7	0.13	6.50E-04	2	6	0.25	9.03E-01*
19	G	T -> G	SNP (transversion)	1347	103	13	Yes	0.13	2	6	0.25	2.88E-03	3	5	0,375	3.09E-08
20	A	T -> A	SNP (transversion)	1384	101	37	Yes	0.37	1	7	0.13	5.30E-10	2	6	0.25	1.09E-02
21	G	C -> G	SNP (transversion)	1424	80	9	No	0.11	1	7	0.13	8.66E-01*	1	7	0.13	8.66E-01*
22	A	C -> A	SNP (transversion)	1437	84	10	Yes	0.12	1	7	0.13	1.00E+00*	2	6	0.25	5.12E-03
23	TT	AA -> TT	Substitution	1481	62	7	No	0.11	1	7	0.13	1.00E+00*	1	7	0.13	1.00E+00*
24	G	A -> G	SNP (transition)	1527	106	90	Yes (duplication)	0.85								
25	C	T -> C	SNP (transition)	1540	139	86	Yes (duplication)	0.62								
26	A	T -> A	SNP (transversion)	1584	253	235	Yes (duplication)	0.93								
27	A	G -> A	SNP (transition)	1638	247	39	Yes (duplication)	0.16								

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

Genetic Mapping

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

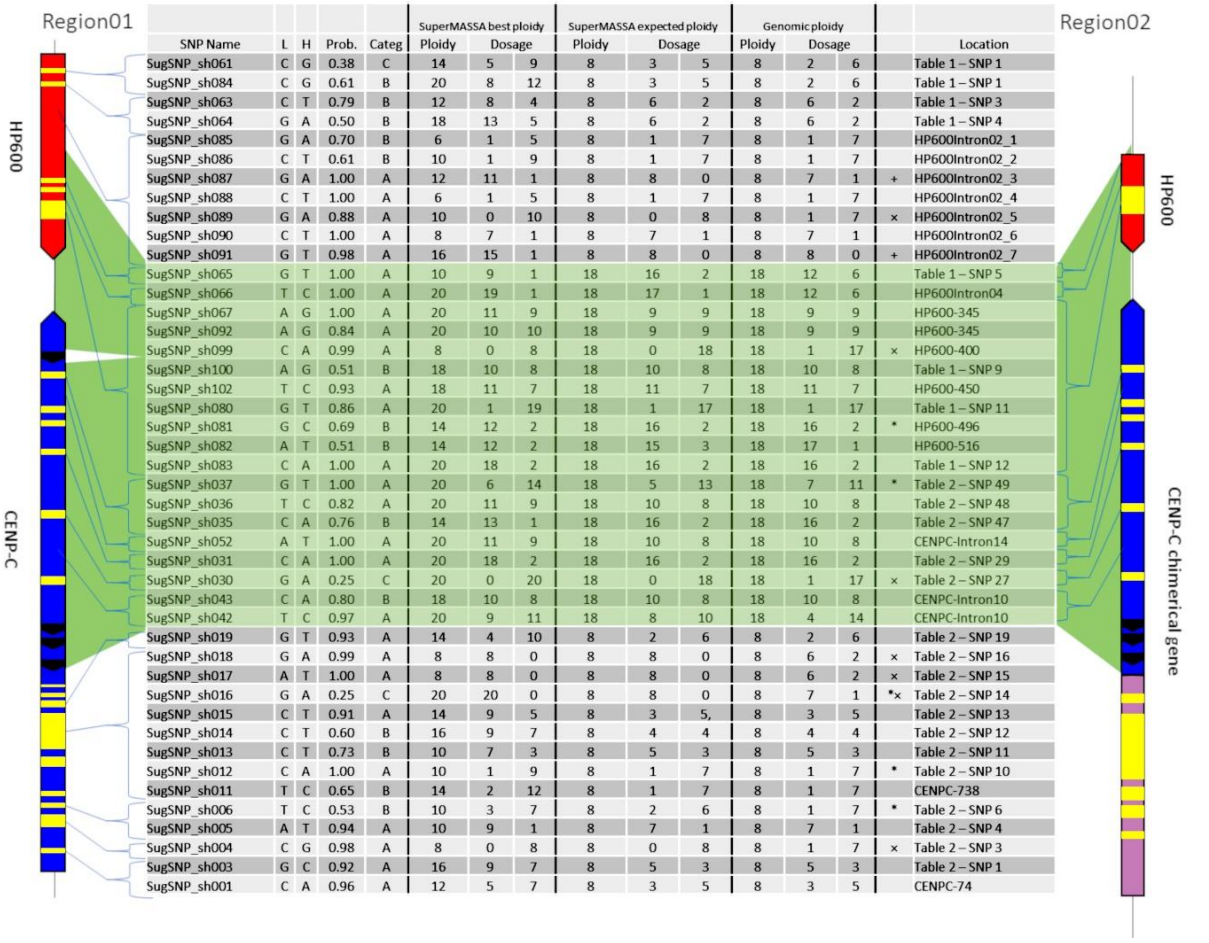
The markers from introns and exons were drawn along Region01 (Figure 5, "Location" column), including the duplicated region found in Region02. Among them, seven exhibited no variants presence in genotyping (Figure 5 – "×" marked), but five were detected in the RNAseq reads. Two markers (Figure 5 – "+"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

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

For the genetic mapping, ten markers were used according to the SuperMASSA best ploidy results. First, attempts were made to add each marker to the existing linkage groups published by Balsalobre et al. (2017), but none of the markers could

be linked to the groups. Then, the markers were tested for linkage among themselves. Two linkage groups could be created (Figure 6 – panel A) with 27.4 cM and 32.7 cM. The SugSNP_sh065 and SugSNP_sh099 markers were physically located in Region01 and Region02, respectively. It was unexpected that duplicated markers were linked to a linkage group, even weakly (see Supplementary Figure 8).

Using all the physical information, the duplicated markers (SugSNP_sh065 and SugSNP_sh099) were excluded (Figure 6 – Panel B). Then, attempts were made to add the remaining markers to the groups again, and the SugSNP_sh005 marker was inserted into Linkage group 02 (Figure 6 – Panel C). The markers that were in the wrong positions according to the physical map (BACs) were also excluded, and the SugSNP_sh005 marker was excluded from Linkage group 01 but remained in Linkage group 02 (Figure 6 – Panel C). Then, an attempt was made to form one linkage group with the remaining markers by forcing OneMap to place the markers in a single group. Again, the size of the group was too large (60.3 cM - Figure 6 – Panel D). Therefore, the best representation of the region was two linkage groups, with Linkage group 01 at 2.1 cM, and Linkage group 02 at 0 cM (Figure 6 – Panel E).



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

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

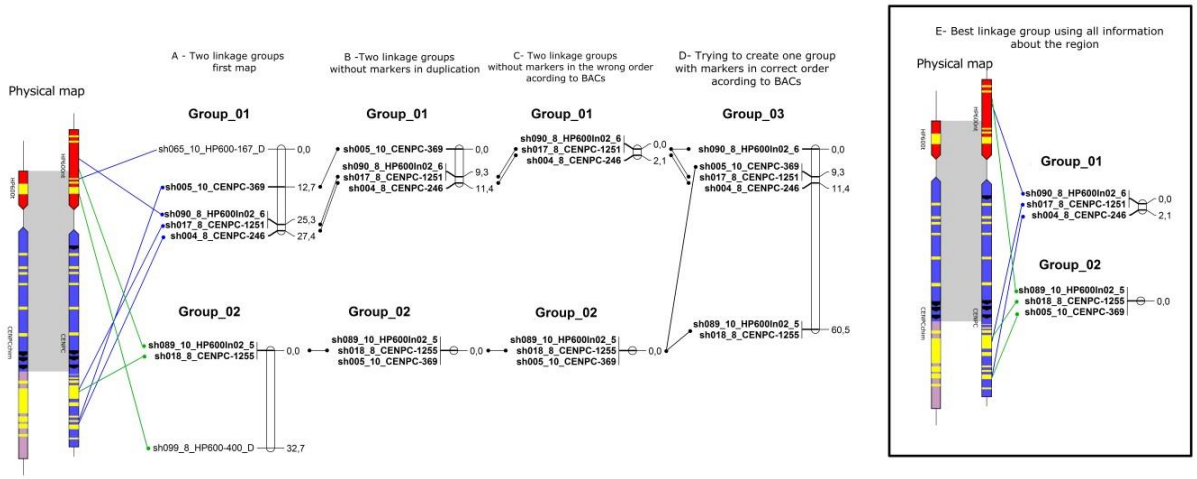


Figure 6. Schematic representation of the sugarcane linkage map. The sugarcane variety SP80-3280 SNPs were used to create multiple linkage maps with information about the sugarcane genome (BACs).

Discussion

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

The approach of mapping the BES in the *S. bicolor* genome, already performed for other libraries (Figueira et al., 2012; Kim et al., 2013; Visendi et al., 2016), revealed high synteny with the *S. bicolor* genome and a large number of TEs in the sugarcane genome. Kim et al. (2013) showed BES anchorage of approximately 6.4%, and Figueira et al. (2012) showed anchorage of approximately 22%. Our data showed 10% BES anchorage in the sorghum genome for both libraries constructed. These results are more similar to those of Kim et al. (2013), since they used only BES ≥ 300 bp and we used BES ≥100 bp.

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

The genomic SNP variation in sugarcane coding regions has been estimated to be one SNP every 50 bp (Cordeiro et al., 2006) and one every 86 bp (Cardoso-Silva et al., 2014). For coding Region01, one SNP was found per 70 bases. Feltus et al. (2004) showed that different ratios of SNPs occur across the genome. When we compared Region01 and Region02, one SNP was found per 12 bases using only the

data for the SP-803280 sugarcane variety. In light of the possible existence of at least one more haplotype, this number could be underestimated.

The chromosome number in the main Brazilian varieties was determined. The chromosome number determination showed an equal number of chromosomes (2n = 112, range: 2n = 98-118). The aneuploid nature of sugarcane hybrid cultivars (D'Hont, 2005; Piperidis et al., 2010) means that they contain different numbers of homeologous chromosomes. A number of differences in the CMA/DAPI patterns were found among the different varieties analyzed in this study, suggesting differences in chromosome contents, i.e., differences in homeologous arrangement.

The hypothetical gene *HP600* and the *CENP-C* gene were used in this work as a case study. The function of *HP600* is unknown, but an ortholog of this gene is present in the genomes of rice (LOC_Os01g43060), maize (GRMZM2G114380) and sorghum (Sobic.003G221600). Sobic.003G221600 (ortholog of *HP600*) was also found in a QTL for BRIX (sugar accumulation) that was mapped by Murray et al. (2008) and based on the sorghum consensus map reported by Mace and Jordan (2011). The *CENP-C* protein is a kinetochore component (Kato et al., 2013; Sandmann et al., 2017) located next to *HP600*. Here, we have demonstrated the existence of paralogous genes for *HP600* and *CENP-C* that are localized in two different homeologous sugarcane chromosome groups. The BAC haplotypes could be separated into two sugarcane homeologous groups as follows: (1) Region01 contained the collinearity region between sorghum and sugarcane *HP600* and *CENP-C* genes and (2) Region02 contained their paralogs.

Region01 is a recurrent case of high gene conservation and collinearity among sugarcane homeologs and the *S. bicolor* genome as reported by other authors (Jannoo et al., 2007; Garsmeur et al., 2011; de Setta et al., 2014; Vilela et al., 2017; Mancini et al., 2018). Region02 contains parts of the *HP600* and *CENP-C* (paralogs) genes. No synteny was found between the sugarcane Region02 and the sorghum genome. In Region02, a third partial gene (ortholog of Sobic.003G299500) was also found next to *CENP-C*, and transcriptome analysis revealed the fusion of partial *CENP-C* exons with the partial exons from the sugarcane ortholog of Sobic.003G299500 to form a chimeric gene. Region02 is a scrambled sugarcane sequence that was possibly formed from different noncollinear ancestral sequences, but the exonic structure of the genes was retained. The phylogenetic analysis of

gene haplotypes from *HP600* and *CENP-C* provided evidence that the multiple genes found in maize are the result of specific duplications in the maize taxa, as expected.

The nature of sugarcane hybrid cultivars, especially the processes of polyploidization (Daniels and Roach, 1987; Paterson et al., 2013) and nobilization (Bremer, 1961), are the main reason for the genomic variability, gene pseudogenization and increases in new genes (McClintock, 1984). It is possible that the structure found in Region02 could be a result of the polyploidization and domestication of sugarcane (Grivet and Arruda, 2002; Cuadrado et al., 2004; D'Hont, 2005; Piperidis et al., 2010). However, the presence of a set of sugarcane homeologs with very similar gene structures leads us to speculate that the occurrence of an ancestral event prior to polyploidization resulted in this structure. Rearrangement events can also be caused by TEs, but they are normally caused by the formation of a pseudogene (Ilic et al., 2003; Lai et al., 2004). In the case of Region02, multiple events resulted in this region, but the number and types (TE, translocations) of events could not be determined with our data.

BAC-FISH hybridization was used to indicate the ploidy of each region. Eight signals were found for Region01 and 10 for Region02. These results are highly consistent with the BAC haplotype and suggest that at least one BAC haplotype is missing in each region. Casu et al. (2012), Xue et al. (2014) and Sun and Joyce (2017) reported different methods to quantify the copy number of endogenous genes, some of which resulted in odd copy numbers. Sun and Joyce (2017) reported that the low or odd numbers could be explained by the contribution of only the *S. spontaneum* or *S. officinarum* genome. The presence of genes without collinearity among the sugarcane homeologs could also explain the result as verified for the orthologs Sobic.003G221800 and Sobic.008G134700.

The genetic, genomic and transcriptome interactions among sugarcane homeologs remain obscure. Several works have attempted to understand these interactions (Jannoo et al., 2007; Wang et al., 2010; Garsmeur et al., 2011; Casu et al., 2012; Figueira et al., 2012; Garcia et al., 2013; de Setta et al., 2014; Xue et al., 2014; Sun and Joyce, 2017; Vilela et al., 2017; Mancini et al., 2018), as well as others. The high polyploidy in sugarcane cultivars make the detection of the ploidy of a locus a challenge (Casu et al., 2012; Garcia et al., 2013; Xue et al., 2014; Sun and Joyce, 2017). Once established, the polyploidy might now fuel evolution by virtue of

its polyploid-specific advantages. Vegetative propagation can lead to the retention of genes. Meiosis may or may not play a role in either the origin or maintenance of a polyploid lineage (Freeling, 2017). Vegetative propagation is widely used to propagate sugarcane (even for non-domesticated sugarcanes) and could explain the high variation in sugarcane (number of SNPs located) and the high level of gene retention. However, it is not the only factor, with sugarcane polyploidization and nobilization also contributing to these effects.

The homologous gene expression in polyploids can be affected in different ways, i.e., the homologous genes may retain their original function, one or more copies may be silenced, or the genes may diversify in function or expression (Ohno, 1970; Lynch and Force, 2000; Hegarty et al., 2006; Buggs et al., 2011). In complex polyploids, the roles of ploidy and genome composition in possible changes in gene expression are poorly understood (Shi et al., 2015). Even in diploid organisms, this task is difficult, as different interactions can affect the expression of a gene, and not all homologs are guaranteed to contribute to a function (Birchler et al., 2005). The expression of the *HP600* and *CENP-C* haplotypes in Region01 could be confirmed. In Region02, the haplotypes of *HP600* were not found in the transcriptome dataset (Cardoso-Silva et al., 2014; Mattiello et al., 2015), but at least two haplotypes of the *CENP-C* gene were expressed.

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

Numerous molecular mechanisms are involved in the creation of new genes, such as exon shuffling, retrotransposons and gene duplications (reviewed in Long et al. (2003)). Gene fusions allow the physical coupling of functions, and their occurrence in the genome increases with the genome size (Snel et al., 2000). Sandmann et al. (2017) describe the function of the protein KNL2, which uses *CENP-C-k* motifs to bind DNA sequences independently and interact with the centromeric transcripts. The *CENP-C* motif is characteristic of *CENP-C*. The *CENP-C* motif in the

rat *CENP-C* protein can bind directly to a chimeric H3/cenH3 nucleosome *in vitro*, suggesting that this motif binds to cenH3 nucleosomes *in vivo*. Consequently, it is directly involved in cell division (Kato et al., 2013; Sandmann et al., 2017). The CENP-C motifs described by Sandmann et al. (2017) were compared with those of *CENP-C* genes in *A. thaliana*, *O. sativa*, *Z. mays* and *S. bicolor* (see Supplementary Figure 9, Supplementary Material). The *CENP-C* haplotypes from Region02 (chimeric gene) have the same *CENP-C* motif as that in sorghum. The *CENP-C* haplotypes from Region01 have one variation in the second residue of the *CENP-C* motif, which is a glycine in sorghum and a valine in *CENP-C* haplotypes from Region01. This result suggests that the chimeric gene retained the ancestral residue at this site, whereas a mutation occurred in *CENP-C* haplotypes from Region01. Therefore, the mutation could have occurred in sorghum and in the haplotypes from Region02, but this is unlikely. This result suggests that the *CENP-C* haplotypes from Region01 and Region02 are able to bind to cenH3 nucleosomes.

The presence of the motif in the *CENP-C* haplotypes from the Region02 proteins could indicate a chimeric protein with a similar function specific to sugarcane, which is involved in the organization of centromeric regions. Moreover, the presence of large LTR retrotransposons in the intronic region of the *CENP-C* haplotypes in Region01 does not influence the gene expression. Furthermore, two studies (Saze et al., 2013; Wang et al., 2013) identified the inactivation of the same gene, IBM2/ANTI-SILENCING 1 (ASI1), which causes gene transcripts with methylated intronic transposons that terminate within the elements. The complete mechanisms that control LTR retrotransposon methylation require further clarification (Kim and Zilberman, 2014).

When we compare *HP600* and *CENP-C* found in SP80-3280 BACs with the *S. spontaneum* genome (Zhang et al., 2018b), we confirmed (i) the presence of the duplication region found in Region02 in one chromosome allele (Chr02D); (ii) the existence of a chimeric gene formed by *CENP-C* and Sobic.003G299500 located in two alleles (Chr02D and Chr02A); and (iii) evidence that the duplication found in Region02 occurred after the separation of Sorghum and before the formation of *Saccharum* genus.

These results have several implications for the integration of the transcriptome and genomic data. First, for example, a gene such as *HP600* that demonstrates

single-copy behavior in the transcriptome data and the genomic behavior of a duplicated gene can cause bias in genetic mapping. Second, a chimeric gene such as the *CENP-C* haplotypes in Region02 can result in different levels of expression of the duplicated and nonduplicated gene regions in the transcriptome data. Using the *CENP-C* gene as an example, if the gene expression quantification probe recovers the nonduplicated portion of the *CENP-C* gene, it will give an expression level only for the *CENP-C* haplotypes in Region01. In contrast, as this probe quantifies the duplicated region of *CENP-C*, it will result in the quantification of *CENP-C* from both Region01 and Region02 and thus overestimate the expression of *CENP-C*. Consequently, analyses of the expression of the gene for functional studies for evaluating the balance of gene expression will be biased.

Molecular markers were also used to compare the ploidy found in BACs with the results from the SuperMASSA software (Garcia et al., 2013). SuperMASSA uses segregation ratios to estimate ploidy, which is not the same as estimating ploidy by chromosome counting because of the differences in the estimation and the real ploidy visualized. The SNPs present in a duplication were mapped in linkage groups (Figure 6 – Panel A) and demonstrated a high distance between the markers in the linkage map. The size of a genetic map is a function of the recombination fraction, with the following two factors influencing the map size: (I) the number of recombinations found between two markers and (II) genotyping errors. In this case, the mapping of duplicated markers is an error and is interpreted by OneMap in a recombination fraction, which inflates the map.

Two markers classified with a ploidy of 10 and one with a ploidy of 8 formed Linkage group 02 (Figure 6 – Panel E). The ploidy is not a determinant for the OneMap construction of a linkage group, but the recombination fraction is. In other words, recombination fractions can still be computed between single-dose markers classified in different ploidy levels. In fact, most nulliplex, simplex and duplex individuals will have the same dosage call using either 8 or 10 as the ploidy level. Additionally, the genome data (BACs and BAC-FISH) demonstrated that all markers had the same ploidy of eight and that the physical distances among the markers were too small and thus probably resulted in the lack of recombination. The fact that we obtained two linkage groups can be explained by the possibility that single-dose markers may be linked in repulsion, and insufficient information is available to

assemble all of the markers into one group. Trying to calculate the recombination fraction between markers D1 and D2 (according to the nomenclature of Wu et al. (Wu et al., 2002)) in diploids presents the same obstacle.

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

The combination of divergent genomes within a hybrid can lead to immediate, profound and highly varied genome modifications, which could include chromosomal and molecular structural modifications (Shen et al., 2005; Doyle et al., 2008; Soltis and Soltis, 2009; Jiang et al., 2011) as well as epigenetic changes (Chen et al., 2010) and global transcriptomic changes (Hegarty et al., 2006; Buggs et al., 2011). The integration of the genetic, genomic and transcriptomic data was used to explain the interaction of the two regions in sugarcane. *HP600* is a hypothetical gene that is next to the *CENP-C* gene, a kinetochore component responsible for the initiation of nucleosomes. The sugarcane *HP600* gene haplotypes in Region01 and the *CENP-C* haplotypes in Region01 were duplicated in another group of homeologous chromosomes. The duplication of the *HP600* haplotypes in Region01 resulted in a paralog pseudogene in the *HP600* haplotypes in Region02. The duplication of *CENP-C* in the haplotypes of Region02 resulted in fusion with another gene, which contained the first five exons of the orthologous Sobic.003G299500 gene and exons eight to fourteen from *CENP-C*. The region where this duplication was inserted (Region02) contained at least three more genes that probably arose due to duplication, which indicates that multiple duplication events occurred in this region.

The *HP600* and *CENP-C* duplication described in this work occurred sometime after the separation of sugarcane and sorghum and before the polyploidization of the *Saccharum* genus. This result is supported by the following information: (I) the molecular clock time; (II) the genes are present in a homeologous group of chromosomes; (III) the *CENP-k* motifs from the *CENP-C* haplotypes in Region02 are more similar to sorghum than to its paralog in sugarcane; and (IV) the duplication was observed in the *S. spontaneum* genome (Zhang et al., 2018b). The formation of a chimeric gene and the scrambled Region02 exhibited a specific moment of

formation before *Saccharum* polyploidization, which makes us wonder which genomic event could be the result of this formation. Multiple events could be the most possible answer.

The transcripts from SP80-3280 revealed full expression of the *HP600* haplotypes in Region01 (in an unbalanced manner) and the lack of expression of the *HP600* haplotypes in Region02. The expression of the *HP600* haplotypes in Region01 can be considered a single-copy gene despite the presence of the duplication. The *CENP-C* gene can be considered fully expressed despite the low coverage of the transcriptome data. The *CENP-C* haplotypes in Region02 have four haplotypes that are considered expressed.

Genetic mapping remains a successful method to improve the production of crop plants. Sugarcane represents one of the crops with difficulties for producing accretive genetic maps, and this impacts the improvement of breeding programs. The variation in ploidy level among the loci and the duplicated genes play a special role in this problem. We used different approaches to show molecular events that affect the genetic mapping as well as the problems associated with defining the ploidy level and dosage among its alleles. Future attention should be given to the relationship between transcription and genomics, as exemplified by the *HP600* gene, which has a single-copy gene behavior in the transcriptome but shows a duplicated region in the genome.

Particular emphasis should be given to the determination studies of the ploidy level and of the duplication loci with the intention of better understanding complex polyploids. These studies remain the most original and challenging in terms of understanding the sugarcane genome. This study sheds light on the influence of the genome arrangement on transcriptome and genetic map analyses in the sugarcane polyploid genome. The integration of genomic sequence arrangements, transcription profiles, cytogenetic organization and the genetic mapping approach might help to elucidate the behavior of gene expression, the genetic structure and successful sequence assembly of the sugarcane genome. Future integrated studies will undoubtedly help to enhance our understanding of complex polyploid genomes including the sugarcane genome.

References

Balsalobre, T.W., da Silva Pereira, G., Margarido, G.R., Gazaffi, R., Barreto, F.Z., Anoni, C.O., et al. (2017). GBS-based single dosage markers for linkage and QTL mapping allow gene mining for yield-related traits in sugarcane. BMC Genomics 18, 72.

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

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

Bremer, G. (1961). Problems in breeding and cytology of sugar cane. Euphytica 10, 59-78.

Buggs, R.J., Zhang, L., Miles, N., Tate, J.A., Gao, L., Wei, W., et al. (2011). Transcriptomic shock generates evolutionary novelty in a newly formed, natural allopolyploid plant. Curr. Biol. 21, 551-556.

Cardoso-Silva, C.B., Costa, E.A., Mancini, M.C., Balsalobre, T.W., Canesin, L.E., Pinto, L.R., et al. (2014). De novo assembly and transcriptome analysis of contrasting sugarcane varieties. PLoS One 9, e88462.

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

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

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

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

Conover, W.J. (1971). Practical Nonparametric Statistics. New York, NY: John Wiley & Sons.

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

Costa, E.A., Anoni, C.O., Mancini, M.C., Santos, F.R.C., Marconi, T.G., Gazaffi, R., et al. (2016). QTL mapping including codominant SNP markers with ploidy level information in a sugarcane progeny. Euphytica 211, 1-16.

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

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

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

D'Hont, A., and Glaszmann, J. (2001). Sugarcane genome analysis with molecular markers, a first decade of research. Proc. Int. Soc. Sugar. Technol. 24, 556-559.

Daniels, J., and Roach, B. (1987). "Taxonomy and evolution," in Sugarcane Improvement through Breeding, ed. H. D. J. (Amsterdam, NL: Elsevier), 7-84.

de Setta, N., Monteiro-Vitorello, C.B., Metcalfe, C.J., Cruz, G.M., Del Bem, L.E., Vicentini, R., et al. (2014). Building the sugarcane genome for biotechnology and identifying evolutionary trends. BMC Genomics 15, 540.

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

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

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

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

Figueira, T.R., Okura, V., da Silva, F.R., da Silva, M.J., Kudrna, D., Ammiraju, J.S., et al. (2012). A BAC library of the SP80-3280 sugarcane variety (*saccharum sp.*) and its inferred microsynteny with the sorghum genome. BMC Res. Notes 5, 185.

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

Garcia, A.A., Mollinari, M., Marconi, T.G., Serang, O.R., Silva, R.R., Vieira, M.L., et al. (2013). SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. Sci. Rep. 3, 3399.

Garsmeur, O., Charron, C., Bocs, S., Jouffe, V., Samain, S., Couloux, A., et al. (2011). High homologous gene conservation despite extreme autopolyploid redundancy in sugarcane. New Phytol. 189, 629-642.

Garsmeur, O., Droc, G., Antonise, R., Grimwood, J., Potier, B., Aitken, K., et al. (2018). A mosaic monoploid reference sequence for the highly complex genome of sugarcane. Nat. Commun. 9, 2638.

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

Gonthier, L., Bellec, A., Blassiau, C., Prat, E., Helmstetter, N., Rambaud, C., et al. (2010). Construction and characterization of two BAC libraries representing a deep-coverage of the genome of chicory (*Cichorium intybus L., Asteraceae*). BMC Res. Notes 3, 225.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., et al. (2012). Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 40, D1178-D1186.

Gopalakrishnan, S., Sullivan, B.A., Trazzi, S., Della Valle, G., and Robertson, K.D. (2009). DNMT3B interacts with constitutive centromere protein CENP-C to modulate DNA methylation and the histone code at centromeric regions. Hum. Mol. Genet. 18, 3178-3193.

Gordon, D. (2003). "Viewing and editing assembled sequences using consed," in Current Protocols in Bioinformatics, eds. D. Baxevanis and D. Davison (New York, NY: John Wiley), 1-43.

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

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

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

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

Guerra, M., and Souza, M.J. (2002). Como Observar Cromossomos: Um Guia de Técnicas em Citogenética Vegetal, Animal e Humana. Ribeirão Preto, Brazil: FUNPEC.

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

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

Hollander, M., Wolfe, D.A., and Chicken, E. (1973). Nonparametric Statistical Methods. New York, NY: John Wiley & Sons.

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

Irvine, J.E. (1999*). Saccharum species* as horticultural classes. Theor. Appl. Genet. 98, 186-194.

Jannoo, N., Grivet, L., Chantret, N., Garsmeur, O., Glaszmann, J.C., Arruda, P., et al. (2007). Orthologous comparison in a gene-rich region among grasses reveals stability in the sugarcane polyploid genome. Plant J. 50, 574-585.

Jiang, B., Lou, Q., Wu, Z., Zhang, W., Wang, D., Mbira, K.G., et al. (2011). Retrotransposon- and microsatellite sequence-associated genomic changes in early generations of a newly synthesized allotetraploid *Cucumis x hytivus* Chen & Kirkbride. Plant Mol. Biol. 77, 225-233.

Jukes, T.H., and Cantor, C.R. (1969). "Evolution of protein molecules," in Mammalian Protein Metabolism, ed. H. Munro (New York, NY: Academic Press), 21-132.

Kato, H., Jiang, J., Zhou, B.R., Rozendaal, M., Feng, H., Ghirlando, R., et al. (2013). A conserved mechanism for centromeric nucleosome recognition by centromere protein CENP-C. Science 340, 1110-1113.

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

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

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

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

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

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

Kohany, O., Gentles, A.J., Hankus, L., and Jurka, J. (2006). Annotation, submission and screening of repetitive elements in repbase: repbase submitter and censor. BMC Bioinformatics 7, 474.

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33, 1870-1874.

Lai, J., Ma, J., Swigonova, Z., Ramakrishna, W., Linton, E., Llaca, V., et al. (2004). Gene loss and movement in the maize genome. Genome Res. 14, 1924-1931.

Landell, M.G.A., Campana, M.P., Figueiredo, P., Vasconcelos, A.C.M., Xavier, M.A., Bidoia, M.A., et al. (2005). Variedades de Cana-de-Açúcar para o Centro-Sul do Brasil: 15a Liberação do Programa Cana IAC (1959-2005). Technical Bulletin IAC 197. Campinas, Brazil: IAC.

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

Le Cunff, L., Garsmeur, O., Raboin, L.M., Pauquet, J., Telismart, H., Selvi, A., et al. (2008). Diploid/polyploid syntenic shuttle mapping and haplotype-specific chromosome walking toward a rust resistance gene (Bru1) in highly polyploid sugarcane (2n approximately 12x approximately 115). Genetics 180, 649-660.

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

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

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

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

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

Manechini, J.R.V., da Costa, J.B., Pereira, B.T., Carlini-Garcia, L.A., Xavier, M.A., Landell, M.G.A., et al. (2018). Unraveling the genetic structure of Brazilian commercial sugarcane cultivars through microsatellite markers. PLoS One 13, e0195623.

Mattiello, L., Riaño-Pachón, D.M., Martins, M.C., da Cruz, L.P., Bassi, D., Marchiori, P.E., et al. (2015). Physiological and transcriptional analyses of developmental stages along sugarcane leaf. BMC Plant Biol. 15, 300.

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

Ming, R., Liu, S.C., Lin, Y.R., da Silva, J., Wilson, W., Braga, D., et al. (1998). Detailed alignment of *Saccharum* and *Sorghum chromosomes*: comparative organization of closely related diploid and polyploid genomes. Genetics 150, 1663-1682.

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

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

Ohno, S. (1970). Evolution by Gene Duplication. New York, NY: Springer-Verlag.

Paiva, J.A., Prat, E., Vautrin, S., Santos, M.D., San-Clemente, H., Brommonschenkel, S., et al. (2011). Advancing *Eucalyptus* genomics: identification and sequencing of lignin biosynthesis genes from deep-coverage BAC libraries. BMC Genomics 12, 137.

Paterson, A.H., Moore, P.H., and Tew, T.L. (2013). "The gene pool of *Saccharum* species and their improvement," in Genomics of the *Saccharinae,* ed. A. H. Paterson (New York, NY: Springer New York), 43-71.

Paux, E., Sourdille, P., Salse, J., Saintenac, C., Choulet, F., Leroy, P., et al. (2008). A physical map of the 1-gigabase bread wheat chromosome 3B. Science 322, 101-104.

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

Piperidis, G., and D'Hont, A. (2001). "Chromosome composition analysis of various *Saccharum* interspecific hybrids by genomic in situ hybridization (GISH)", in: Proceedings of the XXIV Congress, International Society of Sugar Cane Technologists, (Brisbane, QLD), 565-566.

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

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

Riaño-Pachón, D.M., and Mattiello, L. (2017). Draft genome sequencing of the sugarcane hybrid SP80-3280. F1000Research 6, 861.

Roselli, S., Olry, A., Vautrin, S., Coriton, O., Ritchie, D., Galati, G., et al. (2017). A bacterial artificial chromosome (BAC) genomic approach reveals partial clustering of the furanocoumarin pathway genes in parsnip. Plant J. 89, 1119-1132.

RStudio Team (2015). Reproducible Research with R and R Studio. Boston, MA: RStudio, Inc.

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

Sandmann, M., Talbert, P., Demidov, D., Kuhlmann, M., Rutten, T., Conrad, U., et al. (2017). Targeting of *Arabidopsis* KNL2 to centromeres depends on the conserved CENPC-k motif in its C Terminus. Plant Cell 29, 144-155.

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

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

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

Schwarzacher, T., and Heslop-Harrison, P. (2000). Practical *in situ* Hybridization. Didcot, UK: BIOS Scientific Publishers.

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

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

Shi, X., Zhang, C., Ko, D.K., and Chen, Z.J. (2015). Genome-wide dosage-dependent and -independent regulation contributes to gene expression and evolutionary novelty in plant polyploids. Mol. Biol. Evol. 32, 2351-2366.

Snel, B., Bork, P., and Huynen, M. (2000). Genome evolution. Gene fusion versus gene fission. Trends Genet. 16, 9-11.

Soltis, P.S., and Soltis, D.E. (2009). The role of hybridization in plant speciation. Annu. Rev. Plant Biol. 60, 561-588.

Souza, G.M., Berges, H., Bocs, S., Casu, R., D'Hont, A., Ferreira, J.E., et al. (2011). The sugarcane genome challenge: strategies for sequencing a highly complex genome. Trop. Plant Biol. 4, 145-156.

Sun, Y., and Joyce, P.A. (2017). Application of droplet digital PCR to determine copy number of endogenous genes and transgenes in sugarcane. Plant Cell Rep. 36, 1775-1783.

Sylvain, F., Jerome, G., Stephane, R., Catherine, M., Joelle, A., Lieven, S., et al. (2008). Genome annotation in plants and fungi: EuGene as a model platform. Curr. Bioinform. 3, 87-97.

Talbert, P.B., Bryson, T.D., and Henikoff, S. (2004). Adaptive evolution of centromere proteins in plants and animals. J. Biol. 3, 18.

Tomkins, J.P., Yu, Y., Miller-Smith, H., Frisch, D.A., Woo, S.S., and Wing, R.A. (1999). A bacterial artificial chromosome library for sugarcane. Theor. Appl. Genet. 99, 419-424.

Vettore, A.L., da Silva, F.R., Kemper, E.L., Souza, G.M., da Silva, A.M., Ferro, M.I., et al. (2003). Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. Genome Res. 13, 2725-2735.

Vieira, M.L.C., Almeida, C.B., Oliveira, C.A., Tacuatia, L.O., Munhoz, C.F., Cauz-Santos, L.A., et al. (2018). Revisiting meiosis in sugarcane: chromosomal irregularities and the prevalence of bivalent configurations. Front. Genet. 9, 213.

Vilela, M.M., Del Bem, L.E., Van Sluys, M.A., de Setta, N., Kitajima, J.P., Cruz, G.M., et al. (2017). Analysis of three sugarcane homo/homeologous regions suggests independent polyploidization events of *Saccharum officinarum* and *Saccharum spontaneum*. Genome Biol. Evol. 9, 266-278.

Visendi, P., Berkman, P.J., Hayashi, S., Golicz, A.A., Bayer, P.E., Ruperao, P., et al. (2016). An efficient approach to BAC based assembly of complex genomes. Plant Methods 12, 2.

Wang, J., Roe, B., Macmil, S., Yu, Q., Murray, J.E., Tang, H., et al. (2010). Microcollinearity between autopolyploid sugarcane and diploid *Sorghum* genomes. BMC Genomics 11, 261.

Wang, X., Duan, C.G., Tang, K., Wang, B., Zhang, H., Lei, M., et al. (2013). RNA-binding protein regulates plant DNA methylation by controlling mRNA processing at the intronic heterochromatin-containing gene IBM1. Proc. Natl. Acad. Sci. U. S. A. 110, 15467-15472.

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

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

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 35, W265-W268.

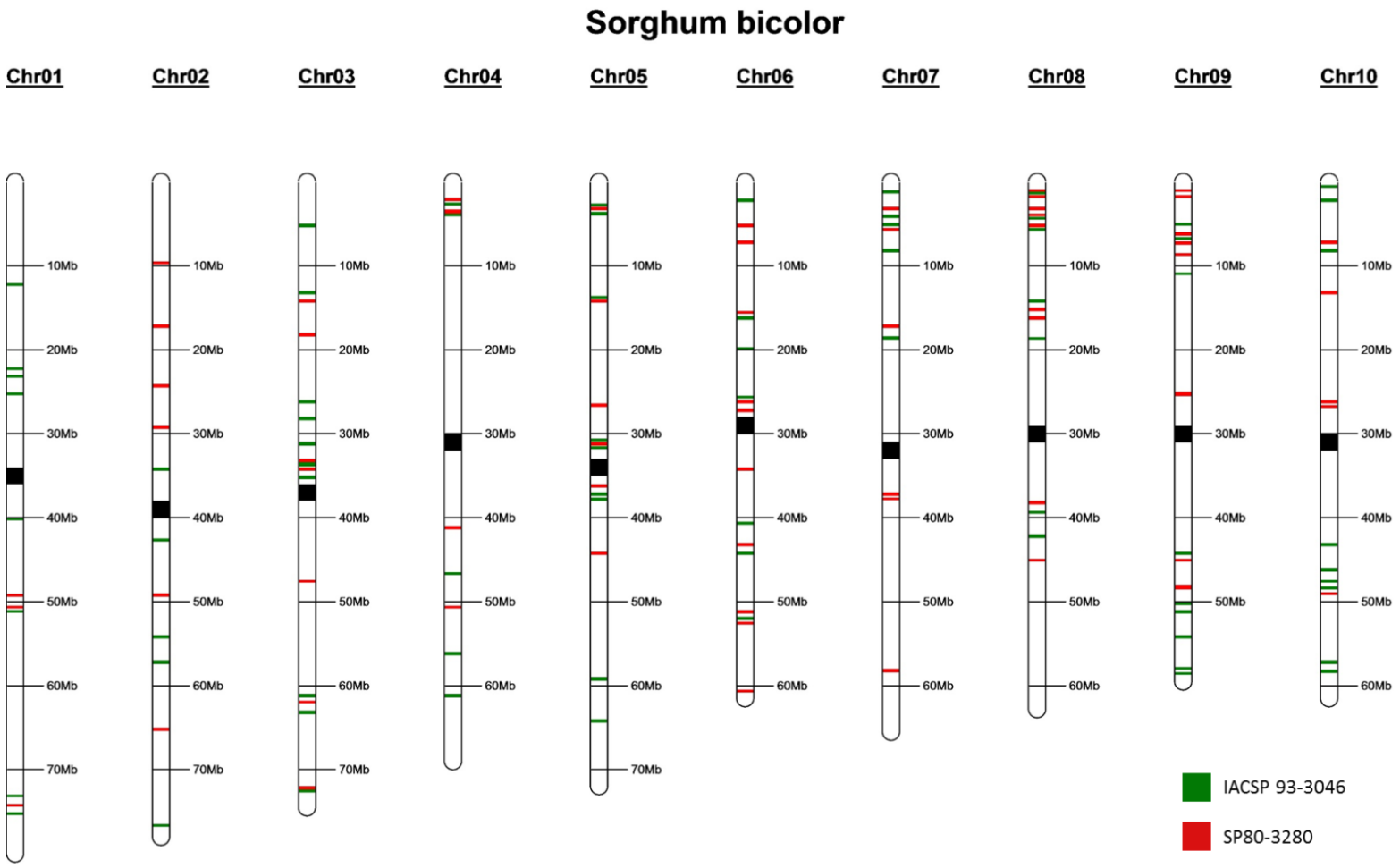
Xue, B., Guo, J., Que, Y., Fu, Z., Wu, L., and Xu, L. (2014). Selection of suitable endogenous reference genes for relative copy number detection in sugarcane. Int. J. Mol. Sci. 15, 8846-8862.

Zhang, J., Zhang, Q., Li, L., Tang, H., Zhang, Q., Chen, Y., et al. (2018a). Recent polyploidization events in three *Saccharum* founding species. Plant Biotechnol. J. doi: 10.1111/pbi.12962

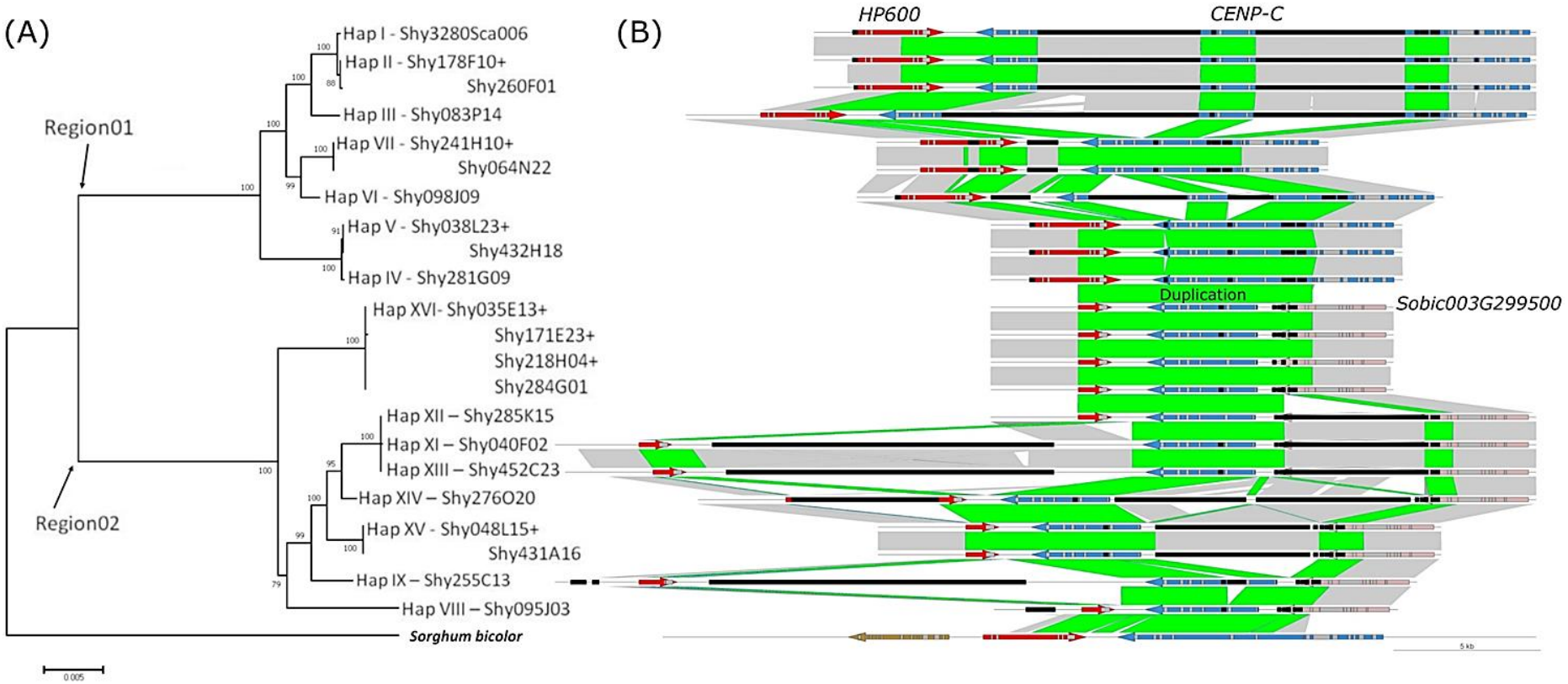
Zhang, J., Zhang, X., Tang, H., Zhang, Q., Hua, X., Ma, X., et al. (2018b). Publisher correction: allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum L*. Nat. Genet. 50, 1754.

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

Supplementary Figures

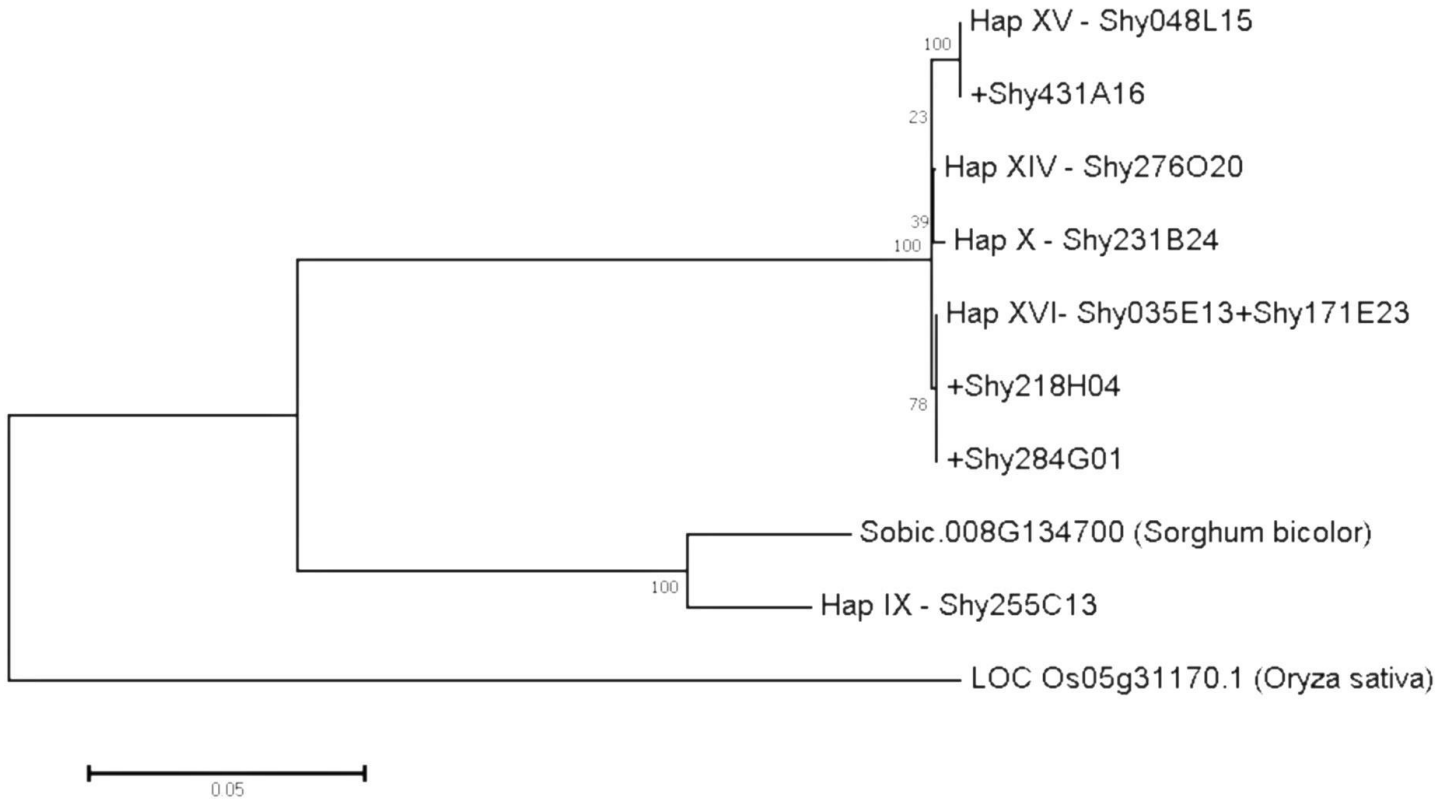


Supplementary Figure 1. BAC-end locations in the *Sorghum* genome according to BLASTn analysis. Schematic representation of the *Sorghum bicolor* genome with 10 chromosomes. The red (sugarcane variety SP80-3280) and green (sugarcane variety IACSP 93-3046) lines show the locations of the paired BAC-end sequences. Black indicates the approximate the position of the *Sorghum bicolor* centromeres.



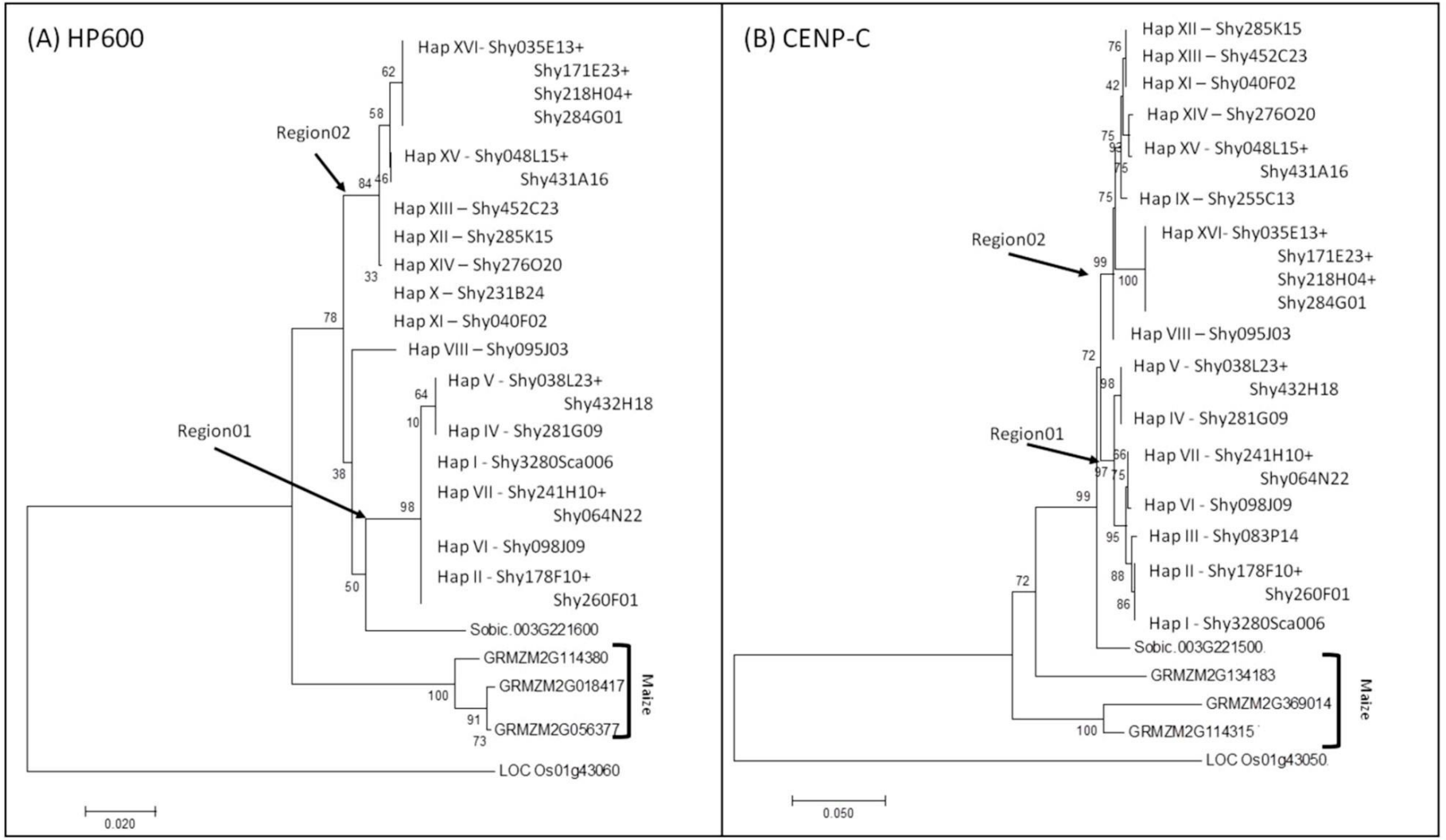
Supplementary Figure 2. Schematic representation of phylogenetics and physical duplications. Panel A: Evolutionary relationships between the duplications found in the 22 sugarcane BACs compared with the same region in *Sorghum bicolor*. Sugarcane BAC Shy231B24 was not included in this analysis because the BAC ends lie in the middle of the duplication. The evolutionary history was inferred using the neighbor-joining method (Saitou and Nei, 1987). The optimal tree with a total branch length = 0.12671921 is shown. The percentages of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches (Felsenstein, 1985). The tree is drawn to scale, with branch lengths in the

same units as the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Kimura 2-parameter method (Kimura, 1980) and are given in units of the number of base substitutions per site. The analysis involved 23 nucleotide sequences. All positions containing gaps and missing data were eliminated. The final dataset included a total of 7025 positions. Evolutionary analyses were conducted in MEGA7 (Kumar et al., 2016). Panel B: Physical representation of the duplications from each BAC according to evolutionary relationships. Green represents the duplications; red represents the HP600 gene; and light blue represents the CENP-C gene. Light pink represents a partial ortholog of the sorghum gene Sobic003G299500. Light gray represents the relationships among BACs outside of the duplicated region. The arrows at the ends of genes HP600, CENP-C and Sobic003G299500 indicate the direction of translation.



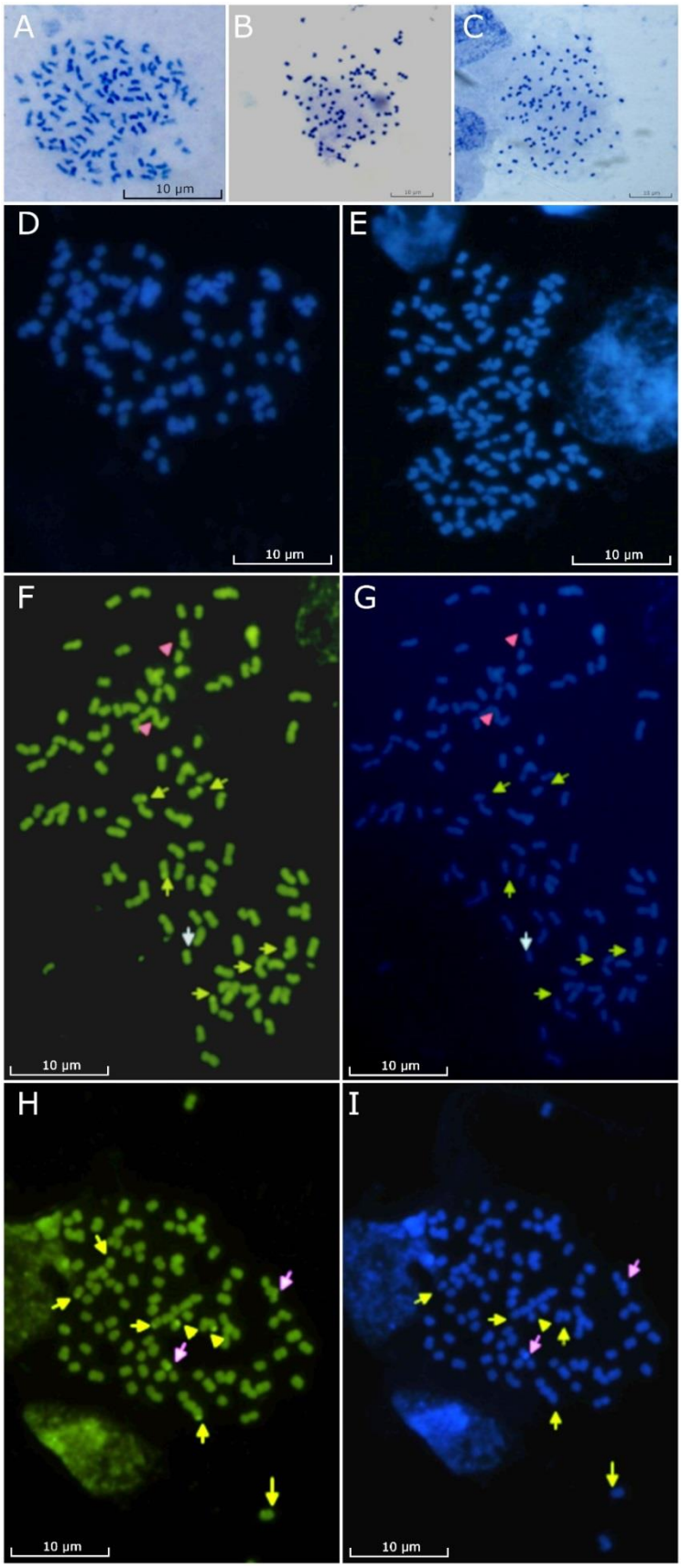
Supplementary Figure 3. Evolutionary relationships of the Sobic.008G134700 gene. The evolutionary history was inferred using the neighbor-joining method (Saitou and Nei, 1987). The optimal tree with a total branch length = 0.47067278 is shown. The percentages of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches (Felsenstein, 1985). The tree is drawn to scale, with branch lengths in the same units as the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Kimura 2-parameter method (Kimura, 1980) and are in units of the number of base substitutions per site. The analysis involved 10 nucleotide sequences. The codon

positions included were the 1st+2nd+3rd+Noncoding positions. All positions containing gaps and missing data were eliminated. The final dataset included a total of 1296 positions. Evolutionary analyses were conducted in MEGA7 (Kumar et al., 2016).



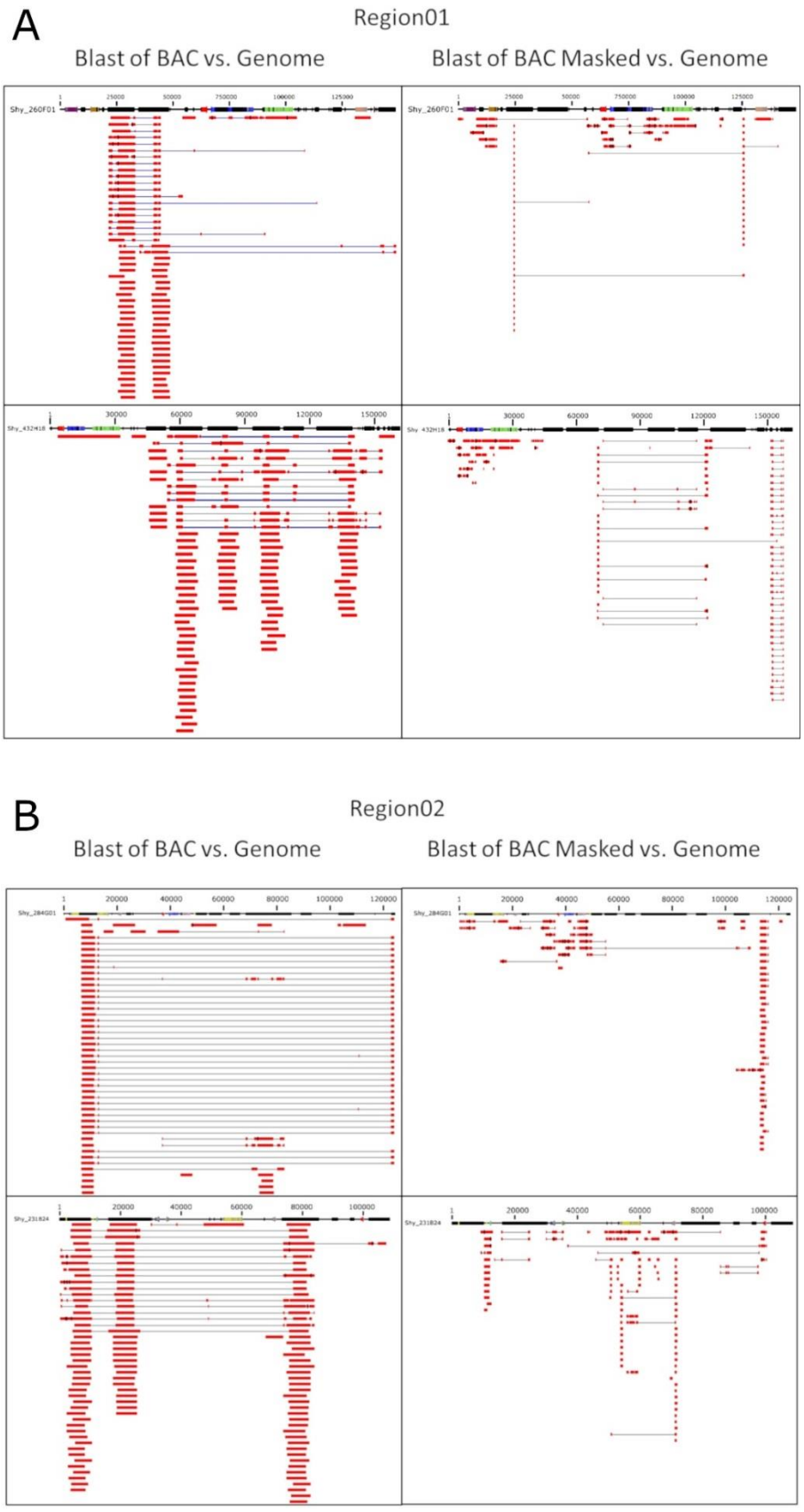
Supplementary Figure 4. Evolutionary relationships of HP600 and CENP-C. Panel A: HP600 evolutionary relationships among the HP600 sugarcane haplotypes in both regions in sorghum, maize (with paralogs) and rice. The haplotypes from BACs

Shy083P14 and Shy255C13 were not used in the analyses because both exhibited a frame shift. There was a total of 100 positions in the final dataset. The optimal tree with a total branch length = 0.48797240 is shown. Panel B: Evolutionary relationships of the CEMP-C haplotypes in both regions in sorghum, maize (with paralogs) and rice. There was a total of 608 positions in the final dataset. The optimal tree with a total branch length = 0.70555298 is shown. The evolutionary history was inferred using the neighbor-joining method (Saitou and Nei, 1987). The percentages of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches (Felsenstein, 1985). The tree is drawn to scale, with the branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. Evolutionary distances were computed using the Kimura 2-parameter method (Kimura, 1980) and are in units of the number of base substitutions per site. The codon positions included were the 1st+2nd+3rd+Noncoding positions. All positions containing gaps and missing data were eliminated. Evolutionary analyses were conducted in MEGA7 (Kumar et al., 2016).



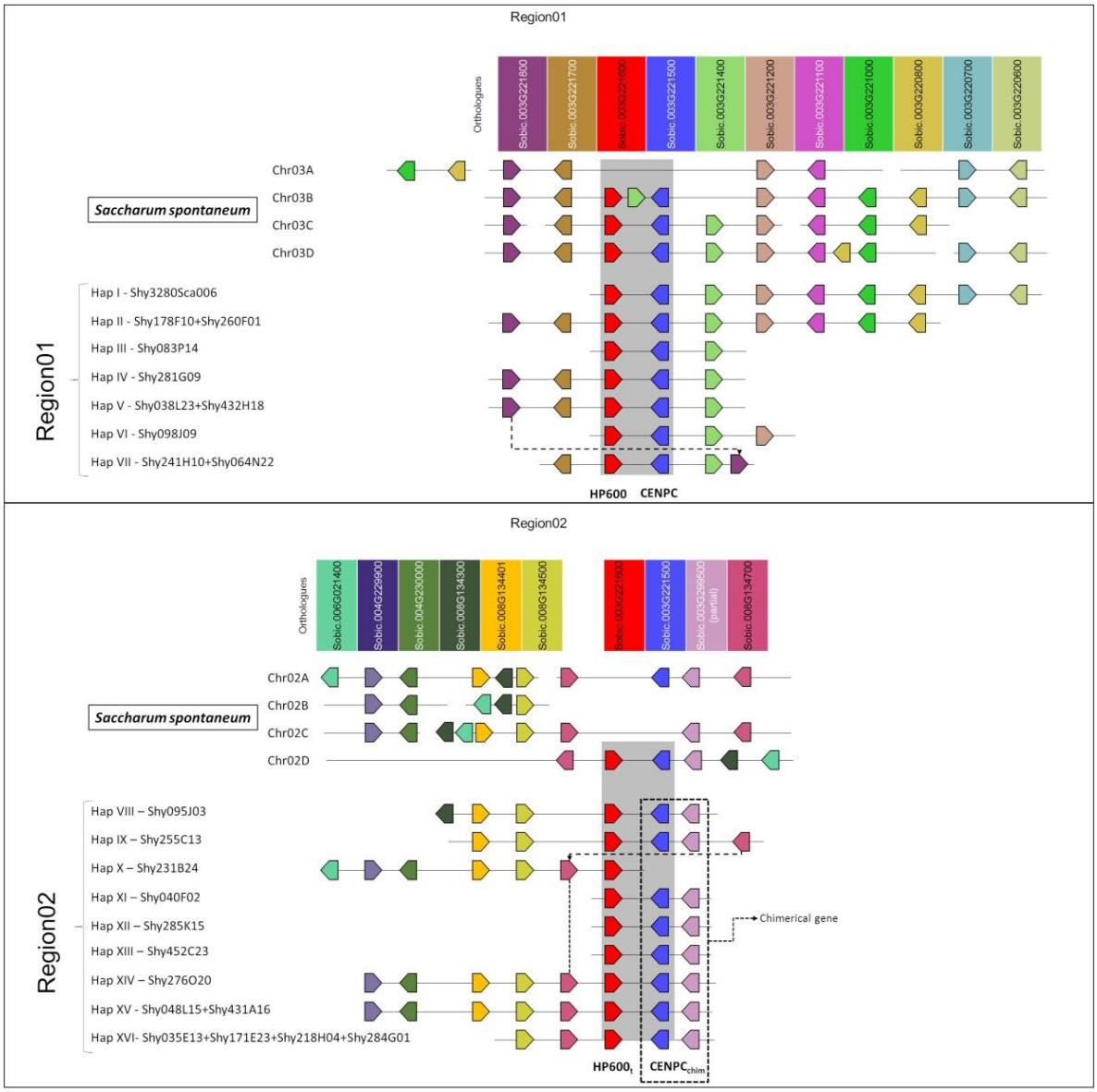
Supplementary Figure 5. Mitotic metaphases of the sugarcane varieties. Panel A: Variety RB835486 with approximately 2n = 112 chromosomes. Giemsa staining.

Panel B: Variety IACSP95-3018 with approximately 2n = 112 chromosomes. Giemsa staining. Panel C: Variety IACSP93-3046 with approximately 2n = 112 chromosomes. Giemsa staining. Panel D: SP803280 with approximately 2n = 110 chromosomes. DAPI staining. Panel E: SP81-3250 with approximately 2n = 114 chromosomes. DAPI staining. Panels F and G: CMA/DAPI banding in the sugarcane variety IACSP 93-3046. The yellow arrows indicate the six CMA^+ (F) and $DAPI^-$ (G) terminal sites. The pink arrows indicate adjacent CMA^+ (F) and $DAPI^+$ (G) sites on the same chromosome. The light blue arrow indicates a CMA^+ (F) and $DAPI^-$ (G) site. Panels H and I: CMA/DAPI banding in the sugarcane variety IACSP 95-3018. The yellow arrows indicate the seven CMA^+ (H) and $DAPI^-$ (I) terminal sites. The pink arrows indicate the chromosomes with adjacent CMA^+ (H) and $DAPI^+$ (I) sites; one site is located at the intercalary position and the other is located at the terminal position.

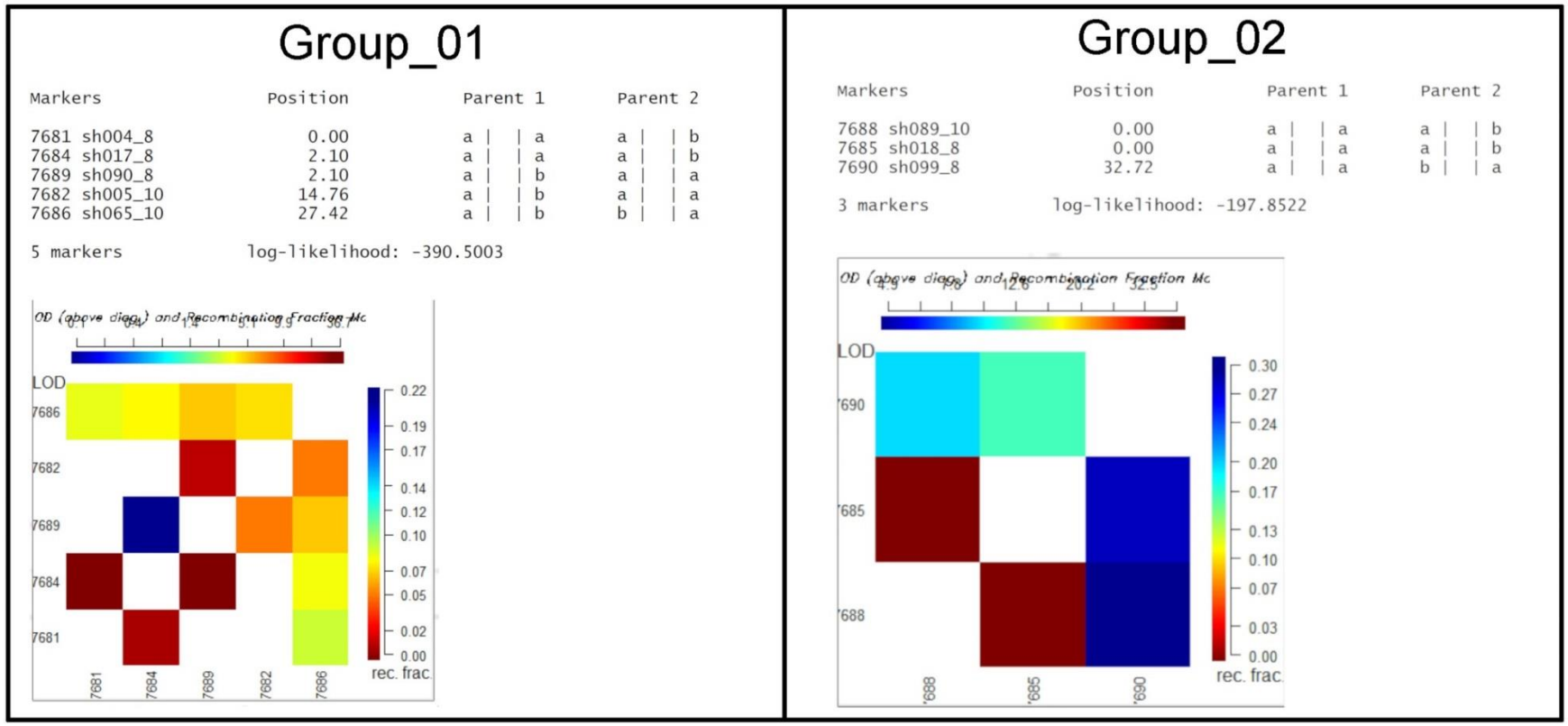


Supplementary Figure 6. BAC BLASTn analysis against sugarcane genome contigs. BLASTn analysis of the sugarcane genome (GCA_002018215.1 – 199.028

sequences) against sugarcane BAC clones. Panel A: BAC Shy_260F01 and BAC Shy_432H18 (Region01) BLAST results. On the left, the repeat regions are not masked. On the right, the repeat regions are masked. Panel B: BAC Shy_284G01 and BAC Shy_231B24 (Region02) BLAST results. On the left, the repeat regions are not masked. On the right, the repeat regions are masked.



Supplementary Figure 7. Schematic comparison with the *S. spontaneum* genome. A schematic comparison between *S. spontaneum* and Region01 and Region02 BACs.



Supplementary Figure 8. Genetic map. The genetic map without the physical information. The markers SugSNP_sh065 and SugSNP_sh099 are weakly linked, but were physically located in both Region01 and Region02.

```
NP_173018.2_centromere_protein_C_Arabidopsis_thaliana     GGVRRSTRIKSRPLEYWRGERFLYGRIHESLTTV
LOC_Os01g43050                                            AGVRRSTRIRSKPLQHWLGERFIYGRIHGTMATV
GRMZM2G114315                                             PGVRKSSRTRSRPLEYWLGERLLYGPIHDNLHGA
Sobic.003G221500                                          SGVRRSTRTRSRPLEHWLGERLLYGPINDTLPAV
Sh035E13_Chimerical_Gene                                  SGVRRSTRTRSRPLEHWLGERLLYGPINDTLPAV
Sh040F02_Chimerical_Gene                                  SGVRRSTRTRSRPLEHWLGERLLYGPINDTLPAV
Sh048L15_Chimerical_Gene                                  SGVRRSTRTRSRPLEHWLGERLLYGPINDTLPAV
Sh171E23_Chimerical_Gene                                  SGVRRSTRTRSRPLEHWLGERLLYGPINDTLPAV
Sh218H04_Chimerical_Gene                                  SGVRRSTRTRSRPLEHWLGERLLYGPINDTLPAV
Sh255C13_Chimerical_Gene                                  SGVRRSTRTRSRPLEHWLGERLLYGPINDTLPAV
Sh276O20_Chimerical_Gene                                  SGVRRSTRTRSRPLEHWLGERLLYGPINDTLPAV
Sh284G01_Chimerical_Gene                                  SGVRRSTRTRSRPLEHWLGERLLYGPINDTLPAV
Sh285K15_Chimerical_Gene                                  SGVRRSTRTRSRPLEHWLGERLLYGPINDTLPAV
Sh431A16_Chimerical_Gene                                  SGVRRSTRTRSRPLEHWLGERLLYGPINDTLPAV
Sh452C23_Chimerical_Gene                                  SGVRRSTRTRSRPLEHWLGERLLYGPINDTLPAV
Sh432H18_cds0040                                          SVVRRSTRTRSRPLEHWLGERLLYGPINDTLPAV
Sh281G09_cds0120                                          SVVRRSTRTRSRPLEHWLGERLLYGPINDTLPAV
Sh260F01_cds0260                                          SVVRRSTRTRSRPLEHWLGERLLYGPINDTLPAV
Sh241H10_cds0090                                          SVVRRSTRTRSRPLEHWLGERLLYGPINDTLPAV
Sh178F10_cds0030                                          SVVRRSTRTRSRPLEHWLGERLLYGPINDTLPAV
Sh098J09_cds0220                                          SVVRRSTRTRSRPLEHWLGERLLYGPINDTLPAV
Sh083P14_cds0190                                          SVVRRSTRTRSRPLEHWLGERLLYGPINDTLPAV
Sh064N22_g0100                                            SVVRRSTRTRSRPLEHWLGERLLYGPINDTLPAV
Sh038L23_cds0080                                          SVVRRSTRTRSRPLEHWLGERLLYGPINDTLPAV
ShPool5c_cds113                                           SVVRRSTRTRSRPLEHWLGERLLYGPINDTLPAV
                                                            **:*:* :*:**::* ***::** *: .:  .
```

Supplementary Figure 9. CENP-C motifs alignment. Alignment of the CENP-C motifs in *Arabidopsis thaliana*, *Oryza sativa*, *Zea mays*, *Sorghum bicolor* and sugarcane BACs.

Supplementary Tables

Supplementary Table 1. BAC assembly and annotation. Summary of the assembled sugarcane BACs.

BAC Name	Assembly							Region	Length w/ Gaps	Transposable Elements (TEs)				Genes	Average Distance by Gene
	Sequencing Technology	Reads	Read Mean Size	Total Bases	Length w/o Gaps	Coverage	Gaps			Annotated		Predicted			
										Bases	%	Bases	%		
Shy038L23	Roche 454	17,577	406	7,131,876	84,182	85	1	Region01	84,282	23,814	28%	37,309	44%	5	16,856
Shy064N22	Roche 454	10,412	413	4,297,624	91,701	47	1	Region01	91,801	20,718	23%	40,179	44%	4	22,950
Shy083P14	Roche 454	4,877	394	1,920,945	99,905	19	1	Region01	100,005	35,056	35%	53,290	53%	3	33,335
Shy098J09	Roche 454	6,157	402	2,474,363	98,874	25	1	Region01	98,974	26,863	27%	35,883	36%	4	24,744
Shy178F10	Roche 454	15,961	445	7,104,720	111,364	64	-	Region01	111,364	59,807	54%	50,537	45%	7	15,909
Shy241H10	Roche 454	31,589	290	9,146,219	134,894	68	1	Region01	134,994	40,990	30%	66,474	49%	5	26,999
Shy260F01	Roche 454	4,346	450	1,954,785	148,093	13	-	Region01	148,093	69,918	47%	74,438	50%	7	21,156
Shy281G09	Roche 454	34,579	298	10,298,703	130,914	79	-	Region01	130,914	47,495	36%	69,302	53%	5	26,183
Shy432H18	Roche 454	15,860	424	6,727,971	162,512	41	-	Region01	162,512	104,933	65%	108,974	67%	3	54,171
Shy3280Sca006	PacBio	60,844	1,274	77,515,256	135,057	574	-	Region01	135,057	44,047	33%	55,205	41%	9	15,006
Shy035E13	Roche 454	6,188	398	2,462,177	105,606	23	1	Region02	105,706	43,471	41%	54,550	52%	5	21,141
Shy040F02	Roche 454	7,402	399	2,955,331	89,075	33	1	Region02	89,175	48,587	54%	61,155	69%	3	29,725
Shy048L15	Roche 454	4,909	384	1,884,792	83,194	23	1	Region02	83,294	29,467	35%	37,594	45%	6	16,659
Shy095J03	Roche 454	5,315	412	2,190,264	90,786	24	-	Region02	90,786	36,285	40%	43,174	48%	6	18,157
Shy171E23	Roche 454	6,349	448	2,845,515	48,796	58	1	Region02	48,896	27,884	57%	27,875	57%	3	16,299
Shy218H04	Roche 454	20,668	299	6,172,319	68,037	91	1	Region02	68,137	14,042	21%	28,951	42%	5	13,627
Shy231B24	Roche 454	22,282	296	6,588,557	107,057	62	2	Region02	107,257	48,126	45%	64,756	60%	7	17,876
Shy255C13	Roche 454	15,045	440	6,615,510	151,415	44	2	Region02	151,615	59,481	39%	73,712	49%	6	30,323
Shy276O20	Roche 454	16,919	443	7,490,042	105,869	71	-	Region02	105,869	45,134	43%	64,911	61%	8	15,124
Shy284G01	Roche 454	12,607	447	5,630,527	122,961	46	1	Region02	123,061	51,904	42%	56,501	46%	5	24,612
Shy285K15	Roche 454	14,393	436	6,273,629	99,815	63	3	Region02	100,115	37,220	37%	58,277	58%	3	33,372
Shy431A16	Roche 454	10,993	392	4,308,085	132,490	33	1	Region02	132,590	58,443	44%	70,437	53%	8	18,941
Shy452C23	Roche 454	16,361	283	4,637,892	100,514	46	2	2	100,714	53,185	53%	61,428	61%	3	33,571
Mean		15,723	429	8,201,178	108,831	71			108,922	44,647	40%	56,301	51%	5	23,771

Supplementary Table 2. Orthologous genes from Region01. Region01 orthologous genes found in sugarcane BACs. "+" indicates that the chromosome has partial genes and "*" indicates more than one copy.

SORGHUM	GENE	BAC GENES	SUCEST-FUN	S. SPONTANEUM [Zhang et al., 2018]	RICE	MAIZE
SOBIC.003G221800	Putative uncharacterized protein (C5XR27) - Probable aldo-keto reductase 5	Sh241H10_g0150, Sh260F01_g0020, Sh281G09_g0030	SCMCLV1032B11	Chr3A, Chr3B* and +, Chr3C* and +, Chr3D* and +,	LOC_Os01g43090	GRMZM2G024315
SOBIC.003G221700	Putative uncharacterized protein (B6U3Y6) – Similar to calcium ion binding related to photosystems II	Sh038L23_g0290, Sh064N22_g0030, Sh241H10_g0030, Sh260F01_g0060, Sh281G09_g0060	SCQSAM1030F07	Chr3A, Chr3B, Chr3C, Chr3D	LOC_Os01g43070	GRMZM2G000256
SOBIC.003G221600	Putative uncharacterized protein (C5XR26)	Sh038L23_g0100, ShPool5c_g118, Sh064N22_g0080, Sh083P14_g0220, Sh098J09_g0230, Sh178F10_g0010, Sh241H10_g0060, Sh260F01_g0210, Sh281G09_g0090, Sh432H18_g0030	SCCCSB1004A04	Chr2D+, Chr3B, Chr3C, Chr3D* and +	LOC_Os01g43060	GRMZM2G114380
SOBIC.003G221500	CENP-C1 (Q66LH0) - CENTROMERE PROTEIN C CENP-C2 (Q66LH1) - CENTROMERE PROTEIN C	Sh064N22_g0100, Sh083P14_g0190, ShPool5c_g113, Sh098J09_g0220, Sh178F10_g0030, Sh241H10_g0090, Sh260F01_g0260 Sh038L23_g0080, Sh281G09_g0120, Sh432H18_g0040	SCSGFL4190C08	Chr2A+, Chr2D+, Chr3B, Chr3C, Chr3D* and +, Chr7B+	LOC_Os01g43050	GRMZM2G114315
SOBIC.003G221400	Putative uncharacterized protein (C5XR25) - Similar to acetyltransferase 1-like	Sh038L23_g0060, Sh064N22_g0120, Sh083P14_g0160, ShPool5c_g090, Sh098J09_g0140, Sh178F10_g0080, Sh241H10_g0120, Sh260F01_g0280, Sh281G09_g0150, Sh432H18_g0080	SCJFRZ2031F10	Chr2D+, Chr3B, Chr3C, Chr3D*	LOC_Os01g43030	GRMZM2G047093
SOBIC.003G221200	CTP synthase (C5XR23)	Sh178F10_g0170, ShPool5c_g051, Sh260F01_g0360	SCCCAM1C03H03	Chr3A* and +, Chr3B+, Chr3C, Chr3D	LOC_Os01g43020	GRMZM2G153058
SOBIC.003G221100	Putative uncharacterized protein (C5XR21)	Sh178F10_g0190, ShPool5c_g043, Sh260F01_g0380	SCSFFL8044H03	Chr3A*, Chr3B, Chr3C, Chr3D	LOC_Os02g37140	AC209364.3_FGP009
SOBIC.003G221000	Putative uncharacterized protein (C5XR20) - Similar to membrane-associated salt-inducible protein-like	Sh178F10_g0260, Shpool5c_g028	SCEZHR1054H09	Chr3A, Chr3B, Chr3C, Chr3D	LOC_Os01g42990	GRMZM2G369931
SOBIC.003G220800	Putative uncharacterized protein (C5XR18) - RNA recognition motif (RRM)-containing protein-like	ShPool5c_g022, Sh178F10_g0290	No hit	Chr3A, Chr3B, Chr3C, Chr3D	LOC_Os08g23120	GRMZM2G152526
SOBIC.003G220700	Putative uncharacterized protein (A0A1B6Q4P3) - Zinc finger protein 3-like	Shpool5c_g018	No hit	Chr3A*, Chr3B* and +, Chr3D* and +	LOC_Os01g42970	GRMZM2G039889
SOBIC.003G220600	Putative uncharacterized protein (C5XR17) - Charged multivesicular body protein 4b	Shpool5c_g005	No hit	Chr1A, Chr1B*, Chr3A*, Chr3B* and +, Chr3D* and +,	LOC_Os09g09480	GRMZM2G107757

Supplementary Table 3. Orthologous genes from Region02. Region02 orthologous genes found in the sugarcane BACs. “+” indicates that the chromosome has partial genes and “*” indicates more than one copy.

SORGHUM	GENE	BAC GENES	SUCEST-FUN	S. SPONTANEUM [Zhang et al., 2018]	RICE	MAIZE
SOBIC.006G021400	Putative uncharacterized protein (A0A1Z5RBR0) - Histidine kinase/Protein kinase (histidine)	Sh231B24_g0030	No hit	Chr2A, Chr2B, Chr2C+, Chr2D, Chr5A*, Chr5B*, Chr5C*, Chr5D*	LOC_Os04g13480	GRMZM2G308046
SOBIC.004G229900	Putative uncharacterized protein (Q651V4) - Ribosomal protein-like	Sh276O20_g0210, Sh431A16_g0040, Sh231B24_g0060	SCQGST1031H03	Chr1A, Chr1B, Chr1D, Chr2A*, Chr2B, Chr2C*, Chr3D, Chr4A, Chr5B*, Chr6A*, Chr6B*, Chr6C, Chr7B, Chr7D, Chr8C*	LOC_Os01g40070	GRMZM2G076892
SOBIC.004G230000	Putative uncharacterized protein (A0A194YRF1) - Similar to NAM-like protein	Sh231B24_g0090, Sh276O20_g0160, Sh431A16_g0060	SCQGLR1086C07	Chr1A, Chr2A*, Chr2B, Chr2C * and +, Chr3D, Chr4A+, Chr5B*, Chr6A, Chr6B * and +, Chr6C, Chr7D, Chr8C	LOC_Os06g02710	GRMZM5G887243
SOBIC.008G134300	Uncharacterized protein (C5YPX1) - Protein tyrosine kinase	Sh095J03_g0010	SCJLST1024G01	Chr2A* and +, Chr2B* and +, Chr2C* and +, Chr2D* and +	LOC_Os12g37980	GRMZM2G332280
SOBIC.008G134401	Uncharacterized protein (A0A1Z5R7M4)	Sh048L15_g0030, Sh431A16_g0090, Sh095J03_g0030, Sh255C13_g0020, Sh231B24_g0110, Sh276O20_g0150	No hit	Chr2A+, Chr2C*	No hit	No hit
SOBIC.008G134500	Putative uncharacterized protein (A0A1Z5R6L9) - Protein-tyrosine kinase	Sh035E13_g0170, Sh095J03_g0210, Sh231B24_g0120, Sh255C13_g0040, Sh284G01_g0080, Sh431A16_g0140, Sh048L15_g0060, Sh218H04_g0010, Sh276O20_g0140	SCSBAM1086F04	Chr2A* and +, Chr2B* and +, Chr2C* and +, Chr2D* and +,	LOC_Os12g37980	GRMZM2G332280
SOBIC.003G221600	Putative uncharacterized protein (C5XR26)	Sh035E13_g0220, Sh040F02_g0030, Sh048L15_g0140, Sh095J03_g0310, Sh171E23_g0010, Sh218H04_g0060, Sh231B24_g0190, Sh255C13_g0080, Sh276O20_g0060, Sh284G01_g0110, Sh285K15_g0090, Sh431A16_g0220, Sh452C23_g0010	SCCCSB1004A04	Chr2D+, Chr3B, Chr3C, Chr3D* and +	LOC_Os01g43060	GRMZM2G056377
SOBIC.003G221500	CENP-C2 (Q66LH1) - CENTROMERE PROTEIN C	Sh035E13_g0230, Sh040F02_g0060, Sh048L15_g0150, Sh095J03_g0330, Sh171E23_g0030, Sh218H04_g090, Sh255C13_g0120, Sh276O20_g0090, Sh284G01_g0120, Sh285K15_g0010, Sh431A16_g0250, Sh452C23_g0030	SCSGFL4190C08	Chr2A+, Chr2D+, Chr3B, Chr3C, Chr3D* and +, Chr7B+	LOC_Os01g43050	GRMZM2G114315

SOBIC.003G299500	Uncharacterized protein (A0A1W0VZP0)	Sh035E13_g0240, Sh040F02_g0090, Sh048L15_g0170, Sh095J03_g0360, Sh171E23_g0060, Sh218H04_g0120, Sh255C13_g0130, Sh276O20_g0120, Sh284G01_g0130, Sh285K15_g0030, Sh431A16_g0280, Sh452C23_g0060	SCJLAM1062D01	Chr2A+, Chr2C* and +, Chr2D+, Chr3A, Chr3C* and +, Chr3D* and +	LOC_Os01g55094	GRMZM2G309660
SOBIC.008G134700	Putative uncharacterized protein (C5YPX8) – Similar to aspartyl protease	Sh035E13_g0190, Sh048L15_g0130, Sh218H04_g0030, Sh231B24_g0160, Sh276O20_g0110, Sh284G01_g0090, Sh431A16_g0190, Sh255C13_g0380	SCBFAD1048G09	Chr2A* and +, Chr2C* and +, Chr2D* and +	LOC_Os05g31170	GRMZM2G060680

Supplementary Table 4. Number of SNPs found in CENP-C and HP600. Summary of the sugarcane SNP counts by gene in the duplications in Region01 and Region02.

		HP600			CENPC			HP600 + CENPC		
		Base length	SNPs	Bases/ SNP	Base length	SNPs	Bases/ SNP	Base length	SNPs	Bases/ SNP
Exonic Region	Region01	419	4	105	695	12	58	1114	16	70
	Region02	419	10	42	695	16	43	1114	26	43
	Region01 + Region02	419	24	17	695	38	18	1114	62	18
	Specific to Region	419	10	42	695	10	70	1114	20	56
Intronic Region	Region01	630	4	158	4703	131	36	5333	135	40
	Region02	630	17	37	4703	160	29	5333	177	30
	Region01 + Region02	630	36	18	4703	352	13	5333	388	14
	Specific to Region	630	15	42	4703	61	77	5333	76	70
Intronic + Exonic	Region01	1049	8	131	5398	143	38	6447	151	43
	Region02	1049	27	39	5398	176	31	6447	203	32
	Region01 + Region02	1049	60	17	5398	390	14	6447	450	14
	Specific to Region	1049	25	42	5398	71	76	6447	96	67

Supplementary Table 5. Number of SNPs found in duplicated regions. Summary of the sugarcane SNP counts in duplications in Region01 and Region02.

	Intergenic			**Whole duplication**		
	Base length	SNPs	Bases/ SNP	Base length	SNPs	Bases/ SNP
Region01	2513	80	31	8960	232	39
Region02	2513	81	31	8960	284	32
Region01 + Region02	2513	269	9	8960	719	12
Specific per region	2513	108	23	8960	203	44

Supplementary Table 6. Chromosome counts. Chromosome counts by sugarcane variety.

CHROMOSOMES	IACSP95-3018	IACSP93-3046	RB835486	SP80-3280	SP81-3250
98	0	0	0	0	6
99	0	0	0	0	3
100	0	0	3	0	6
101	0	0	0	0	0
102	2	0	5	0	6
103	1	0	1	0	1
104	7	0	0	0	1
105	0	0	4	0	0
106	5	2	5	0	5
107	1	3	1	0	2
108	8	3	8	2	3
109	2	7	0	0	0
110	9	11	7	2	6
111	4	2	0	1	4
112	11	13	15	5	7
113	2	0	1	2	0
114	1	3	3	1	2
115	0	1	0	4	0
116	0	1	3	2	0
117	0	0	0	3	0
118	0	0	0	2	0
SUM	53	46	56	24	52

Supplementary Table 7. Sequenom iPLEX MassARRAY® primers. SNPs derived from the HP600 and CENP-C duplicated regions genotyped in the population and the three primers used for genotyping on the SEQUENOM platform.

SNP ID	*Forward Primer ID*	*Forward Primer Sequence*	*Reverse Primer ID*	*Reverse Primer Sequence*	*Extended Primer ID*	*Extended Primer Sequence*
SugSNP_sh081	SugSNP_Sh_081_W1_F	ACGTTGGATGGTCTGACAAAGATAATAAATG	SugSNP_Sh_081_W1_R	ACGTTGGATGCTTTTATTGGGCTTCTTTCC	SugSNP_Sh_081_W1_E	ACCCATCTCCGCGTCAT
SugSNP_sh099	SugSNP_Sh_099_W1_F	ACGTTGGATGTCATTCAGAGCAAGCTGTGG	SugSNP_Sh_099_W1_R	ACGTTGGATGACTATCATCTTCCGAGTCAG	SugSNP_Sh_099_W1_E	TTCCGAGTCAGATGAGC
SugSNP_sh086	SugSNP_Sh_086_W1_F	ACGTTGGATGTTCAATGGTGCAGTCAGCAG	SugSNP_Sh_086_W1_R	ACGTTGGATGGGGAGCTTGTTGGACATTTG	SugSNP_Sh_086_W1_E	AGTCAGCAGCTCTTCCTT
SugSNP_sh090	SugSNP_Sh_090_W1_F	ACGTTGGATGATCTGCTCAAGTGTCGGTTC	SugSNP_Sh_090_W1_R	ACGTTGGATGCTAAGATCTTTTTCAGTGGC	SugSNP_Sh_090_W1_E	TTAAACCTTGTTTCCGTG
SugSNP_sh061	SugSNP_Sh_061_W1_F	ACGTTGGATGCTTTACAGGAGCACCATGGG	SugSNP_Sh_061_W1_R	ACGTTGGATGGATGAAGGAGGCGGGAGGC	SugSNP_Sh_061_W1_E	GGAGCACCATGGGAGAGCC
SugSNP_sh015	SugSNP_Sh_015_W1_F	ACGTTGGATGGCAGGCCATATTCTTGATCC	SugSNP_Sh_015_W1_R	ACGTTGGATGAACCAACTGAGGAACCTCTG	SugSNP_Sh_015_W1_E	TGATCCTGAACCATGCTTGC
SugSNP_sh043	SugSNP_Sh_043_W1_F	ACGTTGGATGAGCACTTGAGCAGCAATGC	SugSNP_Sh_043_W1_R	ACGTTGGATGTTGCTGCTTCACCTATGCTC	SugSNP_Sh_043_W1_E	GTTTGGTTCACTTAGTGGTAC
SugSNP_sh037	SugSNP_Sh_037_W1_F	ACGTTGGATGGGCCAAGATGGCAAGAGAAC	SugSNP_Sh_037_W1_R	ACGTTGGATGGGATTTAGCGACAAGATCTG	SugSNP_Sh_037_W1_E	AGAACATTGAAAGTGAAATCT
SugSNP_sh005	SugSNP_Sh_005_W1_F	ACGTTGGATGAGCAAACCGATGCCTGTTG	SugSNP_Sh_005_W1_R	ACGTTGGATGAAAGTATTCAACTGGATCCG	SugSNP_Sh_005_W1_E	CCTGTTGTGGATCAATCTAAGTT
SugSNP_sh064	SugSNP_Sh_064_W1_F	ACGTTGGATGACAAGGAGGGGAAGCGTAAG	SugSNP_Sh_064_W1_R	ACGTTGGATGAGAGCTAGTTCAACAGTACC	SugSNP_Sh_064_W1_E	GAGCTAGTTCAACAGTACCTTGGCT
SugSNP_sh066	SugSNP_Sh_066_W2_F	ACGTTGGATGCAATTGCAGAACAAGCCTCC	SugSNP_Sh_066_W2_R	ACGTTGGATGTCTCTCTCACCATCTCAATG	SugSNP_Sh_066_W2_E	TACTCCTCGTTCCCAGT
SugSNP_sh035	SugSNP_Sh_035_W2_F	ACGTTGGATGTGCCTGCAGTTATTGGCATC	SugSNP_Sh_035_W2_R	ACGTTGGATGTCTGAATACTGGTCAGGCAC	SugSNP_Sh_035_W2_E	TGGCATCAAAGCATACT
SugSNP_sh013	SugSNP_Sh_013_W2_F	ACGTTGGATGACCAGATATTGTGATGGGTG	SugSNP_Sh_013_W2_R	ACGTTGGATGGGTCTCCCTGTCAATTTCAC	SugSNP_Sh_013_W2_E	TGTGATGGGTGAACCAT
SugSNP_sh016	SugSNP_Sh_016_W2_F	ACGTTGGATGGCAGGCCATATTCTTGATCC	SugSNP_Sh_016_W2_R	ACGTTGGATGCACAAACCAACTGAGGAACC	SugSNP_Sh_016_W2_E	TGAGGAACCTCTGGATTG
SugSNP_sh083	SugSNP_Sh_083_W2_F	ACGTTGGATGGATAATAAATGTAAAGGTCC	SugSNP_Sh_083_W2_R	ACGTTGGATGTCTGTCTTTTATTGGGCTTC	SugSNP_Sh_083_W2_E	TATTGGGCTTCTTTCCTTT
SugSNP_sh019	SugSNP_Sh_019_W2_F	ACGTTGGATGGATGTGCCAATAGACTATCC	SugSNP_Sh_019_W2_R	ACGTTGGATGGCCCTCCAGATGATGAGAAG	SugSNP_Sh_019_W2_E	ACTATTGGCAGATCTACTAG
SugSNP_sh100	SugSNP_Sh_100_W2_F	ACGTTGGATGTCTGACTCGGAAGATGATAG	SugSNP_Sh_100_W2_R	ACGTTGGATGTCTTTGTCAGACACGATAGG	SugSNP_Sh_100_W2_E	GGAAGATGATAGTGATGACA
SugSNP_sh088	SugSNP_Sh_088_W2_F	ACGTTGGATGCTGCGATATCACATCTGCTC	SugSNP_Sh_088_W2_R	ACGTTGGATGCACGGAAACAAGGTTTAAAG	SugSNP_Sh_088_W2_E	TGCTCAAGTGTCGGTTCTCCTT
SugSNP_sh092	SugSNP_Sh_092_W2_F	ACGTTGGATGGATCTTCATTCAGAGCAAGC	SugSNP_Sh_092_W2_R	ACGTTGGATGATCTTCCGAGTCAGATGAGC	SugSNP_Sh_092_W2_E	TTCATTCAGAGCAAGCTGTGGA
SugSNP_sh085	SugSNP_Sh_085_W2_F	ACGTTGGATGAACGCTACATGCAACTCTGG	SugSNP_Sh_085_W2_R	ACGTTGGATGAGCTGCTGACTGCACCATTG	SugSNP_Sh_085_W2_E	CATTGAAAAAATTCTTTTGGTAAG
SugSNP_sh031	SugSNP_Sh_031_W3_F	ACGTTGGATGATGAATCTAGCCATGCACTG	SugSNP_Sh_031_W3_R	ACGTTGGATGCATCATTATGAGGTTGATTC	SugSNP_Sh_031_W3_E	ACTGGAAATACCCCAAG
SugSNP_sh067	SugSNP_Sh_067_W3_F	ACGTTGGATGGATCTTCATTCAGAGCAAGC	SugSNP_Sh_067_W3_R	ACGTTGGATGGGACCTTTACATTTATTATC	SugSNP_Sh_067_W3_E	TCAGAGCAAGCTGTGGA
SugSNP_sh004	SugSNP_Sh_004_W3_F	ACGTTGGATGGGCTAAAATGGTGCTGAAGG	SugSNP_Sh_004_W3_R	ACGTTGGATGTGCTCTTTGCTGCCATTTGC	SugSNP_Sh_004_W3_E	GATGGAGTGAAAGCGAG
SugSNP_sh030	SugSNP_Sh_030_W3_F	ACGTTGGATGCACCTGATTTGTGCAATG	SugSNP_Sh_030_W3_R	ACGTTGGATGTGCCCCTCTTTGACTGTTTC	SugSNP_Sh_030_W3_E	ACTGTTTCTTCATTTTCCC
SugSNP_sh042	SugSNP_Sh_042_W3_F	ACGTTGGATGAGCACTTGAGCAGCAATGC	SugSNP_Sh_042_W3_R	ACGTTGGATGTTGCTGCTTCACCTATGCTC	SugSNP_Sh_042_W3_E	ACCACTAAGTGAACCAAAC
SugSNP_sh003	SugSNP_Sh_003_W3_F	ACGTTGGATGACTCCTTCCGCGGACCCTT	SugSNP_Sh_003_W3_R	ACGTTGGATGACTGCAATGGCCTCGAGGA	SugSNP_Sh_003_W3_E	GCCTCGAGGAGCGCGTCGC
SugSNP_sh017	SugSNP_Sh_017_W3_F	ACGTTGGATGGAGGTTCCTCAGTTGGTTTG	SugSNP_Sh_017_W3_R	ACGTTGGATGCACAGAAATATTGCTCCTCC	SugSNP_Sh_017_W3_E	GTTTGTGCAGAGACACAGA
SugSNP_sh052	SugSNP_Sh_052_W3_F	ACGTTGGATGCTGACATAATCACATAACAC	SugSNP_Sh_052_W3_R	ACGTTGGATGGGCAAAGCTGGAAAAGATAC	SugSNP_Sh_052_W3_E	AACACTATTCAAAGTTCTGTT
SugSNP_sh012	SugSNP_Sh_012_W3_F	ACGTTGGATGGCCAGGATCGAATTTCAGAG	SugSNP_Sh_012_W3_R	ACGTTGGATGTCTGGTGAACCTTCAGATTC	SugSNP_Sh_012_W3_E	TAGAGGTGCTGCAAATGCTAA
SugSNP_sh091	SugSNP_Sh_091_W3_F	ACGTTGGATGCTGATTTGTGATACAGGGAG	SugSNP_Sh_091_W3_R	ACGTTGGATGTGCTGATGCAAGACAACCTG	SugSNP_Sh_091_W3_E	TACAGGGAGGACCGTGAAAACA
SugSNP_sh001	SugSNP_Sh_001_W4_F	ACGTTGGATGACTCCTTCCGCGGACCCTT	SugSNP_Sh_001_W4_R	ACGTTGGATGACTGCAATGGCCTCGAGGA	SugSNP_Sh_001_W4_E	GACCCTTGGCCCCGCCC
SugSNP_sh084	SugSNP_Sh_084_W4_F	ACGTTGGATGCTTTACAGGAGCACCATGGG	SugSNP_Sh_084_W4_R	ACGTTGGATGCAAGCCTACCGATGAAGGAG	SugSNP_Sh_084_W4_E	AGCACCATGGGAGAGCC
SugSNP_sh036	SugSNP_Sh_036_W4_F	ACGTTGGATGTGCCTGCAGTTATTGGCATC	SugSNP_Sh_036_W4_R	ACGTTGGATGTCTGAATACTGGTCAGGCAC	SugSNP_Sh_036_W4_E	TCTCTTGCCATCTTGGCC
SugSNP_sh087	SugSNP_Sh_087_W4_F	ACGTTGGATGTTCAATGGTGCAGTCAGCAG	SugSNP_Sh_087_W4_R	ACGTTGGATGGGGAGCTTGTTGGACATTTG	SugSNP_Sh_087_W4_E	TGTTGGTGAGTTTTGGAA
SugSNP_sh065	SugSNP_Sh_065_W4_F	ACGTTGGATGGGTATTCTTTCGTGTGCCAG	SugSNP_Sh_065_W4_R	ACGTTGGATGCAATCTCCCCCAAGAAATCC	SugSNP_Sh_065_W4_E	AAGAAATCCTTGACTTTCC
SugSNP_sh082	SugSNP_Sh_082_W4_F	ACGTTGGATGGATAATAAATGTAAAGGTCC	SugSNP_Sh_082_W4_R	ACGTTGGATGTCTGTCTTTTATTGGGCTTC	SugSNP_Sh_082_W4_E	ATGACGCGGAGATGGGTCC
SugSNP_sh014	SugSNP_Sh_014_W4_F	ACGTTGGATGACCAGATATTGTGATGGGTG	SugSNP_Sh_014_W4_R	ACGTTGGATGGGTCTCCCTGTCAATTTCAC	SugSNP_Sh_014_W4_E	GCATGATTCTTCTGATGTTC
SugSNP_sh018	SugSNP_Sh_018_W4_F	ACGTTGGATGGAGGTTCCTCAGTTGGTTTG	SugSNP_Sh_018_W4_R	ACGTTGGATGCACAGAAATATTGCTCCTCC	SugSNP_Sh_018_W4_E	TGCATGCCTCGTTTTCTTTGG
SugSNP_sh102	SugSNP_Sh_102_W4_F	ACGTTGGATGTCTGACTCGGAAGATGATAG	SugSNP_Sh_102_W4_R	ACGTTGGATGTCTTTGTCAGACACGATAGG	SugSNP_Sh_102_W4_E	TGTCAGACACGATAGGTTTGTC
SugSNP_sh006	SugSNP_Sh_006_W4_F	ACGTTGGATGAGCAAACCGATGCCTGTTG	SugSNP_Sh_006_W4_R	ACGTTGGATGAAAGTATTCAACTGGATCCG	SugSNP_Sh_006_W4_E	ATTCAACTGGATCCGAAATATTC
SugSNP_sh063	SugSNP_Sh_063_W5_F	ACGTTGGATGACAAGGAGGGGAAGCGTAAG	SugSNP_Sh_063_W5_R	ACGTTGGATGAGAGCTAGTTCAACAGTACC	SugSNP_Sh_063_W5_E	AGCGTAAGTCGGGGCCG
SugSNP_sh089	SugSNP_Sh_089_W5_F	ACGTTGGATGCTGCGATATCACATCTGCTC	SugSNP_Sh_089_W5_R	ACGTTGGATGCACGGAAACAAGGTTTAAAG	SugSNP_Sh_089_W5_E	ACAAGGTTTAAAGTATTGGG
SugSNP_sh011	SugSNP_Sh_011_W5_F	ACGTTGGATGTTCATTAGCTGAGAAGGATG	SugSNP_Sh_011_W5_R	ACGTTGGATGCTTCCTCTTCAGATTCATCC	SugSNP_Sh_011_W5_E	AGGTCAGTAAATAAGTCAAATC
SugSNP_sh080	SugSNP_Sh_080_W5_F	ACGTTGGATGTCATTCAGAGCAAGCTGTGG	SugSNP_Sh_080_W5_R	ACGTTGGATGGGACCTTTACATTTATTATC	SugSNP_Sh_080_W5_E	ACATTTATTATCTTTGTCAGACA

RESULTADOS COMPLEMENTARES

O grupo trabalha em parceria com o Prof. Dr. Antonio Augusto Franco Garcia da Esalq/USP (Escola Superior de Agricultura "Luiz de Queiroz" / Universidade de São Paulo) desenvolvendo mapas genéticos para diferentes espécies de plantas cultivadas.

A partir de 2011 o grupo começou a utilizar SNPs no mapeamento de cana-de-açúcar, gerando o trabalho intitulado "*SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids*" de Garcia e colaboradores (2013). Com esse trabalho, abriu-se a possibilidade de fazer a dosagem alélica e a ploidia de cada loco SNP, o que resultou em um trabalho importante para o grupo, publicado no periódico Scientific Report. Esse trabalho discute o comportamento dos grupos de SNPs, utilizando como exemplo três deles (SugSNP_151, SugSNP_382 e SugSNP_715), cada um desenhado a partir de um gene. Com o objetivo de elucidar o comportamento desses três genes/SNPs, os BACs para cada um foi selecionado. Esses SNPs foram escolhidos para servirem de modelo no artigo e o screening para os três foi feito para que seus clones BACs fossem encontrados e validados. Os BACs encontrados serão utilizados em estudos futuros.

Para o *screening* desses locos, foram feitas hibridizações em todas as membranas (os produtos amplificados foram utilizados como sonda) das duas variedades em colaboração com o CNRGV (Figura 16) e utilizado o Pool 3D (metade da biblioteca da variedade SP 80-3280) utilizando qPCR (PCR quantitativo em tempo real), em nosso laboratório (Figura 17).

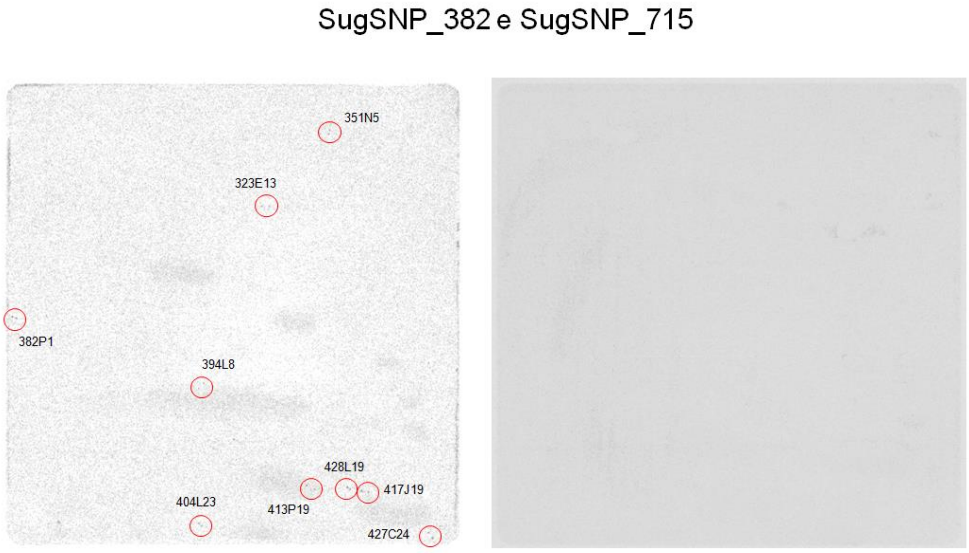


Figura 16. Macroarranjos hibridizados. Esquerda: hibridização do produto do loco SugSNP_382. Direita: hibridização que não funcionou para o loco SugSNP_715. Ambos na variedade SP 80-3280.

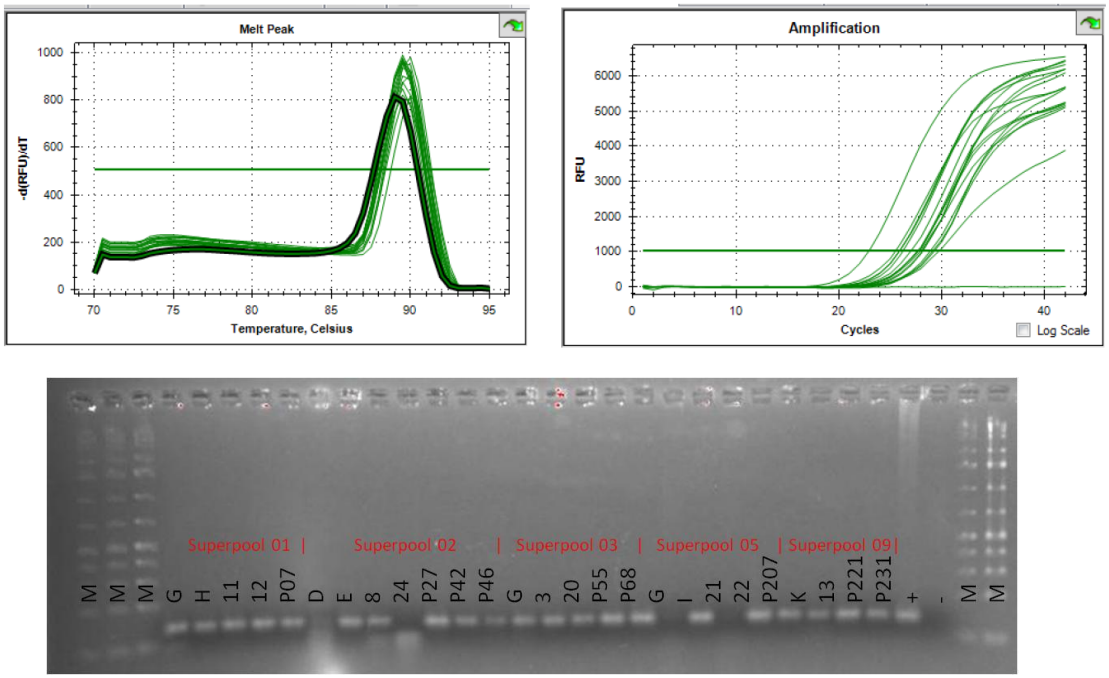


Figura 17. Acima: qPCR mostrando a amplificação (direita) e a curva de *melting* (Esquerda) para o Superpool 01 do loco SugSNP_715 – Temperatura de *melting* 89ºC para controle positivo (DNA genômico da SP 80-3280) e 89,5ºC ou 90ºC para os pools. Abaixo: Mesmo com os picos, os amplificados são colocados em agarose para confirmação do tamanho. Últimas duas pistas são o controle positivo (DNA genômico) e negativo. Marcador 1kb Plus (Invitrogen).

Os clones obtidos através das duas técnicas de seleção de clones BACs foram validados conforme mostra a Tabela 05. As membranas que falharam não foram repetidas (Figura 09 – Direita), devido a limitações orçamentarias. Os clones obtidos poderão ser usados para estudos futuros utilizando esses três SNPs e representam ferramentas fundamentais para o grupo como um todo.

Tabela 07: Total de clones positivos para os 3 SNPs. Alguns dos clones selecionados por Pool 3D se repetiram por Macroarranjos, por isso o total de clones positivos não é igual à soma dos positivos por Pool 3D e Membranas.

Marcador	SP 80-3280			IAC SP 93-3046
	Pool 3D	Membranas	Total	Membranas
SugSNP_151	20	3	23	-0-
SugSNP_382	16	17	26	8
SugSNP_715	26	5	30	16

RESUMO DOS RESULTADOS

Os resultados apresentados pela presente tese alcançaram os objetivos propostos e encontram-se resumidos:

Capítulo I

- Foi possível desenvolver duas bibliotecas de BACs, uma para variedade SP80-3280 e outra para a variedade IACSP93-3046.
- Duas ferramentas de seleção de clones BAC foram construídas: macroarranjos para as variedades SP803280 e IACSP93-3046 e Pool 3D para metade dos clones da variedade SP803280.
- Foi possível implantar novas técnicas para acessar o genoma de cana-de-açúcar, através das bibliotecas de BACs e desenvolver ferramentas para buscar regiões específicas do genoma por meio dos Macroarranjos e Pool3D.
- A técnica de BAC-end foi utilizada para caracterizar as duas bibliotecas de BACs desenvolvidas.
- Todos os clones obtidos por macroarranjos foram validados via PCR e foi possível encontrar 198 clones positivos para nove genes de possível cópia única.
- O gene *HP600*, representado pelo marcador SC-08, foi escolhido para investigações sobre os aspectos genéticos e genômicos de uma região do genoma da cana-de-açúcar. Esse gene encontra-se em um QTL para Brix descrito em sorgo.
- Foram sequenciados 23 BACs que continham o gene *HP600*.
- Foi possível desenvolver um método de anotação de BACs que está continuamente sendo utilizado para outros projetos e permanentemente atualizado.

Capítulo II

- A integração dos dados genéticos, genômicos e transcriptômicos foi utilizada para explicar a interação das duas regiões na cana-de-açúcar.
- *HP600* é um gene hipotético que está ao lado do gene *CENP-C*, um componente responsável pelo início dos nucleossomos. Os haplótipos de genes de cana-de-açúcar de *HP600* na *Region01* e os haplótipos

CENP-C na *Region01* foram duplicados em outro grupo de cromossomos homeólogos.

- A duplicação dos haplótipos *HP600* na *Region01* resultou em um pseudogene parálogo nos haplótipos *HP600* na *Region02*. A duplicação do *CENP-C* nos haplótipos da *Region02* resultou em fusão com outro gene, que continha os primeiros cinco exons do gene ortólogo Sobic.003G299500 e os exons oito a quatorze do *CENP-C*. A região onde esta duplicação foi inserida (*Region02*) continha pelo menos mais três genes que provavelmente surgiram devido à duplicação, o que indica que ocorreram múltiplos eventos de duplicação nesta região.
- A duplicação *HP600* e *CENP-C* descrita neste trabalho ocorreu após a separação da cana-de-açúcar e sorgo e antes da poliploidização do gênero *Saccharum*. Este resultado é suportado pelas seguintes informações: (I) o tempo do relógio molecular, (II) os genes estão presentes em um grupo homólogo de cromossomos; e (III) os motivos *CENP*-k dos haplótipos *CENP-C* na *Region02* são mais semelhantes ao sorgo do que ao seu parálogo na cana-de-açúcar.
- A formação de um gene quimérico e o embaralhamento de genes observados na *Region02* exibiram um momento específico de formação antes da poliploidização do gênero *Saccharum*, o que nos faz pensar qual evento genômico poderia ser o resultado dessa formação: Os TEs que transportam esta região não foram encontrados. Também é possível que TEs tenham sido inseridos nesta região, e as sequências TE foram subsequentemente perdidas. Um evento que resultou em alguma instabilidade genômica também poderia ser um motivo, dentre outros.
- As transcrições do SP80-3280 revelaram a expressão completa dos haplótipos de *HP600* na *Region01* (com expressão diferencial dos haplótipos) e a falta de expressão dos haplótipos *HP600* na *Region02*. A expressão dos haplótipos *HP600* na *Region01* pode ser considerada um gene de cópia única, apesar da presença da duplicação.
- O gene *CENP-C* pode ser considerado totalmente expresso, apesar da baixa cobertura dos dados do transcriptoma. Os haplótipos do *CENP-C* na *Region02* possuem quatro haplótipos considerados expressos.

- Atualmente, apenas marcadores de dose única podem ser usados para construir o mapa genético em cana-de-açúcar, o que é uma limitação do método de mapeamento em poliploides. Tentamos mapear uma região duplicada, o que é uma tarefa difícil para organismos diploides. Novamente, é importante observar que usamos uma variedade de cana-de-açúcar com reprodução assexuada e realizamos o mapeamento genético em progênies artificiais. Não temos ideia de como o genoma da progênie respondeu ao cruzamento, já que a cana é aneuploide.
- O mapeamento genético demonstra que existem obstáculos que ainda precisam ser superados no mapeamento genético de poliploides complexos.
- Foi possível observar a relação entre um mapa de ligação e o mapa físico de uma região em cana-de-açúcar. De fato, é uma região pequena para observar, enquanto a cana-de-açúcar tem um genoma grande, e um mapa de ligação é construído com base na fração de recombinação. No entanto, foi possível observar o que acontece no mapa genético quando um lócus duplicado foi mapeado.

CONCLUSÕES GERAIS

Duas bibliotecas de BACs foram construídas, uma para a variedade SP80-3280 e outra para a variedade SPIAC93-3046, tais variedades são de extrema importância para a cultura nacional de cana-de-açúcar e as bibliotecas obtidas são as duas maiores já construídas. O tamanho das bibliotecas facilitará o entendimento do genoma extremamente complexo da cana-de-açúcar, aumentando a chance de encontrar genes raros e/ou alelos raros.

O desenvolvimento de ferramentas que tornem o *screening* rápido e eficiente é tão importante quanto às bibliotecas. Duas ferramentas foram construídas para a seleção de clones. A plataforma de Pool 3D mostrou-se mais promissora, uma vez que não se utiliza de radioatividade e pode facilmente ser realizada apenas com PCRs. Esta ferramenta foi implantada no laboratório e está sendo utilizada para outros projetos.

A seleção de regiões genômicas contendo genes de interesse se mostrou eficaz, sendo possível recuperar diversas regiões a partir de diferentes genes ou locos. Foi possível recuperar também haplótipos de cada região de interesse, permitindo que estudos diversos possam ser realizados, estabelecendo fisicamente a arquitetura genômica de genes.

O sequenciamento de pontas de BACs, o sequenciamento de BACs completos e as anotações desses BACs são um grande avanço para a pesquisa brasileira. Com elas, foi possível adicionar a pesquisa de cana-de-açúcar regiões genômicas que ajudam a elucidar comportamentos genéticos, genômicos e de expressão gênica.

A integração dos dados genéticos, genômicos e transcriptômicos foi utilizada para explicar a interação das duas regiões na cana-de-açúcar, permitindo um estudo mais profundo da relação genoma x transcrição x expressão de genes. Para tanto, um gene hipotético, de provável cópia única, ligado a um QTL para brix em sorgo, chamado de *HP600*, representado por um marcador genético SC-08.

O gene *HP600* está ao lado do gene *CENP-C*, um gene responsável pela a iniciação dos nucleossomos. As análises dos BACs resultaram na descoberta de duas regiões que continham esse gene: uma com ambos os genes completos e outra com ambos os genes truncados. As análises de expressão de cada haplótipo demonstraram que o gene *HP600* estava sendo expresso, apenas na Região01, provavelmente por isso foi considerado de cópia única nas análises entre os

transcritos de sorgo e arroz. O gene *HP600*, expresso apenas na Região01, apresentaram evidências de expressão diferencial, onde um haplótipo do gene é mais expresso que outro.

O gene *CENP-C* também teve seus haplótipos da região01. A duplicação do *CENP-C* nos haplótipos da Região02 resultou em fusão com outro gene, formado pelos primeiros cinco exons do gene ortólogo Sobic.003G299500 e os exons oito a quatorze do *CENP-C*. A região onde esta duplicação foi inserida (Região02) continha pelo menos mais três genes que provavelmente surgiram devido à duplicação, o que indica que ocorreram múltiplos eventos de duplicação nesta região. A Região01 é uma região sintênica a de sorgo, enquanto a Região02 se mostrou uma região de recombinações e duplicações não sintênica a sorgo.

A formação de um gene quimérico e o embaralhamento de genes observados na Region02 exibiram um momento específico de formação antes da poliploidização do gênero *Saccharum*, o que nos faz pensar qual evento genômico poderia ser o resultado dessa formação. É possível que os genes *HP600* e *CENP-C* tenham sido inseridos através de transposons que foram subsequentemente perdidos. Um evento que resultou em alguma instabilidade genômica também poderia ser um motivo, dentre outros.

Os resultados de BAC-FISH mostraram diferentes ploidias para ambas as regiões e nos levou a concluir que não foram recuperados todos os haplótipos/alelos para a Região01 e Região02. Porém, definiram-se os haplótipos quando detectado diferença significativa entre os BACs. Essa diferença pode ser a presença de sequências repetitivas específicas de cada BAC ou SNPs entre os genes *HP600/CENP-C*. Ainda é necessário considerar a possibilidade de uma haplótipo/alelo ser idêntico ao outro. Apesar dessa possibilidade não poder ser provada com os dados gerados, essa possibilidade deve ser levantada.

Foi possível observar a relação entre um mapa de ligação e o mapa físico de uma região em cana-de-açúcar. De fato, é uma região pequena para observar, enquanto a cana-de-açúcar tem um genoma grande, e um mapa de ligação é construído com base na fração de recombinação. No entanto, foi possível observar o que acontece no mapa genético quando um loco duplicado foi mapeado. Por fim, o projeto como um todo traz não só o domínio de técnicas pouco usadas no Brasil, mas também contribui com ferramentas novas para o estudo do genoma da cana-

de-açúcar e contribui para avanços no estudo de espécies poliploides como um todo.

Este estudo lança luz sobre a influência do arranjo genômico para análises de transcriptoma e mapa genético no genoma poliploide de cana-de-açúcar. A integração de arranjos de sequências genômicas, perfis de transcrição, organização citogenética e abordagem de mapeamento genético pode ajudar a elucidar o comportamento da expressão gênica, a estrutura genética e a montagem bem-sucedida da sequência do genoma da cana-de-açúcar. Tais estudos integrados, sem dúvida, ajudarão a melhorar nossa compreensão de genomas poliploides complexos, incluindo o genoma da cana-de-açúcar.

Ênfase especial deve ser dada aos estudos de determinação do nível de ploidia e dos locais de duplicação com a intenção de melhor compreender os poliploides complexos. Tais estudos continuam sendo os mais originais e desafiadores. Nessa perspectiva, este trabalho apresentou uma abordagem integrada para elucidar a dinâmica alélica em genomas poliploides, tendo a cana-de-açúcar como exemplo.

PERSPECTIVAS

O domínio da tecnologia de bibliotecas de BACs e ferramentas de seleção de clones por um grupo brasileiro abre caminho para o desenvolvimento dos mesmos recursos genômicos para o estudo de outros organismos poliploides com genomas complexos e de interesse para a agricultura brasileira. É o caso de algumas forrageiras (*Brachiaria*, capim colonião e *Paspalum*, entre outras).

As possibilidades de estudos abertas pela disponibilização destes recursos genômicos são incontáveis. Muitos grupos com interesses em evolução, genética, genômica, melhoramento, citogenética e outras áreas, que trabalham no Brasil e no exterior com cana-de-açúcar ou outras gramíneas, poderão rapidamente ter acesso às regiões genômicas de seu interesse. O maior beneficiado com esses conhecimentos será o melhoramento genético da cana-de-açúcar, que poderá contar com informações fundamentais para a obtenção de variedades cada vez mais adaptadas às exigências de uma agricultura produtiva e sustentável.

As aplicações das ferramentas desenvolvidas nesse trabalho contribuíram para iniciar outros projetos explorando a busca de genes de interesse para a agricultura, conforme Anexo I. Outros trabalhos, buscando outras regiões de interesse estão em desenvolvimento, como a busca de região genômica para outros QTLs de *Brix*. Ainda será possível relacionar esses genes com os transcriptomas já publicados e estudar o comportamento da expressão específica de cada haplótipo dos genes alvo.

A expressão de genes em cana-de-açúcar pode ser melhor compreendida, conforme a demonstração do gene *HP600*, que possuí evidencias de expressão diferencial e do gene *CENP-C*, responsável pela formação de um gene quimérico exclusivo de cana-de-açúcar. Ainda, Garsmeur e colaboradores (2018) relataram ausência de colinearidade em 17% dos genes observados, fato também observado na Região02 descrita no Capitulo II. Entender essas regiões podem significar genes novos ou genes com novas funções importantes para o melhoramento da cana-de-açúcar e para a compreensão de genes alvo que representem características.

A técnica de BAC-FISH estabelecida também será utilizada para trabalhos relacionados à ploidia e dosagens de marcadores moleculares, sendo uma ferramenta muito útil para a determinação da ploidia de locos, com informações da localização física desses marcadores. Esta perspectiva poderá ajudar na solução de

problemas de ploidias entre locos, auxiliando na melhoria de mapas genéticos existentes para cana-de-açúcar.

Por fim, espera-se que as novas ferramentas implantadas no laboratório sejam construídas para outros organismos, como aconteceu em uma linhagem de fungo no trabalho de Crucello e colaboradores (2015). Trabalhos como este podem representar grandes avanços para o organismo, sem a necessidade de se montar um genoma completo. Para estudos envolvendo plantas sem genoma, BACs pode representar uma maneira mais rápida de acessar o genoma do que o sequenciamento do genoma completo.

REFERÊNCIAS BIBLIOGRÁFICAS

Aitken K.S., Jackson P.A., McIntyre C.L. (2005). A combination of AFLP and SSR markers provides extensive map coverage and identification of homo(eo)logous linkage groups in a sugarcane cultivar. Theor Appl Genet 110: 789-801.

Aitken KS, Jackson PA, McIntyre CL (2006) Quantitative trait loci identified for sugar related traits in a sugarcane (Saccharum spp.) cultivar x Saccharum officinarum population. Theor Appl Genet 112:1306–1317

Aitken KS, Jackson PA, Mcintyre CL (2007) Construction of a genetic linkage map for Saccharum officinarum incorporating both simplex and duplex markers to increase genome coverage. Genome 50:742-756

Aitken KS, Hermann S, Karno K, Bonnett GD, McIntyre LC, Jackson PA (2008) Genetic control of yield related stalk traits in sugarcane. Theor Appl Genet 117:1191-1203.

Al-Janabi SM, Honeycutt RJ, Mcclelland M, Sobral BWS (1993) A genetic linkage map of *Saccharum spontaneum* L, 'SES 208'. Genetics 134:1249-1260.

Al-Janabi S.M., McClelland M., Petersen C., Sobral B.W.S. (1994). Phylogenetic analysis of organellar DNA sequences in the Andropogoneae: Saccharinae. Theor Appl Genet 88: 933-944.

Al-Janabi SM, Parmessur Y, Kross H, Dhayan S, Saumtally S, Ramdoyal K, Autrey LJC, Dookun-Saumtally A (2007) Identification of a major quantitative trait locus (QTL) for yellow spot (Mycovellosiella koepkei) disease resistance in sugarcane. Mol Breed 19:1-14

Alix K., Baurens F.C., Paulet F., Glaszmann J.C., D'Hont A. (1998). Isolation and characterisation of a satellite DNA family in Saccharum complex. Genome 41(6): 854-864.

Alix K., Paulet F., Glaszmann J.C., D'Hont A. (1999). Inter-Alu like species-specific sequences in the Saccharum complex. Theor Appl Genet 6: 962-968.

Ammiraju, J.S., Luo, M., Goicoechea, J.L., Wang, W., Kudrna, D., Mueller, C., Talag, J., Kim, H., Sisneros, N.B., Blackmon, B., Fang, E., Tomkins, J.B., Brar, D., Mackill, D., Mccouch, S., Kurata, N., Lambert, G., Galbraith, D.W., Arumuganathan, K., Rao, K., Walling, J.G., Gill, N., Yu, Y., Sanmiguel, P., Soderlund, C., Jackson, S., Wing, R.A. (2006). The Oryza bacterial artificial chromosome library resource: construction and analysis of 12 deep-coverage large-insert BAC libraries that

represent the 10 genome types of the genus Oryza. Genome Research, 16(1), 140-147.

Anchordoquy, T.J., Molina, M.C., 2007. Preservation of DNA. Cell Preserv. Technol. 5, 180–188.

Anonymous. 1945. A newly released cane: some notes on NCo310. S. Afr. Sugar J. 30:91.

Asnaghi C., Paulet F., Kaye C., Grivet L., Deu M., Glaszmann J.C., D'Hont A. (2000). Application of synteny across Poaceae to determine the map location of a sugarcane rust resistance gene. Theor Appl Genet 101: 962-969.

Ball AD, Stapley J, Dawson DA, Birkhead TR, Burke T, Slate J (2010) A comparison of SNPs and microsatellites as linkage mapping markers: lessons from the zebra finch (*Taeniopygia guttata*). BMC genomics 11(1):218

Balsalobre, T.W.A., da Silva Pereira, G., Margarido, G.R.A., Gazaffi, R., Barreto, F.Z., Anoni, C.O., Cardoso-Silva, C.B., Costa, E.A., Mancini, M.C., Hoffmann, H.P., de Souza, A.P., Garcia, A.A., and Carneiro, M.S. (2017). GBS-based single dosage markers for linkage and QTL mapping allow gene mining for yield-related traits in sugarcane. BMC Genomics 18: 72.

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

Barbosa, M. H. P., de Silveira, L. C. I., de Oliveira, M. W., de Souza, V. D. F. M., & Ribeiro, S. N. N. (2001). RB867515 Sugarcane cultivar. Crop Breeding and Applied Biotechnology, 1(4).

Besse P., McIntyre C.L., Berding N. (1996). Ribosomal DNA variations in Erianthus, a wild sugarcane relative (Andropogoneae x Saccharinae). Theor Appl Genet 92: 733-743.

Beissinger TM, Hirsch CN, Sekhon RS, Foerster JM, Johnson JM, Muttoni G, et al. (2013) Marker density and read depth for genotyping populations using genotyping-bysequencing. Genetics. 193:1073-81.

Borém A, Santos FR (2004) Biotecnologia simplificada. 2.ed., Viçosa, MG, pp 302.

Bouzidi, M. F., Franchel, J., Tao, Q., Stormo, K., Mraz, A., Nicolas, P., & Mouzeyar, S. (2006). A sunflower BAC library suitable for PCR screening and

physical mapping of targeted genomic regions. Theoretical and Applied Genetics, 113(1), 81-89.

Brandes E.W., Sartoris G.B., Grassl C.O. (1939). Assembling and evaluating wild forms of sugarcane and closely related plants. Proc. Int. Soc. Sugarcane Technol. 6: 128-154.

Brandes E.W. (1956). Origin, dispersal and use in breeding of the Melanesian garden sugarcanes and their derivatives Saccharum officinarum L. Proc. Cong. Int. Soc. Sug. Tecnhol. 9(1): 709-750.

Bremer G. (1924). A cytological investigation of some species and species hybrids within the genus Saccharum. Genetica 5: 97-148; 273-326.

Bremer G. (1930). The cytology of Saccharum. Proc Int Soc Sugar Cane Technol 3: 408-415.

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

Budiman, M. A., Mao, L., Wood, T. C., & Wing, R. A. (2000). A deep-coverage tomato BAC library and prospects toward development of an STC framework for genome sequencing. Genome research, 10(1), 129-136.

Bundock, P. C., & Henry, R. J. (2004). Single nucleotide polymorphism, haplotype diversity and recombination in the Isa gene of barley. Theoretical and Applied Genetics, 109(3), 543-551.

Bundock, P.C., Eliott, F.G., Ablett, G., Benson, A.D., Casu, R.E., Aitken, K.S. and Henry, R.J. (2009) Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploid plant species using 454 sequencing. Plant Biotechnol. J. 7, 347–354.

Burke, D. T., Carle, G. F., Olson, M. V. (1987). Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. Science, 236(4803), 806-812.

Burner D.M. (1987). Cytogenetic analyses of sugarcane relatives (Andropogoneae: Saccharinae). Euphytica 54: 125-133.

Burnquist W.L., Sorrells M.E., Tanksley S. (1992). Characterisation of genetic variability in Saccharum germplasm by means of restriction fragment length polymorphism (RFLP) analysis. Proc. Int. Soc. Sugar Cane Technol 21(2): 355-365.

Byrne S, Czaban A, Studer B, Panitz F, Bendixen C, Asp T. (2013) Genome wide allele frequency fingerprints (GWAFFs) of populations via genotyping by sequencing. PloS One. 8(3).

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

Carneiro M.S., Vieira M.L.C. (2002). Mapas genéticos em plantas. Bragantia 61(2): 89-100.

Chang, Y. L., Henriquez, X., Preuss, D., Copenhaver, G., Zhang, H. B. (2003). A plant-transformation-competent BIBAC library from the Arabidopsis thaliana Landsberg ecotype for functional and comparative genomics. Theoretical and Applied Genetics, 106(2), 269-276.

Chang, Y. L., Tao, Q., Scheuring, C., Ding, K., Meksem, K., Zhang, H. B. (2001). An integrated map of Arabidopsis thaliana for functional analysis of its genome sequence. Genetics, 159(3), 1231-1242.

Chang, Y. L., Chuang, H. W., Meksem, K., Wu, F. C., Chang, C. Y., Zhang, M., Zhang, H. B. (2011). Characterization of a plant-transformation-ready large-insert BIBAC library of Arabidopsis and bombardment transformation of a large-insert BIBAC of the library into tobacco. Genome, 54(6), 437-447.

Chen AH, Lipka AE. (2016) The Use of Targeted Marker Subsets to Account for Population Structure and Relatedness in Genome-Wide Association Studies of Maize (Zea mays L.). G3.

Cho, R. J., Mindrinos, M., Richards, D. R., Sapolsky, R. J., Anderson, M., Drenkard, E., ... & Theologis, A. (1999). Genome-wide mapping with biallelic markers in Arabidopsis thaliana. Nature genetics, 23(2), 203.

CONAB (2018). Companhia Nacional de Abastecimento. Acompanhamento da safra brasileira de cana-de-açúcar. v. 5 - Safra 2018/19, n.1 - Primeiro levantamento, mai. de 2018.

Cordeiro, G. M., Eliott, F., McIntyre, C. L., Casu, R. E., & Henry, R. J. (2006). Characterisation of single nucleotide polymorphisms in sugarcane ESTs. Theoretical and Applied Genetics, 113(2), 331-343.

Cordeiro G.M., Taylor G.O., Henry R.J. (2000). Characterization of microsatellite markers from sugarcane (Saccharum spp.), a highly polyploid species. Plant Science 155: 161-168.

Coyne, C.J., Mcclendon, M.T., Wlling, J.G., Timmeran-Vaughan, G.M., Murray, S., Meksem, K., Lightfoot, D.A., Shultz, J.L., Keller, K.E., Martin, R.R., Inglis, D.A., Rajesh, P.N., Mcphee, K.E., Weeden, N.F., Grusak, M.A., Li, C.M., Storlie, E.W. (2007). Construction and characterization of two bacterial artificial chromosome libraries of pea (Pisum sativum L.) for the isolation of economically important genes. Genome, 50(9), 871-875.

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

Cronquist, A. (1981). An integrated system of classification of flowering plants. Columbia University Press.

Crossa J, Beyene Y, Kassa S, Pérez P, Hickey JM, Chen C, Babu R. (2013) Genomic prediction in maize breeding populations with genotyping-by-sequencing. G3. 3(11):1903–26.

Crucello, A., Sforça, D. A., Horta, M. A. C., dos Santos, C. A., Viana, A. J. C., Beloti, L. L., ... & de Souza, A. P. (2015). Analysis of genomic regions of Trichoderma harzianum IOC-3844 related to biomass degradation. PloS one, 10(4), e0122122.

Daugrois JH, Grivet L, Roques D, Hoarau JY, Lombard H, Glaszmann JC, D'Hont A (1996) A putative major gene for rust resistance linked with a RFLP marker in sugarcane cultivar 'R570'. Theor Appl Genet 92:1059-1064

Daniels J., Roach B.T. (1987). Taxonomy and evolution. In: Heinz D.J. (ed). Sugarcane improvement through breeding. Elsevier Press, Amsterdam, pp 7-84.

Daugrois JH, Grivet L, Roques D, Hoarau JY, Lombard H, Glaszmann JC, D'Hont A (1996) A putative major gene for rust resistance linked with a RFLP marker in sugarcane cultivar 'R570'. Theor Appl Genet 92:1059-1064.

De Setta N, Monteiro-Vitorello C, Metcalfe C, Cruz GM, Del Bem L, Vicentini R, et al. (2014) Building the sugarcane genome for biotechnology and identifying evolutionary trends. BMC Genomics.15:540.

D'Hont A., Lu Y.H., Feldmann P., Glaszmann J.C. (1993). Cytoplasmic diversity in sugarcane revealed by heterologous probes. Sugar Cane 1: 12-15.

D'Hont A., Rao S., Feldman P., Grivet P., Islam-Faridi L., Berding N., Glaszmann J.C. (1995). Identification and characterization of intergeneric hybrids, *Saccharum officinarum* x *Erianthus arundinaceus*, with molecular markers and in situ hybridization. Theor Appl Genet 91: 320-326.

D'Hont, A., Grivet, L., Feldmann, P., Glaszmann, J. C., Rao, S., & Berding, N. (1996). Characterisation of the double genome structure of modern sugarcane cultivars (Saccharum spp.) by molecular cytogenetics. Molecular and General Genetics MGG, 250(4), 405-413.

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

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

Donato M, Peters SO, Mitchell SE, Hussain T, Imumorin IG. (2013) Genotyping-by-sequencing (GBS): a novel, efficient and cost-effective genotyping method for cattle using nextgeneration sequencing. PloS One. 8(5), e62137.

Dong, F., Song, J., Naess, S. K., Helgeson, J. P., Gebhardt, C., & Jiang, J. (2000). Development and applications of a set of chromosome-specific cytogenetic DNA markers in potato. Theoretical and Applied Genetics, 101(7), 1001-1007.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchellet SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE, vol. 6, no. 5, Article ID e19379.

Ethirajan, A.S. (1987). Sugarcane hybridization techniques. p. 129–148. In: Anonymous (eds.), Copersucar International Sugarcane Breeding Workshop. Copersucar, Brazil.

Evans, H. 1935. Investigation on the root-system of sugarcane varieties. Mauritius Dep. Ag. Res. Sta. Bull. 6, 44pp.

FAO (2018). Food and Agricultural Organization of the United Nations. http://www.fao.org.

Febrer, M., Goicoechea, J. L., Wright, J., McKenzie, N., Song, X., Lin, J., Bevan, M. W. (2010). An integrated physical, genetic and cytogenetic map of Brachypodium distachyon, a model system for grass research. PLoS ONE, 5(10), e13461.

Feldmann P., D'Hont A., Guiderdoni E., Grivet L., Glaszmann C. (1997). La canne à sucre. In: L'amélioration des plantes tropicales. Cirad / Orstom, 1997.

Figueira, T. R. E. S., Okura, V., Rodrigues Da Silva, F., Jose Da Silva, M., Kudrna, D., Ammiraju, J. S. S., Taag, J., Wing, R., Arruda, P. (2012). A BAC library of the SP80-3280 sugarcane variety (saccharum sp.) and its inferred microsynteny with the sorghum genome. BMC Research Notes, 5(1), 185.

Fonsêca, A., Ferreira, J., Dos Santos, T. R. B., Mosiolek, M., Bellucci, E., Kami, J., Gepts, P., Geffroy, V., Schweizer, D., Santos, K.G.B. dos, Pedrosa-Harand, A. (2010). Cytogenetic map of common bean (Phaseolus vulgaris L.). Chromosome Research, 18(4), 487–502.

Francia E, Tacconi G, Crosatti C, Barabaschi D, Bulgarelli D,Dall'Aglio E, Val G (2005) Marker assisted selecion in cropplants. Plant Cell Tissue Organ Cult 82:317–342

Garcia A.A.F., Kido E.A., Meza A.N., Souza H.M.B., Pinto L.R., Pastina M.M., Leite C.S., da Silva J.A.G., Ulian E.C., Figueira A., Souza A.P. (2006). Development of an integrated genetic map of a sugarcane (Saccharum spp.) commercial cross, based on a maximum-likelihood approach for estimation of linkage and linkage phases. Theor Appl Genet 112: 298-314.

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

Garsmeur, O., Charron, C., Bocs, S., Jouffe, V., Samain, S., Couloux, A., Droc, G. et al. (2011) High homologous gene conservation despite extreme autopolyploid redundancy in sugarcane. New Phytol. 189, 629–642.

Garsmeur, O., Droc, G., Antonise, R., Grimwood, J., Potier, B., Aitken, K., Jenkins, J., Martin, G., Charron, C., Hervouet, C., Costet, L., Yahiaoui, N., Healey, A., Sims, D., Cherukuri, Y., Sreedasyam, A., Kilian, A., Chan, A., Van Sluys, M. A., Swaminathan, K., Town, C., Bergès, H., Simmons, B., Glaszmann, J. C., van der Vossen, E., Henry, R., Schmutz, J., D'Hont, A. (2018). A mosaic monoploid

reference sequence for the highly complex genome of sugarcane. Nature communications, 9(1), 2638.

Gazaffi R, Oliveira KM, Souza AP, Garcia AAF (2010) Melhoramento Genético e Mapeamento da Cana-de-açúcar. In: Cortez LAB (Ed.) Bioetanol de cana-de-açúcar: P&D para produtividade e sustentabilidade. 1. ed. [S.l.]: Edgar Blücher Ltda 333-343

Giancola, S., Brunel, D., Colot, V., Prum, B., Quesneville, H., Mézard, C., ... & Bérard, A. (2006). Variation in crossing-over rates across chromosome 4 of.

Glaszmann J.C., Lu Y.H., Lanaud C. (1990). Variation of nuclear ribosomal DNA in sugarcane. J. Genet Breed 44:191-198.

Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q et al. (2014) TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS One. 9: e90346.

Goff, S. A., Ricke, D., Lan, T. H., Presting, G., Wang, R., Dunn, M., ... & Hadley, D. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science, 296(5565), 92-100.

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

Goodwin, S., McPherson, J.D., McCombie, W.R., 2016. Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17, 333–351.

Götz, S., García-Gómez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., ... & Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic acids research, 36(10), 3420-3435.

Guimarães C.T., Sills G., Sobral B. (1997). Comparative mapping of Andropogoneae: Saccharum L. (sugarcane) and its relation to sorgum and maize. Proc. Natl. Acad. Sci. 94: 14261-14266.

Grattapaglia D, Sederoff R (1994) Genetic Linkage Maps of Eucalyptus grandis and Eucalyptus urophylla Using a Pseudo-Testcross: Mapping Strategy and RAPD Markers. Genetics 1137:1121-1137.

Grivet L., Daniels, C., Glaszmann J.C., D'Hont A. (2004). A review of recent molecular genetics evidence for sugarcane evolution and domestication. Ethnobotany Research and Applications 2: 9-17.

Ha, S., Moore, P. H., Heinz, D., Kato, S., Ohmido, N., Fukui, K. (1999). Quantitative chromosome map of the polyploid Saccharum spontaneum by multicolor fluorescence in situ hybridization and imaging methods. Plant molecular biology, 39(6), 1165-1173.

Hamilton, C. M., Frary, A., Lewis, C., Tanksley, S. D. (1996). Stable transfer of intact high molecular weight DNA into plant chromosomes. Proceedings of the National Academy of Sciences, 93(18), 9975-9979.

Hamilton, C. M., Frary, A., Xu, Y., Tanksley, S. D., Zhang, H. B. (1999). Construction of tomato genomic DNA libraries in a binary BAC (BIBAC) vector. The plant journal, 18(2), 223-229.

Han, Y., Zheng, D., Vimolmangkang, S., Khan, M. A., Beever, J. E., Korban, S. S. (2011). Integration of physical and genetic maps in apple confirms whole-genome and segmental duplications in the apple genome. Journal of Experimental Botany, 62(14), 5117–5130.

Harvey M., D'Hont A., Alix K., Huckett B. (1998). Use PCR-based markers for identification of Erianthus genetic material in putative intergeneric hybrids (Saccharum x Erianthus). Proc. S. Afr. Sug. Technol. Ass. 72: 218-320.

Heffelfinger C, Fragoso CA, Moreno MA, Overton JD, Mottinger JP, Zhao H, et al. (2014) Flexible and scalable genotyping-by-sequencing strategies for population studies. BMC Genomics.15:979.

Heinz DJ, Tew TL (1987) Hybridization procedures. In: Heinz DJ (eds) Sugarcane Improvement through Breeding, Elsevier, Amsterdam p 313-342

Henry, R. J. (2008). Applications of the sequenom platform to SNP Analysis in plants.

Heslot N, Rutkoski J, Poland J, Jannink J-L, Sorrells ME. (2013) Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. PLoS One, 8;

Hoarau J.Y., Grivet, L., Offmann B., Raboin L.M., Diorflar J.P., Payet J. Hellmann M., D'Hont A., Glaszmann J.C. (2002). Genetic dissection of a modern sugarcane cultivar (Saccharum ssp): II. Detection of QTLs for yield components. Theor Appl Genet (105): 1027-1037.

Hu, Y., Lu, Y., Ma, D., Guo, W., & Zhang, T. (2010). Construction and characterization of a bacterial artificial chromosome library for the A-genome of cotton (G. arboreum L.). BioMed Research International, 2011.

Jannoo N., Grivet L., Seguin M., Paulet F., Domaingue R., Rao P.S., Dookun A., D'Hont A., Glaszmann J.C. (1999). Molecular investigation of the genetic base of sugarcane cultivars. Theor Appl Genet 99: 171-184.

Jannoo N, Grivet L, Chantret N, Garsmeur O, Glaszmann JC, ArrudaP, D'Hont A (2007) Orthologous comparison in a gene-richregion among grasses reveals stability in the sugarcane polyploidgenome. Plant J 50:574–585.

Jeswiet, J. 1929. The development of selection and breeding of the sugarcane in Java. Proc. Int. Soc. Sugar Cane Technol. 3:44–57.

Jiang J., Gill B.S. (1996). Current status and potential of fluorescence in situ hybridization in plant genome mapping. In: PATERSON, A.H. (Ed.). Genome mapping in plants. Georgetown: RG Landes Company. p. 127-135.

Jiang Z, Wang H, Michal JJ, Zhou X, Liu B, Woods LC, et al. (2016) Genome wide sampling sequencing for SNP genotyping: methods, challenges and future development. Int J Biol Sci. 12:100-8.

Johnson, S. J., Wade-Martins, R. (2011). A BACwards glance at neurodegeneration: molecular insights into disease from LRRK2, SNCA and MAPT BAC-transgenic mice.

Hoarau J.Y., Grivet, L., Offmann B., Raboin L.M., Diorflar J.P., Payet J. Hellmann M., D'Hont A., Glaszmann J.C. (2002). Genetic dissection of a modern sugarcane cultivar (Saccharum ssp): II. Detection of QTLs for yield components. Theor Appl Genet (105): 1027-1037.

Kennedy, A.J., Rao, P.S. (2000). Handbook 2000. West Indies Central Sugar Cane Breeding Stn, Groves, St. George, Barbados. p. 1–10.

Kim, C., Wang, X., Lee, T. H., Jakob, K., Lee, G. J., & Paterson, A. H. (2014). Comparative analysis of Miscanthus and Saccharum reveals a shared whole-genome duplication but different evolutionary fates. The Plant Cell, tpc-114.

Kim, U. J., Birren, B. W., Slepak, T., Mancino, V., Boysen, C., Kang, H. L., ... & Shizuya, H. (1996). Construction and characterization of a human bacterial artificial chromosome library. Genomics, 34(2), 213-218.

Kriventseva, E. V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F. A., & Zdobnov, E. M. (2018). OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic acids research, 47(D1), D807-D811.

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

Landell, M. G. A.; Bressiani, J. A. 2008. Melhoramento genético, caracterização e manejo varietal. In: Dinardo-Miranda, L. L.; Vasconcelos, A. C. M.; Landell, M. G. A. (ed.). Cana-de-açúcar. Campinas: Instituto Agronômico. p.101-155.

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

Li H.W., Price S. (1967). Chromosome numbers of noble sugarcane clones. Proc Int Soc Sugar Cane Technol 12: 884-886.

Lima M.L.A., Garcia A.A.F., Oliveira K.M., Matsuoka S., Souza Jr C.L., Souza A.P. (2002). Analysis of genetics similarity detected by AFLP and coefficient of parentage among genotypes of sugarcane (Saccharum spp.). Theor Appl. Genet. 104: 30-38.

Lin M, Lou X, Chang M, Wu R (2003) A general statistical framework for mapping quantitative trait loci in nonmodel systems: issue for characterizing linkage phases. Genetics 165:901-913

Liu, Y. G., Shirano, Y., Fukaki, H., Yanai, Y., Tasaka, M., Tabata, S., Shibata, D. (1999). Complementation of plant mutants with large genomic DNA fragments by a transformation-competent artificial chromosome vector accelerates positional cloning. Proceedings of the National Academy of Sciences, 96(11), 6535-6540.

Liu H, Bayer M, Druka A, Russell JR, Hackett CA, Poland J, Waugh R. (2014) An evaluation of genotyping by sequencing (GBS) to map the Breviaristatum-e (ari-e) locus in cultivated barley. BMC Genomics. 15(1), 104.

Loannou A. P., Amemiya, C. T., Garnes, J., Kroisel, P. M., Shizuya, H., Chen, C., Batzer, M. A., de Jong, P. J. (1994). A new bacteriophage P1–derived vector for the propagation of large human DNA fragments. Nature genetics, 6(1), 84.

Lu Y.H., D'Hont A., Paulet F., Grivet L., Arnaud M., Glaszmann J.C. (1994). Molecular diversity and genome structure in modern sugarcane varieties. Euphytica 78: 217-226.

Lu F, Lipka AE, Glaubitz J, Elshire R, Cherney JH, Casler MD, Costich DE. (2013) Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. PLoS Genetics, 9(1), e1003215.

Lysak, M. A., Fransz, P. F., Ali, H. B., & Schubert, I. (2001). Chromosome painting in Arabidopsis thaliana. The Plant Journal, 28(6), 689-697.

Lysak, M. A., Berr, A., Pecinka, A., Schmidt, R., McBreen, K., & Schubert, I. (2006). Mechanisms of chromosome number reduction in Arabidopsis thaliana and related Brassicaceae species. Proceedings of the National Academy of Sciences, 103(13), 5224-5229.

Ma, L., Vu, G. T., Schubert, V., Watanabe, K., Stein, N., Houben, A., & Schubert, I. (2010). Synteny between Brachypodium distachyon and Hordeum vulgare as revealed by FISH. Chromosome research, 18(7), 841-850.

Machado F.B.P. (2003). Brasil, a doce terra – História do Setor (www.jornalcana.com.br).

Maliepaard C, Jansen J, Van Ooijen JW (1997) Linkage analysis in a full-sib family of an outbreeding plant species: Overview and consequences for applications. Genetical Research 70:237-250

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

Margarido GRA, Pastina MM, Souza AP, Garcia AAF. (2015) Multi-trait multi-environment quantitative trait loci mapping for a sugarcane commercial cross provides insights on the inheritance of important traits. Mol Breeding. 35:175.

Mascher M, Wu S, Amand PS, Stein N, Poland J. (2013) Application of genotyping-by-sequencing on semiconductor sequencing platforms: a comparison of genetic and reference-based marker ordering in barley. PloS One. 8(10), e76925.

Masouleh, A. K., Waters, D. L., Reinke, R. F., & Henry, R. J. (2009). A high-throughput assay for rapid and simultaneous analysis of perfect markers for important quality and agronomic traits in rice using multiplexed MALDI-TOF mass spectrometry. Plant biotechnology journal, 7(4), 355-363.

Marek, L. F., Shoemaker, R. C. (1997). BAC contig development by fingerprint analysis in soybean. Genome, 40(4), 420-427.

Matsuoka S., Garcia A.A.F., Arizono H. (1999). Melhoramento da cana-de-açúcar. In: Borém A ed, Melhoramento de espécies cultivadas. 2 ed. Viçosa, UFV, p 205-251.

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

Manechini, J. R. V., da Costa, J. B., Pereira, B. T., Carlini-Garcia, L. A., Xavier, M. A., de Andrade Landell, M. G., & Pinto, L. R. (2018). Unraveling the genetic structure of Brazilian commercial sugarcane cultivars through microsatellite markers. PloS one, 13(4).

McIntyre CL, Jackson M, Cordeiro GM, Amouyal O, Hermann S,Aitken KS, Eliott F, Henry RJ, Casu RE, Bonnett GD (2006) The identification and characterisation of alleles of sucrosephosphate synthase gene family III in sugarcane. Mol Breed 18:39–50.

Ming R, Liu SC, Moore PH, Irvine JE, Paterson AH (2001) QTL analysis in a complex autopolyploid: genetic control of sugar content in sugarcane. Genome Res 11:2075-2084

Ming R, Wang W, Draye X, Moore H, Irvine E, Paterson H (2002) Molecular dissection of complex traits in autopolyploids: mapping QTLs affecting sugar yield and related traits in sugarcane. Theor Appl Genet 105:332-345

Ming, R., Moore, P. H., Woo, K. K., D'Hont, A., Glaszmann, J. C., Tew, T. L., Mirkov, T. E., Silva, J. D., Jifon, J., Rai, M., Schnell, R. J., Brumbley, S. M., Lakshmanan, P., Comstock, J. C., Paterson, A. H. (2006) Sugarcane improvement through breeding and biotechnology. Plant Breed. Rev. 27:17–117.

Mogg R, Batley J, Hanley S, Edwards D, O'Sullivan H, Edwards KJ (2002) Characterization of the flanking regions of Zea mays microsatellites reveals a large number of useful sequence polymorphisms. Theor Appl Genet 105:532–543.

Mulcahy, D.G., Macdonald, K.S., Brady, S.G., Meyer, C., Barker, K.B., Coddington, J., 2016. Greater than X kb: a quantitative assessment of preservation conditions on genomic DNA quality, and a proposed standard for genome-quality DNA. PeerJ 4, e2528.

Murray SC, Sharma A, Rooney WL, Klein PE, Mullet JE, Mitchell SE, et al. Genetic improvement of sorghum as a biofuel feedstock: I. QTL for stem sugar and grain nonstructural carbohydrates. Crop Sci. 2008;48:2165-79.

Naidu K.M., Sreenivasan T.V. (1987). Conservation of sugarcane germplasm. In: Copersucar Int Sugarcane Breed Wkshp, Copersucar, São Paulo, Brazil, pp 33-53.

Nair N., Nair S., Sreenivasan T.V., Mohan M. (1999). Analysis of genetic diversity and phylogeny in Saccharum and related genera using RAPD markers. Genet. Res. Crop Evol. 46: 73-79.

Nishiyama, I. (1956). Basic numbers in the polyploidy of Saccharum. Journal of Heredity, 47(2), 91-99.

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

Noir, S., Patheyron, S., Combes, M. C., Lashermes, P., & Chalhoub, B. (2004). Construction and characterisation of a BAC library for genome analysis of the allotetraploid coffee species (Coffea arabica L.). Theoretical and Applied Genetics, 109(1), 225-230.

Nuss, K.J., Brett P.G.C. (1995). The release of cultivar NCo310 in 1945 and its impact on the sugar industry. Proc. S. Afr. Sug. Technol. Assoc. 69:3–8.

Oliveira KM, Pinto LR, Marconi TG, Margarido GRA, Pastina MM, Teixeira LHM, Figueira AV, Ulian EC, Garcia AAF, Souza AP (2007) Functional integrated genetic linkage map based on ESTmarkers for a sugarcane (Saccharum spp.) commercial cross. Mol Breed 20:189-208.

Oliveira KM, Pinto LR, Marconi TG, Mollinari M, Ulian EC, Chabregas SM, Falco MC, Burnquist W, Garcia AAF, Souza AP (2009) Characterization of new polymorphic functional markers for sugarcane. Genome 52:191-209.

Paiva, J. A., Prat, E., Vautrin, S., Santos, M. D., San-Clemente, H., Brommonschenkel, S., Fonseca, P.G.S.; Grattapaglia, D.; Song, X.; Ammiraju, J.S.S.; Kudrna, D.; Wing, R.A.; Freitas, A.T.; Berges, H.; Grima-pettenati, J. (2011). Advancing Eucalyptus genomics: identification and sequencing of lignin biosynthesis genes from deep-coverage BAC libraries. BMC genomics, 12(1), 137.

Panje R.R., Babu C.N. (1960). Studies in Saccharum spontaneum. Distribution and geographical association of the chromosome number. Cytologia 25: 152-172.

Pastina MM, Pinto LR, Oliveira KM, Souza AP, Garcia AAF (2010) Molecular mapping of complex traits. In: Henry R, Kole C (eds) Genetics, genomics and breeding of sugarcane, Science Publishers, Enfield, pp 117–148.

Pastina MM, Malosetti M, Gazaffi R, Mollinari M, Margarido GRA, Oliveira KM, Pinto LR, Souza AP, Eeuwijk FA van, Garcia AAF (2012) A mixed model qtl analysis for sugarcane multiple-harvest-location trial data. Theor Appl Genet 124:835-849.

Paterson, A. H., Damon, S., Hewitt, J. D., Zamir, D., Rabinowitch, H. D., Lincoln, S. E., ... & Tanksley, S. D. (1991). Mendelian factors underlying quantitative traits in tomato: comparison across species, generations, and environments. Genetics, 127(1), 181-197.

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., et al. (2009). The Sorghum bicolor genome and the diversification of grasses. Nature 457, 551–556.

Pedrosa, A., Sandal, N., Stougaard, J., Schweizer, D., Bachmair, A. (2002). Chromosomal map of the model legume Lotus japonicus. Genetics, 161(4), 1661-1672.

Pedrosa-Harand, A., Kami, J., Gepts, P., Geffroy, V., & Schweizer, D. (2009). Cytogenetic mapping of common bean chromosomes reveals a less compartmentalized small-genome plant species. Chromosome Research, 17(3), 405-417.

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

Pinto LR, Oliveira KM, Ulian EC, Garcia AAF, Souza AP (2004) Survey in the sugarcane expressed sequence tag database (SUCEST) for simple sequence repeats. Genome 47:795-804.

Pinto LR, Garcia AAF, Pastina MM, Teixeira LHM, Bressiani JA, Ulian EC, et al. (2010) Analysis of genomic and functional RFLP derived markers associated with sucrose content, fiber and yield QTLs in a sugarcane (Saccharum spp.) commercial cross. Euphytica.172:313-27.

Piperidis G., D'Hont A. (2001). Chromosome composition analysis of various Saccharum interespecific hybrids by genomic in situ hybridization (GISH). Proc. Int. Soc. Technol. 24: 565-566.

Piperidis N, Jackson PA, D'Hont A, Besse P, Hoarau JY, Courtois B,Aitken KS, McIntyre CL (2008) Comparative genetics insugarcane enables structured map enhancement and validation of marker-trait associations. Mol Breed 21:233–247.

Piperidis, G., Piperidis, N. & D'Hont, A. (2010). Molecular cytogenetic investigation of chromosome composition and transmission in sugarcane. Mol. Genet. Genom. 284, 65–73.

Poland JA, Brown PJ, Sorrells ME, Jannink JL. (2012). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PloS One. 7(2).

Price S. (1957). Cytological studies in Saccharum and allied genera. II. Geographic distribution and chromosome numbers in S. robustum. Cytologia 22: 40-52.

Price S. (1965). Cytology of Saccharum robustum and related sympatric species and natural hybrids. U.S. Dep. Agric. Res. Serv., Tech. Bull. 133, 47p.

Price S. (1968). Cytology of Chinese and North Indian sugarcanes. Econ Bot 22: 155-164.

Price S., Daniels J. (1968). Cytology of south Pacific sugarcane and related grasses: with special reference to Fiji. J Hered 59: 141-145.

Rabbi, I. Y., Kulembeka, H. P., Masumba, E., Marri, P. R., & Ferguson, M. (2012). An EST-derived SNP and SSR genetic linkage map of cassava (Manihot esculenta Crantz). Theoretical and Applied Genetics, 125(2), 329–342.

Raboin L-M., Oliveira K.M., Lecunff L., Telismart H., Roques D., Butterfield M., Hoarau J-Y., D'Hont A. (2006). Genetic mapping in the high polyploid sugarcane using a biparental progeny: identification of a gene controlling stalk colour and new rust resistance gene. Theor Appl Genet 112(7): 1382-1391.

Rafalski A. (2002). Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Biol. 5:94–100.

Rafalski, A., Tingey, S., & Henry, R. (2008). SNPs and their use in maize. Plant Genotyping II: SNP Technology. CAB International, Wallingford, UK, 30-43.

Reffay N, Jackson PA, Aitken KS, Hoarau JY, D'Hont A, Besse P, McIntyre CL (2005) Characterisation of genome regions incorporated from an important wild relative into Australian sugarcane. Mol Breed 15:367–381.

Ren, C., Lee, M. K., Yan, B., Ding, K., Cox, B., Romanov, M. N., Price, J.A., Dogson, J.B., Zhang, H. B. (2003). A BAC-based physical map of the chicken genome. Genome Research, 13(12), 2754-2758.

Ren C, Xu ZY, Sun S, Lee M-K, Wu C, Scheuring C, Santos TS, Zhang H-B (2005) Genomic DNA libraries and physical mapping. In: Meksem K, Kahl G (eds) The handbook of plant genome mapping: genetic, physical mapping. Wiley-VCH Verlag GmbH, Weinheim, pp 173–213.

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

Roach B.T. (1972). Nobilisation of sugarcane. Proc Int Soc Sugar Cane Tech 14: 206-216.

Roach, B. T. (1986). Evaluation and breeding use of sugarcane germplasm. In Proc Int. Soc. Sugar Cane Technol. Vol. 19, pp. 492-502.

Roach, B. T. (1989). Origin and improvement of the genetic base of sugarcane. Proc. Austral. Soc. Sugar Cane Technol. 11:34–47.

Roach B.T., Daniels J. (1987). A review of the origin and improvement of sugarcane. In: Copersucar Int Sugarcane Breed Wkshp, Copersucar, São Paulo, Brazil, pp 1-31.

Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, Casstevens TM, Gardner CA. (2013)Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biology. 14(6), R55.

Rural Centro, consulta em 2018. http://www.ruralcentro.com.br

Sato, K., Motoi, Y., Yamaji, N., & Yoshida, H. (2011). 454 sequencing of pooled BAC clones on chromosome 3H of barley. BMC Genomics, 12(1).

Schadt, E.E., Turner, S., Kasarskis, A., 2010. A window into third-generation sequencing. Hum. Mol. Genet. 19.

Shizuya, H., Birren, B., Kim, U. J., Mancino, V., Slepak, T., Tachiiri, Y., & Simon, M. (1992). Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. Proceedings of the National Academy of Sciences, 89(18), 8794-8797.

Sills G., Bridges W., Al-Janabi, S., Sobral B. (1995). Genetic analysis of agronomic traits in a cross between sugarcane (Saccharum officinarum L.) and its presumed progenitor (S. robustum Brandes & Jesw. Ex. Grassl). Molecular Breeding (1): 355-363.

Silva JA, Bressiani JA (2005) Sucrose synthase molecular marker associated with sugar content in elite sugarcane progeny. Genet Mol Biol 28(2):294–298.

Simmonds, N.W. (1976). Sugarcanes. p. 104–108. In: N. W. Simmonds (ed.), Evolution of crop plants. Longmans, London.

Singh RK, Singh SP, Tiwari DK, Srivastava S, Singh SB, Sharma ML, et al. (2013) Genetic mapping and QTL analysis for sugar yield-related traits in sugarcane. Euphytica. 191:333-53.

Smit, AFA, Hubley, R, Green, P. RepeatMasker Open-4.0. 2013-2015 http://www.repeatmasker.org.

Sobral B.W.S., Braga D.P.V., LaHood E.S., Keim P. (1994). Phylogenetic analysis of chloroplast restriction enzyme site mutations in the Saccharinae Griseb. subtribe of the Andropogoneae Dumort. tribe. Theor Appl Genet 87: 843-853.

Sociacana, Associação dos Fornecedores de Cana de Guariba. http://www.socicana.com.br Consultado em 2018.

Sonah H, Bastien M, Iquira E, Tardivel A, Legare G, Boyle B, et al. (2013) An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. PLoS ONE.

Song, J., Bradeen, J. M., Naess, S. K., Helgeson, J. P., Jiang, J. (2003). BIBAC and TAC clones containing potato genomic DNA fragments larger than 100 kb are not stable in Agrobacterium. Theoretical and Applied Genetics, 107(5), 958-964.

Souza, G.M., Berges, H., Bocs, S., Casu, R., D'Hont, A., Ferreira, J.E., Henry, R., Ming, R., Potier, B., Sluys, M.A. van, Vincentz, M., and Paterson, A.H. (2011). The sugarcane genome challenge: strategies for sequencing a highly complex genome. Tropical Plant Biol. 4: 145–156.

Spindel J, Wright M, Chen C, Cobb J, Gage J, Harrington S, Mccouch S. (2013) Bridging the genotyping gap: using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. Theoretical and Applied Genetics. 126(11), 2699–716.

Sreenivasan T.V., Ahloowalia B.S. and Heinz D.J. (1987). Cytogenetics. In Sugarcane Improvement Through Breeding. Edited by Heinz D.J. Amsterdam: Elsevier Press. pp211-253.

Stevenson G.C. (1965). Genetics and breeding of sugarcane. Longmans, London. 284p.

Sudhir Kumar, Glen Stecher, and Koichiro Tamura (2015) MEGA7: Molecular Evolutionary Genetics Analysis version 7.0. Molecular Biology and Evolution

Tang, X., De Boer, J. M., Van Eck, H. J., Bachem, C., Visser, R. G. F., De Jong, H. (2009). Assignment of genetic linkage maps to diploid Solanum tuberosum pachytene chromosomes by BAC-FISH technology. Chromosome Research, 17(7), 899–915.

Tanksley, S. D. (1993). Mapping polygenes. Annual review of genetics, 27(1), 205-233.

Tao, Q., Chang, Y. L., Wang, J., Chen, H., Islam-Faridi, M. N., Scheuring, C., Wang, B., Stelly, D.M., Zhang, H. B. (2001). Bacterial artificial chromosome-based physical map of the rice genome constructed by restriction fingerprint analysis. Genetics, 158(4), 1711-1724.

Tao, Q., Zhang, H. B. (1998). Cloning and stable maintenance of DNA fragments over 300 kb in Escherichia coli with conventional plasmid-based vectors. Nucleic Acids Research, 26(21), 4901-4909.

Temnykh, S., DeClerck, G., Lukashova, A., Lipovich, L., Cartinhour, S., McCouch, S. (2001). Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): frequency, length variation, transposon associations, and genetic marker potential. Genome research, 11(8), 1441-1452.

Tew, T.L. (1987). New varieties. p. 559–594. In: D. J. Heinz (ed.), Sugarcane improvement through breeding. Elsevier, Amsterdam.

Tew, T.L. (2003). World sugarcane variety census – Year 2000. Sugar Cane Int. March/April 2003, p. 12–18.

Tomkins, J. P., Yu, Y., Miller-Smith, H., Frisch, D. A., Woo, S. S., & Wing, R. A. (1999). A bacterial artificial chromosome library for sugarcane. Theoretical and Applied Genetics, 99(3–4), 419–424.

Uitdewilligen JGAML, Wolters A-MA, D'hoop BB, Borm TJA.; Visser RGF, Van Eck HJ. (2013) A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. PloS One. 8(5), e62355. doi:10.1371/journal.pone.0062355.

Venter, J. C., Smith, H. O., & Hood, L. (1996). A new strategy for genome sequencing. Nature, 381(6581), 364–366.

Venter, J. C., Adams, M. D., Sutton, G. G., Kerlavage, A. R., Smith, H. O., & Hunkapiller, M. (1998). Shotgun sequencing of the human genome. Science, 280(5369), 1540–1542.

Vettore, A.L., da Silva, F.R., Kemper, E.L., Souza, G.M., da Silva, A.M., Ferro, M.I., Henrique-Silva, F., Giglioti, E.A., Lemos, M.V., Coutinho, L.L., Nobrega, M.P., Carrer, H., França, S.C., Bacci Júnior, M., Goldman, M.H., Gomes, S.L., Nunes, L.R., Camargo, L.E., Siqueira, W.J., Van Sluys, M.A., Thiemann, O.H., Kuramae, E.E., Santelli, R.V., Marino, C.L., Targon, M.L., Ferro, J.A., Silveira, H.C., Marini, D.C., Lemos, E.G., Monteiro-Vitorello, C.B., Tambor, J.H., Carraro, D.M., Roberto, P.G., Martins, V.G., Goldman, G.H., de Oliveira, R.C., Truffi, D., Colombo, C.A., Rossi, M., de Araujo, P.G., Sculaccio, S.A., Angella, A., Lima, M.M., de Rosa Júnior, V.E., Siviero, F., Coscrato, V.E., Machado, M.A., Grivet, L., Di Mauro, S.M., Nobrega, F.G., Menck, C.F., Braga, M.D., Telles, G.P., Cara, F.A., Pedrosa, G., Meidanis, J., and Arruda, P. (2003). Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. Genome Res. 13: 2725–2735.

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

Vilela MM, Del-Bem LE, Van Sluys MA, de Setta N, Kitajima JP, Cruz GMQ, Sforça DA, de Souza AP, Ferreira PCG, Grativol C, Cardoso-Silva CB, Vicentini R, Vincentz M. (2017) Analysis of three sugarcane homo/homeologous regions suggests independent polyploidization events of Saccharum officinarum and Saccharum spontaneum. Genome Biol Evol 9(2):266–278.

Wai, C. M., Ming, R., Moore, P. H., Paull, R. E., Yu, Q. (2010). Development of chromosome-specific cytogenetic markers and merging of linkage fragments in papaya. Tropical plant biology, 3(3), 171-181.

Walker, D.I.T. (1987). Manipulating the genetic base of sugarcane. p. 321–334. In Anonymous (eds.), Copersucar International Sugarcane Breeding Workshop. Copersucar, Brazil.

Wang, K.; Guan, B.; Guo, W. (2008). Completely distinguishing individual A-genome chromosomes and their karyotyping analysis by multiple bacterial artificial chromosome fluorescence in-situ hybridization. Genetics, Austin, v. 178, p. 1117-1122.

Wei X, Jackson PA, McIntyre CL, Aitken KS, Croft B. (2006) Associations between DNA markers and resistance to diseases in sugarcane and effects of population substructure. Theor Appl Genet.114:155-64.

Wei, F., Stein, J.C., Liang, C., Zhang, J., Fulton, R.S., Baucom, R.S., Paoli, E.D., Zhou, S., Yang, L., Han, Y., Pasternak, S, Narechania, A., Zhang, L., Yeh, X.-T., Ying, K., Nagel, D.H., Collura, K., Kudrna, D., Currie, J., Lin, J., Kim, H.R., Anegelova, A., Scara, G., Wissotski, M., Golser, W., Courtney, L., Kruchowski, S., Graves, T.A., Rock, S.M., Adams, S., Fulton, L.A., Fronick, C., Courtney, W., Kramer, M., Spiegel, L., Nascimento, L., Kalyabaranan, A., Chaparro, C., Deragon, J.M.0., Miguel, P.S., Jiang, N., Wessler, S.R., Green, P.J., Yu, Y., Schwartz, D.C., Meyers, B.C., Bennetzen, J.L., Martienssen, R.A., Mccombie, W.R., Aluru, S., Clifton, S.W., Schnable, P.S., Ware, D., Wilson, R.K., Wing, R.A. (2009). Detailed analysis of a contiguous 22-Mb region of the maize genome. PLoS genetics, 5(11), e1000728.

Wu R, Ma CX, Painter I, Zeng ZB (2002) Simultaneous maximum likelihood estimation of linkage and linkage phases in outcrossing species. Theor Popul Biol 61:349-363

Wu, C., Sun, S., Lee, M. K., Xu, Z., Ren, C., Santos, T. S., Zhang, H. B. (2005). Whole-Genome Physical Mapping: An Overview on Methods for DNA Fingerprinting. The Handbook of Plant Genome Mapping: Genetic and Physical Mapping, 257-283.

Wu C, Xu Z, Zhang H-B (2004) DNA libraries. In: Meyers RA (ed) Encyclopedia of molecular cell biology and molecular medicine, Vol. 3 (2nd ed.). Wiley-VCH Verlag GmbH, Weinheim, Germany, pp 385–425.

Yüksel, B., & Paterson, A. H. (2005). Construction and characterization of a peanut HindIII BAC library. Theoretical and Applied Genetics, 111(4), 630-639.

Zhang, H. B., & Wu, C. (2001). BAC as tools for genome sequencing. Plant Physiology and Biochemistry, 39(3–4), 195–209.

Zhang, H.B. Map-based cloning of genes and quantitative trait loci. (2007) In: In: Kole, C., Abbott, A.G. (Ed.). Principles and practices of plant genomics. New Hampshire: Science Publ. v. 1, p. 229-267.

Zhang, X., Scheuring, C., Tripathy, S., Xu, Z., Wu, C., Ko, A., Tian, S.K., Arredondo, F., Lee, M.K., Santos, F.A., Jiang, R.H., Zhang, H.B, Tyler, B.M. (2006). An integrated BAC and genome sequence physical map of Phytophthora sojae. Molecular Plant-Microbe Interactions, 19(12), 1302-1310.

Zhang, Y., Zhang, X., O'Hare, T. H., Payne, W. S., Dong, J. J., Scheuring, C. F., Zhang, M., Huang, J.J., Lee, M.K., Delany, M.E., Zhang, H.B., Dogson, J.B (2011). A comparative physical map reveals the pattern of chromosomal evolution between the turkey (Meleagris gallopavo) and chicken (Gallus gallus) genomes. BMC genomics, 12(1), 447.

Zhang, W., Zuo, S., Li, Z., Meng, Z., Han, J., Song, J., ... & Wang, K. (2017). Isolation and characterization of centromeric repetitive DNA sequences in Saccharum spontaneum. Scientific reports, 7, 41659.

Zhang J, Zhang X, Tang H, Zhang Q, Hua X, Ma X, Zhu F, Jones T, Zhu X, Bowers J, et al. (2018) Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. Nature Genetics, doi.org/10.1038/s41588-018-0237-2.

ANEXOS

Anexo I

“Targeted Sequencing by Gene Synteny” a New Strategy for Polyploid Species: Sequencing and Physical Structure of a Complex Sugarcane Region

M. C. Mancini, C. B. Cardoso-Silva, D. A. Sforça e A. P. de Souza.

Publicado na revista
Frontiers in Plant Science (9, 397; 2018)

frontiers in Plant Science

ORIGINAL RESEARCH
published: 28 March 2018
doi: 10.3389/fpls.2018.00397

"Targeted Sequencing by Gene Synteny," a New Strategy for Polyploid Species: Sequencing and Physical Structure of a Complex Sugarcane Region

Melina C. Mancini[1†], *Claudio B. Cardoso-Silva*[1†], *Danilo A. Sforça*[1] *and Anete Pereira de Souza*[1,2*]

[1] *Center for Molecular Biology and Genetic Engineering, University of Campinas, Campinas, Brazil,* [2] *Departament Plant Biology, Biology Institute, University of Campinas, Campinas, Brazil*

Sugarcane exhibits a complex genome mainly due to its aneuploid nature and high ploidy level, and sequencing of its genome poses a great challenge. Closely related species with well-assembled and annotated genomes can be used to help assemble complex genomes. Here, a stable quantitative trait locus (QTL) related to sugar accumulation in sorghum was successfully transferred to the sugarcane genome. Gene sequences related to this QTL were identified *in silico* from sugarcane transcriptome data, and molecular markers based on these sequences were developed to select bacterial artificial chromosome (BAC) clones from the sugarcane variety SP80-3280. Sixty-eight BAC clones containing at least two gene sequences associated with the sorghum QTL were sequenced using Pacific Biosciences (PacBio) technology. Twenty BAC sequences were found to be related to the syntenic region, of which nine were sufficient to represent this region. The strategy we propose is called "targeted sequencing by gene synteny," which is a simpler approach to understanding the genome structure of complex genomic regions associated with traits of interest.

Keywords: polyploid, physical map, BAC, *Saccharum hybridum*, sugar accumulation, complex genome

OPEN ACCESS

Edited by:
Hikmet Budak,
Montana State University,
United States

Reviewed by:
Diego Mauricio Riaño-Pachón,
Universidade de São Paulo, Brazil
Thiruvarangan Ramaraj,
National Center for Genome Resources, United States

***Correspondence:**
Anete Pereira de Souza
anete@unicamp.br

[†]*These authors have contributed equally to this work.*

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *21 November 2017*
Accepted: *12 March 2018*
Published: *28 March 2018*

Citation:
Mancini MC, Cardoso-Silva CB, Sforça DA and Pereira de Souza A (2018) "Targeted Sequencing by Gene Synteny," a New Strategy for Polyploid Species: Sequencing and Physical Structure of a Complex Sugarcane Region. Front. Plant Sci. 9:397. doi: 10.3389/fpls.2018.00397

INTRODUCTION

When no previously reported genome is available, genome reconstruction is based on a *de novo* assembly strategy (based on sequence read overlap). This task becomes more complicated when an organism has a large genome with highly abundant repetitive elements.

Polyploid species account for approximately one-third of all plants (Wood et al., 2009), many of which are crops with great economic importance, such as wheat, cotton, potato and sugarcane. Sugarcane (*Saccharum* sp.) is the crop with the most complex genome structure because modern sugarcane varieties are derived from interspecific hybridization between *Saccharum officinarum* (basic chromosome number: $x = 10$; $2n = 8x = 80$) and *Saccharum spontaneum* (basic chromosome number: $x = 8$; $2n = 5x = 40$ to $16x = 128$). The resulting hybrids are highly polyploid and aneuploid, with chromosome numbers ranging from 80 to 128 (D'Hont et al., 1998; Irvine, 1999; Grivet and Arruda, 2001) and an estimated whole-genome size of 10 Gb (D'Hont and Glaszmann, 2001). Previous studies have shown that ∼50% of the sugarcane genome is composed of repetitive sequences (Figueira et al., 2012; Kim et al., 2013; de Setta et al., 2014).

Several studies using bacterial artificial chromosomes (BACs), involving either individual BAC assembly (de Setta et al., 2014; Vilela et al., 2017) or pooled strategies (Okura et al., 2016; Visendi et al., 2016), have been reported. In both cases, the applied sequencing strategies are based on the selection of non-overlapping BAC clones. Moreover, a draft sugarcane genome based on whole-genome shotgun sequencing of the SP80-3280 hybrid has been published (Riaño-Pachón and Mattiello, 2017). However, the main problem lies in reconstructing large and complex regions of the genome to represent a specific region of interest. In the present study, the synteny between related species, sorghum (*Sorghum bicolor*) and sugarcane (*Saccharum* sp.), was explored. Among the grasses that have been studied to date, sorghum is considered the closest ancestor of the *Saccharum* complex. Sugarcane and sorghum shared a common ancestor ∼5 million years ago (Paterson et al., 2004), while sugarcane and its sister genus *Miscanthus* share a common ancestor separated by ∼3.8–4.6 million years (Kim et al., 2014). Using the sorghum genome as a reference for annotation is advantageous because it has been completely sequenced and annotated (Paterson et al., 2009). Additionally, some sorghum varieties, referred to as sweet sorghum [*Sorghum bicolor* (L.) Moench], are capable of storing sugar in their stems (Vietor and Miller, 1990). Here, we propose the "targeted sequencing by gene synteny" strategy of sugarcane BAC selection for the reconstruction of a complex sugarcane genome region linked to a quantitative trait locus (QTL) mapped for sugar accumulation (Brix) (Murray et al., 2008) at a specific position on sorghum chromosome 3 (SB-03), based on the high synteny between the sugarcane and sorghum genomes.

MATERIALS AND METHODS

In Silico Data Sources (Sorghum and Sugarcane)

A QTL for Brix was chosen from a study by Murray et al. (2008), which identified the QTL in the SB-03 genome (see **Data Sheet S1** topic "*In silico* data sources"). The sequences of each molecular marker in this region were employed to locate the chromosome position using the sorghum genome v3.1, available on the Phytozome 12.0 database (http://www.phytozome.net/), as a reference. An alignment between sorghum genes and sugarcane transcripts (Cardoso-Silva et al., 2014) was performed through a BLASTn analysis with a cutoff $E < 1e10$. In this step, we selected the best hit for each query alignment (**Table S1**). We designed primer pairs flanking single and conserved exons predicted by alignments between sugarcane and sorghum genes (**Table S2**).

BAC Library Screening, BAC Pooling, and Sequencing

BAC clones from the Brazilian hybrid sugarcane cultivar SP80-3280 that contained the specific selected genes were chosen through screening of 3D pools (see **Data Sheet S1** topic "BAC library screening"). Positive BAC clones containing the same gene were sequenced in different pools to avoid casual overlap of BACs containing homeologous regions. A total of 68 BAC clones were arranged in nine sequencing pools. SMRTbell libraries for sequencing were prepared using the 20 kb procedure according to the Pacific Biosciences (PacBio) protocol, and sequencing was performed at the Arizona Genomics Institute (AGI; Tucson, USA) using a SMRT DNA sequencing system available from PacBio.

Read Trimming and BAC Assembly

The PacBio long reads were masked for vector sequences (*pIndigo*BAC5) using cross_match (-minmatch 10 -minscore 20 -screen), and *E. coli* str. K-12 genomic DNA was removed. *De novo* assembly was performed with the hierarchical assembly pipeline PBcR (the PacBio Corrected Reads Pipeline), implemented as part of wgs-assembler v8.3rc2 (Berlin et al., 2015) and Celera Assembler (Myers et al., 2000). The minimum length of the sequences for correction was set to 500 bp, and the number of partitions for consensus was set to 200. The contigs obtained with the assembler were subjected to error correction by remapping the reads with pbalign (v0.2). The PacBio reads were aligned using the BLASR algorithm (Chaisson and Tesler, 2012), and we performed assembly polishing with the Quiver tool (Chin et al., 2013). See **Data Sheet S1** topic "BAC assembly" for more details.

BAC Annotation and Synteny Analysis

The BAC sequences were annotated in two steps. First, we used a method to predict long terminal repeat transposons (LTRs) via LTR_FINDER (Xu and Wang, 2007). Homology-based repeat analysis was performed to identify transposable elements (TEs) against Poaceae TEs available in the Repbase database (Kohany et al., 2006) via CENSOR. Second, genes were manually predicted using the sorghum genome annotation as a reference. All annotations were manually curated using Artemis: Genome Browser and Annotation Tools (Rutherford et al., 2000). Additionally, sugarcane CDS genes were translated into protein and were aligned by BLASTp (cutoff $E < 1e\text{-}10$) against the sorghum, maize, and rice proteomes obtained from the Phytozome 12.0 database.

RESULTS

In Silico Data Sources (Sorghum and Sugarcane)

Sequence-based marker information related to the QTL for Brix (Murray et al., 2008) was employed for linkage to the physical location on SB-03 (from Sb3:55,265 kb to Sb3:55,952 kb; sorghum genome v3.1 available on Phytozome 12.0 database), comprising ∼700 kb in length (**Data Sheet S2**). A total of 61 predicted genes were found within this region in the sorghum genome, and these genes were used for alignment against the sugarcane transcriptome described by Cardoso-Silva et al. (2014). Fifty-three sorghum genes showed high similarity to sugarcane transcripts (**Table S1**). One primer pair for each of the 53 selected genes was synthesized using the sugarcane transcriptome as a template (**Table S2**).

BAC Library Screening, Sequencing, and Assembly

The primers showing good amplification were employed in the 3D pool screening method. To increase the chance of recovering the homologous region in the sugarcane genome, BAC clones were only selected if they had at least two positive markers. Based on this strategy, a total of 68 BAC clones were identified, pooled, and further sequenced (see **Data Sheet S1** topic "BAC library screening").

Thus, a total of 1,660,342 trimmed long reads were obtained; the number of reads per pool ranged from 139,394 (Pool 03) to 237,520 (Pool 06), with a mean of 184,482 long reads per pool (**Table 1**). The percentage of reads that represented contamination by the *E. coli* genome was 8.25% on average, ranging from 5% (Pool 01) to 13% (Pool 03).

Assembly was performed individually for each pool. The number of contigs that originated from the pools ranged from 16 (Pools 01 and 05) to 27 (Pool 04), with a mean number of contigs of ∼20. A total of 180 contigs were obtained through Celera assembly, with sizes ranging from 187,285 kb (Pool 09) to 8,050 kb (Pool 05). The total length of all the assembled contigs was 8.94 Mb, with an N50 contig length of 91.5 kb and a GC content of 44.74%. The N50 value was higher than that obtained during wheat BAC sequencing using only long reads generated by PacBio, which exhibited a mean N50 of 80 kb (Visendi et al., 2016).

Most of the assembled contigs (112 contigs, 62.2% of the total) exhibited lengths smaller than 50 kb (**Figure S2**) and/or showed low coverage assembly (**Figure S1**); these contigs were not considered in further analyses. However, 68 of the assembled contigs exhibited suitable lengths and high coverage (**Figure 1**).

BAC Annotation

A total of 68 BACs representing the longest contigs with high coverage (**Figures S3**, **S4**) were selected for gene annotation and repetitive element screening. Approximately 51% of the assembled and annotated BACs were identified as repetitive elements, including 41% of long terminal repeat retrotransposons (LTR), 8% of DNA transposons and 2% of non-LTRs. Within the LTRs, the most common groups were *Gypsy* and *Copia*, representing 58 and 42% of the total, respectively (**Table S3**).

A total of 253 complete coding genes were predicted in 55 sugarcane BAC sequences using the sorghum genome as a reference, 211 of which were unique genes, with the number of genes ranging from one to 13, yielding a gene density of one gene per 23.6 kb (**Table S4**). A total of 245 and 243 of these genes were shared with rice and maize, respectively. Additionally, 134 mobile elements inserted within genes were identified, with 69 genes containing inserted mobile elements ranging from 146 bp (Stowaway) to 11,800 bp (LTR/*Copia*) in size.

Corresponding Region of the Sorghum QTL and Synteny Analyses

Based on the analysis of the physical map, it was possible to define the homeologous chromosomes and gene duplications. In total, 20 BAC sequences were successfully mapped to the corresponding sorghum gene position (**Figure 2**). A total of 74 genes were observed in this interval in sorghum, while 59 were identified in sugarcane.

Using the genes annotated in sugarcane as a reference, a total sequence length of 1.25 Mb was necessary to partially cover the target region in SB-03, which was represented by nine BAC sequences (**Figure 3**) divided into four syntenic blocks. There were three gaps found among the four sugarcane syntenic blocks. In two situations, we found sorghum genes without a corresponding BAC sequence between: shy3280sca001 and shy3280sca002 (Sobic003G217500 to Sobic003G217900) and shy3280sca002 and shy3280sca003 (Sobic003G218700); while between shy3280sca004 and shy3280sca006, there were two consecutive sorghum genes that had different BAC sequences.

DISCUSSION

The regions that control economically important traits are often influenced by several genes, and QTL mapping is typically used to determine the genomic position of *loci* that phenotypically

TABLE 1 | Statistical summary of the sequencing (via PacBio) and assembly of the sugarcane BAC pools from the Brazilian hybrid sugarcane cultivar SP80-3280.

	PacBio sequencing			Celera assembly					
Name	**N. BACs**	**Trimmed reads**	***E. coli* %**	**Contigs**	**Longest contig**	**Smallest contig**	**Contig total length**	**N50**	**GC (%)**
Pool 01	4	178,758	5	16	143,471	9,202	582,340	62,347	43.77
Pool 02	8	202,770	7	17	134,154	8,615	1,034,115	109,126	45.08
Pool 03	8	139,394	13	25	142,211	8,101	800,349	54,726	45.45
Pool 04	8	206,601	8	27	122,448	9,303	882,224	41,554	44.06
Pool 05	8	189,764	9.6	16	175,157	8,050	1,186,577	132,868	45.28
Pool 06	8	237,520	9	21	168,704	8,289	920,150	86,198	44.75
Pool 07	8	143,827	6.8	19	164,848	10,668	1,140,957	128,641	45.41
Pool 08	8	186,873	7.4	19	143,661	10,202	1,129,862	108,955	44.41
Pool 09	8	174,835	8.4	20	187,285	10,664	1,261,846	99,030	44.43
Total	68	1,660,342	–	180	–	–	8,938,420	–	–

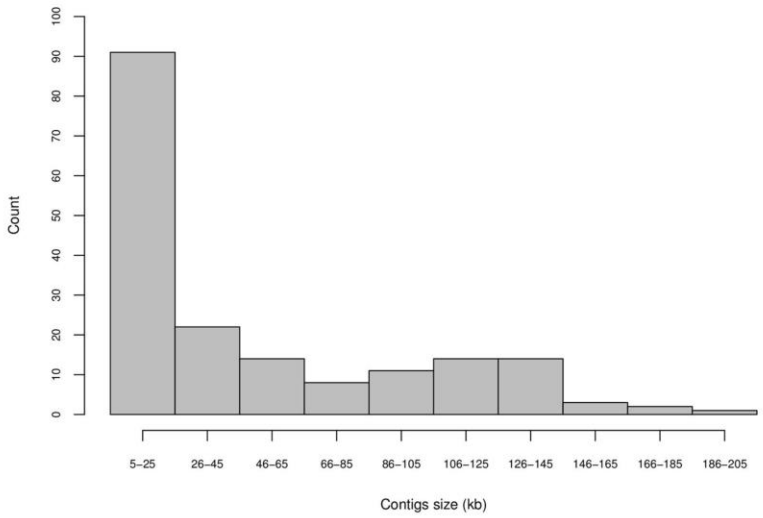


FIGURE 1 | Length distributions of the 180 sugarcane contigs obtained by the assembly of sugarcane BAC pools from the Brazilian hybrid sugarcane cultivar SP80-3280.

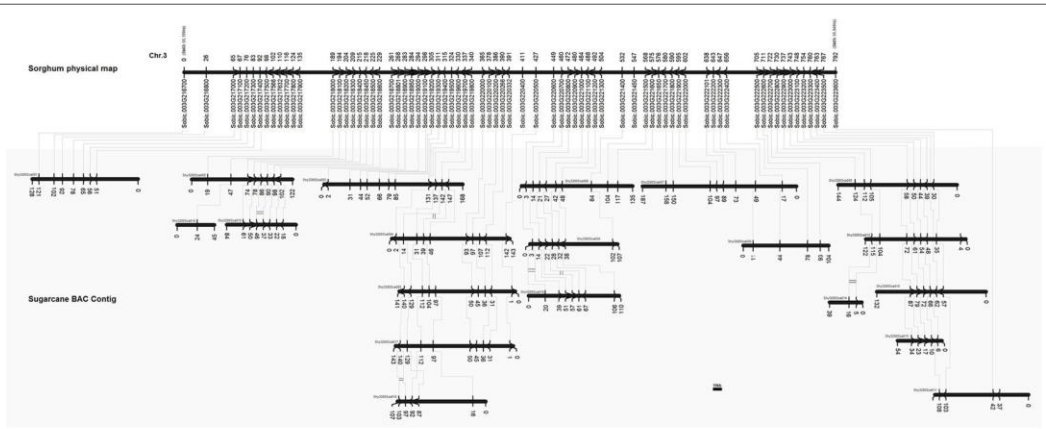


FIGURE 2 | Physical representation of a specific sorghum genome region (Chr3) containing a QTL for sugar accumulation and sugarcane BAC sequences (shaded gray box) from the Brazilian hybrid sugarcane cultivar SP80-3280. The sorghum gene annotation and position (starting in zero kb, representing the beginning of the specific region) were included. The synteny between sorghum and sugarcane genes is represented as is the genomic organization, including the homeologous BAC sequences and tandem duplication in sugarcane genes (double line connection).

influence a desired trait. In sugarcane, these effects are usually low. Genomic characteristics such as a high ploidy level show complex allele dosage and distribution on different homeologous chromosomes, which could explain the lower contribution of individual genes and/or alleles. Therefore, most studies have mapped single allelic doses. The percentage of phenotypic variance explained for various traits ranges from 0.069% (Costa et al., 2016) to 16.2% (Ming et al., 2002). However, such effects are more pronounced in sorghum, ranging from 7.7% (Shiringani et al., 2010) to 25% (Murray et al., 2008).

Sorghum is the most closely related species to sugarcane, with a fully sequenced genome and a large amount of available QTL data. Therefore, this species was used as a reference for selecting a region involved in an important trait, i.e., sugar accumulation, and identifying a homologous set of these genes in sugarcane. If these homologous genes diverged from sorghum

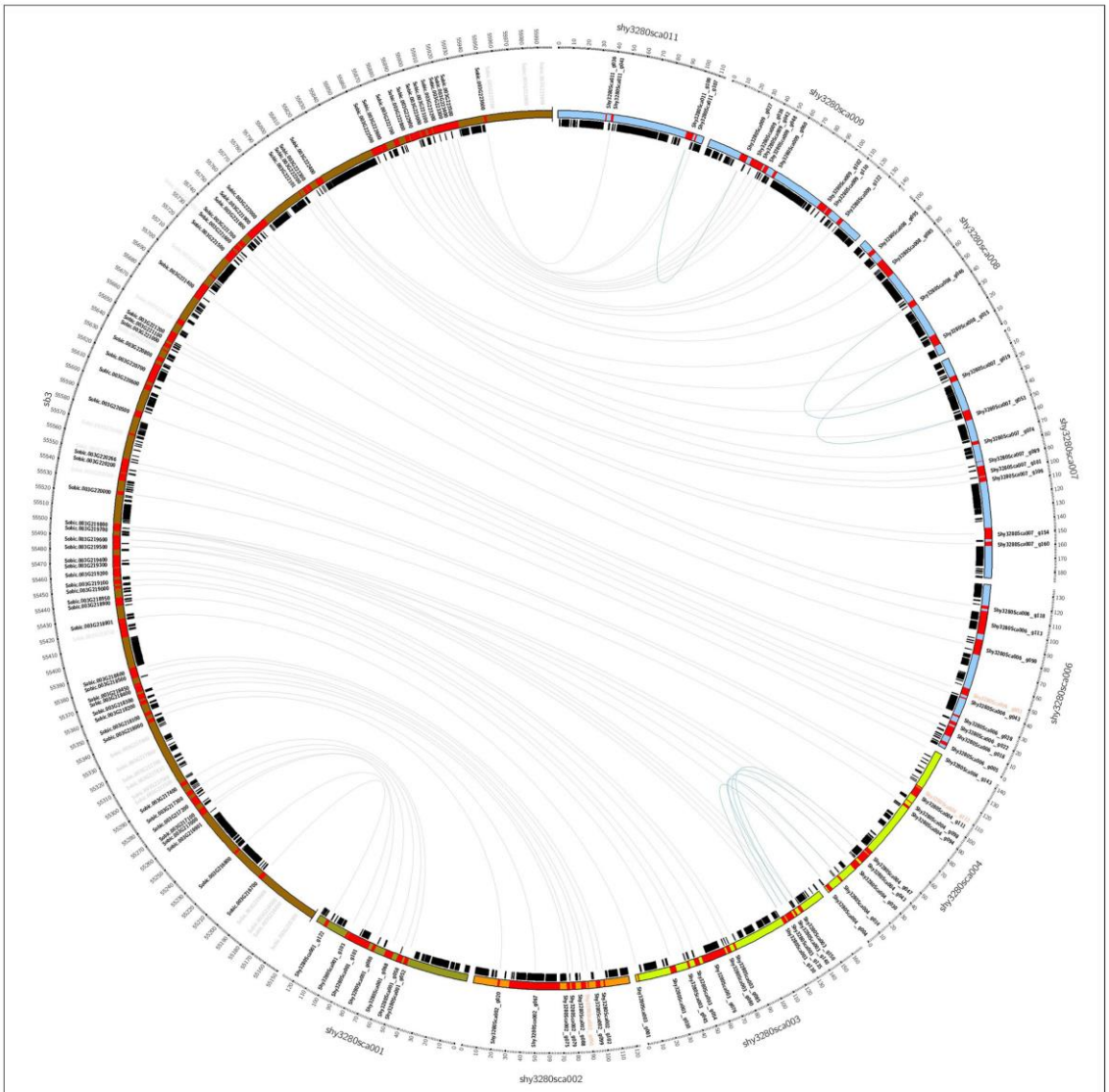


FIGURE 3 | Structural organization relationship among the genes on chromosome 03 in *Sorghum bicolor* (golden brown) and the four blocks of BAC sequences in sugarcane based on synteny. The sugarcane blocks are represented by different colors: green (Shy3280Sca001), orange (Shy3280Sca002), light green (Shy3280Sca003 and Shy3280Sca004), and light blue (Shy3280Sca006, Shy3280Sca007, Shy3280Sca008, Shy3280Sca009, and Shy3280Sca011).

after a speciation event and if they came from the same duplicated group, then they are orthologs (Fitch, 1970) and should have the same function in sugarcane and sorghum. However, further investigation is needed to check if there is evidence of QTLs in this region that are associated with sugar accumulation in sugarcane. This approach, "targeted sequencing by gene synteny," was possible once nearly all the genes were found in the sugarcane transcriptome, as described by Cardoso-Silva et al. (2014). A total of seven genes were not detected in the transcriptome described by Cardoso-Silva et al. (2014); four of these genes were found in a more recently published transcriptome described by Mattiello et al. (2015), and two of these transcripts were shared in a transcriptome described by Hoang et al. (2017). These results showed a high level of synteny in sorghum. More than

100,000 SP80-3280 BAC clones were used as a resource to access the sugarcane genome and recover this complex region. The positive BAC clones for two or more markers were selected for sequencing. A double selection strategy avoided small duplicated regions, pseudogenes and transposable elements carrying gene fragments as well as dramatically reduced the number of BAC clones selected. The advent of third-generation sequencing, and especially technologies resulting in the longest read lengths, such as single molecule real-time (SMRT) DNA sequencing (Eid et al., 2009), may facilitate the assembly process for segmental duplication problems caused by repetitive elements in complex genomes (English et al., 2012).

Large inserts of repetitive elements were observed between genes, but few large repetitive sequences were observed in intron sequences. Such large repetitive sequences in introns have been previously reported in other plants and do not necessarily affect the function of the gene (Kim and Zilberman, 2014). The high level of collinearity between the sorghum and sugarcane genes was utilized to identify the sugarcane homeologous regions associated with the absence of collinearity for repetitive regions (Jannoo et al., 2007; Garsmeur et al., 2011).

According to the comparative analysis with sorghum, at least 1.25 Mb, which was represented by nine sugarcane BAC sequences, was necessary to provide almost total coverage of the QTL region employing the "targeted sequencing by gene synteny" strategy. Some BAC sequences showed overlapping potential clustering in four syntenic blocks, with a highly conserved level of gene collinearity. For BAC clones that showed synteny with sorghum regions, there were two possibilities: complete overlap between BAC sequences suggested that the BACs came from the same homeologous chromosome, whereas total gene collinearity between BAC sequences and unaligned intergenic regions suggested that the BAC sequences came from different homeologous chromosomes. The choice of sorghum QTL stable and rich genes enabled these results to represent estimates for a small region of the sugarcane genome, ensuring the non-randomness of the results. Six of these sugarcane genes presented tandem duplications and could be attributed to the whole-genome duplication and polyploidization process (Alix et al., 2017). Additionally, these genes were inserted in an important biological region for sugarcane, and some hypotheses can be put forward to explain how these genes have maintained their original functions: if the original locus is disabled by mutation, the second gene can supply the necessary functional redundancy, or if both copies are maintained, they could increase the production of a gene product (Ohno, 1970).

These results represent an important step in understanding the genome structure of sugarcane and elucidating the complex architecture of the genomic region. This region should be associated with sugar accumulation. In addition, we propose a sequencing strategy for genome studies in polyploid species or diploid species originating via polyploidization, which present a huge challenge for obtaining the whole-genome sequence. The "targeted sequencing by gene synteny" approach can be applied to such species with complex genomes, especially those that have closely related diploid species with sequenced whole genomes. Furthermore, the use of BACs represents a powerful tool for recovering loci linked to important traits and determining homeologous regions associated with specific loci. Adding syntenic information to sequencing of non-random genome regions enables improving our understanding of genetic structure and identifying molecular markers physically linked to genes of interest in complex species. This strategy is very efficient and useful for the sequencing of regions enriched in genes. These advantages may allow important applications of sequencing results in plant breeding programs of polyploid species, particularly if the whole-genome sequence is not yet available for the species of interest.

DATA ACCESS

All the assembled and annotated BAC sequences were deposited in NCBI GenBank under accession numbers MF737006 to MF737073, and each sequencing pool was deposited in NCBI GenBank under SRA numbers SRR6760342 to SRR6760350. All the data can be found under Bioproject PRJNA398673.

AUTHOR CONTRIBUTIONS

MM and DS: Conducted the experiments; CC-S: Analyzed the sequencing data; MM, CC-S, DS, and AP: Wrote the manuscript. All authors discussed the data, interpreted the results, read and edited the manuscript and approved the final version.

FUNDING

This study was supported by the São Paulo Research Foundation (FAPESP) (2008/52197-4) and Coordination for the Improvement of Higher Education Personnel (CAPES, Computational Biology Program). The first two authors were supported by FAPESP PD fellowships (MM 2014/11482-9 and CC-S 2015/16399-5) and FAPESP-BEPE fellowship (MM 2017/05014-0); DS received a Ph.D. fellowship from FAPESP (2010/50119-6); AP received a research fellowship from the National Council for Scientific and Technological Development (CNPq).

ACKNOWLEDGMENTS

The authors thank Dr. David A. Kudrna from Arizona Genomics Institute, University of Arizona, School of Plant Sciences, University of Arizona, for helping us during optimization of the PacBio sequencing strategy.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.00397/full#supplementary-material

Data Sheet S1 | This file contains Supplementary Information including *in silico* data source, BAC screening, assembly and annotation description.

Data Sheet S2 | File containing the FASTA format sequence of the sorghum target region.

Table S1 | BLAST results for the sugarcane transcript selection.

Table S2 | Primers used to screen the sugarcane BAC library.

Table S3 | Distribution of the predicted transposable elements among the annotated BAC sequences.

Table S4 | Summary of the sugarcane predicted gene annotation and alignment against sorghum, maize and rice proteomes.

Figure S1 | Number of reads mapped onto non-annotated BAC contigs.

Figure S2 | Lengths of the sugarcane non-annotated contigs representing the lowest contigs from each pool.

Figure S3 | Distribution of the number of reads mapped onto each annotated BAC sequence.

Figure S4 | Contig-length distribution of the sugarcane annotated BAC sequences.

REFERENCES

Alix, K., Pierre, R., Géard, P. R., Schwarzacher, T., and Heslop-Harrison, J. S. (2017). Polyploidy and interspecific hybridization: partners for adaptation, speciation and evolution in plants. *Ann. Bot.* 120, 183–194. doi: 10.1093/aob/mcx079

Berlin, K., Koren, S., Chin, C. S., Drake, J., Landolin, J. M., and Phillippys, A. M. (2015). Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.* 33, 623–630. doi: 10.1038/nbt.3238

Cardoso-Silva, C. B., Costa, E. A., Mancini, M. C., Balsalobre, T. W. A., Canesin, L. E. C., Pinto, L. R., et al. (2014). *De novo* assembly and transcriptome analysis of contrasting sugarcane varieties. *PLoS ONE* 9:e88462. doi: 10.1371/journal.pone.0088462

Chaisson, M. J., and Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* 13:238. doi: 10.1186/1471-2105-13-238

Chin, C. S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* 10, 563–569. doi: 10.1038/nmeth.2474

Costa, E. A., Anoni, C. O., Mancini, M. C., Santos, F. R. C., Marconi, T. G., Gazaffi, R., et al. (2016). QTL mapping including codominant SNP markers with ploidy level information in a sugarcane progeny. *Euphytica* 221, 1–16. doi: 10.1007/s10681-016-1746-7

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

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

de Setta, N., Monteiro-Vitorello, C. B., Metcalfe, C. J., Cruz, G. M. Q., Del Bem, L. E., Vicentini, R., et al. (2014). Building the sugarcane genome for biotechnology and identifying evolutionary trends. *BMC Genomics* 15:540. doi: 10.1186/1471-2164-15-540

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., et al. (2009). Real-time DNA sequencing from single polymerase molecules. *Science* 323, 133–138. doi: 10.1126/science.1162986

English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., et al. (2012). Mind the gap: upgrading genomes with pacific biosciences RS long-read sequencing technology. *PLoS ONE* 7:e47768. doi: 10.1371/journal.pone.0047768

Figueira, T. R. S., Okura, V., da Silva, F. R., da Silva, M. J., Kudrna, D., Ammiraju, J. S. S., et al. (2012). A BAC library of the SP80–3280 sugarcane variety (*Saccharum* sp.) and its inferred microsynteny with the sorghum genome. *BMC Res. Notes* 5:185. doi: 10.1186/1756-0500-5-185

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Syst. Biol.* 19, 99–113. doi: 10.2307/2412448

Garsmeur, O., Charron, C., Bocs, S., Jouffe, V., Samain, S., Couloux, A., et al. (2011). High homologous gene conservation despite extreme autopolyploid redundancy in sugarcane. *N. Phytol.* 189, 629–642. doi: 10.1111/j.1469-8137.2010.03497.x

Grivet, L., and Arruda, P. (2001). Sugarcane genomics: depicting the complex genome of an important tropical crop. *Curr. Opin. Plant Biol.* 5, 122–127. doi: 10.1016/S1369-5266(02)00234-0

Hoang, N. V., Furtado, A., Mason, P. J., Marquardt, A., Kasirajan, L., Thirugnanasambandam, P. P., et al. (2017). A survey of the complex transcriptome from the highly polyploid sugarcane genome using full-length isoform sequencing and *de novo* assembly from short read sequencing. *BMC Genomics* 18:395. doi: 10.1186/s12864-017-3757-8

Irvine, J. E. (1999). *Saccharum* species as horticultural classes. *Theor. Appl. Genet.* 98, 186–194. doi: 10.1007/s001220051057

Jannoo, N., Grivet, L., Chantret, N., Garsmeur, O., Glaszmann, J. C., Arruda, P., et al. (2007). Orthologous comparison in a gene-rich region among grasses reveals stability in the sugarcane polyploid genome. *Plant J.* 50, 574–585. doi: 10.1111/j.1365-313X.2007.03082.x

Kim, C., Lee, T.-H., Compton, R. O., Robertson, J. S., Pierce, G. J., and Paterson, A. H. (2013). A genome-wide BAC end-sequence survey of sugarcane elucidates genome composition, and identifies BACs covering much of the euchromatin. *Plant Mol. Biol.* 81, 139–147. doi: 10.1007/s11103-012-9987-x

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

Kim, M. Y., and Zilberman, D. (2014). DNA methylation as a system of plant genomic immunity. *Trends Plant Sci.* 19, 320–326. doi: 10.1016/j.tplants.2014.01.014

Kohany, O., Gentles, A. J., Hankus, L., and Jurka, J. (2006). Annotation, submission and screening of repetitive elements in Repbase: Repbase Submitter and Censor. *BMC Bioinformatics* 7:474. doi: 10.1186/1471-2105-7-474

Mattiello, L., Riaño-Pachón, D. M., Martins, M. C. M., Cruz, L. P., Bassi, D., Marchiori, P. E. R., et al. (2015). Physiological and transcriptional analyses of developmental stages along sugarcane leaf. *BMC Plant Biol.*. 15:300. doi: 10.1186/s12870-015-0694-z

Ming, R., Wang, W., Draye, X., Moore, H., Irvine, E., and Paterson, H. (2002). Molecular dissection of complex traits in autopolyploids: mapping QTL affecting sugar yield and related traits in sugarcane. *Theor. Appl. Genet.* 105, 332–345. doi: 10.1007/s00122-001-0861-5

Murray, S. C., Sharma, A., Rooney, W. L., Klein, P. E., Mullet, J. E., Mitchell, S. E., et al. (2008). Genetic improvement of Sorghum as a biofuel feedstock: I. QTL for stem sugar and grain nonstructural carbohydrates. *Crop Sci.* 48, 2165–2179. doi: 10.2135/cropsci2008.01.0016

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., et al. (2000). A whole-genome assembly of Drosophila. *Science* 287, 2196–2204. doi: 10.1126/science.287.5461.2196

Ohno, S. (1970). *Evolution by Gene Duplication*. Berlin; Heidelberg; New York, NY: Springer-Verlag.

Okura, V. K., Souza, R. S. C., Tada, S. F. S., and Arruda, P. (2016). BAC-Pool sequencing and assembly of 19 Mb of the complex sugarcane genome. *Front. Plant Sci.* 7:342. doi: 10.3389/fpls.2016.00342

Paterson, A. H., Bowers, J. E., and Chapman, B. A. (2004). Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl. Acad. Sci. U.S.A.* 101, 9903–9908. doi: 10.1073/pnas.0307901101

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., et al. (2009). The *Sorghum bicolor* genome and the diversification of grasses. *Nature* 457, 551–556. doi: 10.1038/nature07723

Riaño-Pachón, D. M., and Mattiello, L. (2017). Draft genome sequencing of the sugarcane hybrid SP80-3280 [version 2; referees: 2 approved]. *F1000Research* 6:861. doi: 10.12688/f1000research.11859.2

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A., et al. (2000). Artemis: sequence visualization and annotation. *Bioinformatics* 16, 944–945. doi: 10.1093/bioinformatics/16.10.944

Shiringani, A. L., Frisch, M., and Friedt, W. (2010). Genetic mapping of QTLs for sugar-related traits in a RIL population of *Sorghum bicolor* L. *Moench. Theor. Appl. Genet.* 121, 323–336. doi: 10.1007/s00122-010-1312-y

Vietor, D. M., and Miller, F. R. (1990). Assimilation, partitioning, and nonstructural carbohydrate in sweet compared with grain sorghum. *Crop Sci.* 30, 109–1115. doi: 10.2135/cropsci1990.0011183X003000050030x

Vilela, M. M., Del-Bem, L. E., Sluys, M. A. V., de Setta, N., Kitajima, J. P., Cruz, G. M. Q., et al. (2017). Analysis of three sugarcane homo/homeologous regions suggests independent polyploidization events of *Saccharum officinarum* and *Saccharum spontaneum*. *Genome Bio Evol.* 9, 266–278. doi: 10.1093/gbe/evw293

Visendi, P., Berkman, P. J., Hayashi, S., Golicz, A. A., Bayer, P. E., Ruperao, P., et al. (2016). An efficient approach to BAC based assembly of complex genomes. *Plant Methods* 12:2. doi: 10.1186/s13007-016-0107-9

Wood, T. E., Takebayashi, N., Barker, M. S., Mayrose, I., Greenspoon, P. B., and Rieseberg, L. H. (2009). The frequency of polyploid speciation in vascular plants. *Proc. Natl. Acad. Sci. U.S.A.* 106, 13875–13879. doi: 10.1073/pnas.0811575106

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 35, W265–W268. doi: 10.1093/nar/gkm286

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Anexo II

Declaração Bioética e/ou Biossegurança

COORDENADORIA DE PÓS-GRADUAÇÃO
INSTITUTO DE BIOLOGIA
Universidade Estadual de Campinas
Caixa Postal 6109. 13083-970, Campinas, SP, Brasil
Fone (19) 3521-6378. email: cpgib@unicamp.br

UNICAMP

IB

DECLARAÇÃO

Em observância ao **§5º do Artigo 1º da Informação CCPG-UNICAMP/001/15**, referente a Bioética e Biossegurança, declaro que o conteúdo de minha Tese de Doutorado, intitulada "***VARIAÇÃO GENÉTICA EM POLIPLOIDES COMPLEXOS: DESVENDANDO A DINÂMICA ALÉLICA EM CANA-DE-AÇÚCAR***", desenvolvida no Programa de Pós-Graduação em Genética e Biologia Molecular do Instituto de Biologia da Unicamp, não versa sobre pesquisa envolvendo seres humanos, animais ou temas afetos a Biossegurança.

Assinatura: _____
Nome do(a) aluno(a): Danilo Augusto Sforça

Assinatura: _____
Nome do(a) orientador(a): Anete Pereira de Souza

Data: Campinas, 05 de fevereiro de 2018

Anexo III

Declaração Direitos Autorais

Declaração

As cópias de artigos de minha autoria ou de minha co-autoria, já publicados ou submetidos para publicação em revistas científicas ou anais de congressos sujeitos a arbitragem, que constam da minha Dissertação/Tese de Mestrado/Doutorado, intitulada **VARIAÇÃO GENÉTICA EM POLIPLOIDES COMPLEXOS: DESVENDANDO A DINÂMICA ALÉLICA EM CANA-DE-AÇÚCAR**, não infringem os dispositivos da Lei n.° 9.610/98, nem o direito autoral de qualquer editora.

Campinas, 05 de Fevereiro de 2019

Assinatura : _____
Nome do(a) autor(a): **Danilo Augusto Sforça**
RG n.° 43.496.998-9

Assinatura : _____
Nome do(a) orientador(a): **Anete Pereira de Souza**
RG n.° 8.680.325