



MELINA CRISTINA MANCINI

CONSTRUÇÃO DE UM MAPA FUNCIONAL E DETECÇÃO DE QTLs DE IMPORTÂNCIA  
ECONÔMICA EM UMA POPULAÇÃO DERIVADA DE CRUZAMENTO BI-PARENTAL  
ENTRE DUAS VARIEDADES COMERCIAIS EM CANA-DE-AÇÚCAR

*FUNCTIONAL GENETIC MAP CONSTRUCTION AND QTL OF ECONOMIC IMPORTANCE  
DETECTION IN A DERIVED BI-PARENTAL CROSS BETWEEN TWO COMMERCIAL  
SUGARCANE VARIETIES*

Campinas

2014





UNIVERSIDADE ESTADUAL DE CAMPINAS  
Instituto de Biologia



MELINA CRISTINA MANCINI  
CONSTRUÇÃO DE UM MAPA FUNCIONAL E DETECÇÃO DE QTLs DE IMPORTÂNCIA  
ECONÔMICA EM UMA POPULAÇÃO DERIVADA DE CRUZAMENTO BI-PARENTAL  
ENTRE DUAS VARIEDADES COMERCIAIS EM CANA-DE-AÇÚCAR

*FUNCTIONAL GENETIC MAP CONSTRUCTION AND QTL OF ECONOMIC IMPORTANCE  
DETECTION IN A DERIVED BI-PARENTAL CROSS BETWEEN TWO COMMERCIAL SUGARCANE  
VARIETIES*

Tese apresentada ao Instituto de Biologia da  
Universidade Estadual de Campinas como parte dos  
requisitos exigidos para a obtenção do título de Doutora  
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*

Orientadora/ Supervisor: Profa. Dra. Anete Pereira de Souza

Co-orientador/ Co-supervisor: Dr. Antonio Augusto Franco Garcia

Este exemplar corresponde à versão final da tese  
defendida pela aluna Melina Cristina Mancini e orientada  
pela Profa. Dra. Anete Pereira de Souza.

Profa. Dra. Anete Pereira de Souza

CAMPINAS

2014

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

M312c Mancini, Melina Cristina, 1983-  
Construção de um mapa funcional e detecção de QTLs de importância econômica em uma população derivada de cruzamento bi-parental entre duas variedades comerciais em cana-de-açúcar / Melina Cristina Mancini. – Campinas, SP : [s.n.], 2014.

Orientador: Anete Pereira de Souza.  
Coorientador: Antonio Augusto Franco Garcia.  
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Biologia.

1. Mapeamento de QTL. 2. Marcadores moleculares. 3. *Saccharum*. 4. Melhoramento genético. I. Souza, Anete Pereira de, 1962-. II. Garcia, Antonio Augusto Franco. III. Universidade Estadual de Campinas. Instituto de Biologia. IV. Título.

Informações para Biblioteca Digital

**Título em outro idioma:** Functional genetic map construction and QTL of economic importance detection in a derived bi-parental cross between two commercial sugarcane varieties

**Palavras-chave em inglês:**

QTL mapping

Molecular markers

*Saccharum*

Genetic breeding

**Área de concentração:** Genética Vegetal e Melhoramento

**Titulação:** Doutora em Genética e Biologia Molecular

**Banca examinadora:**

Anete Pereira de Souza [Orientador]

Gabriel Rodrigues Alves Margarido

Mariângela Cristofani Yaly

Eugênio Cesar Ulian

Jurandir Vieira de Magalhães

**Data de defesa:** 04-07-2014

**Programa de Pós-Graduação:** Genética e Biologia Molecular

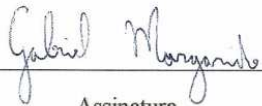
Campinas, 04 de Julho de 2014

Banca Examinadora


Profa. Dra. Anete Pereira de Souza (orientadora)

  
Assinatura

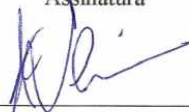
Prof. Dr. Gabriel Rodrigues Alves Margarido

  
Assinatura

Profa. Dra. Mariângela Cristofani Yaly

  
Assinatura

Prof. Dr. Eugênio Cesar Ulian

  
Assinatura

Prof. Dr. Jurandir Vieira de Magalhães

  
Assinatura

Prof. Dr. Michel Georges Albert Vincentz

\_\_\_\_\_  
Assinatura

Profa. Dra. Mirian Perez Maluf

\_\_\_\_\_  
Assinatura

Dr. Hermann Paulo Hoffmann

\_\_\_\_\_  
Assinatura



## Resumo

A crescente busca por variedades de cana-de-açúcar com maior produtividade e resistentes às principais doenças consiste em um importante objetivo para o sucesso de um programa de melhoramento. Assim, a utilização de marcadores moleculares na identificação de locos que controlam características quantitativas (QTLs – *Quantitative Trait Loci*) vêm ganhando cada vez mais destaque no programas de melhoramento genético. A presente tese teve como objetivo contribuir para o conhecimento básico sobre a genética e a biologia molecular da cana-de-açúcar através da detecção de marcadores ligados a características quantitativas. Foi utilizado uma população de cana-de-açúcar contendo 240 indivíduos F<sub>1</sub> derivada do cruzamento entre as variedades comerciais SP81-3250 e RB925345. Para detectar os QTLs foi necessário realizar estudos fenotípicos e genotípicos. Foram coletados dados fenotípicos para as características de produção (altura, diâmetro, número e peso dos colmos) e de qualidade (sólidos solúveis, teor de sacarose do caldo e do colmo, pureza do caldo, teor de fibra) por três anos (2011, 2012 e 2013) nos municípios de Araras e Ipaussu, estado de São Paulo. Através de modelos mistos foi estimada a média, matriz de variância e covariância (VCOV), herdabilidade e a correlação fenotípica entre as características. Os resultados apresentados mostraram um ótimo controle ambiental, com menor valor de herdabilidade para pureza (0,77), além de 30 correlações fenotípicas significativas, confirmando que estes dados podem ser utilizados na detecção dos QTLs. Os dados genotípicos foram obtidos através da análise das regiões contendo microssatélites e de variações genéticas de único nucleotídeo, pelos marcadores SSR (*Simple Sequence Repeat*) e SNP (*Single Nucleotide Polymorphism*), respectivamente. A genotipagem dos SNPs foi realizada por espectrometria de massa pela Plataforma Sequenom MassARRAY® (Sequenom Inc., San Diego, California, USA). A análise foi realizada utilizando o programa SuperMASSA, que possibilitou estimar a ploidia dos locos. Assim, as marcas SNPs foram utilizadas na detecção dos QTLs para as características de produção e de qualidade. Por regressão linear foram encontradas 17 evidências de associação de QTL entre diâmetro dos colmos (quatro evidências), número de colmos (uma evidência), peso dos colmos (uma evidência), conteúdo de sólidos solúveis (duas evidências), teor de sacarose do caldo (três evidências), pureza (duas evidências), toneladas de cana por hectare (duas evidências) e toneladas de Pol por hectare (duas evidências). A proporção da variação fenotípica explicada pelo genótipo variou de 1,6% a 11,1%.

Todos os SNPs que apresentaram associações com as características mencionadas tiveram os níveis de ploidia variando de hexaploide a dodecaploide. Por correlação genotípica-fenotípica, foi detectado sete evidências de associação de QTL entre diâmetro dos colmos (uma evidência), conteúdo de sólidos solúveis (duas evidências), teor de sacarose da cana (uma evidência), teor de sacarose do caldo (duas evidências) e pureza (uma evidência). Os SNPs detectados com correlações genotípica-fenotípica significativas apresentaram níveis de ploidia variando tetradecaploide a icosaploide. As diferentes ploidias permitiu a detecção de QTLs em multi-dose e podem ser usadas como informações prévias sobre os prováveis QTLs, contribuindo para o avanço do conhecimento da genética da cana-de-açúcar.



## Abstract

The increasing search for sugarcane varieties with higher productivity and resistant to major diseases is an important goal for the success of Sugarcane Breeding Program. Thus, molecular markers can be used to identify Quantitative Trait Loci (QTLs) and have become a powerful tool in Breeding Programs. This thesis aimed to contribute for the basic knowledge of genetics and molecular biology in sugarcane through detection of markers linked to quantitative traits. Was used a sugarcane population consisted of 240 F<sub>1</sub> individuals derived from a cross between SP81-3250 and RB925345. To detect the QTLs it was necessary to perform phenotypic and genotypic studies. The phenotypic data were made for cane yield (stalk diameter, stalk height, stalk number, stalk weight and tons of cane per hectare) and quality traits (soluble solid content, sucrose content, juice sucrose content, purity, fiber and tons of Pol per hectare) for three harvest years (2011, 2012 and 2013) in Araras and Ipaussu cities, located in the state of São Paulo. The average, variance and covariance matrix (VCOV), heritability and phenotypic correlation was estimated via mixed models. All results showed a great environmental control, the lowest heritability was purity (0.77), besides 30 significant phenotypic correlations, confirming that these data can be used for QTLs detection. The genotypic data were obtained analyzing the regions containing microsatellites and single nucleotide genetic variants, by the markers SSR (Simple Sequence Repeat) and SNP (Single Nucleotide Polymorphism), respectively. The SNPs genotyping were performed via mass spectrometry by Sequenom MassARRAY<sup>®</sup> platform (Sequenom Inc., San Diego, California, USA). The analysis was performed using the SuperMASSA software allowing to estimate the loci ploidy. The SNPs markers were used for QTL detection for cane yield and quality traits. By linear regression 17 QTL association evidences were found for stalk diameter (four evidences), stalk number (one evidence), stalk weight (one evidence), soluble solid content (two evidences), juice sucrose content (three evidences), purity (two evidences), tons of cane per hectare (two evidences) and tons of Pol per hectare (two evidences). The phenotypic variation explained by genotype ranged from 1.6% to 11.1%. The SNPs associated with the traits mentioned had ploidy levels ranging from hexaploid to dodecaploide. Via genotypic-phenotypic correlation, it was detected seven QTL evidence of association for stalk diameter (one evidence), soluble solid content (two evidences), sucrose content (one evidence), juice sucrose content (two evidences) and purity (one

evidence). The SNPs detected significant genotypic-phenotypic correlations showed ploidy levels ranging from tetradecaploide to icosaploide. The different ploidies allowed the detection of QTLs in multi-dose and can be used as prior information about QTL mapping, contributing to the advancement of the sugarcane genetics knowledge.

## Sumário

---

Dedicatória .....	Xiii
Agradecimentos .....	Xv
Organização da tese .....	Xvii
Introdução .....	1
Objetivos .....	13
Objetivo geral .....	13
Objetivos específicos .....	13
Capítulo I .....	15
“Uso de um modelo estatístico sofisticado para cana-de-açúcar usando características fenotípicas”	
Capítulo II .....	35
“Análise de marcas individuais utilizadas na detecção de QTLs considerando diferentes ploidias em cana-de-açúcar”	
Considerações gerais .....	57
Resumo dos resultados .....	63
Conclusões .....	65
Perspectivas .....	67
Literatura citada .....	69
Anexo I .....	77
Anexo II .....	83
Anexo III .....	111
Anexo IV .....	121
Anexo V .....	131
Anexo VI .....	135



À minha querida avó, Nunciata, que sempre estará comigo.

Aos meus pais, Solange e José Carlos, com todo meu amor,

Dedico



## **Agradecimentos**

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo apoio financeiro no projeto e pela bolsa de doutorado concedida.

Aos meus pais, Solange e José Carlos, exemplos para toda minha vida. Com eles aprendi que nada na vida acontece sem dedicação, disciplina, esforço e perseverança.

À minha irmã Thalita, cunhado Andre, sobrinho Luiz Francisco e avô Annibal, por estarem ao meu lado e aguentarem meu mau humor.

À Prof Dra. Anete Pereira de Souza, por me orientar, profissionalmente e pessoalmente, pela confiança, oportunidade e apoio. Obrigada professora, pelo exemplo e por investir em mim. Com a senhora aprendi a dar o exato valor às coisas boas e não me abalar com aquelas que não saíram como planejei.

Ao meu co-orientador Prof. Dr. Antonio Augusto Franco Garcia, do Departamento de Genética da ESALQ-USP, pela contribuição no crescimento científico, e à seus alunos, em especial, Carina Anoni, Guilherme Pereira e Marcelo Molinari.

Ao Programa de Melhoramento Genético da Cana-de-Açúcar/UFSCar (PMGCA), da Rede Interuniversitária para o Desenvolvimento do Setor Sucroalcooleiro (RIDESA) e a usina Raízen, Unidade Ipaussu, que proporcionaram todo o suporte técnico necessário para realização de um experimento dessa grandeza, sempre cumprindo todos os prazos previamente estipulados.

À Prof Dra. Monalisa Carneiro Sampaio pela disponibilidade do material vegetal.

Aos membros da pré-banca e da banca de defesa, pela contribuição na melhoria deste trabalho.

Ao Prof Dr. Dilermando Perecin, por ter me apresentado este organismo tão complexo e intrigante que é a cana-de-açúcar e a Dra. Luciana Rossini Pinto por ter paciência de me ensinar todas as diferenças genéticas existentes entre a cana-de-açúcar com os outros organismos.

À Aline, Carlão, Danilo, Juverlande e Pathy, funcionários do LAGM, pelo carinho e respeito.

À Tânia e Sandra secretárias do Centro de Biologia Molecular e Engenharia Genética (CBMEG) da UNICAMP, obrigada pela prontidão meninas.

Aos amigos do Lab, turma do café, animadores do Espaço Gourmet, galera do Kabana, obrigada pelos momentos de descontração. Vocês não sabem quanto foram importantes para mim.

Ao grupo cana-de-açúcar, Benício, Danilo, Estela e Thiago, obrigada pelas discussões e valiosa ajuda.

Ao Fabão, vulgo Fábio Matos, pelas inúmeras PCR's que guardou para mim.

Ao Carlos e Elisa, Livia, Guilherme, Patrícia e Rafaela, vocês são a família que escolhi. Com vocês passei bons e maus momentos. É impossível descrever em palavras para o quanto são importantes na minha vida.

Ao gtalk, Skype e Facetime por me fazer ficar mais próxima às pessoas que amo.

Dizem que o importante na vida é fazer o que gosta para se divertir, acima de tudo, então Alininha, Benny, Cacazinha, Dandanzinho, Elisa e Carlos, Estelinha, Gui-Gu Gu-Gui, Juver, Lilizinha, Maumau, Patinha, Pathy, Pripri e Rafases, obrigada minha gente!!!



## Organização da Tese

---

A presente tese teve como objetivo contribuir para o melhor entendimento genético da cana-de-açúcar através da detecção de marcadores moleculares ligados à características de importância econômica em cana-de-açúcar (*Saccharum spp*). Além de sua importância econômica, é amplamente utilizada em estudos científicos devido a sua complexa organização genômica. O fato das cultivares modernas serem derivadas entre cruzamentos interespecíficos, o organismo resultante apresenta alto nível de ploidia e aneuploidia. O trabalho aqui realizado foi de caráter multi-institucional e está inserido em um programa de pesquisa em genética e melhoramento de cana-de-açúcar. Os resultados obtidos durante o período de doutoramento estão apresentados no formato de dois artigos, ainda não publicados.

O capítulo I descreve os resultados das análises de fenotipagem procedida na população de mapeamento entre as variedades SP81-3250 e RB925345 em dois locais (Araras e Ipaussu, ambos no estado de São Paulo) durante três anos (2011, 2012 e 2013). O coeficiente de variação, a herdabilidade e a correlação fenotípica foram calculados através de um modelo misto para 11 diferentes características (diâmetro, altura, número e peso dos colmos, sólidos solúveis (Brix), teor de sacarose da cana e do caldo, pureza, fibra, TCH (Toneladas de Cana por Hectare) e TPH (Toneladas de Pol por Hectare)). Os valores encontrados para herdabilidade foram altos, variando entre 77% (pureza) e 96% (TCH), comprovando a excelência dos dados para a posterior utilização na detecção de QTLs (*Quantitative Trait Loci*). A este artigo serão futuramente agregados dados de outra população de mapeamento, derivada do cruzamento entre as variedades SP80-3280 e RB835486, que foi avaliada seguindo os mesmos critérios e locais, a qual ainda encontra-se em fase de análise. O capítulo II utiliza a análise de todos os dados referentes a características fenotípicas para a detecção de QTLs por marcas individuais. Foram genotipados 290 marcadores SNPs (*Single Nucleotide Polymorphisms*), e um total de 17 evidências de associação de QTL foi encontrado através de regressão linear e sete SNPs foram associados através da correlação genotípica-fenotípica, significando que os marcadores SNPs podem ser promissores quando aplicados aos programas de melhoramento de cana-de-açúcar. Para realizar a genotipagem dos SNPs foi necessário efetuar uma série de ajustes na técnica, que encontram-se discutidos no Anexo I.

Durante todo o período de doutorado, foi realizado um estudo paralelo de mapeamento genético e mapeamento de QTL em cana-de-açúcar de um cruzamento bi-parental pertencente ao IAC (Instituto Agrônomo de Campinas), Centro de Cana, localizado na cidade de Ribeirão Preto, estado de São Paulo. Foi construído um mapa genético, utilizando 634 marcas segregando em dose única, com cobertura total de 4370 cM. O mapa genético foi utilizado para mapear os QTLs, através de mapeamento por intervalo composto. Foram mapeados 19 QTLs para diâmetro e peso dos colmos, fibra, quantidade de sólidos solúveis (Brix) e teor de sacarose (Pol). Esses resultados encontram-se no Anexo II, em forma de artigo, intitulado em “*Mapping Quantitative Trait Loci for yield components in a bi-parental cross between two commercial sugarcane (Saccharum spp.) varieties*” a ser submetido à revista *Molecular Breeding*.

Os genitores desta população também foram alvo de estudos envolvendo sequenciamento do transcriptoma de folhas, o que resultou no artigo científico “*De Novo Assembly and Transcriptome Analysis of Contrasting Sugarcane Varieties*” publicado no periódico *Plos One* (DOI: 10.1371/journal.pone.0088462). Este artigo encontra-se no Anexo III e refere-se à análise do transcriptoma de folhas de seis variedades de cana-de-açúcar envolvidos em cruzamentos bi-parentais (IACSP96-3046 e IACSP95-3018, SP81-3250 e RB925345, e SP80-3280 e RB835486). Através de sequenciamento em larga escala de cDNA e montagem *de novo*, foi possível obter genes únicos para cana-de-açúcar além de identificar um grande número de marcadores moleculares microssatélites (SSR) e polimorfismo de base única (SNP).

O grupo esteve inserido em um trabalho realizado com genotipagem de cana-de-açúcar por meio de SNPs, ajudando no conhecimento sobre as otimizações e principalmente na análise dos dados da genotipagem dos SNP. Os resultados foram publicados no periódico *Scientific Reports* (DOI: 10.1038/srep03399), intitulado “*SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids*”, anexo IV.

A tese ainda conta com uma discussão envolvendo todos os resultados obtidos ao longo do desenvolvimento do trabalho, resumo dos resultados e uma conclusão geral. A grande complexidade do genoma da espécie representa um estimulante desafio para geneticistas de espécies poliploides, em especial quando é alcançado conhecimento inédito, como é o caso da presente tese. Muito ainda deve ser feito para que as informações sejam realmente aplicadas aos programas de melhoramento genético de cana-de-açúcar, que serão discutidas nas perspectivas para a continuidade do trabalho.

## Introdução

---

### O gênero *Saccharum* e seu melhoramento ao longo da história

A cana-de-açúcar (*Saccharum spp*) tem como centro de origem a Ásia, destacando Nova Guiné, China e Índia como centros de maior diversidade (Roach e Daniels 1987). Ao longo de anos se disseminou para África e Europa. No apogeu da navegação portuguesa (século XV) ocorreu uma fase de expansão para ilhas do Atlântico, figurando como centro difusor para as Américas (Figura 1). O cultivo da cana-de-açúcar no Brasil iniciou cerca de 30 anos após sua descoberta, expandindo-se rapidamente ao Nordeste do país, devido ao clima favorável e solo fértil, dando início a um monopólio mundial de produção de açúcar ao Brasil (Figueiredo 2008).

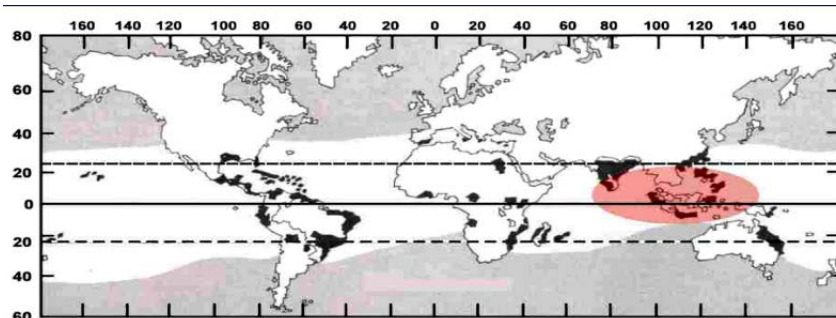


Figura 1. Áreas de cultivo de cana-de-açúcar no mundo destacadas em preto. Centro de origem, destacado em vermelho (Fonte: Ometto 1982).

É uma planta perene, alógama, autopoliploide, isto é, resultado da duplicação de um único genoma, cultivada principalmente em regiões tropicais e subtropicais e apresenta hábito de perfilhamento, característica que influencia todo o manejo da cultura e confere sua propagação vegetativa. Pertence à divisão Magnoliophyta, classe Liliopsita, ordem Cyperales, família Poaceae, tribo Andropogoneae, subtribo Saccharininae, gênero *Saccharum*, contendo seis espécies: *Saccharum officinarum* ( $2n=80$ ), *S. barberi* ( $2n=81-124$ ), *S. robustum* ( $2n=60-250$ ), *S. spontaneum* ( $2n=40-128$ ), *S. sinensis* ( $2n=111-120$ ) e *S. edule* ( $2n=60-80$ ) (Cronquist 1981). Híbridões interespecíficos entre *S. officinarum* e *S. spontaneum* seguido de sucessivos retrocruzamentos com *S. officinarum*, processo denominado “nobilização” ou domesticação, mudaram o cenário do cultivo de cana-de-açúcar, já que antes, apenas a espécie *S. officinarum* era utilizada devido ao alto acúmulo

nos níveis de sacarose. Este processo ocasionou um grande avanço no melhoramento genético da cultura, pois gerou maior variabilidade para a cana-de-açúcar. Assim surgiram as variedades modernas ainda cultivadas atualmente, que concentram as características de alto acúmulo de sacarose e resistência a pragas e doenças (Roach 1972, Ming *et al.* 2006). Essas variedades apresentam a maior parte de seu genoma, cerca de 80%, originado do *S. officinarum* (D'Hont *et al.* 1996).

Historicamente, o processo de melhoramento genético da cana-de-açúcar é antigo, tendo seu início no próprio centro de origem, baseado em domesticação e seleção visual. Ao longo dos anos passou a sofrer maior interferência do homem ao realizar hibridações entre as espécies, em especial entre *S. officinarum* e *S. spontaneum*. Um programa de melhoramento de cana-de-açúcar tem início na obtenção de variabilidade genética através do cruzamento, geralmente, entre cultivares comerciais ou clones pré-comerciais que sejam bastante distintos entre si e com bons atributos agronômicos (Cruz *et al.* 2004), seguido por diversos níveis de seleção até identificar e selecionar o genótipo superior (Souza Júnior 1989, Souza Júnior 1995, Bressiani 2001). Deste modo, a obtenção de uma variedade de cana-de-açúcar exige longo tempo, normalmente 12 a 15 anos, desde a escolha dos parentais até o plantio em escala comercial (Landell *et al.* 1999). Em razão da cana-de-açúcar ser uma espécie alógama, autopoliploide e de propagação vegetativa, existe uma grande variabilidade genética já na primeira geração de cruzamento ( $F_1$ ), originando progênes altamente heterozigóticas. Devido à propagação vegetativa, há uma fixação do potencial genético individual do clone, não havendo alteração genética ao longo das gerações (Borém e Miranda 2005). Entretanto, a autopoliploidia associado à hibridização interespecífica entre *S. officinarum* e *S. spontaneum*, aneuploidia e variação dos eventos de recombinação resulta em uma grande complexidade de seu genoma (Grivet e Arruda 2001).

A cana-de-açúcar é considerada uma cultura de grande valor econômico, assim como seus principais derivados açúcar e etanol, que geram milhões de reais por ano e colocam o país como o maior produtor destes produtos. No Brasil, o agronegócio da cana-de-açúcar constitui um dos setores de maior geração de empregos diretos e indiretos. Segundo dados da Companhia Nacional de Abastecimento (CONAB), estima-se que a área colhida destinada a atividade sucroalcooleira seja de 9,5 milhões de hectares (Unica 2014). Aliado a isso, a previsão total de cana que será moída na safra 2013/2014 é de 625 milhões

de toneladas, sendo a produtividade média brasileira estimada em 80 toneladas/ha (Unica 2014).

### **Marcadores moleculares**

Marcadores moleculares são 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). As causas da variação genética entre os indivíduos são atribuídas às recombinações gênicas durante a meiose e pelas mutações, sendo a causa mais comum, a troca de um nucleotídeo por outro. Uma vez conhecida as causas das variações alélicas elas podem ser classificadas quanto às diferenças no número de sequências repetidas em tandem em um determinado loco (regiões de microsatélites ou SSR - *Simple Sequence Repeat*), pelas inserções ou deleções (InDels) de parte do genoma e também no polimorfismo de base única ou SNP (*Single Nucleotide Polymorphism*).

Nos últimos anos houve um grande avanço dos marcadores moleculares, contribuindo para o desenvolvimento de estudos realizados pela comunidade científica. Os marcadores moleculares podem ser divididos em três grupos: (1) os de baixo rendimento, que são aqueles baseados em hibridização incluindo o RFLP (*Restriction Fragment Length Polymorphism*); (2) os que apresentam rendimento médio, constituído pelos marcadores baseados em PCR, incluindo o RAPD (*Random Amplified Polymorphic DNA*), AFLP (*Amplified Fragment Length Polymorphism*) e SSRs; (3) e os marcadores com alto rendimento baseados em sequências, incluindo os SNPs.

Dentre suas aplicações no melhoramento genético vegetal, está a seleção assistida de fenótipos de importância agrônômica. Uma vez mapeados e identificados, tais fenótipos podem ser selecionados indiretamente por marcadores moleculares diretamente ligados a eles, através da seleção assistida por marcadores (SAM). A identificação da ligação entre marcador e a característica de interesse é um pré-requisito para a aplicação da SAM (Grupta *et al.* 1999, Morgante e Salamini 2003, Charcosset e Moreau 2004, Pinto *et al.* 2009). Considerando que um programa de melhoramento de cana-de-açúcar leva no mínimo 10 anos para lançar uma nova cultivar, os marcadores moleculares podem representar uma importante ferramenta, já que reduzem o tempo de lançamento de novas

cultivares e podem reunir diversas características agronômicas desejáveis na mesma planta (Pinto *et al.* 2009).

Em genomas eucariotos, as sequências de DNA contendo microssatélites são muito frequentes e distribuídas ao acaso, além de serem locos genéticos altamente polimórficos (Ferreira e Grattapaglia 1998). Algumas de suas aplicações são observadas em estudos de diversidade genética e desequilíbrio de ligação (Hao *et al.* 2011, Chen *et al.* 2012), na construção de mapas genéticos e na identificação de QTLs (Ting *et al.* 2012, Yu *et al.* 2011) e em estudos de transferibilidade entre espécies correlacionadas (Wang *et al.* 2005). A exploração dos marcadores com alvo nas sequências repetitivas apresenta uma limitação, o espaçamento médio entre estas sequências em diferentes espécies vegetais é de aproximadamente 6 a 7 Kb (Cardle *et al.* 2000). Porém, esta limitação é suprida por sua natureza multialélica, o que torna estes locos altamente informativos.

Existem dois tipos de marcadores microssatélites: (1) os genômicos, obtidos de sequências randômicas do DNA, (2) e aqueles oriundos de sequências expressas, denominados de EST-SSRs (*Expressed Sequence Tags*) ou marcadores funcionais. Os EST-SSRs são menos informativos devido a sua origem, porém há uma maior probabilidade de estarem geneticamente associados a uma característica fenotípica avaliada, além de apresentarem homologia com genes candidatos, facilitando o mapeamento de QTLs. Desta forma, são considerados marcadores ideais para SAM em programas de melhoramento genético (Liu *et al.* 2012).

Os SNPs são variações na sequência de DNA que ocorrem quando um único nucleotídeo na sequência do genoma é alterado. São considerados a forma mais abundante de variação, e encontram-se dispersos ao longo de todo o genoma (Brookes 1999). A primeira observação das variações pontuais no DNA foi constatada com o sequenciamento do genoma humano ao se comparar segmentos correspondentes. Em humanos, estima-se que ocorra um a cada 1.000 nucleotídeos (Collins *et al.* 1998). Estudos posteriores mostram que a frequência de SNPs em algumas plantas pode ser maior do que a encontrada no genoma humano, como é o caso da batata, que apresenta um SNP a cada 21 pb (Rickert *et al.* 2003), do trigo, com um SNP a cada 370 pb (Khlestina e Salina 2006), do sorgo (planta mais próxima geneticamente da cana-de-açúcar) com pelo menos três SNPs a cada Kb (Feltus *et al.* 2004). Por apresentarem distribuição abundante, os SNPs vêm sendo

aplicados na construção de mapas genéticos de alta resolução (Ganal *et al.* 2011), na identificação de cultivares (Cabezas *et al.* 2011), em estudos de mapeamento por associação com características de interesse econômico (Poland *et al.* 2011), entre outras.

Existem dois tipos de mutação, as classificadas como de transição, que são as mais comuns, em que há troca de uma purina por outra purina (A/G) ou de uma pirimidina por outra pirimidina (C/T), e as do tipo transversões, menos frequentes e acontecem quando há troca de uma purina por uma pirimidina, ou pirimidina por purina (C/G, C/A, T/G ou T/A). Teoricamente qualquer um dos quatro nucleotídeos pode estar envolvido na variação em uma posição do genoma de uma espécie. Entretanto na prática, observa-se que os SNPs se comportam como marcadores bialélicos, e as mutações com maior incidência ocorrem entre apenas dois nucleotídeos (Brookes 1999). Este comportamento faz com que um único loco SNP apresente conteúdo informativo menor se comparado com marcadores multialélicos (Gupta *et al.* 2001), deste modo é necessária a utilização de um maior número de SNPs para assegurar a cobertura de todo o genoma.

Os SNPs ocorrem em regiões codificadoras e não codificadoras nos genomas. Quando ocorre uma substituição de aminoácido na sequência proteica em regiões codificadoras, ela pode ser classificada como não-sinônima. Nesse caso há mudança de aminoácido e pode ocorrer modificações estruturais e funcionais na proteína. Se a substituição não alterar o aminoácido formado, é considerada sinônima. Embora SNPs sinônimos não alterem a sequência proteica, eles podem modificar a estrutura e a estabilidade do RNA mensageiro (Kwok 1999).

Existem diferentes metodologias utilizadas na identificação de SNPs. Entre elas estão o sequenciamento direto do DNA ou RNA, a busca *in silico* em banco de dados disponíveis, e também através de métodos bioquímicos que investigam a presença de variantes. Após a identificação de SNPs é necessário determinar sua frequência em um grupo de indivíduos, certificando-se de sua natureza polimórfica. O rápido desenvolvimento dos métodos de genotipagem de SNPs vem acontecendo devido a redução dos custos assim como o aumento da eficiência dos métodos. Destacam-se a utilização da espectrometria de massas (MALDI-TOF MS- *Matrix Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry*) (Bundock *et al.* 2009), de PCR em tempo real e de microarranjos (Syvänen 2001), além de metodologias para genotipar dezenas de milhares

de SNPs em um único ensaio utilizando as tecnologias de nova geração (Davey *et al.* 2011).

O grande diferencial dos marcadores SNPs é sua aplicação nas espécies poliploides. A maioria dos trabalhos envolvendo estas espécies utilizam marcadores moleculares em dose única SDRF (*Single Dose Restriction Fragment*) (Wu *et al.* 1992). Estes marcadores baseiam-se apenas na presença ou ausência do alelo. Desta forma, mesmo os marcadores co-dominantes, funcionam como dominantes (da Silva *et al.* 1995), significando perdas de informações sobre outras dosagens. Neste cenário, os marcadores SNPs são mais informativos, já que não só determinam a existência de polimorfismo de um loco SNP, como também são capazes de determinar sua frequência em diferentes genótipos. No caso especial da cana-de-açúcar a genotipagem de locos SNPs envolve a identificação do polimorfismo e o número de cópias alélicas (frequência de cada alelo que possui o loco SNP). Desta forma, um SNP será polimórfico entre dois genótipos quando diferir quanto à base analisada ou ao número de cópias alélicas presente em cada um dos genótipos.

Os métodos de espectrometria de massa têm sido geralmente utilizados para detectar e quantificar com precisão alelos com frequências muito baixas (Oberacher 2008). A técnica de espectrometria de massa MALDI-TOF vem sendo aplicada em estudos envolvendo DNA e se mostrou uma metodologia poderosa na genotipagem de SNPs. Apresenta algumas vantagens sobre os métodos tradicionais de análise de DNA: (1) é um método extremamente preciso, pois os resultados gerados não se baseiam no comprimento dos alelos, mas sim a partir do peso molecular de cada molécula, uma propriedade física intrínseca da molécula, excluindo a utilização de padrões de tamanho; (2) os experimentos também excluem a necessidade da utilização de géis, permitindo que a separação e detecção de variações no DNA seja significativamente mais rápida se comparada aos métodos tradicionais de eletroforese (Tost e Gut 2003). A velocidade da aquisição de um sinal completo leva cerca de 100 microsegundos; (3) e, finalmente, a técnica possibilita a análise de múltiplas reações (multiplex), além da automação completa de todos os procedimentos, desde a preparação das amostras até a aquisição e processamentos dos dados. Esta característica assegura uma análise dos ácidos nucléicos em larga escala (Griffin e Smith 2000), conferindo um alto rendimento, o que significa tempo e custos reduzidos.



A resolução da atual geração de espectrômetros de massa MALDI-TOF permite a fácil distinção da substituição de nucleotídeos com variação de massa de 1-7 Daltons, o que corresponde ao tamanho de DNA de 3-25 nucleotídeos (Gut 2004). Isto significa que em um indivíduo heterozigoto é possível distinguir as seis possibilidades bialélicas do primer de extensão: A/C, A/G, A/T, C/G, C/T e G/T. A variação mais fácil de ser resolvida se dá entre os nucleotídeos G e C, pois a diferença é de 40 Daltons. O maior desafio da genotipagem bialélica é a configuração A/T, porque os nucleotídeos A e T diferem-se em apenas 9 Daltons. Enquanto os SNPs bialélicos de maiores frequências de ocorrência são A/G e C/T, diferindo de 16 e 15 Daltons, respectivamente (Haff *et al.* 2001).

A identificação de marcadores SNPs em plantas com genoma complexo apresenta uma série de obstáculos. O maior deles é a alta repetitividade de regiões do genoma (Meyers *et al.* 2001) e como evitar tais porções. Algumas abordagens vêm sendo utilizadas, entre elas o re-sequenciamento pelo método Sanger (Wright *et al.* 2005), que é caro e laborioso, assim como o desenvolvimento *in silico*, através da mineração de banco de dados de EST (Batley *et al.* 2003). Como muitas culturas poliploides apresentam muitas regiões do genoma duplicadas, a mineração em banco de dados de EST resulta em um grande número de SNPs não polimórficos, representando sequências parálogas, consideradas pouco úteis para o melhoramento molecular. Em ambas as estratégias a frequência de desenvolvimento de SNPs é baixa por usar regiões codificantes do genoma.

### **Mapeamento genético com ênfase em cana-de-açúcar**

Os mapas genéticos, ou mapas de ligação, identificam a posição de genes ou marcadores correspondente à sua ordem linear nos cromossomos, tornando-os uma ferramenta fundamental para estudos genéticos, com importante aplicação no melhoramento genético. Os primeiros mapas publicados foram baseados em marcadores morfológicos e citológicos, seguido por aqueles construídos através das isoenzimas e por fim, pelos construídos por marcadores de DNA (Carneiro e Vieira 2002). A disponibilidade de marcadores moleculares juntamente com eficientes métodos de análise dos dados permitiu uma popularização dos mapas genéticos entre a grande maioria das espécies vegetais, incluindo as que apresentam ciclo longo de vida, visto que antes era restrito às espécies com marcadores morfológicos disponíveis (Carneiro e Vieira 2002). Novos

marcadores moleculares foram surgindo ao longo dos anos, fazendo com que o rendimento da genotipagem aumentasse. 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), permitindo um melhor entendimento de fenômenos importantes para a genética e o melhoramento.

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 F<sub>2</sub>, 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 *et al.* 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: (1) ao alto nível de ploidia associado a aneuploidia, resultando em um complexo padrão de segregação cromossômica durante a meiose (Heinz e Tew 1987), (2) 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 (3) as fases de ligação entre os marcadores são desconhecidas (Pastina *et al.* 2012). Contornando esta situação, Wu *et al.* (1992) propuseram a utilização de marcadores SDRF 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 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 *et al.* (2006) propôs 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 entre eles (Maliepaard *et al.* 1997). Esta estratégia de mapa integrado foi utilizado por Oliveira *et al.* (2007). Os mapas genéticos integrados apresentam algumas vantagens, tais como mapas mais saturados, além de estimar uma melhor localização dos QTLs assim como as fases de ligação de maneira mais eficiente se comparado com a estratégia do *pseudo-testcross*. Porém, a utilização da abordagem de marcadores em dose única consegue acessar apenas uma sub amostragem do genoma, resultando em mapas pouco saturados, além de inviabilizar o estudo de dosagem alélica. 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).

Na literatura encontra-se diversos mapas genéticos disponíveis, destacando-se alguns desenvolvidos para as espécies *S. spontaneum* (Al-Janabi *et al.* 1993, da Silva *et al.* 1993, Ming *et al.* 1998, Ming *et al.* 2002), *S. officinarum* (Al-Janabi *et al.* 1993, Mudge *et al.* 1996, Guimarães *et al.* 1999, Ming *et al.* 1998) e variedades comerciais (D'Hont *et al.* 1994, Grivet *et al.* 1996, Hoarau *et al.* 2001, Aitken *et al.* 2005, Reffay *et al.* 2005, Raboin *et al.* 2006, Garcia *et al.* 2006, Aitken *et al.* 2007, Oliveira *et al.* 2007, Anoni *et al.* 2014 (dado não publicado)).

### **Detecção e mapeamento de QTLs**

Uma das aplicações de maior impacto dos mapas genéticos é a localização de genes que controlam características de importância econômica. Essas características apresentam heranças complexas e poligênicas, sendo que os genes ou locos cromossômicos que as controlam são denominados de QTLs (Falconer e Mackay 1996, Lynch e Walsh 1998). O princípio básico do mapeamento de QTLs é de fácil compreensão. Consiste em uma associação entre fenótipo e genótipo, isto é, baseia-se na relação entre a expressão fenotípica de uma dada característica quantitativa, com o resultado da avaliação dos marcadores moleculares, podendo estar associado a um efeito positivo ou negativo desta característica. Porém, é necessário entender suas limitações e dificuldades quando aplicado

à cana-de-açúcar. Sabe-se que o tamanho da população de mapeamento compromete a detecção de QTLs de efeito menor (Lych e Walsh 1998). Porém, esta dificuldade aumenta porque os métodos estatísticos para mapeamento de QTLs foram desenvolvidos para diploides, o que significa que são baseados em marcadores em dose única, e assim, como no mapeamento genético continuamos estudar uma sub amostragem do genoma da cana-de-açúcar. Em outras palavras, a utilização apenas dos marcadores segregando em dose única ocasiona a inutilização de todas as outras segregações diferentes de 1:1 e 3:1, exigindo que a genotipagem da população ocorra com maior número de marcadores moleculares. A detecção dos QTLs em cana-de-açúcar ainda apresenta outro fator agravante atribuído ao alto nível de ploidia, resultando em baixo efeito individual do QTL (Aitken *et al.* 2008). Uma vez que o mapeamento é realizado observando as descendências do cruzamento, a herdabilidade adquire grande importância neste contexto. Por definição, herdabilidade revela o grau de correspondência entre o fenótipo e o genótipo, e é a porção genotípica transmitida aos descendentes (Falconer e Mackay 1996). Assim, valores de herdabilidade altos para uma determinada característica fenotípica significa que tal característica pode ser mapeada mais facilmente.

Para o mapeamento de QTLs é necessário o uso de sofisticadas metodologias estatísticas, além de um grande suporte computacional devido à complexidade das análises (Pastina *et al.* 2010). Os modelos estatísticos mais usados em cana-de-açúcar no mapeamento de QTLs são (1) análises de marcas individuais (*Single Marker – SM*): a ideia central baseia-se na comparação entre as médias fenotípicas com as diferentes classes genotípicas de um determinado marcador, calculada através do teste *t*, análise de variância, regressão linear simples e múltipla e também pelo método de máxima verossimilhança. É o método mais simples que dispensa a necessidade de um mapa genético, porém a posição do QTL não pode ser inferida e existe uma confusão entre efeito do QTL com distância da marca, ou seja, não é possível diferenciar um QTL de um pequeno efeito situado próximo ao marcador de um QTL de grande efeito (Liu, 1998). Essas limitações conferem à análise baixo poder de detecção de QTLs (Doerge 2002); (2) mapeamento por intervalo (*Interval Mapping – IM*) (Lander e Botstein 1989): através de um par de marcadores adjacentes faz inferências sobre a presença de um provável QTL dentro deste intervalo utilizando informações de um mapa genético. Apresenta maior poder estatístico comparado ao SM,

porém por utilizar um par de marcas por vez, não permite analisar a interação entre QTLs presentes em diferentes intervalos de mapeamento dentro do genoma, podendo resultar no mapeamento de QTLs falsos positivos (Doerge 2002); (3) mapeamento por intervalo composto (*Composite Interval Mapping* – CIM) (Zeng 1993, Zeng 1994): combina o IM com outros marcadores utilizados como cofatores, o que permite controlar os efeitos de outros QTLs em diferentes intervalos de mapeamento. Por utilizar um único QTL por vez, o risco de assumir muitas marcas como cofatores é alto, (4) e recentemente Gazaffi (2009) desenvolveu uma metodologia utilizando CIM específico para população de irmãos completos, que não só permite a localização do QTL, fase de ligação e segregação entre os genitores, mas também a análise do QTL com diferentes segregações dos genitores. Esta abordagem foi utilizada por Souza *et al.* (2012) e Anoni *et al.* (2014 dados não publicado).

Diversos estudos relacionados a mapeamento de QTLs ligados à características de produção e produção de açúcar foram publicados em cana-de-açúcar, tais como Ming *et al.* (2001, 2002a, 2002b), Hoarau *et al.* (2002), McIntyre *et al.* (2005), da Silva e Bressiani (2005), Reffay *et al.* (2005), Aitken *et al.* (2006; 2008), Piperidis *et al.* (2008), Pastina *et al.* (2012), Shing *et al.* (2013). Todos contribuíram para aprimorar o conhecimento da estrutura genética da cana-de-açúcar, porém, até o presente momento não foi possível obter um mapa altamente saturado utilizando apenas marcadores SRDF.



## Objetivos

---

### Objetivo Geral

Contribuir para o melhor entendimento genético da cana-de-açúcar por meio da detecção de marcadores moleculares ligados a características de importância econômica em população biparental de mapeamento, considerando na análise, a ploidia e dosagem alélica variável presente em cana-de-açúcar.

### Objetivos específicos

- Avaliar as características fenotípicas de interesse econômico relacionadas à produção (altura, diâmetro, número e peso dos colmos) e qualidade (sólidos solúveis, teor de sacarose do caldo e do colmo, pureza do caldo, teor de fibra) em diferentes locais e anos (colheita)
- Utilizar um modelo estatístico que permita a análise da interação locais e anos (colheita)
- Estabelecer uma metodologia de genotipagem dos SNPs, utilizando a plataforma Sequenom iPLEX MassARRAY® (Sequenom Inc., San Diego, California, USA)
- Genotipar a população de mapeamento com microssatélites ESTs e SNPs polimórficos
- Integrar ambos os marcadores utilizados (EST-SSRs e SNPs) ao mapa genético preliminar construído a partir do cruzamento SP81-3250 x RB925345
- Detectar os QTLs ligados às características fenotípicas avaliadas





### Uso de um modelo estatístico sofisticado para cana-de-açúcar usando características fenotípicas

<sup>a</sup> Centro de Biologia Molecular e Engenharia Genética (CBMEG), Departamento de Genética e Evolução, Universidade Estadual de Campinas (UNICAMP), Cidade Universitária Zeferino Vaz, CP 6010, 13083-875 Campinas, SP, Brazil

<sup>b</sup> Departamento de Genética, Escola Superior de Agricultura Luiz de Queiroz (ESALQ), Universidade de São Paulo (USP), CP 83, 13400-970 Piracicaba, SP, Brazil

<sup>c</sup> Centro de Ciências Agrárias, Universidade Federal de São Carlos, Rodovia Anhanguera, Km 174, Araras - São Paulo - Brazil

---

#### Resumo

A população segregante de cana-de-açúcar contendo 240 indivíduos F<sub>1</sub> derivada do cruzamento entre as variedades SP81-3250 e RB925345 foi avaliada por três anos (2011, 2012 e 2013) para as características de produção (altura, diâmetro, número e peso dos colmos e toneladas de cana por hectare) e características de qualidade (sólidos solúveis, teor de sacarose do caldo e do colmo, pureza do caldo, teor de fibra e toneladas de pol por hectare) nos municípios de Araras e Ipaussu, ambos no estado de São Paulo. Através de modelos mistos foi estimada a média, matriz de variância e covariância (VCOV), herdabilidade e a correlação fenotípica entre todas as características avaliadas. Mostraram um ótimo controle ambiental. Para as características de produção, a herdabilidade variou entre 0,83 (altura dos colmos) a 0,96 (toneladas de cana por hectare). Entre as características de qualidade, o menor valor de herdabilidade foi 0,77 (pureza) enquanto que o maior foi 0,88 (teor de fibra). Ao considerar todas as características, foi encontrado 30 correlações fenotípicas significativas, importantes para uma possível seleção indireta. Os resultados apresentados podem ser úteis ao melhoramento da cana-de-açúcar no intuito de diminuir o tempo da liberação necessária para novas cultivares.

Palavras-chave: poliploide, locos de características quantitativas, múltiplos ambientes

---

## 1. Introdução

A cana-de-açúcar (*Saccharum* spp.) é uma espécie autopoliploide complexa e é considerada uma das culturas mais cultivadas mundialmente (ver <http://www.fao.org>). Cultivares comerciais de cana-de-açúcar são resultados do cruzamento interespecífico entre a espécie domesticada *Saccharum officinarum* ( $2n = 80$ ) e a espécie selvagem *S. spontaneum* ( $2n = 40-120$ ), seguido de diversos retrocruzamentos com a *S. officinarum* (Irvine, 1999; Ha *et al.*, 1999). A complexidade de seu genoma pode ser atribuído ao elevado número de cromossomos, variando de 100 a 130 (D'Hont *et al.*, 1998; Irvine, 1999), ao tamanho do genoma, aproximadamente 10 Gb (D'Hont e Glaszmann, 2001; D'Hont, 2005; Piperidis *et al.*, 2010) e a variação do número de cromossomos que pode ser encontrada nos grupos hom(e)ólogos, caracterizando a aneuploidia (Grivet e Arruda, 2001).

Um dos objetivos principais de um programa de melhoramento consiste em aumentar a produção (Cox *et al.*, 1994) e pode ser alcançado através do conhecimento acumulado sobre o melhoramento de plantas. Desde o início da agricultura, a domesticação das plantas foi baseada na seleção visual do melhor fenótipo. Gregor Mendel em 1900 formulou as teorias que originaram as leis da genética e os mecanismos de herança das características. A partir destes conhecimentos, o fenótipo passou a ser definido como resultado da ação de um ou mais genes com o ambiente. Assim, o entendimento dos componentes genéticos e relações entre as características de interesse com o ambiente são essenciais para desenvolver as estratégias de melhoramento (Cruz *et al.*, 2004). A porção da variância genética também é importante para calcular a herdabilidade e conhecer a porção da variância fenotípica atribuída aos efeitos genéticos hereditários (Falconer e Mackay, 1996), podendo ser utilizada em estudos de resposta a seleção.

Várias características relacionadas à produção de cana-de-açúcar apresentam variação quantitativa e podem ser correlacionadas entre si. Por exemplo, os componentes de produção de açúcar dependem da combinação entre o diâmetro, altura, número e peso dos colmos e Brix (Hoagarth, 1971). Alto teor de fibra afeta o teor de açúcar e a eficiência da moagem, pois reduz a quantidade de caldo extraído, exigindo mais energia para moer a cana-de-açúcar (Ming *et al.*, 2002). Por outro lado o baixo teor de fibra diminui a queima do bagaço, resultando em baixa eficiência da energia recuperada (Hoagarth, 1971; Ming *et*

*al.*, 2002). Através dos exemplos citados fica comprovado a existência de uma complexa relação entre as características, dificultando a seleção de variedades superiores. Assim, o mapeamento dos locos de características quantitativas (QTL - *Quantitative Traits Loci*) tornou-se uma ferramenta importante para buscar um melhor entendimento da arquitetura genética dos locos que as controlam.

Devido à complexidade destas características, a estimativa eficiente de parâmetros genéticos está condicionada à escolha de modelos estatísticos mais refinados. Respondendo a essa demanda é observada diferentes estratégias disponíveis na literatura científica. Nos modelos tradicionais, como a análise de variância conjunta, todos os efeitos são considerados fixos (com exceção do erro residual) o que limita o poder da análise. A abordagem de modelos lineares mistos (Henderson 1984) tem proporcionado algumas vantagens em comparação aos modelos lineares comuns principalmente em experimentos de cana-de-açúcar que normalmente são conduzidos em vários locais e anos (colheitas), chamados de METs (*Multi-Environment Trial*). Entre as vantagens, destacam-se a capacidade de considerar algumas variáveis como aleatórias, ao invés de fixas, além de utilizar diferentes estruturas de variância-covariância de efeitos aleatórios para verificar a presença de heterocedasticidade e correlações entre os fatores. Esta abordagem permite a análise de dados desbalanceados (Smith *et al.*, 2005; Pastina *et al.*, 2012), além de utilizar modelos mais realistas para a variação de erro (blocos incompletos, a correlação espacial) e assumir alguns conjuntos de efeitos (por exemplo, genótipos) como aleatório (Smith *et al.*, 2005).

A escolha em estabelecer um efeito como sendo fixo ou aleatório depende do objetivo da análise (Smith *et al.*, 2005). Se o objetivo da análise for seleção, isto é, identificar as melhores variedades entre outras em seleção, os efeitos devem ser considerados como aleatórios. Na análise em que o objetivo é a diferença entre pares de variedades, o mais indicado é considerar o efeito como fixo. De acordo com o tipo de efeito, existem dois procedimentos de estimação, o melhor estimador linear não viesado (BLUE - *Best Linear Unbiased Estimation*) para efeitos fixos e o melhor preditor linear não viesado (BLUP - *Best Linear Unbiased Prediction*, Henderson, 1950) para efeitos aleatórios.

O objetivo deste trabalho foi propor o uso de modelos lineares mistos através da modelagem adequada das matrizes de variância-covariância (VCOV) para os efeitos genéticos (G) e não genéticos, obter os componentes de variância para estimar a herdabilidade e as correlações fenotípicas entre as características de produção e de qualidade de uma população segregante de cana-de-açúcar. Estas análises devem ser consideradas o primeiro passo para o mapeamento de QTLs.

## **2. Material e métodos**

### 2.1 Material vegetal

A população segregante usada para coletar dados fenotípicos foi desenvolvida pelo Programa de Melhoramento Genético da Cana-de-Açúcar da UFSCar (Universidade Federal de São Carlos), que faz parte da RIDESA (Rede interinstitucional de Desenvolvimento do Setor sucroalcooleiro). Foi constituída por 240 indivíduos provenientes do cruzamento entre as variedades comerciais SP81-3250 (genitor feminino) e RB925345 (genitor masculino). A variedade SP81-3250 apresenta resistência à ferrugem marrom, enquanto RB925345 é suscetível. Ambas as variedades apresentam alta produtividade, teor de sacarose e teor de fibra.

### 2.2 Dados fenotípicos

A população segregante foi plantada em 2010 em dois locais (Araras e Ipaussu, ambos no Estado de São Paulo, Brasil) e avaliados no primeiro, segundo e terceiro anos de colheita (2011, 2012 e 2013) para diâmetro dos colmos (DC, em mm), altura dos colmos (AC, em m), número de colmos (NC), peso dos colmos (PC, em kg), teor de sólidos solúveis (Brix), teor de sacarose da cana-de-açúcar (Pol%Cana), teor de sacarose do caldo (Pol%Caldo), pureza (PUR), fibra (FIB), toneladas de cana por hectare (TCH) e toneladas de Pol por hectare (TPH). O delineamento experimental foi em blocos aumentados de Federer, composto por blocos incompletos com até 27 genótipos do cruzamento bi-parental

e três padrões de checagem. Os padrões foram as variedades SP80-3150, RB925345, RB867515.

Todos os dados fenotípicos foram coletados e avaliados 12 meses após o plantio e 12 meses após o corte, nos três anos, de acordo com a metodologia descrita por CONSECAN (2006). Para Brix, Pol%Cana, Pol%Caldo, PUR e FIB apenas duas repetições de cada local foram avaliadas, enquanto que o restante das características foram avaliadas em todas as repetições dos dois locais. Todas as características foram avaliadas em um conjunto de 10 colmos por parcela. Ao compor o peso foi adicionado o peso total de cada parcela. O número de colmos foi estimado por contagem direta dos mesmos no campo. O valor de TCH foi determinado multiplicando o peso por metro linear pela constante 6.666,67. Através dos valores para TCH e Pol% Cana foi estimado TPH.

### 2.3 Análise dos dados

O modelo misto para diferentes locais e anos (colheita) foi usado para calcular as médias ajustadas conjuntas, para obter correlações genéticas entre as características e classificação de genótipos para a seleção. As análises foram realizadas no programa GenStat 16<sup>a</sup> edição (Payne *et al.*, 2009), baseado na máxima verossimilhança (REML), utilizando o seguinte modelo linear:

$$y_{ijkmn} = \mu + l_n + h_m + r_{k(mn)} + b_{j(kmn)} + t_{imn} + \varepsilon_{ijkmn}$$

onde  $y_{ijkmn}$  é o fenótipo do genótipo  $i$  no bloco  $j$  na repetição  $k$  do local  $n$  e colheita  $m$ ,  $\mu$  é a média,  $l_n$  é o efeito fixo do local  $n^{\text{th}}$  ( $n = 1, N = 2$ ),  $h_m$  é o efeito fixo da colheita  $m^{\text{th}}$  ( $m = 1, \dots, M$ ;  $M = 2$  ou  $M = 3$  dependendo do local),  $r_{k(mn)}$  é o efeito fixo da repetição  $k^{\text{th}}$  ( $k = 1, \dots, K$ ;  $K = 2$  or  $K = 3$  dependendo da característica) no local  $m$  e colheita  $n$ ,  $b_{j(kmn)}$  é o efeito aleatório do bloco  $j^{\text{th}}$  na repetição  $k$  no local  $n$  e colheita  $m$ ,  $t_{imn}$  é o efeito aleatório tratamento  $i^{\text{th}}$  ( $i = 1, \dots, I$ ;  $I = 243$ ) no local  $n$  e colheita  $m$ , e  $\varepsilon_{ijkmn}$  é o erro residual. Os tratamentos foram separados em dois grupos, sendo  $g_{imn}g_{ij}$  o efeito genético aleatório do genótipo  $ij^{\text{th}}$  ( $I = 1, \dots, I_g$ ;  $I_g = 240j = 1, \dots, J_g = 380$ ) no local  $n$  e colheita  $m$ , e  $c_{imn}$  o efeito fixo do padrão  $ij^{\text{th}}$  ( $i = I_g + 1, \dots, I_g + I_c$ ;  $I_c = 3j = J_g +$

1, ...,  $J_g + J_c = 383$  and  $I_g + I_c = 243$ ) no local  $n$  e colheita  $m$ . Para genótipos, assumiu-se que o vetor  $g = (g_{111}, \dots, g_{IMN})'$  possui distribuição normal multivariada com vetor de média zero e matriz VCOV genética  $G = G_P \otimes I_{I_g}$ , isto é,  $g \sim N(0, G)$ , onde  $P$  é o número de combinações de local-colheita,  $\otimes$  representa o produto direto de Kronecker entre as matrizes genética  $G_P$  e de identidade  $I_{I_g}$  com respectivas dimensões  $P \times P$  e  $I \times I$  número de combinações de local-colheita e genótipos, respectivamente. Diversas estruturas de variância e covariância foram examinadas para matriz  $G_P$  (Tabela 1) e comparadas pelos critérios de informação Akaike (*AIC*; Akaike, 1974) e Bayesiano (*BIC*; Schwarz, 1978) (Pastina *et al.*, 2012). Os modelos de 1–6 usaram a combinação fatorial local-colheita como ambientes diferentes (*E*), isto é,  $G_P = G_{P \times P}^E$ , enquanto que os modelos de 7–12 usaram produtos diretos da matriz VCOV matriz para local (*L*) e colheita (*H*), isto é,  $G_P = G_{M \times M}^L G_{N \times N}^H$ . Para os resíduos, assumiu-se  $\varepsilon \sim N(0, R)$ , onde  $\varepsilon = (\varepsilon_{11111}, \dots, \varepsilon_{IJKMN})'$  e  $R$  é a matriz VCOV residual. Similarmente à matriz  $G$ ,  $R = R_P \otimes I_{I_r}$ , onde  $P$  é o número de combinações de colheita-local-repetição,  $R_P$  é uma matriz  $P \times P$  e  $I_{I_r}$  é o número de combinações entre locais-cortes-repetição e blocos. A matriz  $R_P$  foi examinada e comparada por *AIC* e *BIC* para várias estruturas de locais, colheita e repetição após a seleção para  $G_P$ . Para cada característica, os efeitos fixos para efeitos de interação entre local, colheita e padrões foram testados pelo teste de Wald e mantidas no modelo se forem estatisticamente significativas ( $P < 0.05$ ). As correlações genéticas entre as características foram calculadas utilizando-se as médias ajustadas de cada característica por meio do coeficiente de correlação de Pearson, assumindo nível de significância de  $\alpha = 0.05$  realizado no programa R (<http://www.cran.r-project.org>).

A herdabilidade individual a nível de plantas ( $H_{plants}^2$ ) foi calculada com base nas estimativas dos componentes de variância assumindo estrutura de identidade para a matriz  $G_P$  (modelo 1) usando a razão  $\sigma_G^2 / \sigma_P^2$ , onde  $\sigma_G^2$  é o componente da variância genotípica e  $\sigma_P^2$  é a variação fenotípica total para cada característica. Por meio do cálculo da média harmônica do número de ambientes utilizada como numerador da estimativa da variância da interação genótipo-ambiente; e da média harmônica do número de parcelas utilizada como numerador da estimativa da variância residual (Holland *et al.* 2003), foi possível calcular a razão entre  $\sigma_G^2 / \sigma_P^2$  com o objetivo de fornecer estimativas apropriadas do cálculo

da herdabilidade no sentido amplo com base na média dos genótipos ( $H_{means}^2$ ), onde  $\sigma_p^2$  é a variância fenotípica entre as médias dos genótipos para cada característica.

Tabela 1 - Descrição e número de parâmetros ( $n_{PAR}$ ) dos modelos examinados para a matriz genética de variância-covariância  $G_P$ .

Matrix $G_P$	$n_{PAR}$	Descrição
$G_P = G_{P \times P}^E$		
(1) ID	1	Identidade (variância genética homogênea)
(2) UNIF	2	Uniforme
(3) DIAG	$P$	Diagonal (ou variância genética heterogênea)
(4) $CS_{Het}$	$P + 1$	Simetria Composta
(5) FA1	$2P$	Fator Analítico de primeira ordem
(6) UNST	$\frac{P(P+1)}{2}$	Não-estruturada
$G_P = G_{N \times N}^L \otimes G_{M \times M}^H$		
(7) UNST $\otimes$ ID	$\frac{N(N+1)}{2} + 1$	Não estruturada e identidade para locais e colheitas; respectivamente
(8) UNST $\otimes$ UNIF	$\frac{N(N+1)}{2} + 2$	Não-estruturada e uniforme para locais e colheitas; respectivamente
(9) UNST $\otimes$ DIAG	$\frac{N(N+1)}{2} + M$	Não-estruturada e diagonal para locais e colheitas; respectivamente
(10) UNST $\otimes$ AR1	$\frac{N(N+1)+2(M+1)}{2} - 1$	Não-estruturada e auto-regressiva de primeira ordem para locais e colheitas; respectivamente
(11) UNST $\otimes$ $CS_{Het}$	$\frac{N(N+1)}{2} + M + 1$	Não-estruturada e simetria composta para locais e colheitas; respectivamente
(12) UNST $\otimes$ UNST	$\frac{N(N+1)+M(M+1)}{2} - 1$	Não-estruturada para ambos locais e colheitas.

### 3. Resultados

Na tabela 2 encontram-se os modelos selecionados para a matriz G considerando cada característica. Como é possível notar diferentes modelos foram selecionados. Em geral, observa-se que os modelos selecionados para as características de produção consideraram as estruturas de VCOV  $G_P = G_{N \times N}^L \otimes G_{M \times M}^H$ , com exceção para diâmetro dos colmos, que utiliza a combinação de cortes-locais como ambiente. Nas características de qualidade os modelos selecionados consideraram as estruturas de VCOV  $G_P = G_{P \times P}^E$ , porém este modelo necessita estimar um número maior de parâmetros quando comparados à estrutura  $G_P = G_{N \times N}^L \otimes G_{M \times M}^H$ . Os modelos selecionados para: TCH, TPH e PC apresentaram diferentes componentes de variância e covariância genéticos não correlacionados tanto entre locais como entre colheita; DC, POL%Caldo e PUR

apresentaram variâncias genéticas diferentes (heterogêneas) e covariâncias genéticas iguais (homogêneas) para as diferentes combinações de ambientes; Brix, Pol%Cana e FIB indicaram as mesmas variâncias e covariâncias genéticas para as diferentes combinações de ambientes; AC assumiu que variâncias e covariâncias genéticas foram heterogêneas para locais enquanto que as covariâncias entre cortes são correlacionas, conferindo menor correlação a medida que aumenta o número de cortes; NC mostrou variância e covariância genética heterogênea entre locais e cortes, embora as variâncias genéticas diferiram, as covariâncias genéticas são comuns. Baseados em tais modelos foi possível a obtenção dos BLUPs para cada característica individual e posterior obtenção dos parâmetros genéticos

Tabela 2 – Modelos selecionados para a matriz  $G_p$  considerando cada característica separadamente. Os critérios AIC e BIC foram utilizados para comparar as estruturas de variâncias-covariâncias da matriz.

Característica	Modelo selecionado*	Critério AIC	Critério BIC
Diâmetro dos colmos	4	20155,60	20213,70
Altura dos colmos	10	790,01	828,74
Número de colmos	11	45439,25	45490,97
Peso dos colmos	12	47036,89	47101,53
Brix	2	7818,48	7841,98
Pol%Cana	2	7691,9	7715,41
Pol%Caldo	4	8656,28	8656,28
PUR	4	11899,07	11946,07
FIB	2	6695,41	6718,91
TCH	12	45321,05	45385,65
TPH	12	15464,73	15523,41

\* Modelo selecionado para a matriz  $G_p$  como descrito na Tabela 1.

Os resultados das médias e herdabilidades para as 11 características nos três anos (colheita) e nos dois locais estão resumidos na Tabela 3. A média fenotípica para as características de produção variou de 21,07 a 31,67 mm para DC, 1,99 a 2,67 m para AC, 57,5 a 203,3 para NC, 76,2 a 239,1 kg para PC e 90,2 a 210 para TCH, enquanto que as características de qualidade variaram de 21 a 19,62 para BRIX, 16 a 14,07 para Pol%Cana, 19 para 16,66 para Pol%Caldo, 90,19 a 84,44 para PUR, 11,97 a 12,91 para FIB e 22,99 a 16,32 para TPH. As médias gerais das características estudadas entre genitores não apresentaram grandes diferenças (Figura 1 e 2). No geral, foi encontrada uma boa precisão experimental além de alta herdabilidade (Tabela 3). A média fenotípica encontrada para DC foi 25,62, com coeficiente de variação residual (CV\_R) de 7,15. AC apresentou média de 2,34 e valor de CV\_R de 10,06. Os valores do CV\_R para NC e PC foram de 22,06 e 18,23,



respectivamente, com médias de 117,72 para NC e 168,86 para PC, e variância genética (529,90 para NC e 734,20 para PC). Os altos valores de CV\_R podem ser explicados pela influência direta da variância, que por sua vez está relacionado com a magnitude da característica alvo de estudo. BRIX teve sua média em 20,94 e CV\_R em 4,05. As características de qualidade Pol%Cana e Pol%Caldo mostraram valores similares, com médias de 15,53 e 18,58, respectivamente, e mesmo valor de CV\_R (5,48). PUR teve média de 88,52 e CV\_R 2,26. FIB mostrou média de 12,65 e 5,66 para CV\_R. Finalmente, TCH e TPH tiveram as médias em 149,48 e 22,24, e CV\_R de 6,8 e 18,29, respectivamente.

Tabela 3. Médias, estimativas dos componentes de variância genotípica ( $\hat{\sigma}_G^2$ ) e fenotípica ( $\hat{\sigma}_P^2$ ), coeficiente de variação genotípica (CV\_G) e residual (CV\_R), e herdabilidade no sentido amplo entre a média ( $H_{means}^2$ ) e individual ( $H_{plants}^2$ ) para diâmetro, altura, número de peso dos colmos, TCH, Brix, Pol%Cana, Pol%Caldo, PUR, FIB e TPH dos 240 indivíduos da população segregante de cana-de-açúcar derivados do cruzamento entre as variedades SP81-3250 e RB925345, avaliados em dois locais e três anos (colheita).

Trait	Médias	$\hat{\sigma}_G^2$	$\hat{\sigma}_P^2$	CV_G	CV_R	$H_{means}^2$	$H_{plants}^2$
Diâmetro dos colmos	25,62	3,55	7,30	7,35	7,15	0,94	0,49
Altura dos colmos	2,34	0,02	0,08	5,87	10,06	0,83	0,23
Número dos colmos	117,72	529,90	1296,20	19,56	22,06	0,92	0,41
Peso dos colmos	168,86	734,20	1884,30	16,05	18,23	0,90	0,39
Brix	20,94	0,65	1,65	3,86	4,05	0,85	0,39
Pol%Cana	15,53	0,53	1,47	4,68	5,48	0,83	0,39
Pol%Caldo	18,58	0,81	2,18	4,84	5,48	0,84	0,37
PUR	88,52	1,79	6,51	1,51	2,26	0,77	0,27
FIB	12,65	0,52	1,13	5,73	5,66	0,88	0,46
TCH	149,48	543,10	772,90	15,59	6,80	0,96	0,70
TPH	22,24	11,540	29,90	15,27	18,29	0,85	0,39

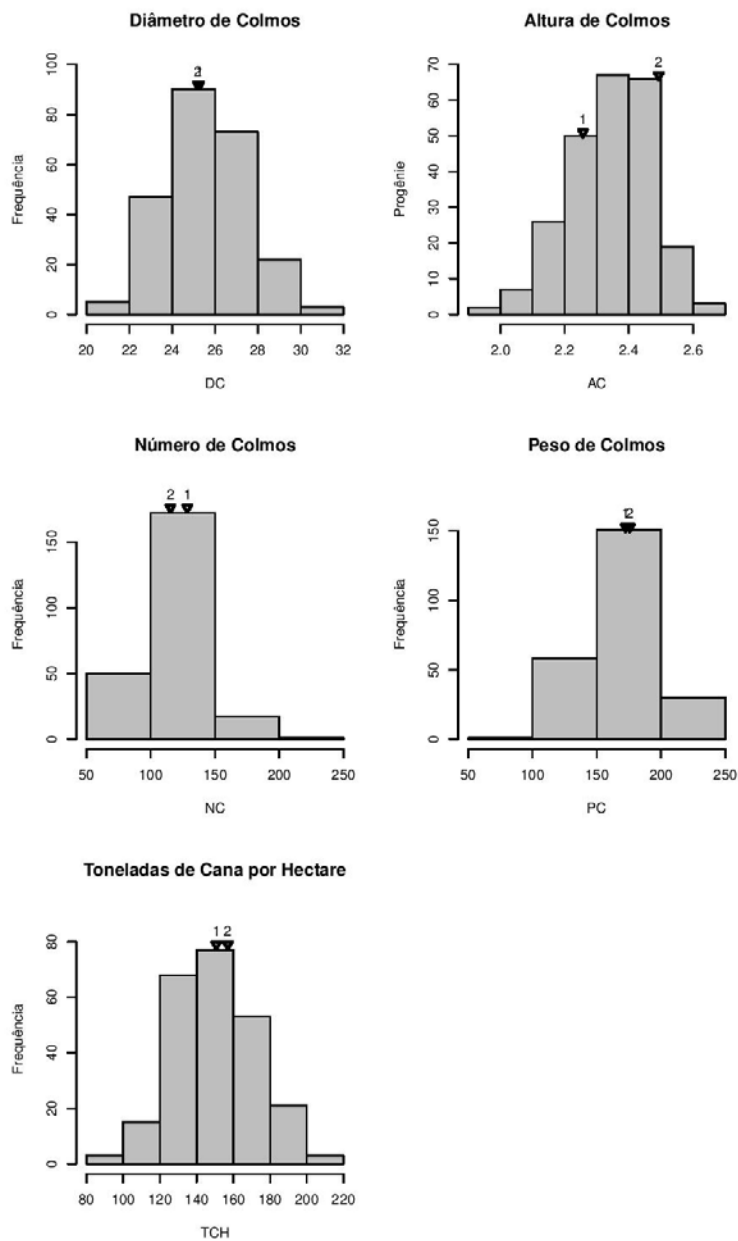


Figura 1. Frequência de distribuição dos valores fenotípicos para altura, número, peso e diâmetro dos colmos e TCH dos 240 indivíduos da população segregante de cana-de-açúcar e dos genitores SP81-3250 (1) e RB925345 (2), avaliados em dois locais e três anos (colheita).

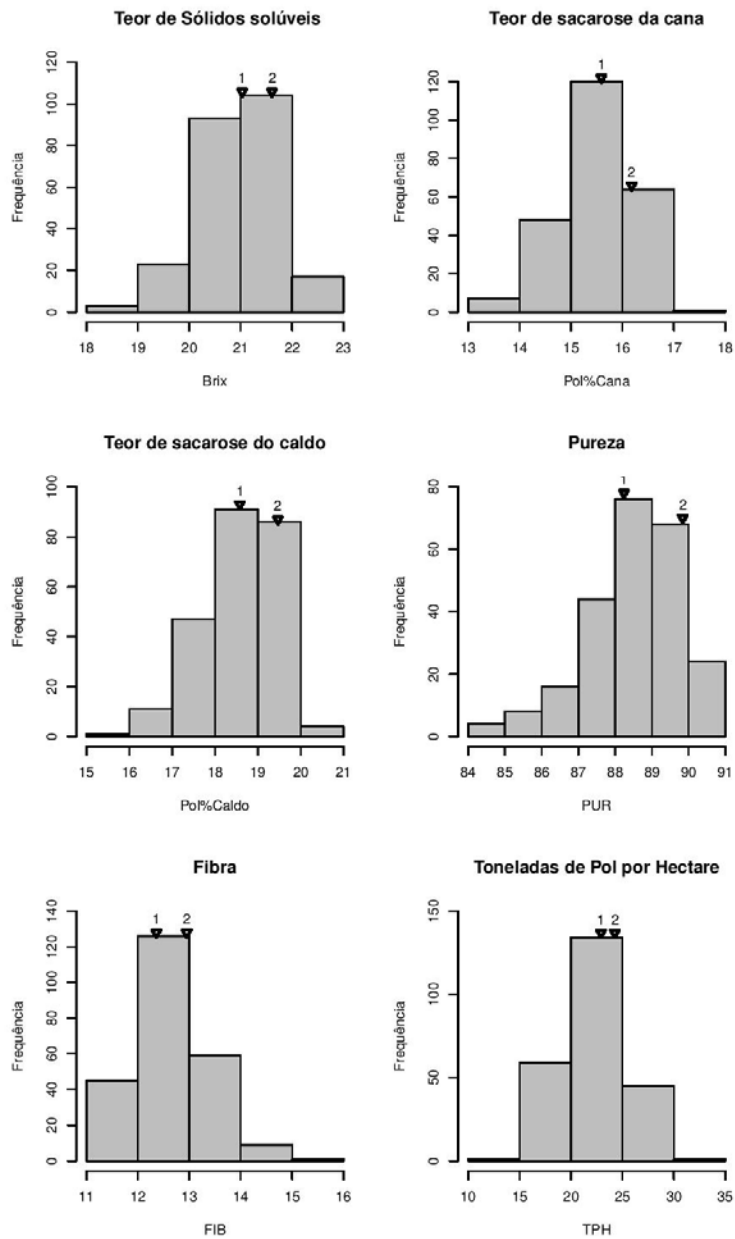


Figura 2. Frequência de distribuição dos valores fenotípicos para Brix, Pol%Cana, Pol%Caldo, Pureza, Fibra e TPH dos 240 indivíduos da população segregante de cana-de-açúcar e dos genitores SP81-3250 (1) e RB925345 (2), avaliados em dois locais e três anos (colheita).

Os maiores valores para herdabilidade foram aqueles encontrados pela herdabilidade média (Tabela 3). TCH apresentou a maior herdabilidade (0,94), seguido de DC (0,94), NC (0,92) e PC (0,90), enquanto PUR mostrou o valor mais baixo (0,77). Para AC e Pol%Cana a herdabilidade foi a mesma (0,83), tal como para Brix e TPH (0,85). Pol%Caldo apresentou valor de 0,88. Já a herdabilidade individual de plantas, por ter sido calculada adicionando a variância genética, bem como a interação entre local e número de repetições, apresentou os menores valores. Constatou-se que TCH também apresentou o maior valor (0,70), seguido de DC (0,49), FIB (0,46) e NC (0,41), enquanto AC e PUR apresentaram os valores mais baixos (0,23 e 0,27, respectivamente). PC, Brix, Pol%Cana e TPH a herdabilidade foi a mesma (0,39). O valor encontrado para Pol%Caldo foi de 0,37.

A correlação fenotípica para todas as características medidas foi realizada com base no *p valor* ( $P < 0.05$ ) (Figura 3). Foram encontradas 30 correlações significativas separadas em correlações positivas e negativas. Entre as correlações fenotípicas positivas houve a separação em forte (PC-TCH, PC-TPH, Brix-Pol%Cana, BRIX-Pol%Caldo, Pol%Cana-Pol%Caldo, Pol%Cana-PUR, Pol%Caldo-PUR e TCH-TPH); moderada (AC-PC, AC-TCH, AC-TPH, NC-PC, NC-TCH, NC-TPH, Brix-PUR e Brix-FIB); e baixa (DC-AC, DC-PC, DC-TCH, DC-TPH, Brix-TPH, Pol% Cana-TPH, Pol% Caldo-FIB, Pol% Caldo-TPH e PUR-TPH). As correlações negativas e significativas foram encontradas em: DC-NC, DC-FIB, PC-FIB, FIB-TCH e FIB-TPH.

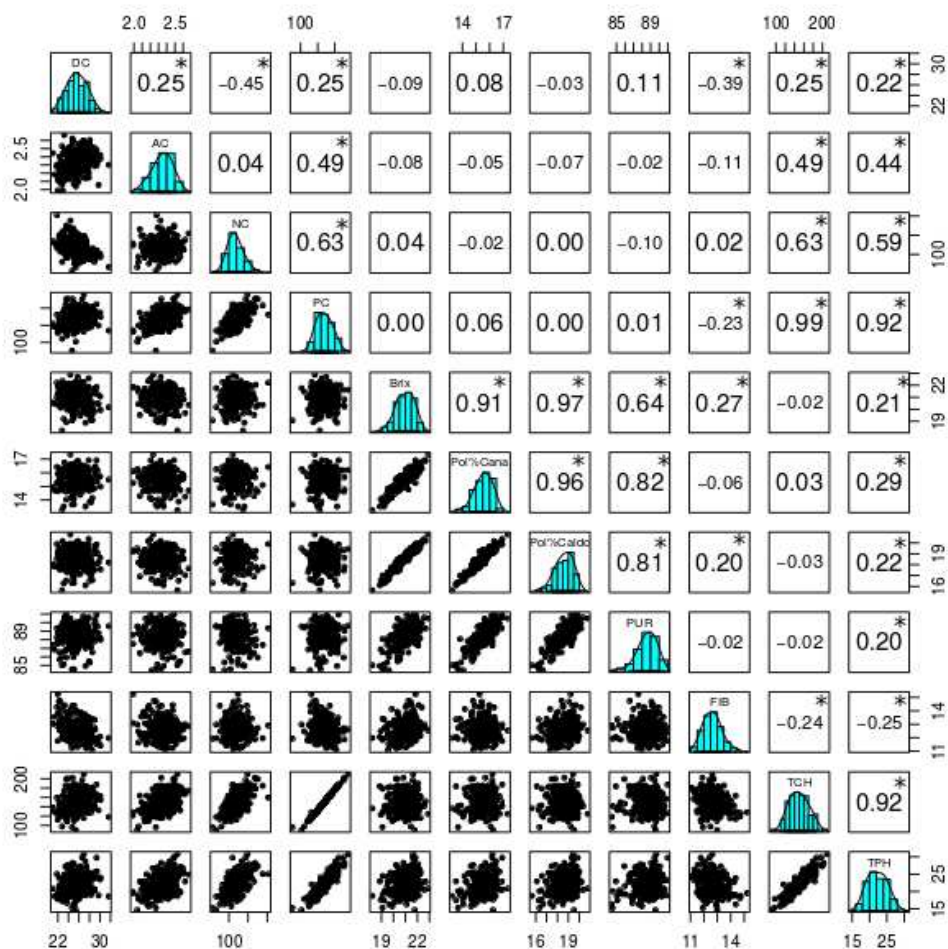


Figura 3. Estimativas da correlação fenotípica entre DC, AC, NC, PC, Brix, Pol%Cana, Pol%Caldo, PUR, FIB, TCH e TPH para os 240 indivíduos da população segregante de cana-de-açúcar e dos genitores SP81-3250 e RB925345, avaliados em dois locais e três anos (colheita).

\* Significativo ao nível de 5% ( $P < 0.05$ )

#### 4. Discussão

A cana-de-açúcar é uma das culturas mais importantes ao redor do mundo e sua importância é atribuída principalmente a seus derivados, açúcar e etanol. A produção de cana-de-açúcar colocou o Brasil como maior produtor mundial (Unica 2014), com constante busca por aumento de produção. No entanto, aumentar a produção gera a necessidade de criar uma forma adequada e sustentável para a modernização do cultivo da cana-de-açúcar. Neste sentido, a agricultura de alta tecnologia associada à biologia

molecular vem se tornando uma ferramenta poderosa para aumentar a produtividade sem aumentar a área plantada de cana-de-açúcar.

O aumento da produtividade exige o conhecimento da base genética das características que estão diretamente ligadas à produção de açúcar, e assim conseguir manusear corretamente um conjunto de características simultaneamente. Vários estudos que envolvem características de produção e de qualidade foram conduzidos em cana-de-açúcar, no entanto, utilizando uma abordagem estatística com uma série de limitações (Lin *et al.*, 1993; Gallacher, 1997), especialmente em dados desbalanceados (Smith *et al.*, 2005; Piepho e Möhring, 2007; Pastina *et al.*, 2012), ao assumir homogeneidade das variâncias e a ausência de correlações genéticas entre anos (colheita) e locais para estimar os valores genéticos (Balzarini, 2002) e a falta de dados fenotípicos. Assim, o uso de modelos mistos permitiu a heterogeneidade de variâncias e correlações genéticas (Malosetti *et al.*, 2013) e a análise dos dados passou a ser mais adequado ao experimento, principalmente por considerar a interação anos (colheita) e local (Pastina *et al.*, 2012).

Embora o ajuste das estruturas de variância e covariância tenha sido ideal para os dados do presente estudo, é importante ressaltar que o uso de maior número de locais e cortes muito provavelmente possibilitaria o ajuste de outras estruturas de variância e covariância. Por exemplo, considerando dados de cana-de-açúcar provenientes de cortes da mesma planta ao longo dos anos, espera-se que os dados apresentem algum grau de correlação pelo simples fato das medidas serem tomadas no mesmo indivíduo. Este fato pode ser mais pronunciado em populações de programas de melhoramento, onde o número de cortes e locais é maior em relação aos experimentos utilizando populações de mapeamento. Devido ao foco do estudo, a estruturação da matriz VCOV foi primeiramente realizada considerando a matriz de efeitos genéticos (matriz  $G$ ) para posteriormente ser considerada a matriz de efeitos não genéticos (matriz  $R$ ) (Material Suplementar). Assim, houve a possibilidade de definir estruturas da matriz  $G$  mais verossímeis, o que possibilitou a estimativa mais eficiente de componentes de variância genéticos e estimativa dos parâmetros genéticos.

O experimento contou com uma boa precisão experimental e controle da variação ambiental, medidos através do coeficiente de variação. Sabe-se que quanto menor é a estimativa do CV maior será a precisão e qualidade do experimento e consequentemente as

pequenas diferenças entre as estimativas de médias serão significativas (Filho Cargnelutti e Storck, 2007). A maioria dos resultados encontrados atendeu a recomendação descrita para alta precisão experimental com valores de CV inferiores a 10% (Pimentel-Gomes, 1985; Costa *et al.*, 2002; Perecin *et al.*, 2004), com exceção para número e peso dos colmos e TPH. Provavelmente, o maior valor do CV para número e peso dos colmos é porque os colmos da cana-de-açúcar continuam se alongando enquanto novos ainda estão sendo formados (Hoarau *et al.*, 2002), dificultando a fenotipagem.

A variação fenotípica encontrada entre os 240 indivíduos já era esperada, pois são considerados híbridos interespecíficos, além de serem derivados de um cruzamento que não sofreu nenhuma seleção. Observou-se que a amplitude fenotípica encontrada entre os indivíduos foi maior que entre os genitores, caracterizando segregação transgressiva, também observada por Hoarau *et al.* (2002) e Mancini *et al.* (2012). O peso e diâmetro dos colmos entre os genitores apresentaram quase os mesmos valores, que pode refletir anos de seleção até terem sido lançados como variedades do Programa de Melhoramento Genético de Cana da UFSCar.

Definida como a porção herdável transmitida em um cruzamento (Falconer e Mackay, 1996), herdabilidade é um parâmetro importante, uma vez que determina a resposta à seleção e pode ajudar na escolha da melhor estratégia para ser aplicada em um programa de melhoramento (Piepho e Möhring, 2007). Todos os valores de herdabilidade apresentados foram classificados como alto segundo Resende (2002) ( $h^2 > 0,50$ ), variando de 0,77 a 0,96, indicando que a variação fenotípica observada é devido a variação genotípica desta população. Os valores de herdabilidade foram mais elevados em relação aos apresentados por Hoarau *et al.* (2002), Aitken *et al.* (2006), Liu *et al.* (2007), Aitken *et al.* (2008), Pinto *et al.* (2009), Ahmed *et al.* (2012) e Mancini *et al.* (2012), lembrando que todos estes estudos não consideraram a interação entre os anos (colheita). A superioridade dos valores é atribuída não apenas ao baixo CV do experimento, mas também ao modelo estatístico utilizado, que permitiu a integração dos dados entre os anos (colheita) e locais e assumindo heterogeneidade das variâncias, o que tornou as estimativas de herdabilidade mais precisas. Este fato ganha um foco importante no melhoramento de plantas ao selecionar os genótipos através das características com alta herdabilidade, atuando na seleção de gerações precoces.

Outra forma de seleção de genótipos é através da correlação fenotípica que pode combinar mais de uma característica desejável na mesma planta, permitindo uma seleção indireta (Ram *et al.*, 1997). Por exemplo, o principal objetivo de um programa de melhoramento genético da cana-de-açúcar é aumentar o rendimento de açúcar. Porém, atualmente vem se notado pequenos aumentos nesta taxa. Isto pode ser explicado por anos de seleção para açúcar, o que dificulta o aumento neste índice. Uma forma de contornar essa dificuldade é realizar a seleção em outras características correlacionadas. Por meio dos resultados alcançados, observou-se um grande número de correlações fenotípicas significativas entre as características avaliadas, incluindo aquelas relacionadas à produção de açúcar e produtividade, medida em TCH. Este fato favorece a seleção indireta entre estas características o que pode contribuir no aumento da produtividade da cana-de-açúcar.

Usando os resultados apresentados neste estudo, podemos expandir o conhecimento para análise e detecção de QTLs. Estas informações são importantes para entendimento da arquitetura genética da espécie (Liu *et al.*, 2007) tornando-se uma ferramenta poderosa se aplicada em programas de melhoramento genético da cana-de-açúcar através da seleção assistida por marcadores.

## **5. Conclusão**

O uso de modelos lineares mistos permitiu calcular a herdabilidade e correlação fenotípica entre as características de produção e de qualidade para uma população segregante de cana-de-açúcar, assumindo heterogeneidade das variâncias e interação local e anos (colheita), resultando em valores mais precisos e realistas. Os resultados apresentados são confiáveis para a utilização no mapeamento e detecção de QTL.

## **Referências**

- 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, 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



- Ahmed AO, Obeid A (2012) Investigation on variability, broad sensed heritability and genetic advance in Sugarcane (*Saccharum spp*). International Journal of AgriScience 2: 839-844
- Akaike H (1974) A new look at the statistical model identification. IEEE Trans Automat Contr AC 19:716-723
- Balzarini M (2002) Applications of mixed models in plant breeding. In: Kang MS (ed.) Quantitative genetics, genomics and plant breeding. New York: CABI Publishing, 353-363p
- Consecana – Conselho nacional dos produtores de cana-de-açúcar, açúcar e álcool do estado de São Paulo (2006) Manual de instruções – CONSECANAS-SP. Piracicaba: CONSECANAS, 112p
- Costa NHAD, Seraphin JC, Zimmermann FJP (2002) Novo método de classificação de coeficientes de variação para a cultura do arroz de terras altas. Pesquisa Agropecuária Brasileira 37:243-249
- Cox MC, Hogarth DM, Hansen PB (1994) Breeding and selection for high early season sugar content in a sugarcane (*Saccharum spp*. hybrids) improvement program. Aust J Agric Res 45:1569-1575
- Cruz CD, Regazzi AJ, Carneiro PCS (2004) Modelos Biométricos aplicados ao melhoramento genético. 3 ed. Viçosa: UFV, 460p
- D'Hont A, Ison D, Alix K, Roux C, Glaszmann JC (1998) Determination of basic chromosome numbers in the genus *Saccharum* by physical mapping of ribosomal RNA genes. Genome 41:221–225
- D'Hont A, Glaszmann JC (2001) Sugarcane genome analysis with molecular markers, a first decade research. Proc Int Soc Sugarcane Technol 24:556-559
- D'Hont A (2005) Unravelling the genome structure of polyploids using FISH and GISH; examples in sugarcane and banana. Cytogenet Genome Res 109:27–33
- Falconer DS, Mackay TF (1996) Introduction to quantitative genetics. 4 ed Londres: Longman Group, 464p
- Federer WT (1956) Augmented (or hoonuiaku) designs. Hawaiian Planters' Record 55: 191-208
- Filho Cargnelutti A, Storck L (2007) Estatísticas de avaliação da precisão experimental em ensaios de cultivares de milho. Pesquisa agropecuária brasileira 42:17-24
- Gallacher DJ (1997) Evaluation of sugarcane morphological descriptors using variance component analysis. Aust J Agric Res, 48:769-774
- Grivet L, Arruda P (2001) Sugarcane genomics: depicting the complex genome of an important tropical crop. Curr Opin Plant Biol 5:122–127

- Ha S, Moore PH, 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 Mol Biol* 39:1165-1173
- Henderson CR (1950) Estimation of genetic parameters (abstract). *Annals of Mathematical Statistics* 21:309-310
- Henderson CR (1984) Applications of linear models in animal breeding. Ontario: University of Guelph, 462 p
- Hogarth DM (1971) Quantitative inheritance studies in sugar-cane: II. Correlations and predicted responses to selection. *Aust J Agric Res* 22:103-109
- Holland JB, Nyquist WE, Cervantes-Martinez CT (2003) Estimating and interpreting heritability for plant breeding: an update. *Plant Breed Rev* 22:9-113
- Irvine JE (1999) *Saccharum* species as horticultural classes. *Theor Appl Genet* 98:186–194
- Lin JF, Chen RK, Lin YQ, (1993) The inheritance of sugar characters in sugarcane. *J. Fujian Agric Coll* 22:392-397 (in Chinese)
- Liu GF, Zhou Hk, Hu H, Zhu ZH, Hayat Y, Xu HM, Yang J (2007) Genetic analysis for brix weight per stool and its component traits in sugarcane (*Saccharum officinarum*). *Journal of Zhejiang University* 8:860-866
- Malosetti M, Ribaut JM, van Eeuwijk FA (2013) The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. *Frontiers in Physiology*. *Front Physio* 4:44
- Mancini MC, Leite DC, Perecin D, Bidóia MAP, Xavier MA, Landell, Pinto LR MGA (2012) Characterization of the Genetic Variability of a Sugarcane Commercial Cross Through Yield Components and Quality Parameters. *Sugar Tech* 14:119–125
- 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
- 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
- Payne RW, Murray DA, Harding SA, Baird DB, Soutar DM (2009) *GenStat for Windows* (12th Edition) Introduction. VSN International, Hemel Hempstead
- Perecin D, Marques DG, Landell MGA(2004) Selo de qualidade para ensaios de melhoramento de cana-de-açúcar. Reunião anual da região brasileira da sociedade internacional de biometria, Uberlândia – MG 382-384
- Piepho HP, Möhring JM (2007) Computing heritability and selection response from unbalanced plant breeding trials. *Genetics* 177:1881–1888

Pimentel-Gomes F (1985) Curso de Estatística Experimental. Livraria Nobel 467p

Pinto LR, Garcia AAF, Pastina MM, Teixeira LHM, Bressiane JA, Ulian EC, Bidoia MAP, Souza AP (2009) 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: 313–327

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

Ram B, Chaudhary BS, Singh S (1997) Response to indirect selection in ratoon of sugarcane seedlings. *Aust J Agric Res* 48:207-213

Resende MD (2002) Genética biométrica e estatística no melhoramento de plantas perenes. Embrapa Informação Tecnológica: Brasília, 975p

Schwarz GE (1978) Estimating the dimension of a model. *Annals of Statistics* 6:461-464

Smith AB, Cullis BR, Thompson R (2005) The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. *Journal of Agricultural Science* 143, 449–462

## Material suplementar

Modelos selecionados para a matriz  $R$  considerando cada característica separadamente. Critérios AIC e BIC corresponde ao modelo selecionado especificado na segunda coluna.

Característica	Matriz $R$	Critério AIC	Critério BIC
	$R = R_{P \times P}^E \otimes R_{K \times K}^B$		
Diâmetro dos colmos	$R = CS_{\text{Het}} \otimes \text{UNS}$	19358,01	19487,12
Brix	$R = \text{DIAG} \otimes \text{ID}$	7794,62	7841,62
Pol%Caldo	$R = \text{ID} \otimes \text{ID}$	8656,28	8656,28
Pol%Cana	$R = \text{DIAG} \otimes \text{ID}$	7657,58	7704,59
PUR	$R = \text{ID} \otimes \text{ID}$	11899,07	11946,07
FIB	$R = CS_{\text{Het}} \otimes \text{UNIF}$	6504,71	6563,46
	$R = R_{N \times N}^L \otimes R_{M \times M}^H \otimes R_{K \times K}^J$		
Altura dos colmos	$R = \text{ID} \otimes CS_{\text{Het}} \otimes \text{ID}$	591,57	649,68
Peso dos colmos	$R = \text{DIAG} \otimes \text{UNS} \otimes \text{UNS}$	46065,27	45929,51
Números de colmos	$R = \text{ID} \otimes \text{UNS} \otimes CS_{\text{Het}}$	45083,84	45187,28
TCH	$R = \text{DIAG} \otimes CS_{\text{Het}} \otimes \text{UNIF}$	44740,68	44837,57
TPH	$R = \text{DIAG} \otimes CS_{\text{Het}} \otimes \text{ID}$	15291,42	15361,84

## Capítulo II

---

### **Análise de marcas individuais utilizadas na detecção de QTLs considerando diferentes ploidias em cana-de-açúcar**

<sup>a</sup> Centro de Biologia Molecular e Engenharia Genética (CBMEG), Departamento de Genética e Evolução, Universidade Estadual de Campinas (UNICAMP), Cidade Universitária Zeferino Vaz, CP 6010, 13083-875 Campinas, SP, Brazil

<sup>b</sup> Departamento de Genética, Escola Superior de Agricultura Luiz de Queiroz (ESALQ), Universidade de São Paulo (USP), CP 83, 13400-970 Piracicaba, SP, Brazil

<sup>c</sup> Centro de Ciências Agrárias, Universidade Federal de São Carlos, Rodovia Anhanguera, Km 174, Araras - São Paulo - Brazil

---

### **Resumo**

A maioria das espécies vegetais é poliploide e representa aproximadamente 70% das angiospermas, sendo que muitas apresentam importância econômica. A presença de vários conjuntos de cromossomos no genoma destas plantas confere grande complexidade em sua estrutura genômica, e no caso especial da cana-de-açúcar, o nível de complexidade é aumentado devido à aneuploidia. A maioria dos modelos utilizados para estudos genéticos e genômicos foi desenvolvida para organismos diploides. Um exemplo prático em cana-de-açúcar é a interpretação do polimorfismo dos marcadores microssatélites como marcas dominantes e o fato de detectar apenas o polimorfismo entre os marcadores segregando em dose única. Neste contexto, os marcadores de polimorfismo de nucleotídeo único, por serem co-dominantes em cana-de-açúcar, foram utilizados com intuito de detectar indícios de QTLs associados a características de produção (diâmetro, altura, número e peso dos colmos e Toneladas de Cana por Hectare) e de qualidade (Brix, Pol%Cana, Pol%Caldo, pureza, fibra e Toneladas de Pol por Hectare). Um total de 240 indivíduos derivados do cruzamento entre as variedades comerciais de cana-de-açúcar SP81-3250 e RB925345 foram genotipados com SNPs por meio de espectrometria de massa pela Plataforma Sequenom iPLEX MassARRAY®. A ploidia dos locos SNPs foi estimada com o programa SuperMASSA. Pelo método de regressão linear foram encontradas 17 evidências de associação de QTL entre diâmetro dos colmos (quatro evidências), número de colmos (uma evidência), peso dos colmos (uma evidência), conteúdo de sólidos solúveis (duas evidências), teor de sacarose do caldo (três evidências), pureza (duas evidências), toneladas

de cana por hectare (duas evidências) e toneladas de Pol por hectare (duas evidências). A proporção da variação fenotípica explicada pelo genótipo variou de 1,6% a 11,1%. Todos os SNPs que apresentaram associações com as características mencionadas tiveram os níveis de ploidia variando de hexaploide a dodecaploide. Por correlação genotípica-fenotípica, foi detectado sete evidências de associação de QTL entre diâmetro dos colmos (uma evidência), conteúdo de sólidos solúveis (duas evidências), teor de sacarose da cana (uma evidência), teor de sacarose do caldo (duas evidências) e pureza (uma evidência). Os SNPs detectados com correlações genotípica-fenotípica significativas apresentaram níveis de ploidia variando tetradecaploide a icosaploide. O conhecimento de diferentes ploidias permitiu a detecção de QTLs em multi-dose que podem ser usadas como informações prévias sobre os prováveis QTLs para esta população de mapeamento, contribuindo para o avanço do conhecimento da genética da cana-de-açúcar.

Palavras-chave: poliploidia, marcadores moleculares

---

## Introdução

O processo de evolução que tornou a cana-de-açúcar (*Saccharum spp*) uma espécie poliploide teve início com a combinação de ao menos dois conjuntos de cromossomos originados do mesmo genoma, caracterizando-a como um organismo autopoliploide (Acquaah 2007; Chen 2010). A principal limitação ao realizar estudos genéticos em cana-de-açúcar é atribuída ao alto nível de ploidia e aneuploidia de seu genoma. A aneuploidia é devido ao cruzamento interespecífico entre *S. officinarum* ( $2n = 80$ ) e *S. spontaneum* ( $2n = 40-120$ ), resultando na variação cromossomal entre 100 a 130 (D'Hont *et al.* 1998). Outro fator que dificulta o estudo genético em cana-de-açúcar é atribuído ao fato que a maior parte das ferramentas moleculares disponíveis foi desenvolvida para espécies diploides e para aplicá-las em espécies poliploides faz-se necessário realizar ajustes para a correta interpretação dos dados.

Dentre as principais aplicações das técnicas moleculares destacam-se a construção de mapas genéticos e detecção de QTL (*Quantitative Trait Loci*) através do uso de

marcadores moleculares. Essas ferramentas são de extrema importância, pois aumentam o conhecimento das estruturas genéticas e da arquitetura genética das características quantitativas. Estudos envolvendo mapeamento de QTLs podem ser aplicados em programas de melhoramento de plantas para identificar marcadores moleculares ligados às características de importância agrônômica (Pinto *et al.* 2009).

Os marcadores moleculares que mais foram empregados para os estudos genéticos em cana-de-açúcar são o RFLP (*Restriction Fragment Length polymorphism*), RAPD (*Random Amplified Polymorphic DNA*), AFLP (*Amplified Fragment Length Polymorphism*) e SSR (*Simple Sequence Repeat*) por originarem um grande número de marcas polimórficas (Alwala e Kimbeng 2010).

A construção de mapas genéticos e a identificação de QTLs em organismos poliploides é baseada em marcadores de dose única (Wu *et al.* 1992), fato que representou um avanço nas últimas décadas. Marcadores em dose única representam alelos presentes em cópia única em apenas um dos genitores, segregando na proporção mendeliana de 1:1, ou em uma cópia presente nos dois genitores, segregando na proporção mendeliana de 3:1 (Wu *et al.* 2002). Existem poucos mapas genéticos de cana-de-açúcar que incluíram locos em multi-doses, ou seja, alelos que estão presentes em mais de uma cópia em um dos genitores. As multi-doses mais comuns são as duplas-doses e as triplas-doses. O uso de locos com alta dosagem alélica (multi-doses) poderá aumentar a cobertura do genoma da cana-de-açúcar (Aitken *et al.* 2007; Edemé *et al.* 2006), contudo, ainda não existem estudos para detecção de QTLs baseado em multi-doses para cana-de-açúcar.

Apesar dos marcadores RFLP, RAPD, AFLP e SSR gerarem, via de regra, muitas marcas polimórficas, eles não permitem estimar o número de alelos presentes em um loco em organismos poliploides (Garcia *et al.* 2013). Este fato evidencia a necessidade de utilizar marcadores moleculares que permitam a distinção de dosagem alélica desses organismos. Assim, os SNPs (*Single Nucleotide Polymorphisms*) são marcadores moleculares adequados para contornar esta limitação. Especialmente por serem marcadores amplamente distribuídos ao longo do genoma, bialélicos, co-dominantes e de alto rendimento (Giancola *et al.* 2006; Masouleh *et al.* 2009), são indicados para construção de mapas genéticos de alta resolução (Batley *et al.* 2003).

Atualmente, diferentes técnicas de genotipagem de SNPs estão disponíveis, entre elas destacam-se a plataforma Sequenom iPLEX MassARRAY® (Sequenom Inc., San Diego, California, USA), Illumina GoldenGate Genotyping Assay™ (Illumina Inc., San Diego, California, USA), Genotipagem-por-sequenciamento (*Genotyping by Sequencing – GBS*; Elshire *et al.* 2011) e RADseq (Baird *et al.* 2008). Todas estas tecnologias permitem a construção de mapas genéticos e a detecção de QTLs com maior precisão. Como consequência, o conhecimento da arquitetura genética das características quantitativas torna-se maior e sua aplicação em programas de melhoramento genético mais eficiente. Um grande número de estudos científicos envolvendo a associação entre marcador molecular e característica fenotípica está disponível para cana-de-açúcar (Al-Janabi *et al.* 2007; Piperidis *et al.* 2008; Aitken *et al.* 2008). Entre os modelos estatísticos para estabelecer uma relação entre o genótipo e fenótipo do indivíduo, a análise de marcas individuais (*Single Marker - SM*) é um método simples e amplamente utilizado para mapeamento de QTLs em cana-de-açúcar, aplicada por Al-Janabi *et al.* (2007), Piperidis *et al.* (2008), Aitken *et al.* (2008), Pinto *et al.* (2009). Esta abordagem pode ser implementada através de testes *t*, análise de variância, regressão linear simples e múltipla (Pastina *et al.* 2010). Entre as principais vantagens deste método destacam-se a simplicidade e rápida velocidade de execução das análises; disponibilidade de programas amplamente utilizados, tais como SAS (SAS Institute 1989) e pacotes do R (Team R Development Core 2008); o mapa genético não é necessário; permite a inclusão de marcadores não ligados nos mapas genéticos e pode ser estendido para modelos de múltiplos locos (Pinto *et al.* 2009; Pastina *et al.* 2010).

Utilizando modelo estatísticos de SM, marcadores moleculares SNPs foram usados para a detecção de QTLs associados às características de produção (diâmetro, altura, número e peso dos colmos e Toneladas de Cana por Hectare - TCH) e de qualidade (Brix, Pol%Cana, Pol%Caldo, pureza, fibra e Toneladas de Pol por Hectare - TPH). A associação marcador molecular e característica fenotípica foi realizada por meio de regressão linear e análise de correlação genotípica-fenotípica considerando uma marca por vez entre todos os indivíduos de uma população de mapeamento de cana-de-açúcar, derivada de um cruzamento bi-parental entre variedades comerciais.



## **Material e métodos**

### ***População segregante e dados fenotípicos***

Foi utilizada uma população segregante composta por 240 indivíduos F<sub>1</sub> derivada do cruzamento entre as variedades comerciais SP81-3250 (genitor feminino) e RB925345 (genitor masculino), ambas do Programa de Melhoramento de Cana-de-Açúcar da UFSCar (Universidade Federal de São Carlos), inserido nos trabalhos realizados pela RIDESA (Rede Interinstitucional de Desenvolvimento do Setor Sucroalcooleiro). A população segregante foi plantada em 2010 em dois locais (Araras e Ipaussu, ambos no estado de São Paulo, Brasil), e avaliada no primeiro, segundo e terceiro cortes anuais (2011, 2012 e 2013) para as características de produção (diâmetro, altura, número e peso dos colmos e Toneladas de Cana por Hectare - TCH) e de qualidade (Brix, Pol%Cana, Pol%Caldo, pureza, fibra e Toneladas de Pol por Hectare - TPH). O experimento foi instalado em Blocos de Federer (Federer, 1956), com três repetições. A coleta de dados, avaliação e análise dos dados fenotípicos foram descritos por Mancini *et al.* (2014 - dados não publicados).

### ***Extração de DNA***

O DNA genômico foi extraído do meristema apical da cana-de-açúcar, usando o protocolo CTAB modificado por Al-Janabi *et al.* (1999). A qualidade e concentração do DNA foi verificada por meio de eletroforese em gel de agarose 1%, no NanoDrop® 8000 Espectrofotômetro (Thermo Fisher Scientific Inc., Waltham, Massachusetts, USA) e Quantifluor® (Promega, Fitchburg, Wisconsin, USA).

### ***Marcadores moleculares***

Foram genotipados 290 locos SNPs desenvolvidos a partir do projeto SUCEST (Projeto de Sequenciamento de EST de Cana-de-Açúcar – *Sugarcane Expressed Sequence Tag*) e descritos por Garcia *et al.* (2013). O método de genotipagem dos SNPs utilizado foi

espectrometria de massa com ionização por dessorção a laser auxiliada por matrix e análise por tempo de voo (MALDI-TOF - *Matrix-Assisted Laser Desorption/ Ionization-Time of Flight*), através da plataforma Sequenom iPLEX MassARRAY® (Sequenom Inc., San Diego, Califórnia, USA). A genotipagem seguiu o guia descrito pelo fabricante (iPLEX *Gold Application Guide* - Sequenom) para baixo nível de plex por reação (*low iPLEX Gold Reactions* – Sequenom). Ambos os genitores da população segregante foram genotipados 20 vezes para cada loco SNP. A análise do SNP é baseada em informações alelo-específicas e por extensão de uma única base do *primer* seguida pela espectrometria de massa para detectar polimorfismos. Foi assumida a mesma eficiência de ionização para todos os alelos, com as intensidades de massa proporcionais à abundância de cada alelo.

### ***Classificação dos locos SNPs***

Os dados de genotipagem originados pelo Sequenom iPLEX MassARRAY® foram representado por gráficos de dispersão bi-dimensional e representam intensidades dos alelos de cada indivíduo. Todos os locos foram classificados usando o programa SuperMASSA (Serang *et al.* 2012), que determinou a probabilidade *a posteriori* do loco apresentar nível de ploidia variando de dois a 20.

Após todos os locos SNPs serem classificados para a ploidia mais provável, houve uma atribuição em três categorias diferentes, baseada na probabilidade *a posteriori* de cada ploidia por loco e também na dosagem para cada ploidia fixada, utilizando as medianas. Assim, as categorias foram divididas em: (1) categoria A.1: locos SNPs e medianas do conjunto de indivíduos com probabilidade *a posteriori* maior ou igual a 0,8; (2) categoria A.2: locos SNPs com probabilidade *a posteriori* maior ou igual a 0,8 e medianas do conjunto de indivíduos com probabilidade *a posteriori* menor que 0,8 e (3) categoria B: locos SNPs com probabilidade *a posteriori* menor que 0,8. Esta categorização foi necessária em virtude da baixa qualidade nos dados de genotipagem de alguns locos SNPs, com o intuito de não comprometer os resultados finais.

### ***Análises das características quantitativas e de marcas individuais***

Os SNPs pertencentes à categoria A.1 foram submetidos a análises de marcas individuais e foi adotado o modelo de regressão linear:

$$y_j = \mu + bx_j + e_j$$

Onde,  $y_j$  é o fenótipo do  $j$ -ésimo indivíduo  $j$ ,  $\mu$  é a média geral,  $b$  é o efeito aditivo,  $x_j$  é a dosagem alélica do  $j$ -ésimo indivíduo e  $e_j$  é o termo residual. A hipótese de nulidade testada foi a ausência de associação ( $H_0: b = 0$ ). Os critérios utilizados para detectar evidências de associação foram  $p$ -valor menor ou igual a 0,05 e  $p$ -valor menor ou igual a 0,001 (correção de Bonferroni para múltiplos testes) (Province 1999). Para os SNPs com evidência de associação, a proporção da variação fenotípica explicada pelo genótipo foi estimada por meio do coeficiente de determinação ( $R^2$ ). Todas as análises foram realizadas no programa R (<http://www.cran.r-project.org>).

### ***Correlação genotípica-fenotípica***

Os SNPs atribuídos às categorias A.2 e B foram utilizados para o estudo da correlação entre os genótipos e os fenótipos observados. No entanto, devido a baixa qualidade dos dados de genotipagem de tais SNPs, a informação genotípica utilizada para este estudo foi a razão entre as intensidades dos alelos de cada indivíduo, originadas pela genotipagem dos locos SNPs pelo Sequenom Iplex MassARRAY® (Sequenom Inc., San Diego, Califórnia, USA). O coeficiente de correlação de Pearson foi utilizado como medida para o cálculo da correlação entre as razões de intensidade dos alelos e as características em estudo. A hipótese de nulidade testada foi a ausência de correlação ( $H_0: r = 0$ ), com níveis de significância igual a 0,01 e 0,001. Todas as análises foram realizadas no programa R (<http://www.cran.r-project.org>).

## Resultados e discussão

### *Classificação dos locos SNPs*

A ploidia estimada para os 290 locos SNPs, através do programa SuperMASSA (Serang *et al.* 2012), variou de dois a 20, com probabilidade *a posteriori* variando de 0,22 a 1. Estas estimativas confirmam que o número de cromossomos homólogos em cana-de-açúcar não é constante, caracterizando sua condição aneuploide (Heinz e Tew 1987). As estimativas do nível de ploidia em 209 locos SNPs (79%) apresentaram probabilidade *a posteriori* maiores ou iguais a 0,80 e para os 89 locos SNPs (21%) restantes a probabilidade *a posteriori* foi menor que 0,80. (Figura 1 e Tabela 1). As probabilidades menores que 0,35 foram identificadas em locos classificados com nível de ploidia 20, e estes resultados foram associados com dados de baixa qualidade. Entretanto, todos os locos SNPs foram classificados no nível de ploidia com maior probabilidade *a posteriori* de acordo com o método de classificação.

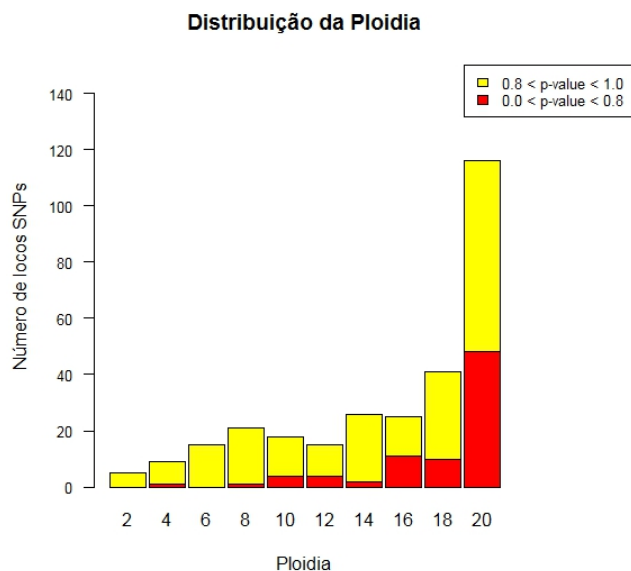


Figura 1. Distribuição dos níveis de ploidia classificados pelo SuperMASSA entre os indivíduos da população segregante derivada do cruzamento entre as variedades comerciais SP81-3250 e RB925345 de cana-de-açúcar. Em amarelo, a probabilidade *a posteriori* maior ou igual a 0,8 e em vermelho a probabilidade *a posteriori* menor que 0,8, de acordo com o nível de ploidia a qual foram classificados.

Tabela 1. Classificação dos 290 locos SNPs e a probabilidade *a posteriori* entre os indivíduos da população segregante derivada do cruzamento entre as variedades comerciais SP81-3250 e RB925345 de cana-de-açúcar.

Nível de ploidia	Quantidade de loco SNP por nível de ploidia	Varição da probabilidade <i>a posteriori</i>
2	5	0,95 - 1
4	9	0,47 - 1
6	15	0,87 - 1
8	20	0,77 - 1
10	18	0,48 - 1
12	15	0,54 - 1
14	26	0,60 - 1
16	25	0,42 - 1
18	41	0,38 - 1
20	116	0,22 - 1
Total	290	

Três locos SNPs (SugSNP0032, SugSNP0553 e SugSNP0467) foram escolhidos para exemplificar o genótipo em relação ao nível de ploidia classificado (Figura 2). Cada gráfico foi representado por 240 pontos que representam os indivíduos genotipados da população segregante. O eixo horizontal indica a intensidade do alelo com menor massa e o eixo vertical a intensidade do alelo com maior massa. As linhas pontilhadas correspondem ao nível de ploidia para cada loco e suas respectivas classes genotípicas (dosagem). Assim, a ploidia do loco é fixa enquanto que a dosagem é variável entre os indivíduos.

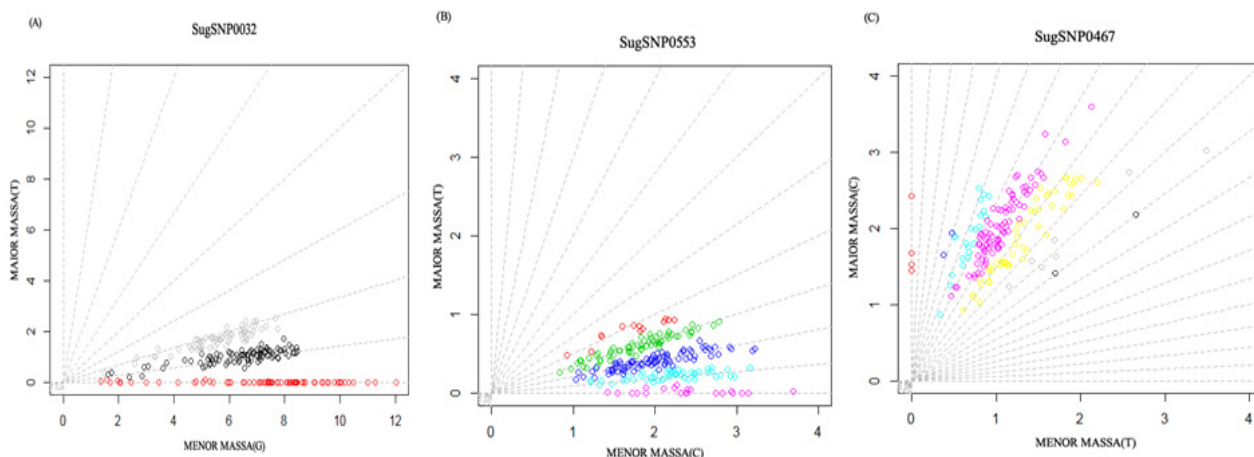


Figura 2. Classificação de três locos SNPs (SugSNP0032, SugSNP0553 e SugSNP0467) utilizando o programa SuperMASSA entre os indivíduos da população segregante de cana-de-açúcar derivada do cruzamento entre as variedades comerciais SP81-3250 e RB925345. O primeiro loco foi classificado como octaploide com três classes genotípicas (A); o segundo foi classificado como dodecaploide com cinco classes genotípicas (B) e o terceiro loco classificado como icosaploide (C).

No SugSNP0032, três nuvens de pontos estiveram claramente definidas, sendo classificado com nível de ploidia oito e probabilidade *a posteriori* igual a um. As três nuvens de pontos representam que para este loco foram encontradas três classe genotípicas (GGGGGGGG, TGGGGGGG e TTGGGGGG). Entre os genitores foram encontrados sete e uma dose para os nucleotídeos T e G, respectivamente, o que significa que uma dose do nucleotídeo T e sete doses do nucleotídeo G estão presentes nos oito cromossomos homólogos. Para o SugSNP0553 foi observado a formação de cinco nuvens e classificado com nível de ploidia 12 com probabilidade *a posteriori* igual a um. O genitor SP81-3250 apresentou 9T:3C (genótipo TTTTTTTTTCCC), ou seja, com nove nucleotídeos T e três nucleotídeos C presentes nos 12 cromossomos homólogos. O genitor RB925345 apresentou 11T:1C (genótipo TTTTTTTTTTTC), representando 11 nucleotídeos T e um nucleotídeo C presentes nos 12 cromossomos homólogos. Finalmente para o SugSNP0467 as nuvens apresentaram grande dispersão e houve confundimento entre as classes genotípicas. Este loco foi classificado com nível de ploidia 20. Apesar de ser possível encontrar modelos com alta probabilidade *a posteriori* com alto nível de ploidia (Garcia *et al.* 2013), para este loco, a probabilidade *a posteriori* foi menor que 0,35. Os genitores apresentaram 7C:13T (genótipo CCCCCCTTTTTTTTTTTTTT), que representa a presença de sete doses do nucleotídeo C e 13 doses do nucleotídeo T entre os 20 cromossomos homólogos.

A classificação dos SNPs nas três categorias de acordo com a probabilidade *a posteriori* resultou em 33 SNPs classificados na categoria A.1, 163 SNPs classificados na categoria A.2 e 94 SNPs classificados na categoria B (Figura 3). O principal motivo que levou a atribuição dos SNPs em categorias diferentes pode ser devido à grande quantidade de dados perdidos para estes locos SNPs, fazendo com que os resultados da classificação realizada pelo programa SuperMASSA sejam menos confiáveis, o que pode comprometer os resultados da análise de correlação. Assim, a estratégia para aumentar a qualidade dos dados foi utilizar as intensidades dos alelos de cada indivíduo, originadas pela genotipagem dos locos SNPs pelo Sequenom Iplex MassARRAY®, através da mediana da probabilidade *a posteriori* do conjunto de indivíduos a serem classificados de acordo com a ploidia mais provável. Desta forma, ao utilizar apenas os locos SNPs classificados na categoria A.1, na análise de marcas individuais, esperou-se evitar a ocorrência de associações falso-positivas.

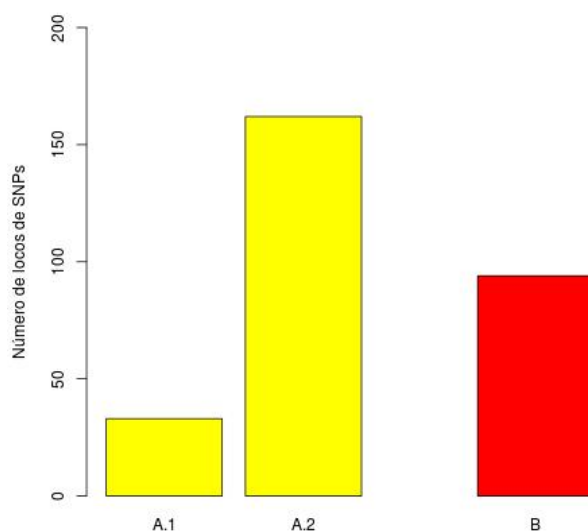


Figura 3. Distribuição dos locos SNPs em três categorias diferentes de acordo com a probabilidade *a posteriori* para a ploidia mais provável classificados pelo SuperMASSA. A.1: locos SNPs e medianas do conjunto de indivíduos com probabilidade *a posteriori* maior ou igual a 0,8; A.2: locos SNPs com probabilidade *a posteriori* maior ou igual a 0,8 e medianas do conjunto de indivíduos com probabilidade *a posteriori* menor que 0,8 e B: locos SNPs com probabilidade *a posteriori* menor que 0,8.

### *Análise de marcas individuais*

Os 33 locos SNPs pertencentes à categoria A.1 foram utilizados na análise de marcas individuais para as características de produção (diâmetro, altura, número e peso dos colmos e TPH) e de qualidade (Brix, Pol%Cana, Pol%Caldo, pureza, fibra e TCH). Considerando que apenas 33 SNPs classificados com boa estimativa de ploidia, optou-se por considerar o nível de significância igual a 5% ( $p < 0,05$ ), para detectar evidência de associação. Além disso, uma vez que a análise por marcas individuais é uma abordagem de baixo poder estatístico, e que a correção de Bonferroni é uma medida muito conservativa, as evidências de associação entre a marca e a característica poderiam não ser detectadas ao considerar correção para múltiplos testes.

No total, foram encontradas 17 (51,5%) evidências de associação entre a marca e a característica e seis possíveis efeito de pleiotropia, ou seja, um único loco influenciando mais de uma característica fenotípica (Figura 4). A magnitude da proporção da variação

fenotípica ( $R^2$ ) explicada pelo genótipo variou de 11,1% para diâmetro dos colmos a 1,6% para Brix (Tabela 2). O baixo valor da variação fenotípica que os marcadores conseguiram detectar pode ser explicado, ao menos em parte, devido ao alto nível de ploidia da cana-de-açúcar, e a existência de vários alelos em cada loco influenciando a mesma característica, fazendo com que o efeito individual de cada associação seja baixo (Hoarau *et al.* 2002). Pinto *et al.* (2009) e Anoni *et al.* (2014 - dados não publicados) utilizando uma população bi-parental derivada de um cruzamento entre variedades comerciais, também detectaram QTLs de pequeno efeito.



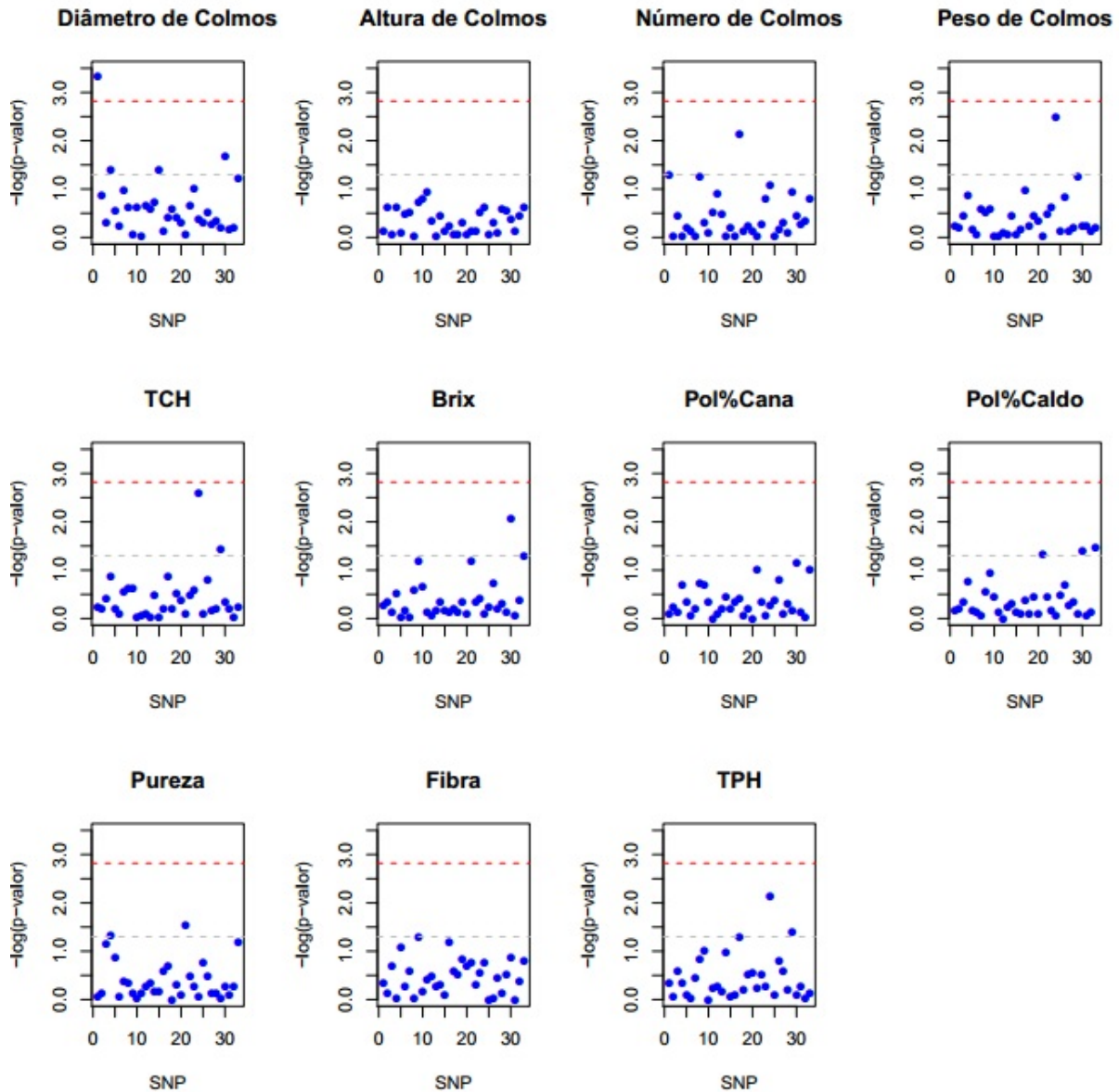


Figura 4. Análise de marcas individuais associados com diâmetro, altura, número e peso dos colmos, TCH, Brix, Pol%Cana, Pol%Caldo, pureza, fibra e TPH. A linha em cinza representa nível de significância de 5% e a vermelha 0,1%, (Correção de Bonferroni), considerando somente os locos SNPs pertencentes a categoria A.1 entre os indivíduos da população segregante derivada do cruzamento entre as variedades comerciais SP81-3250 e RB925345 de cana-de-açúcar.

Tabela 2. Detecção de associações entre a marca e a característica nos indivíduos da população segregante derivada do cruzamento entre as variedades comerciais SP81-3250 e RB925345 de cana-de-açúcar.

Característica	Número de colmos		Peso dos colmos		Diâmetro dos colmos		Brix		Pol%Caldo		Pureza		TCH		TPH	
	R <sup>2</sup> (%)	-log (p-valor)	R <sup>2</sup> (%)	-log (p-valor)	R <sup>2</sup> (%)	-log (p-valor)	R <sup>2</sup> (%)	-log (p-valor)	R <sup>2</sup> (%)	-log (p-valor)	R <sup>2</sup> (%)	-log (p-valor)	R <sup>2</sup> (%)	-log (p-valor)	R <sup>2</sup> (%)	-log (p-valor)
SugSNP_0756	3,2	2,13														
SugSNP_0689			3,7	2,47									3,9	2,60	3,1	2,12
SugSNP_0464					7,0	3,32										
SugSNP_0470					11,1	1,39					10,4	1,32				
SugSNP_0711					1,9	1,40										
SugSNP_0229					2,3	1,69	3,0	2,08	1,8	1,39						
SugSNP_0393							1,6	1,30	1,9	1,46						
SugSNP_0837									1,8	1,34	2,1	1,55				
SugSNP_0194													2,0	1,43	1,9	1,41

Foi encontrada associação entre o loco SNP SugSNP\_0756 com número de colmos e o loco SNP SugSNP\_0689 com peso dos colmos, com variação fenotípica de 3,2% e 3,7%, respectivamente. Entre as características, o maior número de associações encontradas foi para diâmetro dos colmos, detectando quatro locos SNPs, com variação fenotípica entre 1,9 a 11,1%. Ao observar Pol%Caldo, foram encontradas três associações com locos SNPs, cada uma explicando 1,8, 1,9 e 1,8% da variação fenotípica. Duas associações foram encontradas para Brix (1,6 e 3% da variação fenotípica), pureza (2,1 e 10,4% da variação fenotípica), TCH (2 e 3,9% da variação fenotípica) e TPH (1,9 e 3,1% da variação fenotípica) (Tabela 2).

Os possíveis efeitos pleiotrópicos foram encontrados entre as características peso dos colmos, TCH e TPH através do loco SNP SugSNP\_0689. Esse fato pode ser explicado porque TCH é uma estimativa calculada através do peso dos colmos, que por sua vez também é usada para estimar TPH. O loco SNP SugSNP\_0194 também revelou evidências de influência às características TCH e TPH. Dois locos SNPs detectados em diâmetro também foram associados a outras características, o SugSNP\_0470 relacionados a pureza e o SugSNP\_0229 a Brix e Pol%Caldo. Brix e Pol%Caldo foram influenciados pelos SNP SugSNP\_0393 e Pol%Cana e pureza pelo SNP SugSNP\_0837. Muitas dessas características também apresentaram correlações fenotípicas, como Brix e Pol, Pol e pureza, e TCH e TPH (Mancini *et al.* 2014- dados não publicados). Outros estudos envolvendo sorgo (Shiringani *et al.* 2010), milho (Clark *et al.* 2006) e trigo (Sukhwinder-Singh *et al.* 2012) também encontraram efeitos pleiotrópicos.

Percebe-se que as características ligadas ao teor de açúcar tiveram a influência de mesmas regiões no genoma, indicando a possibilidade de seleção simultânea por marcadores, para mais de uma característica. Muitos estudos realizados em cana-de-açúcar detectaram QTLs ligados a características relacionadas ao açúcar, tais como Brix (Hoarau *et al.* 2002, Aitken *et al.* 2006), Pol (Ming *et al.* 2001, 2002, Aitken *et al.* 2006, Pinto *et al.* 2009) e produção de açúcar (Ming *et al.* 2002).

Todos os SNPs que apresentaram associações com as características citadas acima tiveram os menores níveis de ploidia estimados, variando de hexaploides a dodecaploide, que são mais aceitáveis para cana-de-açúcar. Estes níveis de ploidia estão de acordo com estudos citogenéticos realizados em cana-de-açúcar, sendo proposto como número básico de cromossomos para o gênero *Saccharum* entre  $x=6, 8, 10$  e  $12$  (Sreenivasan *et al.* 1987). Posteriormente as espécies *S.*

*spontaneum* e *S. officinarum* foram descritas com número básico de cromossomos de  $x=8$  e  $x=10$  respectivamente (D'Hont *et al.* 1996, 1998).

### ***Correlação genotípica-fenotípica***

A correlação genotípica-fenotípica para os parâmetros de produção (diâmetro, altura, número e peso dos colmos e TPH) e de qualidade (Brix, Pol%C, Pol%J, pureza, fibra e TCH) foram feitas considerando nível de significância de 0,01% ( $p < 0,001$ ). Foi usado um  $p$ -valor mais conservativo quando comparado com a análise de marcas simples devido a probabilidade *a posteriori* da classificação dos SNPs ser menor que 0,8. Um total de sete correlações genotípica-fenotípicas significativas foram encontradas, em especial entre as características que são ligadas ao teor de açúcar (Figura 5, Tabela 3).

Observa-se que praticamente as mesmas características que tiveram correlações genotípicas-fenotípicas foram detectadas para a análise de marcas individuais, com exceção de Pol%Cana. O SNP SugSNP\_0375 foi correlacionado com Brix e Pol%Caldo, enquanto que o SNP SugSNP\_0197 com Pol%Cana e Pol%Caldo. Estes resultados, assim como os das marcas individuais, reforça a hipótese de que as características ligadas ao teor de açúcar podem ser influenciadas pelas mesmas regiões genômicas.

Todos os locos SNPs em que foram detectadas correlações genotípica-fenotípica significativas, a variação do nível de ploidia foi mais alta comparada a encontrada na análise de marcas individuais, variando de 14 a 20. Estes níveis de ploidia são menos prováveis para cana-de-açúcar (D'Hont *et al.* 1996, 1998), o que pode comprometer a confiabilidade destes resultados.

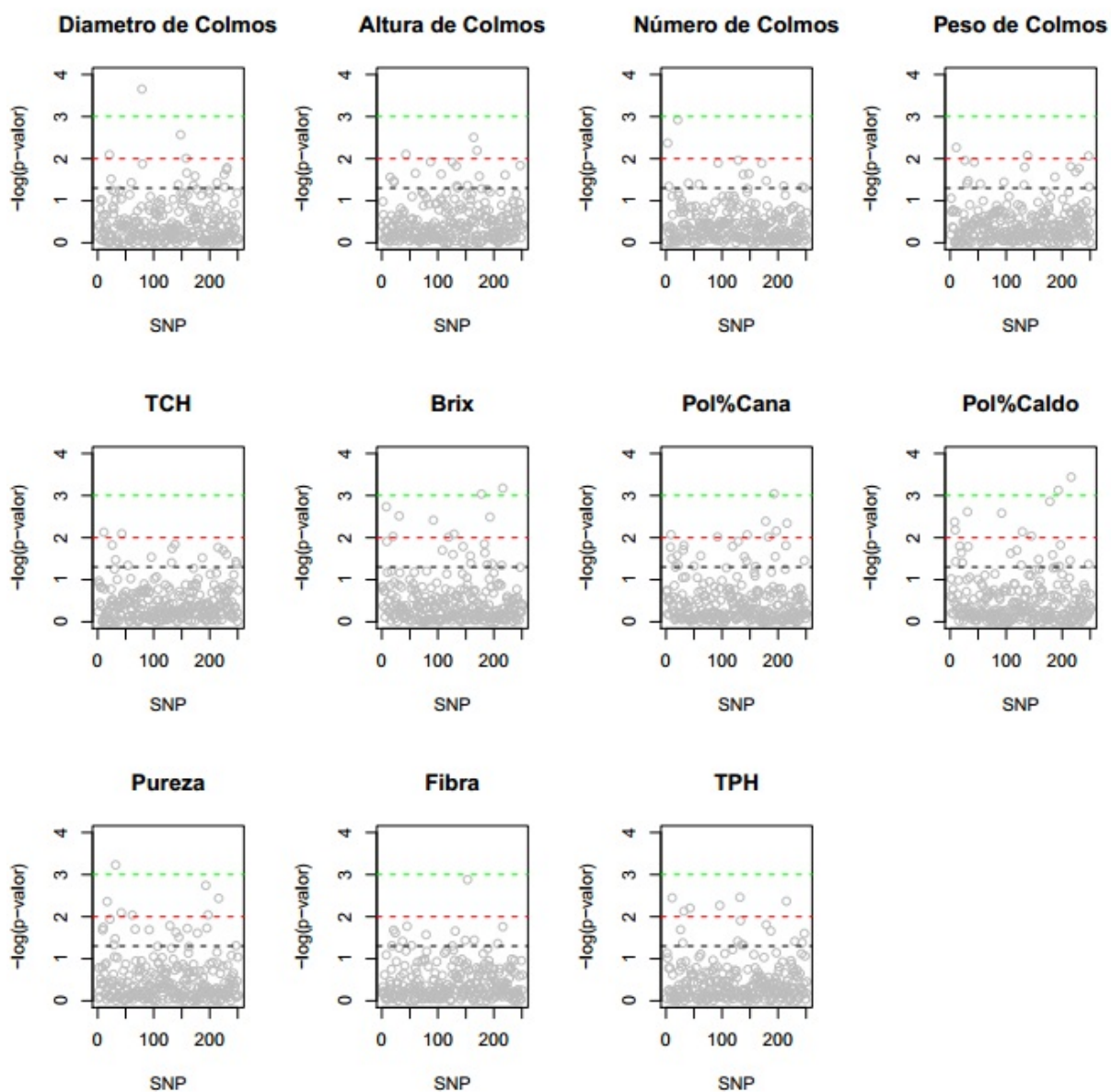


Figura 5. Correlação genotípica-fenotípica associados com diâmetro, altura, número e peso dos colmos, TCH, Brix, Pol%Cana, Pol%Caldo, pureza, fibra e TPH. A linha em preto representa nível de significância de 5%, a vermelha 1% e a verde 0,01%, considerando os locos SNPs pertencentes a categoria A.2 e B entre os indivíduos da população segregante derivada do cruzamento entre as variedades comerciais SP81-3250 e RB925345 de cana-de-açúcar.

Tabela 3. Correlação genotípica-fenotípica entre os indivíduos da população segregante derivada do cruzamento de cana-de-açúcar entre as variedades comerciais SP81-3250 e RB925345.

	Diâmetro dos colmos	Brix	Pol%Cana	Pol%Caldo	Pureza
	$-\log(p \text{ valor})$	$-\log(p \text{ valor})$	$-\log(p \text{ valor})$	$-\log(p \text{ valor})$	$-\log(p \text{ valor})$
SugSNP_0724	3,66				
SugSNP_0132		3,04			
SugSNP_0375		3,18		3,13	
SugSNP_0197			3,04	3,44	
SugSNP_0520					3,23

## Conclusão

O uso de uma população segregante derivada do cruzamento entre variedades comerciais favorece a detecção de QTLs associados com características de interesse econômico por ter passado por muitos ciclos de seleção. Sabe-se que é de grande importância para os Programas de Melhoramento da cana-de-açúcar conhecer essa associação e existem diferentes métodos estatísticos que a possibilitam. O motivo por ter optado pela análise de marcas individuais é devido ao baixo número de marcas polimórficas para a construção do mapa genético. Porém a abordagem usada permitiu a análise de multi-doses originadas dos marcadores SNPs, além de obter informações prévias sobre os prováveis QTLs para esta população segregante, com grande potencial de serem aplicados na seleção assistida por marcadores moleculares.

## Referências

- Acquaah G (2007) Principles of plant genetics and breeding Wiley-Blackwell, Malden.
- 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*, v.50, p.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, Forget L, Dookun A (1999) An improved and rapid protocol for the isolation of polysaccharide and polyphenol-free sugarcane DNA. *Plant Molecular Biology Reporter* 17:1-8.
- 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

Alwala S, Kimbeng CA (2010) Molecular Genetic Linkage Mapping in *Saccharum*: Strategies, Resources and Achievements. In: Genetics, Genomics and Breeding of Sugarcane. Henry, RJ, Kole C. (eds.), 69–96. CRC Press.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* 3, e3376.

Batley J, Barker G, O’Sullivan H, Edwards KJ, Edwards D (2003) Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. *Plant physiology*, v. 132, n. 1, p. 84-91.

Chen L, Lou Q, Zhuang Y, Chen J, Zhang X, Wolukau JN (2007) Cytological diploidization and rapid genome changes of the newly synthesized allotetraploids *Cucumis* × *hytivus*. *Planta* 225:603-614.

Clark RM, Wagler TN, Quijada P, Doebley J (2006) A distant upstream enhancer at the maize domestication gene *tb1* has pleiotropic effects on plant and inflorescent architecture. *Nature Genetics* VOLUME 38, NUMBER 5

Cordeiro GM, Taylor GO, Henry RJ (2000) Characterization of microsatellite markers from sugarcane (*Saccharum* sp.) a highly polyploid species. *Plant Science*, v.155, p.161-168.

Costet L, Le Cunff L, Royaert S, Raboin LM, Hervouet C, Toubi L, Telismart H, Garsmeur O, Rousselle Y, Pauquet J, Nibouche S, Glaszmann JC, Hoarau JY, D’Hont A (2012) Haplotype structure around *Bru1* reveals a narrow genetic basis for brown rust resistance in modern sugarcane cultivars. *Theor Appl Genet.* 125:825–836. DOI 10.1007/s00122-012-1875-x

Creste S., Tulmann Neto A., Figueira A. (2001) Detection of single sequence repeats polymorphisms in denaturing polyacrylamide sequencing gel by silver staining. *Plant Molecular Biology Reporter*, Athens, v.19, n.4, p.299-306.

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

Edemé SJ, Glynn NG, Comstock JC (2006) Genetic segregation of microsatellite markers in *Saccharum officinarum* and *S. spontaneum*. *Heredity*, v.97, p.366-375.

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

Federer WT (1956) Augmented (or hoonuiaku) designs. *Hawaiian Planters' Record*, 55: 191-208.

Garcia AAF, Mollinari M, Marconi TG, Serang OR, Silva RR, Vieira MLC, Vicentini R, Costa EA, Mancini MC, Garcia MOS, Pastina MM, Gazaffi R, Martins ERF, Dahmer N, Sforça DA, Silva CBC, Bundock P, Henry R, Souza GM, van Sluys MA, Landell MGA, Carneiro MS, Vincentz MAG, Pinto LR, Vencovsky R, Souza AP (2013) SNP genotyping allows an in-depth

characterization of the genome of sugarcane and other complex autopolyploids. *Scientific Reports*, 3: 3399.

Giancola S, Mckhann HI, Bérard A, Camilleri C, Durand S, Libeau P, Roux F, Reboud X, Gut IG, Brunel D (2006) Utilization of the three high-throughput SNP genotyping methods, the GOOD assay, Amplifluor and TaqMan, in diploid and polyploid plants. *TAG. Theoretical and applied genetics.*, v. 112, n. 6, p. 1115-1124.

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

Hoarau JY, Grivet L, Offmann B, Raboin LM, Diorflar JP, Payet J, Hellmann M, D’Hont A, Glaszmann JC (2002) Genetic dissection of a modern sugarcane cultivar (*Saccharum spp*) II detection of QTLs for yield components. *Theor Appl Genet* 105:1027–1037

Marconi TG, Costa EA, Miranda HRCAN, Mancini MC, Cardoso-Silva CB, Oliveira KM, Pinto LR, Molinari M, Garcia AAF, Souza AP (2011) Functional markers for gene mapping and genetic diversity studies in sugarcane. *BMC Research Notes* 4:264.

Masouleh AK, Waters DLE, Reinke RF, Henry RJ (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*, v. 7, n. 4, p. 355-363.

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.

Pastina MM, Pinto LR, Oliveira KM, Souza AP, Garcia AAF (2010) *Genetics, Genomics and Breeding of Sugarcane*. Henry, R. J. & Kole, C. (ed.) 117–148 (CRC Press).

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*, v.47, p.795–804.

Pinto LR, Oliveira KM, Marconi T, Garcia AAF, Ulian EC, Souza AP (2006) Characterization of novel sugarcane expressed sequence tag microsatellites and their comparison with genomic SSRs. *Plant Breeding* 125:378–384. DOI: 10.1111/j.1439-0523.2006.01227.x

Pinto LR, Garcia AAF, Pastina MM, Teixeira LHM, Bressiane JA, Ulian EC, Bidoia MAP, Souza AP (2009) Analysis of genomic and functional RFLP derived markers associated with sucrose content, fiber and yield QTLs in a sugarcane (*Saccharum spp.*) commercial cross. *Euphytica* DOI 10.1007/s10681-009-9988-2.



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

Province MA (1999) Sequential methods of analysis for genome scan. In: Rao DC, Province MA (eds) *Dissection of complex traits*. Academic Press, San Diego 583 p

Rossi M, Araujo PG, Paulet F, Garsmeur O, Dias VM, Chen H, Van Sluys MA, D'Hont AD (2003) Genomic distribution and characterization of EST-derived resistance gene analogs (RGAs) in sugarcane. *Molecular Genetics and Genomics*, v.269, p.406-419.

Schuelke M (2000) An economic method for the fluorescent labeling of PCR fragments. *Nature Biotechnology*. vol 18, p.233-234.

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

Shiringani AL, Frisch, M, 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

Sukhwinder-Singh, Hernandez MV, Crossa J, Singh PK, Bains NS, Singh K, Sharma I (2012) Multi-Trait and Multi-Environment QTL Analyses for Resistance to Wheat Diseases. *PLoS ONE* 7(6): e38008. doi:10.1371/journal.pone.0038008

Wu K, Burnquist W, Sorrels M, Tew T, Moore P, Tanksley S (1992) The detection and estimation of linkage in polyploids using single-dose restriction fragments. *Theor Appl Genet* 83:294–300.

Wu K, Burnquist W, Sorrels M, Tew T, Moore P, Tanksley S (1992) The detection and estimation of linkage in polyploids using single-dose restriction fragments. *Theor Appl Genet* 83:294–300.



## Considerações Gerais

---

A população de mapeamento da presente tese é constituída de 240 indivíduos F<sub>1</sub>. Foi desenvolvida pelo Programa de Melhoramento Genético da Cana-de-Açúcar da UFSCar/RIDESA, sendo derivada do cruzamento bi-parental entre as variedades SP81-3250 e RB925345, apresentando segregação para as características agroindustriais tonelada de cana por hectare (TCH), Pol%Cana, tonelada de pol por hectare (TPH) e resistência à ferrugem. A variedade SP81-3250 (CP70-1547 x SP71-1279) apresenta alta produtividade, com altos teores de sacarose e de fibra, resistência à ferrugem e período útil de industrialização (PUI) longo, enquanto a RB925345 (H59-1966 x genitor masculino desconhecido) apresenta alto teor de sacarose, alta produtividade, alto teor de fibra no início de safra, suscetibilidade à ferrugem e PUI curto.

Estudos genotípicos e fenotípicos essenciais para o mapeamento genético e detecção de QTLs foram realizados com esta população, garantindo uma confiável relação genótipo-fenótipo para cada indivíduo. Com o objetivo de confirmar a identidade dos indivíduos F<sub>1</sub> de todas as repetições do experimento plantado em Araras-SP e em Ipaussu-SP, foi realizada uma nova genotipagem em cerca de 1.500 indivíduos que estavam plantados nos respectivos campos, assim como a coleta dos dados fenotípicos dos mesmos. Através do método de UPGMA (método da distância genética média) e similaridade genética de Jaccard (Jaccard 1908), foi calculada a similaridade genética entre todos os indivíduos analisados através do programa NTSYS v.2.1 (Rohlf 2000). Os grupos que apresentaram similaridade inferior a 100% foram considerados como contaminantes não sendo observada nestes casos associação com outros grupos. Apenas no campo experimental de Ipaussu foi observada a ocorrência de quatro trocas entre os indivíduos. Os resultados apresentados indicam que houve erros a uma taxa insignificante (cerca de 1%) durante a implantação do experimento, considerando a magnitude destes campos experimentais.

A abordagem de modelos mistos na análise dos dados fenotípicos permitiu contornar o problema de dados desbalanceados sem interferir na qualidade dos resultados, além de considerar a interação genótipo-local-corte ao assumir a heterogeneidade da variância genética. Isso porque o modelo estatístico empregado representou fielmente o experimento, considerando a média dos 240 diferentes genótipos, em dois locais distintos e avaliados durante os anos de 2011, 2012 e 2013. Todos os resultados encontram-se discutidos no artigo apresentado no Capítulo I, com

exceção da avaliação de isoporização, florescimento e resistência à ferrugem. A exclusão destas características da análise por modelos mistos foi devido ao sistema de avaliação, baseado em escalas de notas, aliado a falta de tempo hábil para maior investigação de um modelo que representasse esse cenário.

Em paralelo às fenotipagens, os 240 indivíduos F<sub>1</sub> foram genotipados através de marcadores SSRs e SNPs. Uma das propostas desta tese foi adequar as metodologias de genotipagem para organismos poliploides, justificando a necessidade de uma série de otimizações para um resultado final robusto, confiável e de alta reprodutibilidade. Desta forma, foram necessários alguns ajustes e padronizações na metodologia de genotipagem dos SSRs para utilizar com maior precisão o sequenciador Li-Cor 4300 DNA Analyser. Assim como foi despendido tempo e esforço para garantir uma análise de qualidade utilizando os marcadores SNPs, processo descrito em maiores detalhes no Anexo I. Vale ainda ressaltar que a genotipagem dos SNPs pela tecnologia Sequenom iPLEX MassARRAY® (Sequenom Inc., San Diego, California, USA) foi estabelecida no LAGM através dos trabalhos desenvolvidos pela presente tese.

O grupo contava com um conjunto prévio de dados de um painel de associação brasileiro de cana-de-açúcar referente à genotipagem de cerca de 1000 locos SNPs. Este painel foi composto por 134 variedades de cana-de-açúcar envolvidas nos programas de melhoramento genético brasileiro da cana-de-açúcar. Também contava com informações sobre a genotipagem de 241 locos SNPs em uma população bi-parental (Marconi 2011). Tais informações foram usadas como norte para o início das otimizações. Nas análises prévias foi constatada uma possível tendência de aumento da variância dentro das classes genotípicas, à medida que as dosagens alélicas aumentavam. Enquanto que os locos SNPs em dose única apresentavam baixa variância e alta qualidade. Lembrando que até então, as análises percorreram somente entre os marcadores SNPs em dose única, portanto tornou-se imprescindível o aprimoramento da técnica de genotipagem para diminuir o ruído que as altas doses apresentavam, e assim garantir reações de amplificação dos marcadores SNPs com alta confiança e qualidade.

De posse dos resultados fenotípicos e genotípicos foi possível realizar a detecção dos QTLs para as características relacionadas as características de produção e de qualidade, utilizando multi-doses originadas dos marcadores SNPs, por meio da análise de marcas

individuais e correlação genotípica-fenotípica. Esses resultados encontram-se discutidos no Capítulo II e mostrou-se um método inovador no estudo de poliploides complexos.

A fim de verificar o comportamento genético dos marcadores moleculares genotipados, aqueles que apresentaram dose única (254 SSRs e 31 SNPs) foram utilizados para a construção do mapa genético prévio, através do programa OneMap (Margarido *et al.* 2007). Adotou-se um LOD mínimo de 4,5 e fração de recombinação 0,4, para definição dos grupos de ligação (GL), adotando a função de Kosambi (REF) para determinar as distâncias em centimorgans (cM) entre as marcas estimadas. Mesmo com a utilização de SNPs apresentando segregação 1:2:1, o mapa preliminar resultante obteve baixa cobertura. Um total de 163 marcas foram ligadas a 50 grupos de ligação, com cobertura total de 2147,6 cM (Anexo V). O mapa genético aqui apresentado ficou aquém de todos os mapas disponíveis na literatura científica (Hoarau *et al.* 2001, Aitken *et al.* 2005, Reffay *et al.* 2005, Raboin *et al.* 2006, Garcia *et al.* 2006, Aitken *et al.* 2007, Oliveira *et al.* 2007, Anoni *et al.* 2014 (dados não publicados). Tal resultado comprova a necessidade de uma maior saturação do mapa por marcadores moleculares e inviabiliza qualquer conclusão mais precisa em relação a este mapa genético preliminar.

Um fato curioso foi notado no GL4, o maior grupo de ligação formado (611cM), que agrupou todos os marcadores SNPs (20) que foram ligados ao mapa genético (Figura 2). Por análise de dois pontos, foi estimada a fase de ligação entre todos os marcadores ligados ao GL4. Constatou-se que os marcadores SNPs estavam em fase de repulsão em ambos os genitores. Notou-se que nas posições 284 e 588,4 cM do GL4 os marcadores não estavam ligados, o que resultaria em dois GLs de menor comprimento. Este cenário pode ser alterado quando mais marcas forem adicionadas ao mapa genético. No entanto, as sequências dos *clusters* de onde foram originadas os 20 SNPs ligados ao GL4 foram alinhadas contra o genoma do sorgo (*Sorghum bicolor*) (Tabela 1), por ser considerado o organismo mais próximo evolutivamente da cana-de-açúcar. Esta análise evidenciou similaridade entre cana-de-açúcar e sorgo nesta região analisada, contudo a ordenação do genoma não foi mantida.

No mesmo GL foram alinhados fragmentos de seis diferentes cromossomos de sorgo (cromossomos 1, 2, 3, 8, 9 e 10). Uma possível hipótese para explicar esta diferença reside em assumirmos que durante o processo de poliploidização, o genoma da cana-de-açúcar passou por uma série de rearranjos até chegar na qual é cultivada atualmente, não sendo mantida a arquitetura genômica do ancestral diploide. Tal hipótese poderá ser investigada em futuros

estudos de genômica comparativa. Outra abordagem para este aspecto é considerar que este resultado provavelmente deve-se ao número reduzido de marcadores utilizados, sugerindo uma nova organização assim que o mapa genético for saturado por mais marcadores moleculares. Um fator a ser lembrado é que as cultivares modernas de cana-de-açúcar são híbridos interespecíficos, apresentando alto nível de ploidia e aneuploidia, dificultando interpretações e hipóteses acerca da estruturação e evolução do genoma dessa espécie.

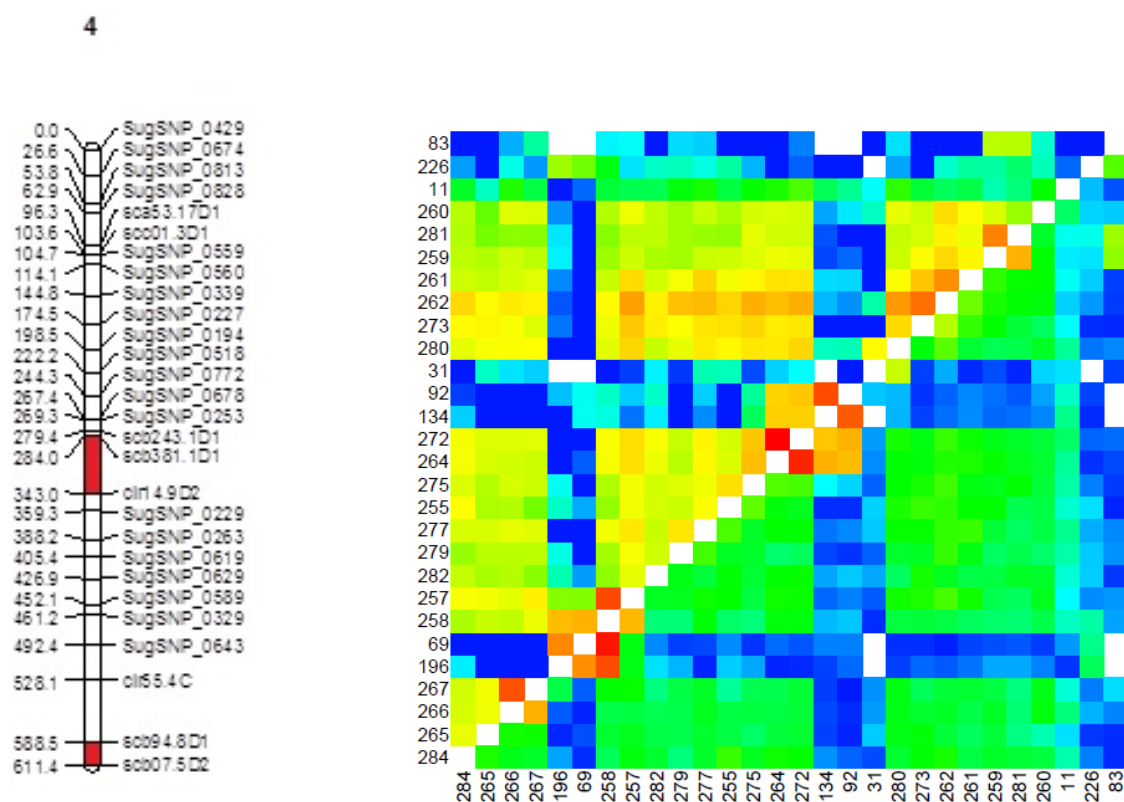


Figura 2. À esquerda, representação do maior grupo de ligação (GL4) formados por marcadores SSRs e SNPs, com cobertura total de 611,4 cM. Em vermelho sinaliza regiões do GL que não estão ligadas ao considerar os critérios estabelecidos. À direita, o *heatmap* para o GL4, sendo que as cores quentes representam ligação forte entre os marcadores moleculares, enquanto as cores frias ligações fracas, resultados embasados pelos valores de LOD (acima da diagonal) e pela fração de recombinação.

Tabela 1. Alinhamento dos SNPs ligados ao GL4 do mapa genético preliminar entre os indivíduos da população segregante derivada do cruzamento entre as variedades comerciais SP81-3250 e RB925345 de cana-de-açúcar, em relação ao genoma do sorgo.

SNP	Cluster/cana	Proteína/sorgo	Crom	Posição	Proteína anotada
SugSNP_0429	Contig2449b2	Sb01g044810.2	1	67.970.813 - 67.981.206	Putative MADS-domain transcription factor
SugSNP_0674	Contig1253b2	Sb03g026370.1	3	53.131.814 - 53.136.957	Putative uncharacterized protein
SugSNP_0813	Contig543b2	Sb09g019930.3	9	48.939.510 - 48.952.130	Pyruvate phosphate dikinase
SugSNP_0828	Contig472b2	Sb09g006050.1	9	8.731.871 - 8.733.975	Putative uncharacterized protein
SugSNP_0559	Contig1674b2	Sb01g000730.1	1	728.960 - 732.969	Putative heat shock transcription factor
SugSNP_0560	Contig1674b2	Sb01g000730.1	1	728.960 - 732.969	Putative heat shock transcription factor
SugSNP_0339	Contig2843b2	Sb02g024270.1	2	58.508.033 - 58.511.011	Putative uncharacterized protein
SugSNP_0227	Contig267b1	Sb10g030840.1	10	60.510.083 - 60.514.690	Catalase isozyme 1
SugSNP_0194	Contig368b1	Sb02g034490.3	2	69.068.354 - 69.073.043	Cathepsin B-like cysteine protease
SugSNP_0518	Contig1949b2	Sb08g020580.1	8	51.607.702 - 51.613.086	Auxin-responsive protein IAA30
SugSNP_0772	Contig652b2	Sb03g036650.1	3	64.710.004 - 64.718.045	Receptor-like protein kinase ARK1
SugSNP_0678	Contig1159b2	Sb10g030910.2	10	60.564.328 - 60.566.485	unkwon protein
SugSNP_0753	Contig747b2	Sb09g024340.1	9	53.900.807 - 53.903.670	Phosphoglycerate kinase
SugSNP_0229	Contig241b1	Sb01g047140.1	1	70.246.505 - 7.024.865	Aquaporin TIP1-1
SugSNP_0263	Contig109b1	Sb03g003220.1	3	3.357.017 - 3.362.737	NADP-dependent malic enzyme
SugSNP_0619	Contig1424b2	Sb01g033020.1	1	56.062.629 - 56.065.118	Putative uncharacterized protein
SugSNP_0629	Contig1293b2	Sb01g028760.2	1	50.183.023 - 50.186.127	Serine/threonine-protein kinase SAPK3
SugSNP_0589	Contig1552b2	Sb08g001480.1	8	1.439.461 - 1.444.900	Putative uncharacterized protein
SugSNP_0329	Contig2962b2	Sb08g006530.1	8	10.179.089 - 10.183.678	PHD finger protein, putative, expressed
SugSNP_0643	Contig1253b2	Sb03g026370.1	3	53.131.814 - 53.136.957	Putative uncharacterized protein

SNP: identificação do SNP genotipado; Cluster/cana: sequências de onde foram selecionados e desenvolvidos os SNPs em cana-de-açúcar; Proteína/sorgo: proteína da sequência de cana-de-açúcar correspondente em sorgo; Crom: cromossomo equivalente em sorgo; Posição: posição física, em pares de base, da sequência de cana-de-açúcar no genoma do sorgo.

O marcadores SNPs SugSNP\_0559 e SugSNP\_0560 foram desenvolvidos a partir do mesmo *cluster*, e informações adquiridas por meio do alinhamento contra o genoma do sorgo indicou que estes locos estão fisicamente ligados e posicionados no cromossomo 1 do sorgo (Tabela 1). Geneticamente estes locos estão a uma distância de aproximadamente 10 cM segundo o mapa genético construído nesta tese, o que pode significar uma distância muito pequena quando comparada com o tamanho total do genoma da cana-de-açúcar. Tal resultado sugere que estes marcadores estão ligados geneticamente e fisicamente, sendo corroborado pela posição em que foram ligados ao mapa genético preliminar. Já os marcadores SNPs SugSNP\_0674 e SugSNP\_0643, foram desenvolvidos a partir do mesmo *cluster* e estão fisicamente ligados no cromossomo 3 do sorgo, porém apareceram ligados ao GL4 de cana-de-açúcar com uma grande distância genética, o que sustenta o fato da necessidade de mais marcas para a construção do mapa genético.





## **Resumo dos resultados**

---

Os resultados apresentados pela presente tese alcançaram os objetivos de maneira satisfatória, e encontram-se aqui resumidos:

### **Capítulo I**

- Os dados fenotípicos para as características de produção (altura, diâmetro, peso de colmos) e de qualidade (sólidos solúveis, teor de sacarose do caldo e da cana, pureza do caldo e teor de fibra) foram coletados com sucesso durante três anos e dois locais
- Utilizando a abordagem por modelos mistos os resultados apresentados mostraram um ótimo controle ambiental para as características avaliadas
- O valor de herdabilidade para as características de produção variou entre 0,83 e 0,96 para peso dos colmos e toneladas de cana por hectare, respectivamente, enquanto que para as características de qualidade variou de 0,77 a 0,88 para pureza e fibra, respectivamente
- Foram encontradas 30 correlações fenotípicas significativas, confirmando que estes dados podem ser utilizados na detecção dos QTLs

### **Capítulo II**

- A metodologia de genotipagem dos SNPs utilizando a plataforma Sequenom iPLEX MassARRAY® (Sequenom Inc., San Diego, California, USA) foi estabelecida com êxito no LAGM
- Utilizando o programa SuperMASSA foi possível estimar a ploidia de cada loco SNP
- Foi identificado 17 evidências de associação de QTL pelo método de regressão linear e sete pela correlação genotípica-fenotípica, utilizando os marcadores SNPs entre as características relacionadas à produção e à qualidade
- A abordagem usada permitiu a análise de multi-doses originadas dos marcadores SNPs, mostrou-se um método inovador no estudo de cana-de-açúcar e com grande potencial de serem aplicados na seleção assistida por marcadores moleculares



## Conclusões

---

A genotipagem com SNPs da população F<sub>1</sub> permitiu a identificação de locos com diferentes ploidias e dosagens alélicas. Tais informações possibilitaram que um modelo genético apropriado para organismos com alto nível de ploidia fosse aplicado à análise dos dados fenotípicos obtidos, fornecendo assim, estimativas mais confiáveis e precisas da variação genética observada.

Concluimos que foi possível contribuir para o maior conhecimento genético da cana-de-açúcar ao detectarmos evidências de QTLs presentes em locos com variação de dosagem alélica, associados a características de importância econômica.

O estabelecimento de metodologias apropriadas para genotipagem permitiu o mapeamento de QTLs em organismos poliploides complexos é, portanto, uma das contribuições desta tese ao estudo de espécies poliploides em geral.



## Perspectivas

---

Os resultados da presente tese abriram novos caminhos no estudo genético da cana-de-açúcar resultando em um grande avanço científico. Porém este estudo necessita de continuidade para que o mapa genético possa ser saturado a tal ponto que permita o mapeamento preciso de QTLs. Neste contexto será aplicada uma nova abordagem, baseada na metodologia de genotipagem-por-sequenciamento ou GBS (*Genotyping-by-Sequencing* - Elshire *et al.* 2011). É uma estratégia bastante robusta e de alto rendimento, que pode ser aplicada em espécies de grande variabilidade genética (Elshire *et al.* 2011).

A fenotipagem desta população de mapeamento foi encerrada em 2013, resultando em dados de excelente qualidade. Ao Capítulo I serão adicionados os dados referentes às mesmas características avaliadas para outra população de mapeamento, derivada do cruzamento entre as variedades SP80-3280 e RB835486, para só então o manuscrito final ser submetido ao periódico *Field Crops Research*. Até o presente momento a avaliação de florescimento, isoporização e resistência à ferrugem foram excluídas das análises para não comprometer a qualidade dos resultados. Esse fato foi atribuído ao tempo hábil disponível para a conclusão do Doutorado. Assim, as três características aqui descritas serão alvo de estudos posteriores.

Vale reforçar que devido à complexidade do genoma da cana-de-açúcar os estudos que visam sua compreensão não devem ser baseados em estudos pontuais e isolados. Por isso outras atividades de pesquisa estão sendo realizadas com o intuito de colaborar com a elucidação da genética e genômica da cana-de-açúcar, destacando-se a análise do transcriptoma de folhas e bibliotecas de BACs (*Bacterial Artificial Chromosome*).



## Literatura citada

---

- Aitken KS, Jackson PA, McIntyre CL (2005) A combination of AFLP and SSR markers provides extensive map coverage and identification of homo(eo)logous linkage groups in a sugarcane cultivar. *Theoretical and Applied Genetics* 110:789-801
- 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 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
- Amorim L, Bergamin Filho A, Sanguino A, Cardoso CON, Moraes VA, Fernandes CR (1987) Metodologia de avaliação da ferrugem da cana-de-açúcar (*Puccinia melanocephala*). *Boletim Técnico Copersucar* 39:13-16
- 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
- Batley J, Barker G, O'Sullivan H, Edwards KJ, Edwards D (2003) Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. *Plant Physiology* 132:84-91
- Borém A, Santos FR (2004) *Biotechnology simplificada*. 2.ed., Viçosa, MG, pp 302
- Borém A, Miranda GV (2005) *Melhoramento de Plantas*. 4a Edição. Viçosa: Editora Universitária, pp 525p
- Bressiani JA (2001) Seleção seqüencial em cana-de-açúcar. Piracicaba. Tese (Doutorado) – Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Brasil
- Brookes AJ (1999) The essence of SNPs. *Gene* 234:177-186
- Bundock PC, Elliott FG, Ablett G, Benson AD, Casu RE, Aitken KS, Henry RJ (2009) Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploidy plant species using 454 sequencing. *Plant biotechnology journal* 7:347-354

Cabezas JA, Ibáñez J, Lijavetzky D, Vélez D, Bravo G, Rodríguez V, Carreño I, Jermakow AM, Carreño J, Ruiz-García L, Thomas MR, Martínez-Zapater JM (2011) A 48 SNP set for grapevine cultivar identification. *BMC Plant Biology* 11:153.

Cardle L, Ramsay L, Milbourne D, Macaulay M, Marsahall D, Waugh R (2000) Computation and experimental characterization of physically clustered simple sequence repeats in plants. *Genetics* 156:847-854.

Carneiro MS, Vieira MLC (2002) Mapas genéticos em plantas. *Bragantia* 61:89-100

Charcosset A, Moreau L (2004) Use of molecular markers for the development of new cultivars and the evaluation of genetic diversity. *Euphytica* 137:81-94

Chen X, Min D, Yasir TA, Hu Y-G (2012) Genetic Diversity, Population Structure and Linkage Disequilibrium in Elite Chinese Winter Wheat Investigated with SSR Markers. *PLoS ONE* 7(9): e44510. doi:10.1371/journal.pone.0044510

Collings FS, Brooks LD, Chakravarti A (1998) A DNA polymorphism discovery resource for research on human genetics variation. *Genome Research* 8:1229-1231

Consecana – Conselho nacional dos produtores de cana-de-açúcar, açúcar e álcool do estado de São Paulo (2006) Manual de instruções – Consecana-SP. Editora Consecana, Piracicaba p 112

Cordeiro GM, Taylor GO, Henry RJ (200) Characterization of microsatellite markers from sugarcane (*Saccharum* sp.) a highly polyploid species. *Plant Science* 155:161-168

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

Cruz CD, Regazzi AJ, Carneiro PCS (2004) Modelos Biométricos aplicados ao melhoramento genético. 3a Edição. Vicososa: Editora UFV p 460

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

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* 12:499-510

D'Hont A, Lu YH, Gonzáles DL, Grivet L, Feldmann P, Lanaud C, Glaszmann JC (1994) A molecular approach to unravelling the genetics of sugarcane, a complex polyploid of the andropogoneae. *Genome* 37:222-230

D'Hont A, Grivet L, Feldmann P, Rao S, Berding N, Glaszmann JC (1996) Characterisation of the double genome structure of modern sugarcane cultivars (*Saccharum* spp.) by molecular cytogenetics. *Molecular & general genetics* 250:405-413



Doerge RW (2002) Mapping and analysis of quantitative trait loci in experimental populations. *Nat Rev Genet* 3:43-52

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6(5): e19379. doi: 10.1371/journal.pone.0019379

Falconer DS, Mackay TF (1996) *Introduction to quantitative genetics*. 4 ed. Londres: Editora Longman Group p 464

Feltus FA, Wan J, Schulze SR, Estill JC, Jiang N, Paterson AH (2004) An SNP Resource for Rice Genetics and Breeding Based on Subspecies *Indica* and *Japonica* Genome Alignments. *Genome Research* 14:1812-1819

Ferreira ME, Grattapaglia D (1998) *Introdução ao uso de Marcadores Moleculares*. 3 ed. Brasília: EMBRAPA-CENARGEN p 220

Figueiredo P (2008) Breve histórico da cana-de-açúcar e do papel do Instituto Agrônomo no seu estabelecimento no Brasil. In: Dinardo-Miranda LL, Vasconcelos ACM, Landell MGA (Eds) *Cana-de-açúcar*. Campinas, Instituto Agrônomo 1:31-44

Ganal MW, Durstewitz G, Polley A, Berard A, Buckler ES, Charcosset A, Clarke JD, Graner E-M, Hansen M, Joets J, Le Paslier MC, McMullen MD, Montalent P, Rose M, Schon CC, Sun Q, Walter H, Martin OC, Falque M. (2011) A large maize (*Zea Mays* L.) SNP genotyping array: Development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. *PLoS ONE* 6(12): e28334. doi: 10.1371/journal.pone.0028334

Garcia AAF, Kido EA, Meza AN, Silva JAGD, Souza AP, Pinto LR, Pastina MM, Leite CS, da Silva JAG, Ulian EC, Figueira A, Souza HMB (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 AAF, Mollinari M, Marconi TG, Serang OR, Silva RR, Vieira MLC, Vicentini R, Costa EA, Mancini MC, Garcia MOS, Pastina MM, Gazaffi R, Martins ERF, Dahmer N, Sforça DA, Silva CBC, Bundock P, Henry R, Souza GM, van Sluys MA, Landell MGA, Carneiro MS, Vincentz MAG, Pinto LR, Vencovsky R, Souza AP (2013) SNP genotyping allows an in-depth characterization of the genome of sugarcane and other complex autopolyploids. *Scientific Reports* 3:3399

Gazaffi R (2009) *Desenvolvimento de modelo genético estatístico para mapeamento de QTLs em progênie de irmãos completos, com aplicação em cana-de-açúcar*. PhD Thesis, Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Brazil

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

- 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
- Griffin TJ, Smith LM (2000) Single-nucleotide polymorphism analysis by MALDI-TOF mass spectrometry. *Trends in Biotechnology* 18:77-84
- Grivet L, D'Hont A, Roques D, Feldmann P, Lanaud C, Glaszmann JC (1996) RFLP mapping in cultivated sugarcane (*Saccharum spp.*): Genome organization in a highly polyploid and aneuploid interspecific hybrid. *Genetics* 142:987-1000
- Grivet L, Arruda P. (2001) Sugarcane genomics: depicting the complex genome of an important tropical crop. *Current Opinion in Plant Biology* 5:122-127
- Guimarães CT, Honeycutt RJ, Sills GR, Sobral BWS (1999) Genetic maps of *Saccharum officinarum* L. and *Saccharum robustum* Brandes and Jew. *Ex. Grassl. Genet Mol Biol* 22:125-132
- Gupta PK, Varshney RK, Sharma PC, Armes B (1999) Molecular markers and their applications in wheat breeding. *Plant Breed* 118:369-390
- Gupta PK, Roy JK, Prasad M (2001) Single nucleotide polymorphisms: a new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. *Current Science* 80:524-535
- Gut IG (2004) DNA analysis by MALDI-TOF mass spectrometry. *Human mutation* 23:437-441
- Haff LA, Belden AC, Hall LR, Ross PL, Smirnov IP (2001) SNP Genotyping by MALDI-TOF Mass Spectrometry. In: Housby JN (Ed.). *Mass Spectrometry and Genomic Analysis*. Kluwer Academic Publishers p 149
- Hao C, Wang L, Ge H, Dong Y, Zhang X (2011) Genetic Diversity and Linkage Disequilibrium in Chinese Bread Wheat (*Triticum aestivum* L.) Revealed by SSR Markers. *PLoS ONE* 6(2): e17279. doi:10.1371/journal.pone.0017279
- Hoarau JY, Offmann B, D'Hont A, Risterucci AM, Roques D, Glaszmann JC, Grivet L (2001) Genetic dissection of a modern sugarcane cultivar (*Saccharum spp.*). I. Genome mapping with AFLP markers. *Theor Appl Genet* 103:84-97
- Heinz DJ, Tew TL (1987) Hybridization procedures. In: Heinz DJ (eds) *Sugarcane Improvement through Breeding*, Elsevier, Amsterdam p 313-342
- Henry RJ (2008) *Plant genotyping II: SNP technology*. [S.l.]: CABI, v. 2
- Jaccard P (1908) Nouvelles recherches sur la distribution florale. *Bulltin de la Société Vandoise des Science Naturelles* 44:223-270

Khlestina E, Salina E (2006) SNP markers: methods of analysis, ways of development and comparison on an example of common wheat. *Russian Journal of Genetics* 42:585-594

Kwok PY, Gu Z (1999) Single nucleotide polymorphism libraries: why and how are we building them? *Molecular Medicine Today* 12:538-543

Landell MGA, Alvarez R, Zimback L, Campana MP, Silva MA, Vila Nova JC, Pereira A, Perecin D, Gallo PB, Martins ALM, Kanthack RAD, Figueiredo P, Vasconcelos ACM (1999) Avaliação final de clones IAC de cana-de-açúcar da serie 1982, em Latossolo Roxo da região de Ribeirão Preto. *Bragantia, Campinas* 58:269-280

Lander E, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121:185-199

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 M, Qiao G, Jiang J, Yang H, Xie L et al.(2012) Transcriptome sequencing and de novo analysis for Ma bamboo (*Dendrocalamus latiflorus* Munro) using the Illumina platform. *PloS One* 7(10): e46766. doi: 10.1371/journal.pone.0046766

Lynch M, Walsh B (1998) *Genetics and analysis of quantitative traits*. Sinauer Associates, Sunderland

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

Margarido GRA, Souza AP, Garcia AAF (2007) OneMap: software for genetic mapping in outcrossing species. *Hereditas* 144:78-79

Meyers BC, Tingey SV, Morgante M (2001) Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Research* 11:1660-1676

Ming R, Liu SC, Lin YR, da Silva J, Wilson W, Braga D, van Deynze A, Wenslaff TF, Wu KK, Moore PH, Burnquist W, Sorrells ME, Irvine JE, Paterson AH (1998) Detailed alignment of saccharum and sorghum chromosomes: comparative organization of closely related diploid and polyploid genomes. *Genetics* 150:1663-1682

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 PH, Wu K-K, D'Hont A, Glaszmann JC, Tew TL, Mirkov TE, da Silva J, Jifon J, Rai M, Schnell RJ, Brumbley SM, Lakshmanan P, Comstock JC, Paterson AH (2006) Sugarcane

Improvement through Breeding and Biotechnology. In: J. Janick (Ed.) Plant Breeding Reviews, Oxford, UK: John Wiley & Sons, Vol 2, pp 118

Morgante M, Salamini F (2003) *From plant genomics to breeding practice*. Current Opinion in Plant Biotechnol 14:214-219

Mudge J, Anderson WR, Kehrer R, Fairbanks DJ (1996) A RAPD genetic map of *Saccharum officinarum*. Crop Sci 36:1362-1366

Oberacher H (2008) On the use of different mass spectrometric techniques for characterization of sequence variability in genomic DNA. Analytical and Bioanalytical Chemistry 391:135-149

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

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

Paterso AH, Tanksley SD, Sorrells ME (1991) DNA markers in plant improvement. Advances in Agronomy, San Diego 46:39-89

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, Bressiane JA, Ulian EC, Bidoia MAP, Souza AP (2009) Analysis of genomic and functional RFLP derived markers associated with sucrose content, fiber and yield QTLs in a sugarcane (*Saccharum spp.*) commercial cross. Euphytica 172:313-327

Poland JA, Bradbury PJ, Buckler ES, Nelson RJ. (2011) Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. PNAS 108:6893-6898

Ometto JC (1982) Bioclimatologia Vegetal. São Paulo: Ceres p 425

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

- 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
- Rickrt AM, Jeong JH,eyer S, Nagel A, Ballvora A, Oefner PJ, Gebhardt C (2003) First generation SNP/InDel markers tagging loci for pathogen resistance in the potato genome. *Plant Biotechnology Journal* 1:399-410
- Roach B (1972) Nobilisation of sugarcane. *Proceedings of the International Society for Sugar Cane Technology. Anais* 14:206–216
- Roach BT, Daniels JA (1987) A review of the origin and improvement of sugarcane. In: *Copersucar Internacional Sugarcane Breeding Workshop*. Piracicaba: Copersucar p 31
- Rohlf FJ (2000) *NTSYS-pc Numerical Taxonomy and Multivariate Analysis System*, version 2.1. Exeter Software, Setauket, New York, NY
- Serang O, Mollinari M, Garcia AAF (2012) Efficient Exact Maximum a Posteriori Computation for Bayesian SNP Genotyping in Polyploids. *PLoS ONE* 7(2): e30906. doi: 10.1371/journal.pone.0030906
- da Silva JAG, Sorrells ME, Burnquist W, Tanksley SD (1993) RFLP linkage map and genome analysis of *Saccharum spontaneum*. *Genome* 36:782-791
- da Silva JAG, Honeycutt RJ, Burnquist W, Al-Janabi SM, Sorrells ME, Tanksley SD, Sobral BWS (1995) *Saccharum spontaneum* L. “SES 208” genetic linkage map combining RFLP- and PCR-based markers. *Molecular Breeding* 1:165-179
- Souza Júnior CL (1989) Componentes da variância genética e suas implicações no melhoramento vegetal. Piracicaba: FEALQ p 134
- Souza Júnior CL (1995) Melhoramento de espécies de reprodução vegetativa. Piracicaba: ESALQ, Departamento de Genética, p.41, 1995. (Publicação Didática). STEVENSON, G.C. Genetics and breeding of sugarcane. London, Longmans p 284
- Syvänen AC (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Reviews Genetics* 2:930-942
- Tanksley SD (1993) Mapping polygenes. *Annual Review of Genetics* 27:205-233
- Ting N-C, Jansen J, Nagappan J, Ishak Z, Chin C-W, et al. (2012) Identification of QTLs Associated with Callogenesis and Embryogenesis in Oil Palm Using Genetic Linkage Maps Improved with SSR Markers. *PLoS ONE* 8(1): e53076. doi:10.1371/journal.pone.0053076
- Tost J, Gut IG (2003) Genotyping single nucleotide polymorphisms by mass spectrometry. *Mass spectrometry reviews* 21:388-418

Unica – União da indústria de can-de-açúcar. Disponível em <http://www.unica.com.br/> Acesso em 24 de Abril de 2014

Wang ML, Barkley NA, Yu JK, Dean RE, Newman ML, Sorrells ME, Pederson GA (2005) Transfer of simple sequence repeat (SSR) markers from major cereal crops to minor grass species for germplasm characterization and evaluation. *Plant Genetic Resources* 3:45-57

Wright SI, Bi IV, Schroeder SC, Yamasaki M, Doebley JF, McMullen MD, Gaut BS (2005) The effects of artificial selection on the maize genome. *Science* 308:1310-1314

Wu KK, Burnquist W, Sorrells ME, Tew TL, Moore PH, Tanksley SD (1992) The detection and estimation of linkage in polyploids using single-dose restriction fragments. *Theoretical and Applied Genetics* 83:294-300

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

Yu H, Xie W, Wang J, Xing Y, Xu C, et al. (2011) Gains in QTL Detection Using an Ultra-High Density SNP Map Based on Population Sequencing Relative to Traditional RFLP/SSR Markers. *PLoS ONE* 6(3): e17595. doi:10.1371/journal.pone.0017595

Zeng ZB (1993) Theoretical basis of separation of multiple linked gene effects on mapping quantitative trait loci. *Proc Natl Acad of Sci* 90:10972-10976

Zeng ZB (1994) Precision mapping of quantitative trait loci. *Genetics* 136:1457-1466

### **Estabelecimento de novas técnicas de genotipagem de marcadores SSRs e SNPs otimizadas para organismos poliploides complexos**

A análise do polimorfismo em organismos poliploides complexos requer o estabelecimento de metodologias apropriadas nas técnicas de genotipagem, o que justificou a série de otimizações realizadas.

Dois métodos de genotipagem de microssatélites (SSR) foram utilizados na população de mapeamento: (1) eletroforese em gel de poliacrilamida desnaturante a 6% revelado em coloração com prata, seguindo o protocolo descrito por Creste *et al.* (2001), que encontra-se estabelecido na rotina de genotipagem para cana-de-açúcar no Laboratório de Análise Genética e Molecular, UNICAMP (LAGM), e (2) pelo método de genotipagem baseado na detecção a laser na eletroforese em gel poliacrilamida desnaturante a 6%, analisado no sequenciador 4300 DNA Analyser (Li-Cor, Lincoln), que foi introduzido na rotina de genotipagem de cana-de-açúcar no LAGM, técnica que exigiu alguns ajustes.

A genotipagem no sequenciador 4300 DNA Analyser também é baseada em PCR (Reação de Polimerização em Cadeia) e consiste no uso de um dos *primers* carregando uma fluorescência, que pode ser *6-carboxyfluorescein* (FAN), *hexachloro-6-carboxy-fluoresceine* (HEX), *6-carboxy-X-rhodamine* (ROX) e *tetrachloro-6-carboxy-fluoresceine* (TET). Porém esses *primers* marcados com a fluorescência apresentam um elevado custo o que torna inviável a genotipagem de uma população de mapeamento, devido ao elevado número de locos necessários para gerar o mapa genético. A fim de tornar o método menos custoso foi utilizada a abordagem descrita por Schuelke (2000), cujo princípio é a utilização de três *primers*: um deles é o *forward* (Figura 3A) que carrega uma cauda M13 em sua terminação 5', composta por 18 bases (CAC GAC GTT GTA AAA CGA), o outro é *reverse* (Figura 3B) com sequência específica e reversa ao *forward* e por fim a fluorescência universal M13 (Figura 3C).

Porém, com a adição da cauda M13 ao *primer forward*, sua temperatura de anelamento tende a sofrer uma queda de 5° da temperatura em relação a sua temperatura otimizada sem a cauda. Assim, alguns ajustes foram necessários nas reações de PCR, desde a quantidade e concentração dos reagentes presentes no ensaio, até adaptações na ciclagem. Devido à

dificuldade de otimizar os *primers* com cauda M13 em cana-de-açúcar, optou-se pela PCR baseada em *touchdown*, com variação de 56 a 51°C. Essas temperaturas foram utilizadas por adequar a temperatura média dos *primers* sem a cauda de forma a abranger todos os marcadores utilizados. Duas etapas compõem a técnica sendo que nos primeiros ciclos da reação de PCR, o *primer forward* com a sequência M13 é incorporado ao produto de amplificação (Figura 3D). Posteriormente, seguem-se os ciclos visando o anelamento da fluorescência M13 (Figura 3E) que é incorporada ao produto final da amplificação (Figura 3F). A grande vantagem do método é a análise simultânea de dois diferentes fragmentos amplificados, através da detecção de distintas fluorescências a partir de dois comprimentos de onda (700 e 800 nanômetros). Além do mais, a eletroforese é visualizada em tempo real dispensando a coloração em prata e conferindo automação e maior sensibilidade à genotipagem dos marcadores microssatélites.

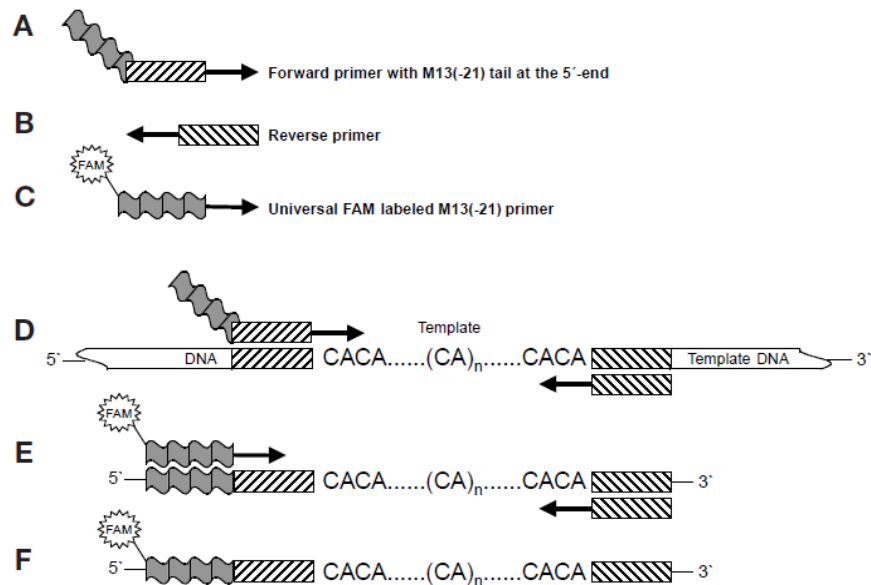


Figura 3: Método de PCR utilizando fluorescência. (A) *primer forward* marcado com a cauda M13 na extremidade 5', (B) *primer reverse*, (C) *primer* com fluorescência universal M13, (D) incorporação do *primer forward* com a cauda M13 ao produto de amplificação, (E) anelamento dos produtos amplificados a fluorescência universal M13 e (F) produto final detectado pelo sistema à laser (Fonte: Schuelke 2000).



Além da utilização dos SSRs foram também utilizados os marcadores SNPs (*Single Nucleotide Polymorphism*) na genotipagem da população de mapeamento em questão. Os SNPs foram escolhidos por serem marcadores codominantes em organismos poliploides, de natureza bialélica e abundantes no genoma. Em geral, os métodos de genotipagem disponíveis para marcadores SNPs não são aplicáveis a organismos complexos como a cana-de-açúcar. Por esta razão foi utilizada a análise por espectrometria de massas, que apresenta alta acurácia dos dados, pois permite a medição direta e rápida de produtos de DNA, ao invés de ler apenas uma marcação (fluorescente ou radioativa), além do alto rendimento na identificação de variações genéticas. As vantagens de utilizar tal metodologia são as possibilidades de incorporar marcadores com segregação 1:2:1 (Wu *et al.* 2002) aos mapas genéticos e identificar locos em multi-doses, aumentando o conhecimento genético da cana-de-açúcar, já que a priori sua ploidia é desconhecida.

Assim, a genotipagem de 290 maradores SNPs foi feita através da plataforma Sequenom iPLEX MassARRAY® (Sequenom Inc., San Diego, Califórnia, USA), e baseia-se na extensão de um *primer* alelo-específico por um terminador de massa modificada (Sequenom 2007). Os produtos dessa reação são analisados usando um espectrômetro de massas MALDI-TOF (*Matrix-Assisted Laser Desorption/ Ionization-Time of Flight*) e cada região polimórfica é detectada pela massa do alelo específico. Cada genitor foi genotipado 20 vezes para cada loco SNP afim de garantir alta confiança nos resultados.

As etapas do processo de genotipagem dos SNPs consistiram de: (1) definição do ensaio: oligonucleotídeos de captura dos SNPs e de extensão de bases únicas foram desenhados a partir das sequências selecionadas com possíveis locos SNPs, utilizando o banco de dados do SucEST (Projeto de sequenciamento de EST de cana-de-açúcar). O desenvolvimento dos oligonucleotídeos foi realizado por meio do programa *MassArray Assay Design* (Sequenom, Inc. San Diego), que também checou a possível formação de dímeros e agrupou conjuntos de 10 locos por reação. Esta etapa foi desenvolvida por Marconi (2011); (2) amplificação dos produtos contendo os SNPs; (3) tratamento com a enzima SAP (*Shrimp Alkaline Phosphatase*): neutralização dos dNTPs não incorporados na amplificação dos produtos contendo os SNPs, tornando-os inviáveis para futuras reações; (4) reação de extensão (*iPLEX Gold Reaction*): utilizando os produtos amplificados contendo os SNPs, o *primer* alelo específico se anelou exatamente adjacente ao sítio do SNP, estendendo apenas uma base; (5) limpeza dos produtos

amplificados: remoção do excesso de íons das reações de *iPLEX*; (6) transferência do produto amplificado para o SpectroCHIP e análise no espectrômetro de massa MALDI-TOF: as amostras foram transferidas para o SpectroCHIP; e (7) análise da intensidade produzida pela espectrometria de massa: através dos produtos estendidos alelo-específicos de massas diferentes, dependendo do nucleotídeo que foi adicionado, ou seja, dependendo da forma alélica presente naquela amostra (Figura 4). Assim, os dados obtidos fornecem dois sinais de intensidade de massa referentes ao loco SNP genotipado, sendo que a intensidade é proporcional a cada um dos alelos analisados. Essa distinção das dosagens alélicas permite observar classes genotípicas que não é possível visualizar com o uso dos marcadores microssatélites, fato que qualifica os marcadores SNPs na excelência de estudos genéticos para organismos poliploides. A classificação da ploidia foi realizada utilizando o programa SuperMASSA (Serang *et al.* 2012).

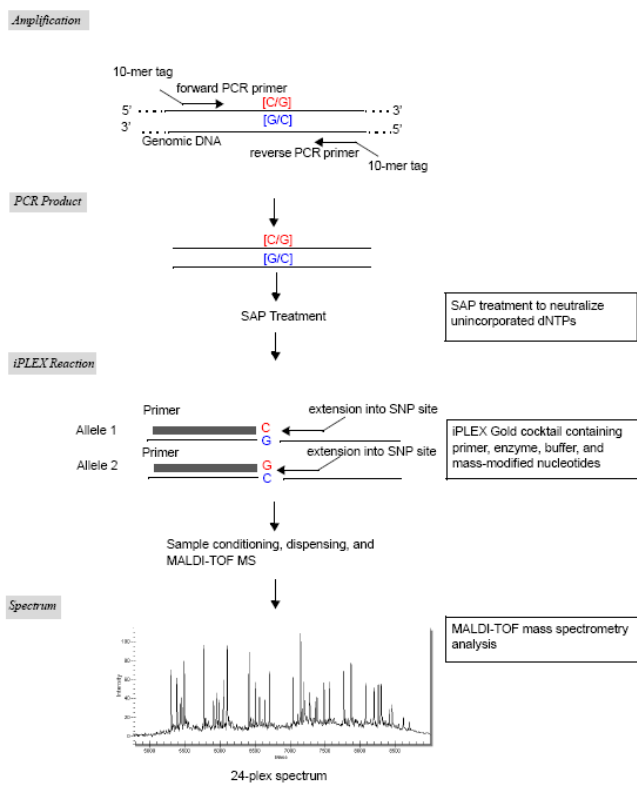


Figura 4. *iPLEX Gold Reaction*: amplificação dos fragmentos contendo os sítios polimórficos, reação de extensão e espectrometria de massas resultando em picos com medidas das massas das formas alélicas presentes nas amostras analisadas.

Foi realizada uma análise prévia a partir de um conjunto de dados disponíveis pelo grupo, que indicou uma possível tendência de aumento da variância dentro das classes genóticas, à medida que as dosagens aumentavam. Isso significa que os dados de loco SNP em dose única tem baixa variância e alta qualidade. Como as análises anteriores sempre foram feitas com marcadores em dose única, não houve necessidade em aperfeiçoar a técnica para doses maiores, já que esses dados sempre eram descartados. Notou-se claramente que os SNPs em dose única estão otimizados e à medida em que a dosagem aumenta o ruído do sinal também aumenta. Porém uma nova abordagem para detectar locos SNPs em altas doses vem sendo proposta através do uso do programa SuperMASSA (Serang *et al.* 2012). E para validar tal proposta, os protocolos foram otimizados para não comprometer a aplicação da metodologia para organismos poliploides.

A ideia inicial foi verificar como um mesmo loco SNP se comporta em diferentes níveis de ensaio multiplex, porém, realizar este tipo de análise torna a técnica economicamente inviável. Por essa razão, como uma primeira abordagem, foram analisados os resultados obtidos a partir da reação de captura da região contendo loco o SNP, em gel de agarose 3% corado com brometo de etídio. Para todos os testes realizados constatou-se a formação de produtos inespecíficos, com uma tendência de não formação de produtos inespecíficos para aqueles que já eram conhecidos como SNPs em dose única. Com o intuito de obter maior controle na especificidade dos produtos amplificados, foram realizados testes de gradiente (56° a 61°C) na reação de captura, e visualizados em gel de agarose 3% corado com brometo de etídio. A quantidade de produtos inespecíficos diminuíram, enquanto que a intensidade das bandas referente aos produtos específicos aumentaram. Também foi testada a interação entre os pares de *primers* que amplificam os fragmentos contendo locos SNPs, pela análise dos multiplex juntos (10 locos) e separados, visualizados em gel de agarose 3% corado com brometo de etídio. Conclui-se que através destes estudos que, aparentemente, a formação de produtos inespecíficos, bem como a quantidade de reações plexadas na mesma reação não interfere no resultado final.

Resultados apresentados em nota científica pelo Sequenom® (Sequenom, Inc. San Diego), “Targeted Genotyping Solutions for Plant and Animal Genomics”, (PAG 2012) demonstraram que uma melhor acurácia dos dados é obtida quando a quantidade de DNA está bem otimizada. A partir desse estudo, foi realizado uma nova abordagem nas otimizações, agora usando diferentes concentrações de DNA. Foram testadas as quantidades de 1, 4, 8 e 10 ng nas

variedades IACSP95-3018, IACSP93-3046 e SP80-3280, cada reação repetida 10 vezes. Para ter maior confiabilidade sobre a concentração ótima de DNA, as amostras envolvidas na otimização foram requantificadas utilizando o aparelho Quantifluor® (Promega, Fitchburg, Wisconsin, USA), método mais confiável por usar fluorescência que só se anela em DNA fita dupla. A partir dos resultados desse experimento constatou-se que a concentração de DNA não interferiu na acurácia dos dados. Porém depois de todas as otimizações houve um grande avanço na interpretação dos resultados, sendo fundamentais nas estratégias de análises em desenvolvimento. Após as etapas de otimizações, os 290 marcadores SNPs foram genotipados na população de mapeamento, através da Plataforma Sequenom iPLEX MassARRAY® seguindo as etapas já descritas acima.

### **Literatura citada**

Creste S, Tulmann Neto A, Figueira A (2001) Detection of single sequence repeats polymorphisms in denaturing polyacrylamide sequencing gel by silver staining. *Plant Molecular Biology Reporter* 19:299-306

Marconi TG (2011) Mapa Funcional em cana-de-açúcar utilizando marcadores moleculares baseados em SSR e SNP. Tese (Doutorado em Genética e Biologia Molecular) – Pós-graduação – Universidade Estadual de Campinas, UNICAMP, Brasil

PAG Plant and Animal Genome (2012) Sequenom - Targeted Genotyping Solutions for Plant and Animal Genomics. Nota científica. Disponível em: <<https://pag.confex.com/pag/xx/webprogram/Session1269.html>>

Schuelke M (2000) An economic method for the fluorescent labeling of PCR fragments. *Nature Biotechnology* 18:233-234

Sequenom (2007) *Typer 4.0 Manual*. San Diego: Sequenom, p 179

Serang O, Mollinari M, Garcia AAF (2012) Efficient Exact Maximum a Posteriori Computation for Bayesian SNP Genotyping in Polyploids. *PLoS ONE* 7(2):1-13

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

## Anexo II

---

### Mapping quantitative trait loci for yield components in a bi-parental cross between two commercial sugarcane (*Saccharum* spp.) varieties

ANONI, C.A.<sup>\*a</sup>; MANCINI, M.C.<sup>\*b</sup>; COSTA, E.A.<sup>\*b</sup>; GAZAFFI, R.<sup>c</sup>; PASTINA, M.M.<sup>d</sup>; PERECIN D.<sup>e</sup>; PINTO, L.R.<sup>f</sup>; SOUZA, A. P.<sup>b,g</sup>; GARCIA, A.A.F.<sup>a</sup>

<sup>a</sup> Departamento de Genética, Escola Superior de Agricultura Luiz de Queiroz (ESALQ), Universidade de São Paulo (USP), CP 83, 13400-970, Piracicaba, São Paulo, Brazil

<sup>b</sup> Centro de Biologia Molecular e Engenharia Genética (CBMEG), Universidade Estadual de Campinas (UNICAMP), Cidade Universitária Zeferino Vaz, CP 400, 13083-875, Campinas, São Paulo, Brazil

<sup>c</sup> Centro de Ciências Agrárias, Universidade Federal de São Carlos, Rodovia Anhanguera, Km 174, Araras, São Paulo, Brazil

<sup>d</sup> Núcleo de Biologia Aplicada, EMBRAPA Milho e Sorgo, Rodovia MG424, km 65, 35701-970, Sete Lagoas, Minas Gerais, Brazil

<sup>e</sup> Universidade Estadual Paulista (UNESP), Departamento de Ciências Exatas, Jaboticabal, São Paulo, Brazil

<sup>f</sup> Centro Avançado da Pesquisa Tecnológica do Agronegócio de Cana, IAC/Apta, CP 206, 14001-970, Ribeirão Preto, São Paulo, Brazil

<sup>g</sup> Departamento de Biologia Vegetal, Instituto de Biologia, Universidade Estadual de Campinas (UNICAMP), Campinas, SP, Brazil

\*These authors contributed equally to this work

---

### Abstract

Quantitative Trait Loci (QTL) mapping is an important tool in sugarcane (*Saccharum* spp.) breeding programs because it provides information about genetic effects, allelic interactions, and the position and number of QTLs and knowledge of the genetic architecture of quantitative traits. Such information combined with marker-assisted selection can help breeders reduce the development time of new sugarcane varieties. In this study, we performed QTL mapping of important agronomic traits in a commercial sugarcane cross via Composite Interval Mapping (CIM). An integrated genetic linkage map was constructed that considered 634 molecular markers (AFLP and SSR) displaying 1:1 and 3:1 segregation ratios. A total of 187 individuals were obtained from a bi-parental cross between IAC95-3018 and IACSP93-3046; these individuals were arranged in two locations across three harvests. The yield components evaluated were stalk diameter (SD), stalk weight (SW), fiber percentage (Fiber), sucrose content (Pol), and the soluble solid content (Brix). The genetic linkage map had a total length of 4370 cM, with 113 linkage groups in 15 putative homology groups. A total of 19 QTLs were detected for SD (four QTLs), SW (three QTLs), Fiber (five QTLs), Pol (four QTLs) and Brix (three QTLs). The proportion of phenotypic variation explained by each QTL ranged from 0.10% to 6.67%. The estimated additive effects and dominance effects for each mapped QTL were used to obtain segregation patterns for the QTLs that varied between 1:1:1:1, 1:2:1, 3:1 and 1:1 segregation ratios. Our results provide assistance for breeding programs by accounting for an improved dissection of complex traits in sugarcane.

*Keywords: polyploidy; linkage analysis; association; QTL*

---

## Introduction

Sugarcane is a complex autopolyploid and outbred species with high levels of heterozygosity, belongs to the Poaceae family and the *Saccharum* genus, of which the following species are often recognized: *S. barbari* ( $2n = 111 - 120$ ), *S. edule* ( $2n = 60 - 80$ ), *S. officinarum* ( $2n = 80$ ), *S. robustum* ( $2n = 60 - 80$ ), *S. sinense* ( $2n = 80 - 140$ ) and *S. spontaneum* ( $2n = 40 - 128$ ) (Mukherjee, 1957). Modern sugarcane varieties are derived from interspecific hybridization between *S. officinarum* and *S. spontaneum*, which results in highly polyploid and aneuploid plants. This hybridization represented a large breakthrough in modern sugarcane breeding. Disease problems were solved, yields increased, ratooning capacity improved, and plants were adapted for growth under several environmental conditions (Roach, 1972).

Genetic linkage map construction and Quantitative Trait Loci (QTL) mapping are important tools that allow for the optimization and improvement of breeding programs. In addition to information concerning the genetic architecture of traits, linkage and pleiotropy (Garcia et al., 2008), genetic maps also provide information regarding gene localization, the quantitative traits associated with specific genomic regions, genomic synteny analyses, and whole-genome quantification. However, the construction of genetic maps for sugarcane is more complicated and laborious compared to work in diploid species for the following reasons: i) the high level of polyploidy and aneuploidy results in a complex chromosomal segregation pattern during meiosis (Heinz and Tew, 1987); ii) the mapping progenies are derived from bi-parental crosses between highly heterozygous outbred parents, in which there are different numbers of alleles per locus, resulting in a mixture of marker segregation ratios in the progeny (Wu et al., 2002; Lin et al., 2003); and iii) the linkage phases between the markers are unknown (Pastina et al., 2012).

Wu et al. (1992) outlined these limitations and proposed the development of genetic linkage maps using only the segregation analysis of single-dose markers (SDMs). These markers represent alleles present in one copy in one of the parents, segregating at a 1:1 ratio in the progeny, or alleles present in one copy in both parents, segregating at a 3:1 ratio (Wu et al., 2002). The majority of the published sugarcane linkage maps were commonly constructed based on a double *pseudo-testcross* strategy (Grattapaglia et al., 1994). This strategy results in two individual maps, one for each parent (Daugrois et al., 1996; Ming et al., 2001; 2002; Aljanabi et al., 2007). Garcia et al. (2006) presented an integrated sugarcane linkage map constructed using the methodology proposed by Wu et al. (2002), in which the information provided by the markers segregating in 1:1 and 3:1 ratios were combined. This arrangement resulted in an integrated linkage map for both parents. Integrated linkage maps have some advantages. Maps with greater saturation are obtained, providing the best estimates concerning QTL location, and the linkage and linkage phases can also be more efficiently estimated compared to maps developed using the double *pseudo-testcross* strategy. This approach was also used for the sugarcane linkage map presented by Oliveira et al. (2007). At this time, none of the published sugarcane genetic linkage maps are considered saturated. The most saturated map presents with a total map length of 9774.4 cM (Aitken et al., 2014).

An important application of genetic linkage maps is QTL mapping, which allows for the identification of loci along the genome that contribute to trait phenotypic variance. Most agronomic traits have a quantitative nature with polygenic inheritance (Falconer and Mackay, 1996; Lynch and Walsh, 1998), such traits are highly influenced by the environment. QTL mapping studies in sugarcane are commonly performed for traits of agronomic importance, such

as brown rust (*Puccinia melanocephala*) resistance, smut (*Ustilago scitaminea*) resistance, yellow spot (*Mycovellosiella koepkei*) resistance, flowering time, sugar yield, stalk length, stalk diameter (SD), stalk number, stalk weight (SW) and fiber content (reviewed by Pastina et al., 2010). In sugarcane, QTL mapping for yield-related traits is usually performed using single-marker analysis, interval mapping and composite interval mapping for individual maps obtained through a double *pseudo-testcross* strategy. The first QTL mapping study in sugarcane, analyzing cane yield and sugar yield, was performed by Sills et al. (1995). This study was followed by other works that made important contributions to QTL mapping research, including Ming et al. (2001; 2002a; 2002b), Hoarau et al. (2002), McIntyre et al., (2005), da Silva and Bressiani, (2005), Reffay et al. (2005), Aitken et al. (2006; 2008), Piperidis et al. (2008), Pastina et al. (2012), and Shing et al. (2013).

The majority of QTL studies in sugarcane are performed based on two individual genetic maps, one for each parent (Daugrois et al., 1996; Ming et al., 2001; 2002; Hoarau et al., 2002; Reffay et al., 2005; Aitken et al., 2006; 2008; Shing et al., 2013). According to Pastina et al. (2010), the *pseudo-testcross* strategy reduces statistical power and does not allow for the estimation of additive and dominant effects separately, which complicates the interpretation of QTL mapping results. The first approach to consider information related to linkage phase, segregation patterns and QTL effects in a full-sib family was developed by Lin et al. (2003). The QTL results are biased when only the intervals of adjacent markers are considered, as there is no control for the interference caused by the QTLs located outside the mapping interval, resulting in false-positive detection. Recently, Gazaffi et al. (2014) developed an approach for Composite Interval Mapping (CIM) that considers full-sib progenies. In this methodology, linkage phase and segregation patterns are determined in addition to QTL location, which allows for the simultaneous analysis of QTLs with different segregation patterns. Additionally, Gazaffi et al. (2014) took cofactors into account during the mapping procedure, enabling more precise estimates of putative QTLs located along the genome. In this study, we constructed an integrated genetic linkage map in a bi-parental cross between two Brazilian commercial sugarcane varieties with Simple Sequence Repeats (SSR) derived from Expressed Sequence Tags (ESTs) and Amplified Fragment Length Polymorphism (AFLP) markers. We also performed QTL mapping using CIM for traits related to cane and sugar yields. We report on the segregation pattern and the additive and dominance effects for each mapped QTL.

## **Material and Methods**

### *Mapping population and field experiments*

Full-sib progeny, a composite of 187 individuals, were obtained from a bi-parental cross between the elite clone IACSP95-3018 (female parent) and the variety IACSP93-3046 (male parent) developed at the Sugarcane Breeding Program at the Instituto Agronômico de Campinas (IAC). IACSP95-3018 is a promising clone used as parent in the IAC Sugarcane Breeding Program. IACSP93-3046 has a high level of sucrose, good tillering and an erect stool habit and is recommended for mechanical harvesting and prone environments.

The full-sib progeny and the sugarcane varieties RB835486 and SP81-3250, which were used as common checks, were planted in 2007 in the Sales de Oliveira region in the state of São Paulo, Brazil. Plants were placed in a randomized complete block design with four replicates, and 2-meter rows were spaced 1.5 meters apart. The same population was also planted in 2011 in the

Ribeirão Preto region in São Paulo, Brazil. This planting followed the same design, however, only three replicates were planted. Both parents and two varieties (SP81-3250 and RB83-5486) were included in each replicate as checks. The field experiment was evaluated in 2008 (plant cane) and 2009 (ratoon cane) in Sales de Oliveira and in 2012 (plant cane) and 2013 (ratoon cane) in Ribeirão Preto.

#### *Molecular marker data and map construction*

To construct the linkage map, 187 individuals from the full-sib progeny were screened. The following two types of molecular markers were used: SSRs, which were divided among genomic (gSSR) and Expressed Sequence Tags (EST-SSR), and AFLPs. Of the 140 SSRs used, 105 were EST-SSR primers developed in the Brazilian Sugarcane EST Project (SUCEST), presented in Pinto et al. (2004; 2006), Oliveira et al. (2007; 2009), and Marconi et al. (2011). The remaining 35 SSRs were gSSRs developed by Cordeiro et al. (2000) and CIRAD (*Centre de Cooperation Internationale em Recherché Agronomique pour le Développement*, Montpellier, France) as described in Rossi et al. (2003). For the AFLPs, DNA was digested with *EcoRI* and *MseI*, and a total of 25 selected primer combinations (*EcoRI/MseI*) were screened according to Vos et al. (1995). The amplification products from the SSRs and AFLPs were separated by electrophoresis on 6% denaturing polyacrylamide gels. The gels were visualized by silver staining according to Creste et al. (2001). Markers were scored based on their presence (1) or absence (0) in the parents and the segregating progeny. Marker segregation (1:1 or 3:1 presence to absence ratio) was verified through a chi-square test and a Bonferroni correction was considered to control for type I error in multiple tests (Province, 1999).

The SSR markers were named according to the origin of their locus and were followed by a number referring to the amplified allele. The AFLP markers were identified by six letters representing the *EcoRI/MseI* selective primer and were followed by a number referring to the amplified fragment. All of these markers received a letter to denote the origin of the parental polymorphism according to the cross type described by Wu et al. (2002). For example, the D<sub>1</sub> locus is heterozygous in IACSP95-3018 and homozygous in IACSP93-3046. The D<sub>2</sub> locus is homozygous in IACSP95-3018 and heterozygous in IACSP93-3046. Both D<sub>1</sub> and D<sub>2</sub> present with 1:1 segregation ratios. Locus C is heterozygous in both parents, with a 3:1 segregation ratio.

The linkage map was constructed using the OneMap package (Margarido et al., 2007) and only SDMs were considered. To avoid false-positive linkages, a stringent LOD score of 5.77 and recombination frequency of 0.40 were used to determine the linkage groups (LGs). The higher LOD value was used due to the large number of evaluated markers and the high number of chromosomes in modern sugarcane cultivars (over 100). To determine the marker order and linkage phases in each LG with a maximum of five markers, all possible orders were compared, and the most likely order was considered. For LGs with more than five markers, all possible orders for the five most informative markers were compared, and then other markers were added sequentially and were arranged in the most likely initial order at positions with greater likelihood. The distances in the genetic map were expressed in centimorgans (cM) based on the Kosambi function (Kosambi, 1944).



### Phenotypic data analysis

The quantitative traits measured corresponded to important economic components in sugarcane production, including soluble solid content (Brix), fiber percentage (Fiber), sucrose content (Pol), stalk weight (SW), and stalk diameter (SD). Each measurement was obtained from a sample of 10 stalks that were harvested from each individual plot according to the methods, for plant cane and ratoon crops, described by Consecana (2006). Brix corresponds to the total dissolved solids in cane juice. Fiber refers to the insoluble mater in water that is present in the stalk.

To evaluate the phenotypic data obtained from different locations and harvests, an appropriate mixed model was adjusted by comparing different structures for the variance-covariance (VCOV) matrices of genetic ( $\mathbf{G}$ ) and non-genetic ( $\mathbf{R}$ ) effects. The statistical model used for each trait separately was as follows (bold terms indicate random variables):

$$Y_{ijrkm} = \mu + L_k + H_m + B_{j(km)} + \mathbf{G}_{i(km)} + \mathbf{e}_{ijrkm}$$

where:  $Y_{ijrkm}$  is the phenotype of the  $i^{th}$  genotype ( $i = 1, \dots, n$ ) in the  $j^{th}$  block ( $j = 1, 2, 3, 4$ ), the  $r^{th}$  replication ( $r = 1, \dots, 4$ ), the  $k^{th}$  location ( $k = 1, 2$ ) and the  $m^{th}$  harvest ( $m = 1, 2$ );  $\mu$  is the trait mean;  $L_k$  is the location effect;  $H_m$  is the harvest effect;  $B_{j(km)}$  is the block effect of the  $j^{th}$  block in the  $k^{th}$  location and the  $m^{th}$  harvest;  $\mathbf{G}_{i(km)}$  is the genetic effect of the  $i^{th}$  genotype in the  $k^{th}$  location and the  $m^{th}$  harvest; and  $\mathbf{e}_{ijrkm}$  is the non-genetic effect.

The VCOV matrix  $\mathbf{G}$  was obtained via the Kronecker direct product of  $\mathbf{G}_M \otimes \mathbf{I}_{ng}$ , where  $\mathbf{G}_M = \mathbf{G}^L \otimes \mathbf{G}^H$ ,  $\otimes$  is the Kronecker direct product of the genetic effects for the matrices  $\mathbf{G}_M$  e  $\mathbf{I}_{ng}$ , and  $\mathbf{I}_{ng}$  is an identity VCOV matrix of the genotypes as proposed by Pastina et al. (2012). Likewise, the  $\mathbf{R}_M$  matrix was obtained via the Kronecker direct product of  $\mathbf{R}^L$  (the residual effects between locations) and the  $\mathbf{R}_M$  (the residual effects between harvests) matrices; therefore,  $\mathbf{R}_M = \mathbf{R}^L \otimes \mathbf{R}^H$ . It was assumed that  $\mathbf{e} \sim N(0, \mathbf{R})$  and  $\mathbf{g} \sim N(0, \mathbf{G})$ , where  $\mathbf{e} = 11111, \dots, e_{ijrkm}$  and  $\mathbf{g} = 111, \dots, g_{ikm}$ , respectively. Different models for VCOV structures were compared via AIC (Akaike Information Criterion) (Akaike, 1974) and BIC (Bayesian Information Criterion) (Schwarz, 1978). For the  $\mathbf{G}_M$  matrix, five different models were compared. The models (a-d) (see Table 1 in the results) consider the combination of the location-harvest as a specific environment for the  $\mathbf{G}_M$  matrix structures. Model (e) (see Table 1 in the results) considers the direct product of location and harvest. Model (a) specifies homogeneous variance and no covariance (correlation) between environments. Model (b) considers heterogeneous variance but not the correlation between environments. Model (c) takes into account an approximation of an unstructured VCOV matrix that had no correlation between environments. Model (d) specifies different variances and no correlation between environments. Model (e) considers different variances and no correlation for locations and harvests separately. The models described above were also compared for the  $\mathbf{R}_M$  matrix. Once  $\mathbf{G}$  and  $\mathbf{R}$  VCOV matrices were modeled, BLUPs (Best Linear Unbiased Predictors) combining information concerning the locations and harvests were obtained for each trait. The variance components were then estimated using Residual Maximum Likelihood. All of the statistical analysis previously described was performed using Genstat 16<sup>th</sup> edition (Payne et al., 2013).

The broad-sense heritability coefficient was calculated for each trait based on the following equation:

$$h^2 = \frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_{G \times E}^2}{km} + \frac{\sigma^2}{r}}$$

where:  $\sigma_G^2$ ,  $\sigma_{G \times E}^2$  and  $\sigma^2$  are the genotypic, the genotype-environment interaction and the residual variance components, respectively;  $r$  is the harmonic mean of the replicates; and  $km$  is the harmonic mean of the factorial combination of locations and harvests in different environments. Genetic correlations were estimated for each pair of traits based on BLUPs using the Pearson correlation coefficient implemented using R software (R Development Core Team 2013).

### *Quantitative trait analysis*

To test the association between phenotype and genotype, each of the five traits related to cane yield and sugar yield were analyzed separately using CIM methodology (Zeng, 1993) according to the extended approach described by Gazaffi et al. (2014). The model is based on three orthogonal contrasts and estimates the additive effects for each parental and intralocus interaction (dominance effect) for the QTL alleles. QTL mapping analyses were performed as described below: (i) conditional multipoint probabilities of the QTL genotype for all positions in a discrete grid of evaluation points with a step size of 1 cM along the genome were estimated via Hidden Markov Models (HMM), which were implemented using the OneMap package (Margarido et al., 2007); (ii) genetic predictors were estimated for each position along the genome, as described in Gazaffi et al. (2014); (iii) the genome scan was performed (in a 1 cM grid) to detect the QTLs; (iv) following the identification of a QTL, the significance levels of the additive ( $\alpha_p$ ,  $\alpha_q$ ) and dominance ( $\delta_{pq}$ ) effects were verified, and the segregation pattern was estimated (Gazaffi et al., 2014). The inclusion of cofactors was based on a multiple linear regression analysis using stepwise selection based on AIC criteria. For this purpose, we considered a maximum of 27 cofactors and a window size of 20 cM. The threshold used for the detection of a QTL was calculated using 1,000 permutations and a significance level of 0.05 (Churchill and Doerge, 1994) and the distribution of the second highest peak for each LG was considered (Chen and Storey, 2006). For positions with evidence of putative QTLs, significant marginal effects for  $\alpha_p$ ,  $\alpha_q$  and  $\delta_{pq}$  were verified through the chi-square statistic converted to LOD Scores, with a significance level of 0.05. In this step, no corrections for multiple tests were considered, as chi-square statistics were performed only in positions with evidence of QTLs and not in the entire genome. The proportions of phenotypic variance ( $R^2$ ) explained by each detected QTL were obtained for all of the effects simultaneously. All analyses were performed in the R environment (R Development Core Team 2013).

## **Results**

### *Linkage map*

In total, 140 SSR and 25 AFLP primer combinations were used. We scored 1102 polymorphic markers in this mapping progeny. Of these markers, 634 (57.5%) were segregating in single dose (SD), according to 1:1 (377) and 3:1 (257) proportions. 420 markers (66.2%) were linked in 113 LGs, and 214 (33.8%) remained unlinked. The LGs varied in length, from 1.0 cM to 142.9 cM, the average length was 39 cM by LG, and the accumulated length was 4370 cM. The average distance between markers was 10.7 cM, with an irregular distribution along the

chromosomes. Seventeen gaps were observed, ranging from 30 cM to 35 cM. These gaps demonstrated that some parts of the genome were partially covered. The final linkage map obtained was smaller than the maps constructed for R570 (Hoarau et al., 2001), Q165 (Aitken et al., 2005; 2014) and SP80-180 x SP80-4966 (Oliveira et al., 2007), which were 5849 cM, 9058.3 cM, 9774.4 cM and 6261.1 cM in length, respectively.

The 113 LGs were used to establish the putative homology groups (HG), which were based on at least one pair of common SSR derived markers due to their co-dominance inheritance. In total, 82 LGs were regrouped into 15 putative HGs. Each HG contained between 2 and 16 LGs (Supplementary Material). The markers were not equally distributed within the different HGs, in agreement with Hoarau et al. (2001), Aitken et al. (2005; 2014), Raboin et al. (2006) and Oliveira et al. (2007). The largest HG grouped 54 markers distributed along 11 LGs. The smallest HG grouped six markers distributed between 2 LGs. The remaining 31 LGs did not contain a pair of common SSR derived markers that allowed for assignment into different HGs. In some cases, the marker order was maintained between LGs. For example, in HG I, the markers scb060 and SMC1047HA on LG 4 were found in the same order on LG 8. In HG III, the marker order cv038 and cir012 on LG 29 was preserved in LG 30 and LG 39. In HG X, the marker order scb007 and scb043 on LG 69 was preserved on LG 71. Gardiner et al. (1993), Aitken et al. (2005; 2014) and Oliveira et al. (2007) also reported that marker order was preserved in some LGs that belonged to the same HG.

### *Phenotypic analysis*

Different structures for the VCOV matrix were compared in the mixed model analysis, the structure with the smaller AIC and BIC values was considered the most appropriate (Table 1). For the traits SD, SW and Fiber, the smallest AIC and BIC values resulted in the selection of different structures for the  $\mathbf{G}_M$  matrix. In this case, the VCOV structure with the smallest number of parameters was selected. For SD, SW and Fiber, the best model for the  $\mathbf{G}_M$  matrix was the Unstructured-Unstructured model based on the direct product of  $\mathbf{G}_{j \times j}^L$  and  $\mathbf{G}_{m \times m}^H$  matrices for the locations and harvests, respectively. For Brix and Pol, the best model was the Unstructured model, this model considered the factorial combination of locations and harvests as a specific environment. The selected model for SD, SW and Fiber considers that each location and each harvest have particular genetic variances and covariances. Likewise, the selected model for Brix and Pol allows for a specific genetic variance and covariance for each environment. For all traits, the adjusted  $\mathbf{R}_M$  matrix (non-genetic effects) was the Identity that takes into account a lack of genetic correlation between the location and the harvest. Other structures were not fitted due to the non-convergence of parameters in the prediction procedure. Thus, BLUPs were obtained for each trait, which allowed for the estimation of genetic parameters (Table 2).

Broad-sense heritability coefficients ranged from moderate to high. The coefficients were as follows: 76.78 for SD, 69.24 for SW, 59.54 for Fiber, 43.74 for Brix and 45.89 for Pol. Heritability estimates were consistent with the coefficient of variation (CV), confirming that the field experiments had good controls. The SW trait had the highest CV value, which was expected, as SW is highly influenced by environmental conditions. For the other traits, the CV values were low. The CV values ranged from 5.40 for Brix to 9.70 for SD. The genetic and phenotypic variability coefficients were low for all traits reported.

Pairwise correlation coefficients were significant for SD-SW, SD-Fiber, SD-Pol, SW-Fiber, SW-Pol, Fiber-Brix and Brix-Pol. The greatest significant phenotypic correlation was

measured for Brix and Pol (0.91). Intermediate significant phenotypic correlations were reported for SD-SW (0.69), SD-Fiber (-0.39), SW-Fiber (-0.25), SW-Pol (0.23), Fiber-Brix (0.22) and SW-Pol (0.23). The lowest significant correlation was observed for SD-Pol (0.17).

Table 1 – Different mixed models for the VCOV  $\mathbf{G}_M$  matrix and the AIC and BIC values (bold values are significant).

Trait	$\mathbf{G}_M$ matrix	$\mathbf{G}_M$ matrix <sup>(1)</sup>	AIC criteria	BIC criteria	<i>n</i> Parameters
SD	$\mathbf{G}_M = \mathbf{G}_{m \times m}^{L-H}$	(a) Id	12480.12	12491.77	1
		(b) Diag	12461.48	12490.59	4
		(c) FA1	12260.74	12313.15	8
		(d) Uns	<b>12185.27</b>	12249.32	10
		(e) Uns x Uns	12194.31	<b>12229.25</b>	5
SW	$\mathbf{G}_M = \mathbf{G}_{j \times j}^L \otimes \mathbf{G}_{m \times m}^H$	(a) Id	11434.12	11445.73	1
		(b) Diag	11385.78	11414.80	4
		(c) FA1	11230.84	11283.08	8
		(d) Uns	<b>11212.33</b>	11276.19	10
		(e) Uns x Uns	11222.14	<b>11256.97</b>	5
Fiber	$\mathbf{G}_M = \mathbf{G}_{m \times m}^{L-H}$	(a) Id	4250.95	4261.46	1
		(b) Diag	4227.33	4253.62	4
		(c) FA1	4126.75	4174.06	8
		(d) Uns	<b>4106.58</b>	4164.40	10
		(e) Uns x Uns	4126.86	<b>4158.40</b>	5
Brix	$\mathbf{G}_M = \mathbf{G}_{j \times j}^L \otimes \mathbf{G}_{m \times m}^H$	(a) Id	4606.56	4617.07	1
		(b) Diag	4586.77	4613.05	4
		(c) FA1	4534.15	4581.46	8
		(d) Uns	<b>4520.26</b>	<b>4578.09</b>	<b>10</b>
		(e) Uns x Uns	-	-	5
Pol	$\mathbf{G}_M = \mathbf{G}_{m \times m}^{L-H}$	(a) Id	4578.34	4588.86	1
		(b) Diag	4551.41	4577.70	4
		(c) FA1	4506.14	4553.45	8
		(d) Uns	<b>4487.21</b>	<b>4545.03</b>	<b>10</b>
		(e) Uns x Uns	-	-	5

(a) Id: Identity model; (b) Diag: Diagonal model; (c) FA1: First-order factor analytic model; (d) Uns: Unstructured model; (e) Uns x Uns: Unstructured model for locations and harvests.

Table 2 - Phenotypic correlations between traits, and parameter estimations considering BLUPs for the following traits: broad-sense heritability coefficient ( $h^2\%$ ); genotypic variance ( $\sigma_G^2$ ); phenotypic variance ( $\sigma_P^2$ ); coefficient of variation ( $CV\%$ ) and mean ( $\mu$ ).

Trait	SD	SW	Fiber	Brix	Pol	Parameter	SD	SW	Fiber	Brix	Pol
SD	1.000	0.69*	-0.39*	0.03	0.17*	$h^2\%$	76.78	69.24	59.54	43.74	45.89
SW		1.000	-0.25*	0.12	0.23*	$\sigma_G^2$	2.294	1.160	0.221	0.179	0.164
Fiber			1.000	0.22*	-0.07	$\sigma_P^2$	9.451	6.482	1.172	1.494	1.452
Brix				1.000	0.91*	$CV\%$	9.703	22.623	7.286	5.407	7.428
Pol					1.000	$\mu$	25.91	9.65	12.92	19.99	14.91

\*Significance level at 5% ( $P = 0.05$ ).

## *QTL mapping*

In total, 420 SDMs (1:1 and 3:1) were considered in the QTL mapping procedure. Based on the information for these markers, 19 QTLs were detected for SD, SW, Fiber, Pol and Brix. These QTLs were based on the CIM approach and the integrated genetic map constructed in this study (Table 3, Figure 1). For all of the traits evaluated in plant and ratoon canes, 1,000 permutation tests were performed, which resulted in LOD Score threshold values of 4.03 for SD, 4.15 for SW, 4.77 for Fiber, 5.8 for Pol, 6.79 for Brix. QTLs were identified along 14 LGs and seven distinct HGs. For all traits, at least one additive or dominance effect was significant. Regardless, the majority of QTLs had significant additive effects in both parents. Overall, the proportion of phenotypic variance ( $R^2$ ) explained by the QTLs was low, which is expected in studies involving sugarcane. The  $R^2$  ranged from 0.10% to 6.67%, and the QTLs segregated according to ratios of 1:1:1:1, 1:2:1, 3:1 and 1:1.

QTL mapping for SD identified four QTLs (Table 3, Figure 1) in different LGs. The LGs were as follows: 48 (3.46 cM), 68 (17 cM), 71(2 cM) and 87 (15.11 cM), which explains 18.34% of the phenotypic variance ( $R^2$ ). The QTL located in LG 87 had the highest observed peak for SD, with an LOD Score of 8.517, which explains 4.14% of the phenotypic variance and the QTL segregates in a 1:1:1:1 fashion. QTLs in LGs 68 and 71 had a 3:1 segregation pattern. The QTL in LG 48 had a 1:1 segregation ratio. The LGs 68, 71 and 48 contributed 5.97%, 2.88% and 6.67% of the phenotypic variance, respectively, for this trait. For SD, the parent IACSP93-3046 made the largest contribution to trait variability. The majority of the QTLs had negative effects, contributing to a decrease in the SD phenotype. From the four mapped SD QTLs, two had significant additive effects for the parent IACSP93-3046 (LG 48 and LG 87), and two had significant additive effects for both parents (LG 68 and LG 71). The LG 87, besides presenting with a significant additive effect for the parent IACSP93-3046, also demonstrated a significant dominance effect. QTLs mapped in LG 68 and LG 71 were located in HG 10. QTLs mapped in LG 48 and LG 87 were not assigned to any homology group.

Three QTLs were detected for SW. The QTLs were in LGs 35 (7 cM), 72 (47 cM) and 87 (15.11 cM) and explained 1.11%, 1.78% and 2.85% of the phenotypic variance, respectively. The QTLs mapped for SW had predominant significant negative effects. The QTL detected in LG 35 (HG3) presented the highest LOD Score (6.862), a 1:2:1 segregation pattern, and negative effects for SW in both parents. In LG 72 (HG11), a QTL was identified with an intermediate LOD Score (6.284). This QTL had a significant positive additive effect for parent IACSP95-3018 and a significant positive dominance effect, segregating in a 1:1:1:1 fashion. The smallest LOD Score (4.738) was related to the QTL detected in LG 87, in the same position of QTL SD.1. This QTL showed only a significant dominance effect, with a 1:1 segregation pattern.

Five putative QTLs were identified for fiber content in LGs 15 (3 cM), 31 (33.35 cM), 59 (14 cM), 72 (85.1 cM) and 110 (9.51 cM). These QTLs account for 6.26% of the phenotypic variance (Table 3). The most significant QTL was mapped into LG 15 (HG2), with a LOD Score of 8.518. The proportion of the phenotypic variance explained by each QTL ranged from 0.10 (LG 110) to 2.23 (LG 31). The QTLs located in LGs 31 and 72 had a 1:2:1 segregation pattern. The QTLs in LGs 15 and 59 segregated in a 1:1 fashion, and the QTL in LG 110 had a 3:1 segregation pattern. The majority of the QTLs identified for fiber content had significant negative effects. The QTL located in LG 15 had a significant additive effect for IACSP95-3018. The QTLs located in LG 31 and LG 72 had significant additive effects for IACSP93-3046. The QTLs in LG 59 and LG 110 had significant additive effects for both parents. The QTLs detected in LGs

31 and 72 also had significant dominance effects. The QTLs mapped in LG 31, LG 59, LG 72 were placed in HG 3, HG 7 and HG 11, respectively. The LG 72 also presented with a QTL for SW in a nearby region.

Table 3 - QTL effects estimated via CIM (SD.; SW.; F.; B. and P. are related to mapped QTLs for SD, SW, Fiber, Pol and Brix, respectively.  $\alpha_p$ ,  $\alpha_q$  are the additive effects for the parents  $P$  (IACSP95-3018) and  $Q$  (IACSP93-3046), and  $\delta_{pq}$  is the dominance effect.  $R^2$  is the proportion of phenotypic variance that is explained by each QTL).

QTL	LG <sup>(1)</sup>	Position (cM) <sup>(1)</sup>	LOD <sup>(2)</sup>	$\alpha_p$ <sup>(3)</sup>	$\alpha_q$ <sup>(3)</sup>	$\delta_{pq}$ <sup>(3)</sup>	Segregation	$R^2$ (%)
SD.1	87	15.11	8.517	0.159	<b>-0.255</b>	<b>-0.58</b>	1:1:1:1	4.14
SD.2	48	3.46	6.202	-	<b>-0.408</b>	-	1:1	6.67
SD.3	71	2	5.079	<b>-0.167</b>	<b>0.555</b>	-	3:1	2.88
SD.4	68	17	3.192	<b>0.939</b>	<b>-0.565</b>	-	3:1	5.97
								18.34
SW.1	35	7	6.862	<b>-0.265</b>	<b>-0.3</b>	-	1:2:1	1.11
SW.2	72	47	6.284	<b>0.315</b>	0.029	<b>0.547</b>	1:1:1:1	1.78
SW.3	87	15.11	4.738	0.116	0.06	<b>-0.348</b>	1:1	2.85
								6.36
F.1	15	3	8.518	<b>-0.154</b>	-	-	1:1	0.56
F.2	110	9.51	7.305	<b>-0.064</b>	<b>-0.182</b>	-	3:1	0.1
F.3	59	14	6.845	<b>0.07</b>	<b>-0.158</b>	-	1:1	1.23
F.4	72	85.1	6.046	-0.028	<b>0.112</b>	<b>-0.115</b>	1:2:1	1.74
F.5	31	33.35	5.75	0.051	<b>-0.092</b>	<b>0.127</b>	1:2:1	2.23
								6.26
P.1	51	12	8.715	<b>-0.068</b>	<b>0.108</b>	<b>0.057</b>	3:1	1.06
P.2	26	8.04	7.113	<b>0.118</b>	-0.032	-	1:1	0.25
P.3	84	22	6.751	<b>-0.113</b>	0.006	0.008	1:1	1.45
P.4	72	57	6.381	<b>0.054</b>	<b>0.056</b>	<b>0.117</b>	1:1:1:1	2.77
								5.17
B.1	72	85.1	12.259	<b>0.052</b>	<b>0.093</b>	<b>0.149</b>	1:1:1:1	2.26
B.2	51	21	8.945	<b>-0.073</b>	<b>0.12</b>	0.039	1:2:1	0.91
B.3	54	14.51	7.14	<b>0.044</b>	-0.02	<b>-0.118</b>	1:1:1:1	0.53
								3.39

<sup>(1)</sup> LG: Linkage Group;

<sup>(2)</sup> LOD Score obtained for each trait separately (4.03 for SD, 4.15 for SW, 4.77 for Fiber, 5.8 for Pol, and 6.79 for Brix);

<sup>(3)</sup> Bold values correspond to significant marginal effects.

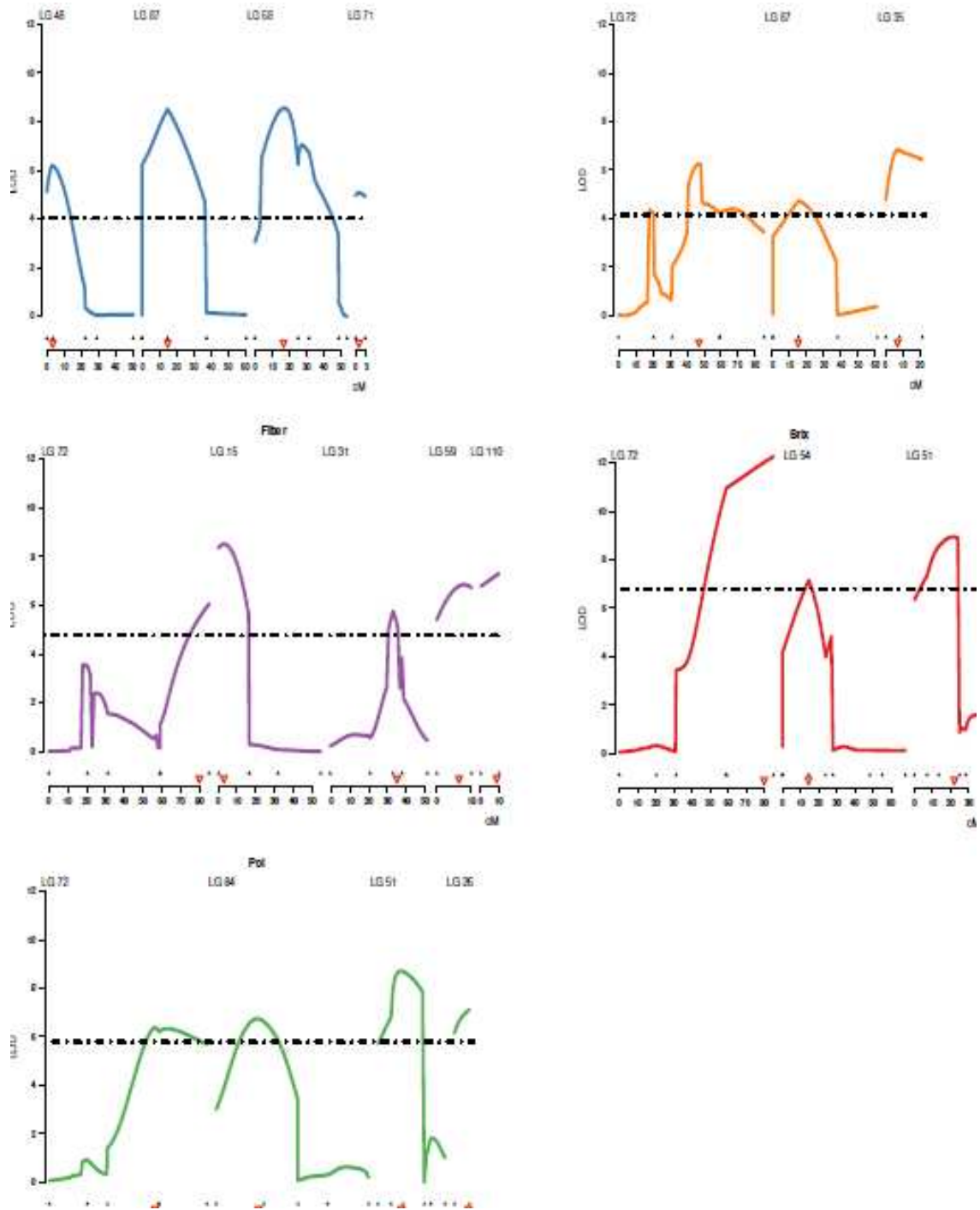


Figure 1. QTLs estimated by CIM associated with SD, SW, Fiber, Pol and Brix in a bi-parental cross between the elite clone IACSP95-3018 and the variety IACSP93-3046.

For Pol, four QTLs (Table 3) were identified in LGs 26 (8.04 cM), 51 (12 cM), 72 (57 cM) and 84 (22 cM). The proportion of the phenotypic variance explained for all QTLs was 5.17%, ranging from 0.25% (LG 26) to 2.77% (LG 72) for each QTL. The highest LOD Score value (8.715) was related to the QTL detected in LG 51, and this QTL accounts for 1.06% of the phenotypic variance. This QTL segregated in a 3:1 fashion and was placed in HG 5. The smallest LOD Score (6.381) was noted for the QTL detected in LG 72 (HG11), this QTL had a 1:1:1:1 segregation pattern. The QTLs mapped in LGs 26 (HG2) and 84 (unknown HG) segregated in a 1:1 fashion. The parents contributed to the trait variance with significant positive effects. The QTLs detected in LG 51 and LG 72 showed significant additive effects for both parents, along with significant dominance effects. The QTLs in LG 26 and LG 84 had significant additive effects for IACSP95-3018.

Three QTLs were identified for Brix. The QTLs were in LGs 51 (21 cM), 54 (14.51 cM) and 72 (85.1 cM) and were placed in HG 5, HG 6 and HG 11, respectively. The proportions of phenotypic variance explained for the QTLs were 0.53% (LG 54), 0.91% (LG 51) and 2.26% (LG 72). The QTL in LG 72 showed the highest LOD Score value (12.259) among all of the mapped QTLs for all of the evaluated traits. This QTL also showed significant positive additive effects for both parents and a dominance effect. The QTL in LG 54 had the smallest LOD Score value (7.14) and had a significant additive effect for parent IACSP95-3018 and a significant dominance effect. The last QTL detected for Brix was located in LG 51 (LOD 8.945), with significant additive effects for both parents. The QTLs in LG 51 and LG 54 had 1:1:1:1 segregation ratios. The QTL in LG 72 segregated in a 1:2:1 fashion. The QTL placed in LG 72 was mapped to the same region as the QTL detected for Fiber, at a distance of almost 30 cM to the QTL mapped for Pol.

## Discussion

QTL mapping is a useful tool to evaluate genotypes of putative importance for breeding programs. The majority of quantitative traits in sugarcane are of complex inheritance. However, the successful implementation of QTL mapping results depends on the construction of reliable genetic linkage maps. Our genetic linkage map contains 113 LGs and is in agreement with the chromosome number expected in modern sugarcane varieties derived from *S. officinarum* ( $2n=80$ ) and *S. spontaneum* ( $2n=40-128$ ). Such modern varieties have an expected genome composition made up of approximately 70 - 80% *S. officinarum*, 10 - 20% *S. spontaneum* and 5 - 17% recombinant chromosomes. However, 214 markers remained unlinked. Combined with the reduced size of the majority of LGs, with an average length of 39 cM, these results indicate that the genetic linkage map is not well saturated. This finding can most likely be attributed to the low level of polymorphism found in some regions of the genome of *S. officinarum*, which corresponds to a large portion of the modern sugarcane genome as a consequence of the nobilization process (Alwala and Kimberg, 2010). Grivet et al. (1996), Ming et al. (1998), Hoarau et al. (2001), Aitken et al. (2005), Oliveira et al. (2007) and Pastina et al. (2012) also identified small LGs with few linked markers. Moreover, saturated linkage maps could be obtained in sugarcane if molecular markers with higher dosages were considered. These multiplex markers, which segregate in 11:3, 13:1, 69:1, 7:1, or 15:1 ratios, are widely distributed along the genome. Such markers would likely increase the coverage of the linkage map.

When only single-dose polymorphisms are considered, gaps are commonly expected. In our linkage map, gaps ranging from 30 cM to 35 cM were observed. These values are smaller



than 40 cM, as reported by Hoarau et al. (2001) and we observed an increase in the number of unlinked markers and a decrease in the number of linked markers per LG. Another possible explanation for the large number of unlinked markers is the fact that progeny derived from a cross between two commercial varieties were used. This type of progeny has complex meiotic behavior, including aneuploidy. In addition, no pairing of chromosomes occurs, which may result in a large proportion of unlinked markers in the genetic map (Garcia et al., 2006). The alleles derived from the same SSR or EST-SSR locus were linked into 113 LGs. In total, 15 HGs were identified, which was significantly greater than the expected number of chromosomes for the genus *Saccharum*, which ranges between  $x = 8$  and  $x = 10$  (D'Hont et al., 1998; Irvine, 1999; Grivet and Arruda, 2001). Four HGs were formed by only two LGs. The small size of LGs may represent chromosome fragments, hindering the correct grouping of HGs. These findings confirm that the map is not saturated and reinforces the need to use multiple dosage markers.

Because the association between genotype and phenotype in sugarcane is based on trial data studies performed at different locations and harvests, it is important to take into account appropriate assumptions of variance and covariance matrices for genotype and residual effects (Smith et al., 2007). For all of the traits in this study, the fitted VCOV structure for the genetic effects matrix was the Unstructured model. This model shows that each interaction, namely the genotype-environment interaction (models a-d) (see Table 1 in results) or the genotype-location-harvest interaction (model e) (see Table 1 in results), is inherent for each combination of the location and harvest. Pastina et al. (2012) obtained similar results in their phenotypic analysis for different genotypes and yield-related traits in sugarcane. By applying mixed models to trial data performed in different environments, it is possible to detect the heterogeneity of genetic variances and the correlation between environments (Malosetti et al., 2013), allowing for a more realistic understanding of the genotype effects. It is important to consider the plausible correlations in the residuals by fitting non-genetic effects via a (co)variance matrix. In this study, only one structure for the  $\mathbf{R}_M$  matrix was fitted: the Identity model, which considers uncorrelated residuals. We believe that the lack of fit is due to the use of only two locations and two harvests. When large numbers of combinations between the locations and harvests are used, the (co)variances of residuals can be better captured. Analyses that account for the predicted means of different environments obtained via BLUPs improve the detection of significant marker-phenotype associations along different environments.

The phenotypic analysis demonstrates through the CV that the field trials had good precision and that the environment was controlled with accuracy. CV values ranged from 5.407 (Brix) to 22.623 (SW). The highest CV among all traits was for SW, a yield related trait that is strongly influenced by the semi-perennial nature of sugarcane, where stalk elongation and stalk formation occur simultaneously. The heritability coefficients ranged from high values (SD, SW and Fiber) to intermediate values (Brix and POL). These results indicate that SD, SW and Fiber had larger proportions of genetic variation compared to Brix and Pol and explain the differences between the genotypes under study. The pairwise phenotypic correlations performed in this study suggest that certain traits, such as Brix-Pol and SW-SD, are highly correlated. Brix and Pol measures are obtained from the proportion of sugar in the sugarcane stalks, therefore, high values are expected for the significance correlations for these traits. We also observed negative significant correlations for Fiber-SD and Fiber-SW. Using the same bi-parental cross described in this study, Mancini et al. (2012) reported similar phenotypic correlations among the above traits, confirmed a real correlation among these traits for this population.

CIM offers several advantages when compared to single-marker and interval mapping approaches (Zeng, 1993; 1994; Jansen and Stam, 1994). However, few studies use CIM in QTL mapping for yield-related traits in sugarcane (Aitken et al., 2008; Gazaffi, 2009; Shing et al., 2013). Considering that the CIM methodology was performed along with an integrated genetic linkage map and that information about the QTL genotypes was obtained via conditional probabilities, the QTL mapping results are more informative. The mapping model presented here, proposed by Gazaffi et al. (2014), has proven to be an excellent approach to QTL mapping in an outcrossing species. This model considers a full-sib progeny obtained from two non-inbred parents. The QTL mapping results provide estimates of the QTL effects, including information about the additive effects for both parents and the dominance effects and segregation patterns for the mapped QTLs. Using this approach, it is possible to obtain the segregation pattern of mapped QTLs, including those QTLs located at less informative regions. Furthermore, these results are very useful in marker-assisted selection. It is possible to identify alleles influencing increasing or decreasing phenotypic trait values. Such information has not been reported in previous studies involving outcrossing species.

The number of detected QTLs was low. Considering the chosen threshold for distinct traits, 19 QTLs were identified. These QTLs included four for SD, three for SW, five for Fiber, four for Pol and three for Brix. Each QTL was located at a different LG, and each trait was considered separately. Intermediate values of estimated heritability coefficients for Brix (43.74) and Pol (45.89) may cause a decrease in QTL detection power. Only three QTLs were detected for SW. This low number of QTLs can be explained by the high CV value (22.623), as SW is strongly influenced by environmental conditions. A better control of environmental variation could increase the QTL detection power for SW. A comparison between the sugarcane QTL mapping results is difficult due to the choice of different parents during the crossing process, different experimental designs, and performance in different environments. However, several studies report the detection of a low number of QTLs associated with cane yield and sugar yield traits (Sills et al., 1995; Ming et al., 2002a; 2002b; Jordan et al., 2004; da Silva and Bressiani, 2005; Aitken et al., 2006; Piperidis et al., 2008). By including additional markers as cofactors in a mapping model, both the variation associated with QTLs located outside the mapping interval and the detection of false-positive QTLs are reduced. It is worth noting that all of the mapped QTLs had good performance considering the average of the environments (location  $\times$  harvest). We used BLUPs in the QTL mapping procedure, which prevents us from asserting QTL stability across different combinations of environments or locations and harvests. Such claims can be based on mapping models that take into account the QTL-location-harvest interaction, as suggested by Pastina et al. (2012).

In the QTL mapping procedure, the genetic predictors were estimated for all of the positions along the linkage map. Because we used only SDMs (in 1:1 or 3:1 segregation patterns), there is a lack of genotypic information to classify QTL genotypic classes, and this limitation that has been previously mentioned by Pastina et al. (2012) and Gazaffi et al. (2014). Some genetic predictors could be obtained as linear combinations of others. In this sense, the singular matrix of genetic predictors could create problems in the estimation of parameters due to the collinearity of the matrix of genetic predictors. For instance, type D<sub>1</sub> markers provide information concerning the additive effects for the IACSP95-3018 parent, while type D<sub>2</sub> markers provide information for IACSP93-3046. Only informative contrasts were included in the QTL mapping analysis. These occurrences clarify the non-estimate of certain QTL effects on SD.2, SD.3, SD.4, SW.1, F.1, F.2, F.3 and P.2 (Table 3).

All of the mapped QTLs demonstrated significant additive and/or dominance effects, with a predominance of additive effects, indicating the presence of alleles that allow for polymorphism in the investigated traits. In particular, these effects were detected in IACSP93-3046, since this parent meets the current requirements of breeding programs. Significant dominance effects were detected on QTLs SD.1, SW.2, SW.3, F.4, F.5, P.1, P.4, B.1 and B.3. These QTLs assist in our understanding of the genetic complexity of sugarcane. The phenotypic variation explained by each QTL remained low and in accordance with the expected results for sugarcane. This variation ranged from 0.10 (F.2) to 6.67 (SD.2). Due to the high level of polyploidy in sugarcane, low effects for the associations between markers and QTLs are expected. Hoarau et al. (2002) reported  $R^2$  values between 3% and 7%. Aitken et al. (2006) reported variation between 2% to 8%, and Aitken et al. (2008) reported variation between 2% a 10%. Ming et al. (2002) reported on phenotypic variation with a greater influence on the expression of the evaluated traits, which ranged from 3.8% to 16.2%. It is likely that the inclusion of more molecular markers, in addition to the use of mapping models that consider multiple traits and multiple environments, may increase the proportion of phenotypic variation explained by each QTL.

Information concerning the segregation pattern was estimated only at positions of mapped QTLs, and then no multiple correction tests were required. Those QTLs that demonstrated 1:1 segregation patterns had only one significant effect ( $\alpha_p$ ,  $\alpha_q$  or  $\delta_{pq}$ ). QTLs with 1:2:1 or 1:1:1:1 segregation patterns had two significant effects. QTLs with 3:1 segregation patterns, besides meeting the assumptions of the hypothesis tests for effects, as described by Gazaffi et al. (2014), had all three significant effects. This information helps us understand the behavior of QTL alleles in progeny.

QTL mapping for correlated traits at nearby or common regions within the same LG is important for future investigations involving linkage and pleiotropy. SD and SW have significant positive correlations, and had two QTLs (SD.1 and SW.3) mapped to the same position at LG 87. However, this LG remained unlinked, with no HG. We believe that these QTLs may be pleiotropic. Brix and Pol showed high positive phenotypic correlations. Two QTLs, P.1 and B.2, were located in the same LG (LG 51). The mapping profiles for these QTLs were similar and these QTLs may be pleiotropic. In LG 72, the QTLs SW.2, F.4, P.4 and B.1 were detected in nearby regions. SW and Pol and Fiber and Pol were positively correlated, providing evidence of pleiotropic QTLs. SW and Fiber had a negative phenotypic correlation. QTLs F.4 and QTL P.4 were detected on LG 72 and deserve special attention in marker-assisted selection. When the positive selection for one of these traits is performed, the other trait will be negatively selected.

Genome dissection by QTL mapping in polyploid species such as sugarcane is hindered by the small number of markers, the exclusive use of SDMs and the large-scale occurrence of unlinked markers, resulting in maps with poor saturation. The absence of inbred lines also reduces mapping accuracy. In contrast with SDM, we use mixed models applied to phenotypic data, which allowed us to model VCOV structures for genetic effects to predict genotype values and avoid the unbalanced data. QTL genotypes were estimated via multipoint conditional probabilities. The advantage of this approach is an increase in the statistical power of QTL mapping. We also performed a QTL mapping procedure in an integrated genetic linkage map. Was considered a CIM approach with the inclusion of additive and dominance effects, enabling us to estimate the segregation patterns and linkage phases for all 19 of the mapped QTLs.

A recent study claims that the ploidy level of sugarcane ranges between six and 14 (a mixed-ploidy). Only 30% of polymorphisms obtained by molecular markers are classified as single-dose (Garcia et al. 2013). When molecular markers such as AFLP or SSR are used, it is

impossible to determine the ploidy level of all polymorphisms due to the dominant nature of these system markers. Using only the SDM to the linkage map construction and QTL mapping in sugarcane is uncovered and underestimated. The linkage map construction and QTL mapping methodology for outcrossing species were developed based on diploid approaches and do not consider information relative to ploidy level and dosage. When information about ploidy and multiple dosages are included, the construction of genetic linkage maps in complex species can be more realistic. With information concerning multiple dosages is possible to obtain information besides obtaining a more saturated genetic linkage map, so the location of QTLs can be more precise, QTL effects and segregation patterns and interactions can also be estimated with more precision. The development of new approaches that include ploidy and dosage information are necessary to better understand the genetic architecture of complex polyploid genomes, such as sugarcane. This information can then be used for assisted selection studies.

## References

Aitken KS, Jackson PA, McIntyre CL (2005) A combination of AFLP and SSR markers provides extensive map coverage and identification of homo(eo)logous linkage groups in a sugarcane cultivar. *Theoretical and Applied Genetics* 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

Akaike H (1974) A new look at the statistical model identification. *IEEE Trans Automat Contr AC* 19:716–723

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

Aitken KS, McNeil MD, Hermann S, Bundock PC, Kilian A, Heller-Uszynska K, Henry RJ, Li J (2014) A comprehensive genetic map of sugarcane that provides enhanced map coverage and integrates high-throughput Diversity Array Technology (DArT) markers. *BMC Genomic* 15:152

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

Consecana – Conselho nacional dos produtores de cana-de-açúcar, açúcar e álcool do estado de São Paulo (2006) Manual de instruções – CONSECANA-SP. Editora CONSECANA, Piracicaba, p 112

Cordeiro GM, Taylor GO, Henry RJ (2000) Characterization of microsatellite markers from sugarcane (*Saccharum* sp.) a highly polyploid species. *Plant Science* 155 161-168

Creste S, Tulmann Neto A, Figueira A (2001) Detection of single sequence repeats polymorphisms in denaturing polyacrylamide sequencing gel by silver staining. *Plant Molecular Biology Reporter* 19:299-306

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

Daugrois JH, Grivet L, Roques D, Hoarau JY, Lombardi 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* 92: 1059-1064

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

Falconer DS, Mackay TF (1996) *Introduction to quantitative genetics*. 4 ed. Londres: Editora Longman Group, p 464

Garcia AAF, Kido EA, Meza AN, Silva JAGD, Souza AP, Pinto LR, Pastina MM, Leite CS, da Silva JAG, Ulian EC, Figueira A, Souza HMB (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 AAF, Wang S, Melchinger AE, Zeng ZB (2008) Quantitative trait loci mapping and genetic basis of heterosis in maize and rice. *Genetics* 180: 1707-1724

Garcia AAF, Mollinari M, Marconi TG, Serang OR, Silva RR, Vieira MLC, Vicentini R, Costa EA, Mancini MC, Garcia MOS, Pastina MM, Gazaffi R, Martins ERF, Dahmer N, Sforça DA, Silva CBC, Bundock P, Henry R, Souza GM, van Sluys MA, Landell MGA, Carneiro MS, Vincentz MAG, Pinto LR, Vencovsky R, Souza AP (2013) SNP genotyping allows an in-depth characterization of the genome of sugarcane and other complex autopolyploids. *Scientific Reports* 3:3399

Grattapaglia D, Sederoff R (1994) Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross: mapping strategy and RAPD markers. *Genetics* 137:1121–1137

Grivet L, D’Hont A, Roques D, Feldmann P, Lanaud C, Glaszmann JC (1996) RFLP mapping in cultivated sugarcane (*Saccharum* spp.): Genome organization in a highly polyploid and aneuploid interspecific hybrid. *Genetics* 142:987-1000

Grivet L, Arruda P (2001) Sugarcane genomics: depicting the complex genome of an important tropical crop. *Curr Opin Plant Biol* 5:122–127

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

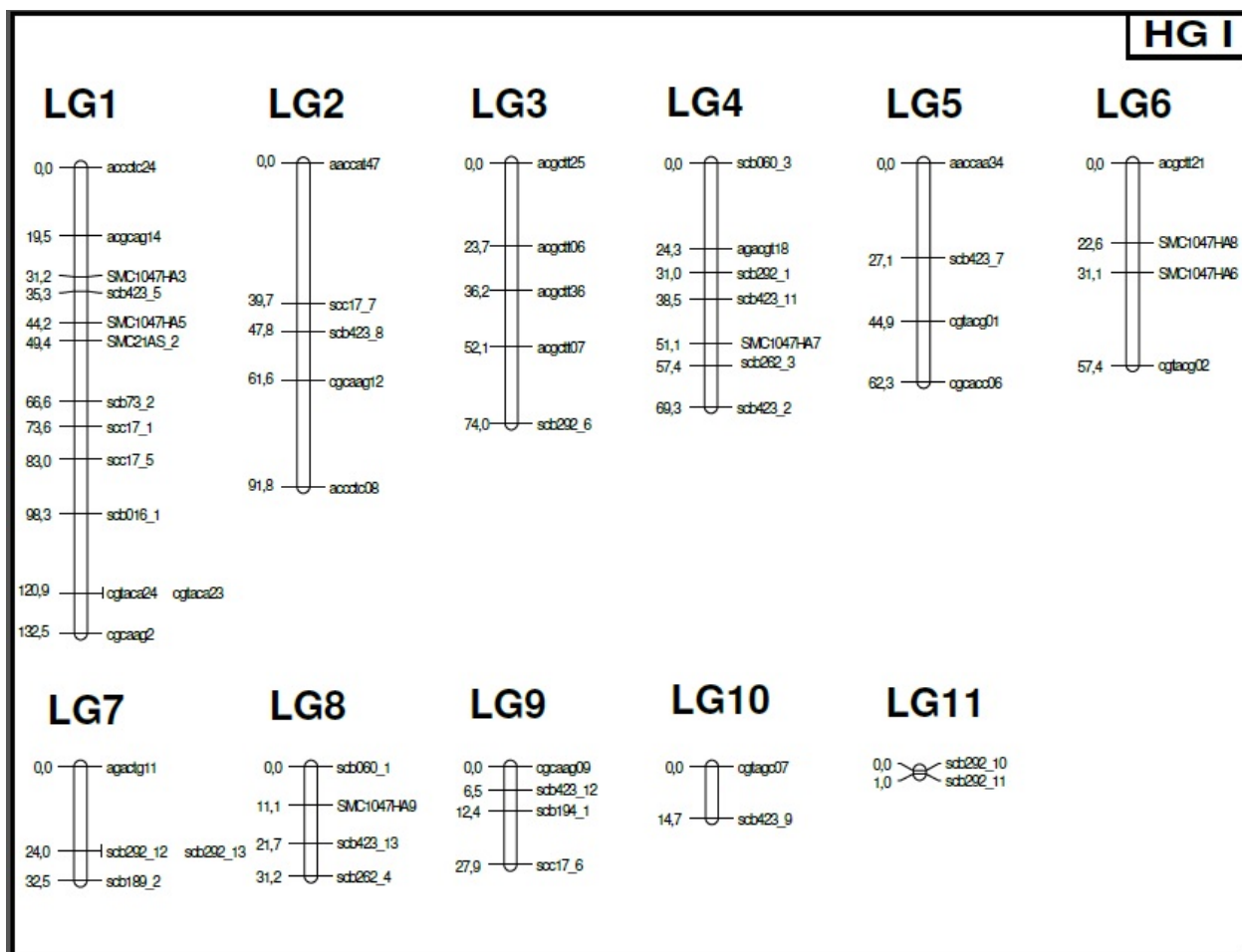
- Hoarau JY, Offmann B, D'Hont A, Risterucci AM, Roques D, Glaszmann JC, Grivet L (2001) Genetic dissection of a modern sugarcane cultivar (*Saccharum* spp.). I. Genome mapping with AFLP markers. *Theor Appl Genet* 103:84-97
- Hoarau JY, Grivet L, Offmann B, Raboin LM, Diorflar JP, Payet J, Hellmann M, D'Hont A, Glaszmann JC (2002) Genetic dissection of a modern sugarcane cultivar (*Saccharum* spp.). II. Detection of QTLs for yield components. *Theor Appl Genet* 105:1027–1037
- Irvine JE (1999) *Saccharum* species as horticultural classes. *Theor Appl Genet* 98:186–194
- Jansen RC, Stam P (1994) High resolution of quantitative traits into multiple loci via interval mapping. *Genetics* 136:1447-1455
- Jordan DR, Casu RE, Besse P, Carroll BC, Berding N, McIntyre CL (2004) Markers associated with stalk number and suckering in sugarcane collocate with tillering and rhizomatousness QTLs in sorghum. *Genome* 47:988–993
- Kosambi DD (1944) The estimation of map distances from recombination values. *Annu Eugene* 12:172–175
- 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
- Lynch M, Walsh B (1998) *Genetics and analysis of quantitative traits*. Sinauer Associates, Sunderland
- Mancini MC, Leite DC, Perecin D, Bidóia MAP, Xavier MA, Landell, Pinto LR MGA (2012) Characterization of the Genetic Variability of a Sugarcane Commercial Cross Through Yield Components and Quality Parameters. *Sugar Tech* 14:119–125
- Marconi TG, Costa EA, Miranda HRCAN, Mancini MC, Cardoso-Silva CB, Oliveira KM, Pinto LR, Molinari M, Garcia AAF, Souza AP (2011) Functional markers for gene mapping and genetic diversity studies in sugarcane. *BMC Research Notes* 4:264
- Margarido GRA, Souza AP, Garcia AAF (2007) OneMap: software for genetic mapping in outcrossing species. *Hereditas* 144:78-79
- McIntyre CL, Whan VA, Croft B, Magarey R, Smith GR (2005) Identification and validation of molecular markers associated with *Pachymetra* root rot and brown rust resistance in sugarcane using map- and association-based approaches. *Mol Breed* 16:151–161
- Ming R, Liu SC, Lin YR, da Silva J, Wilson W, Braga D, van Deynze A, Wenslaff TF, Wu KK, Moore PH, Burnquist W, Sorrells ME, Irvine JE, Paterson AH (1998) Detailed alignment of *saccharum* and *sorghum* chromosomes: comparative organization of closely related diploid and polyploid genomes. *Genetics* 150:1663–1682

- 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, Del Monte TA, Hernandez E, Moore PH, Irvine JE, Paterson AH (2002) Comparative analysis of QTLs affecting plant height and flowering among closely-related diploid and polyploid genomes. *Genome* 45:794–803
- Murkherjee SK (1957) Origin and distribution of *Saccharum*. *Botanic Gas* 119:55-61
- 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
- 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, van Eeuwijk FA, Garcia AAF (2012) A mixed model QTL analysis for sugarcane multiple-harvest-location trial data. *Theor Appl Genet* 124:835–849
- 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
- Piperidis N, Jackson PA, D’Hont A, Besse P, Hoarau JY, Courtois B, Aitken KS, McIntyre CL (2008) Comparative genetics in sugarcane enables structured map enhancement and validation of marker-trait associations. *Mol Breed* 21: 233-247
- Province MA (1999) Sequential methods of analysis for genome scan. In: Rao DC, Province MA (eds) *Dissection of complex traits*. Academic Press, San Diego p 583
- Raboin LM, Oliveira KM, Raboin LM, Lecunff L, Telismart H, Roques D, Butterfield M, Hoarau JY, D’Hont A (2006) Genetic mapping in sugarcane, a high polyploid, using bi-parental progeny: identification of a gene controlling stalk colour and a new rust resistance gene. *Theor Appl Genet* 112:1382-1391

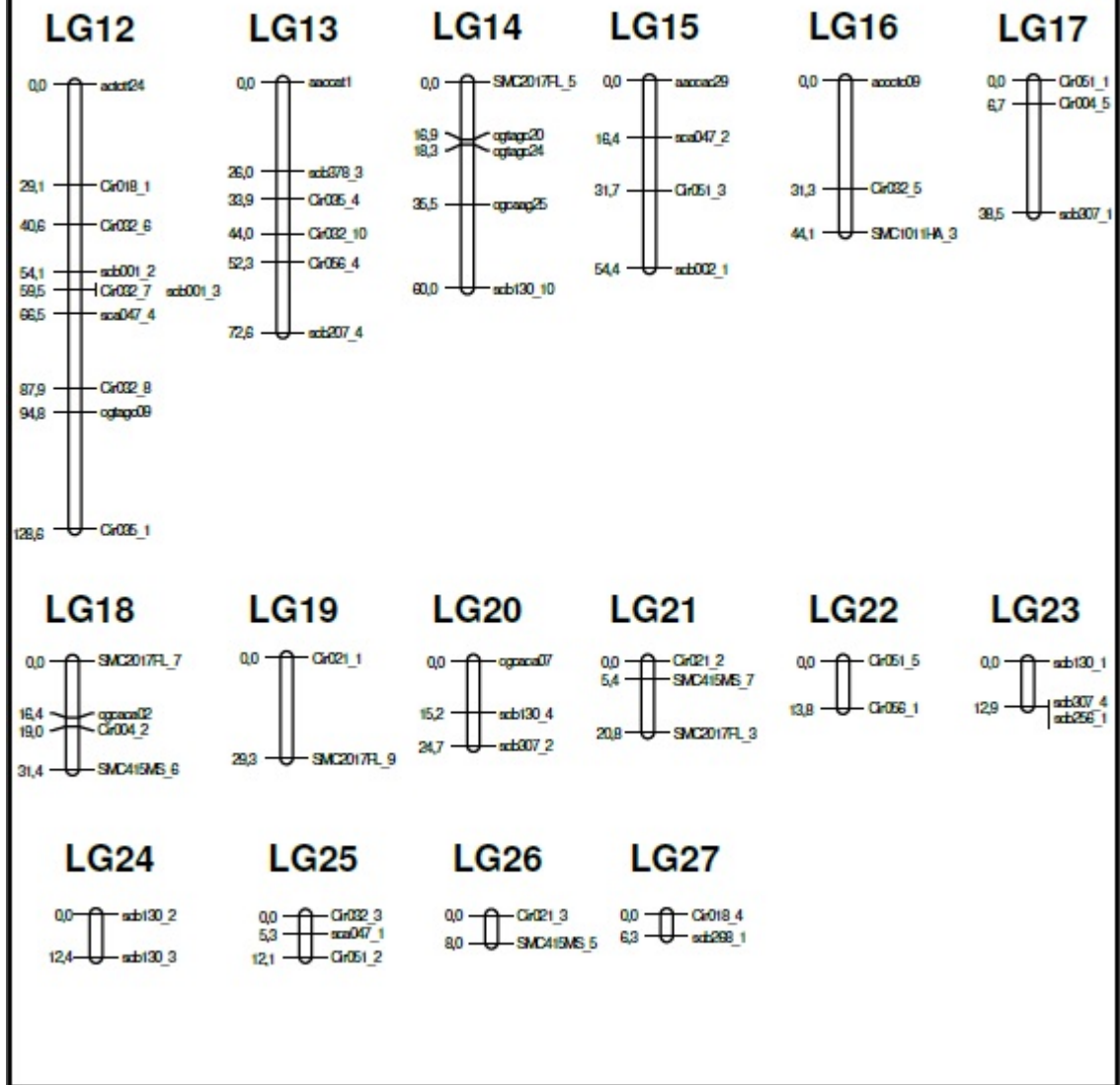
- Reffay N, Jackson PA, Aitken KS, Hoarau JY, D'Hont A, Besse P, McIntyre CL (2005) Characterisation of genome regions incorporated from an important wild relative into Australian sugarcane. *Mol Breed* 15:367–381
- Rossi M, Araujo PG, Paulet F, Garsmeur O, Dias VM, Chen H, Van Sluys MA, D'Hont AD (2003) Genomic distribution and characterization of EST82 derived resistance gene analogs (RGAs) in sugarcane. *Molecular Genetics and Genomics* 269: 406-419
- Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6:461–464
- Sills GR, Bridges W, Al-Janabi SM, Sobral BWS (1995) Genetic analysis of agronomic traits in a cross between sugarcane (*Saccharum officinarum* L.) and its presumed progenitor (*S. robustum* Brandes & Jesw. ex Grassl). *Mol Breed* 1:355–363
- Smith AB, Stringer JK, Wei X, Cullis BR (2007) Varietal selection for perennial crops where data relate to multiple harvests from a series of field trials. *Euphytica* 157:253–266
- Vos P, Hogers R, Bleeker M, Reijans M, Lee TV, Hornes M, Frijters A, Pot J, Peleman J, Kulper M, Zabeau M (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research* 23: 4407-4414
- Wu K, Burnquist W, Sorrels M, Tew T, Moore P, Tanksley S (1992) The detection and estimation of linkage in polyploids using single-dose restriction fragments. *Theor Appl Genet* 83:294–300
- 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
- Zeng ZB (1993) Theoretical basis of separation of multiple linked gene effects on mapping quantitative trait loci. *Proc Natl Acad of Sci* 90:10972-10976
- Zeng ZB (1994) Precision mapping of quantitative trait loci. *Genetics* 136:1457-1468



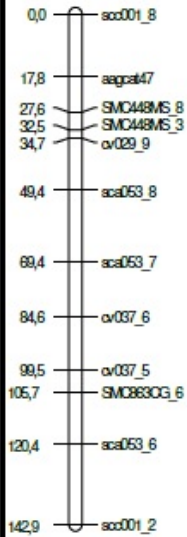
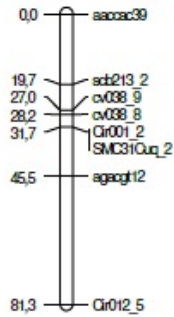
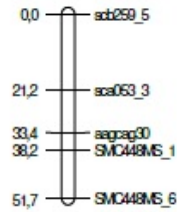
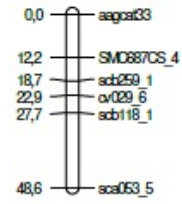
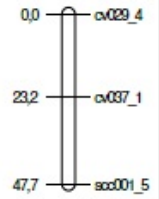
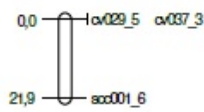
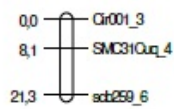
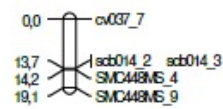
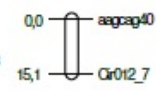
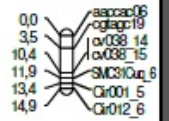
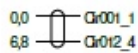
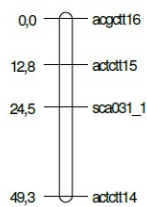
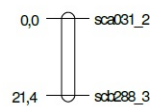
Supplementary Material



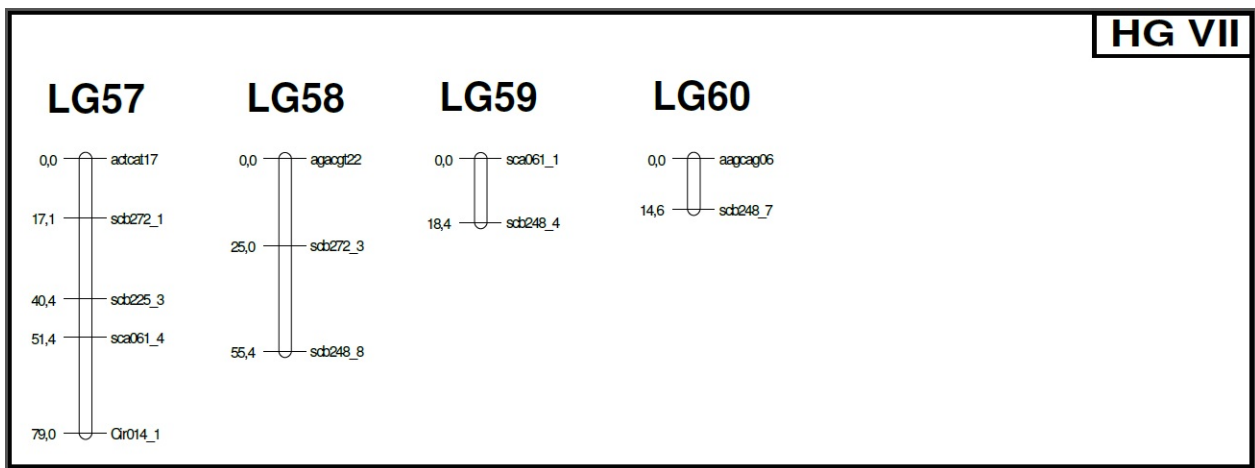
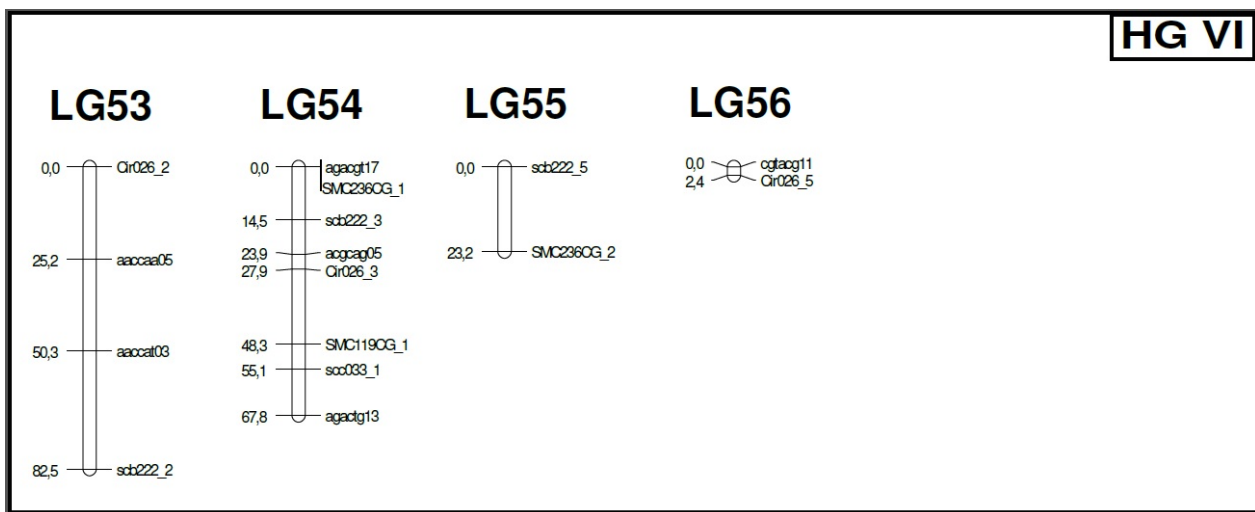
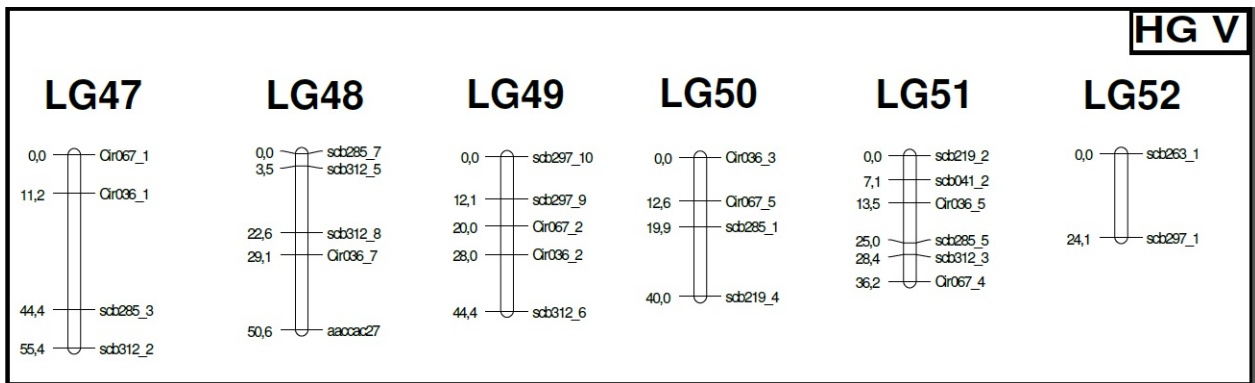
A genetic linkage map in a bi-parental cross between the elite clone IACSP95-3018 (female parent) and the variety IACSP93-3046.

**HG II**

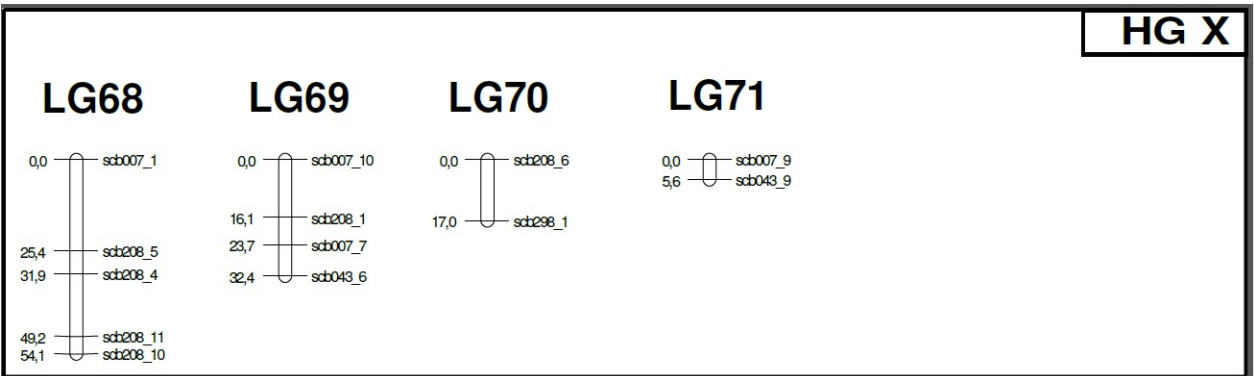
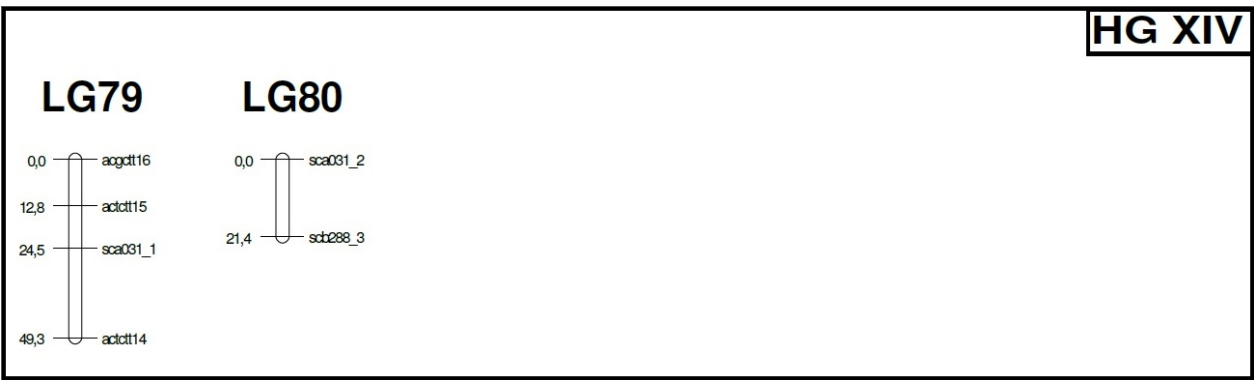
Continued

**HG III****LG28****LG29****LG30****LG31****LG32****LG33****LG34****LG35****LG36****LG37****LG38****LG39****LG40****HG XIV****LG79****LG80**

Continued



Continued

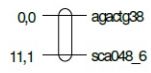
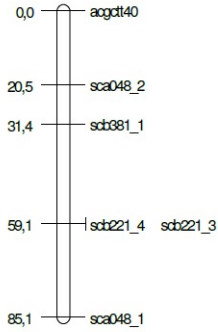


Continued

**HG XI**

**LG72**

**LG73**

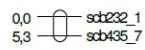
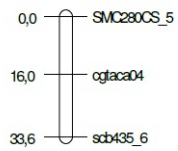
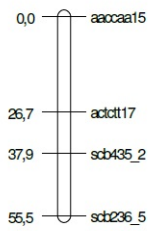


**HG XII**

**LG74**

**LG75**

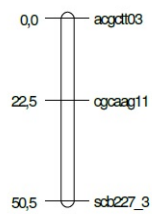
**LG76**



**HG XIII**

**LG77**

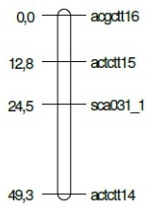
**LG78**



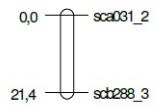
Continued

**HG XIV**

**LG79**

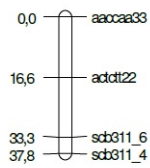


**LG80**

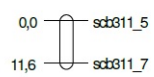


**HG XV**

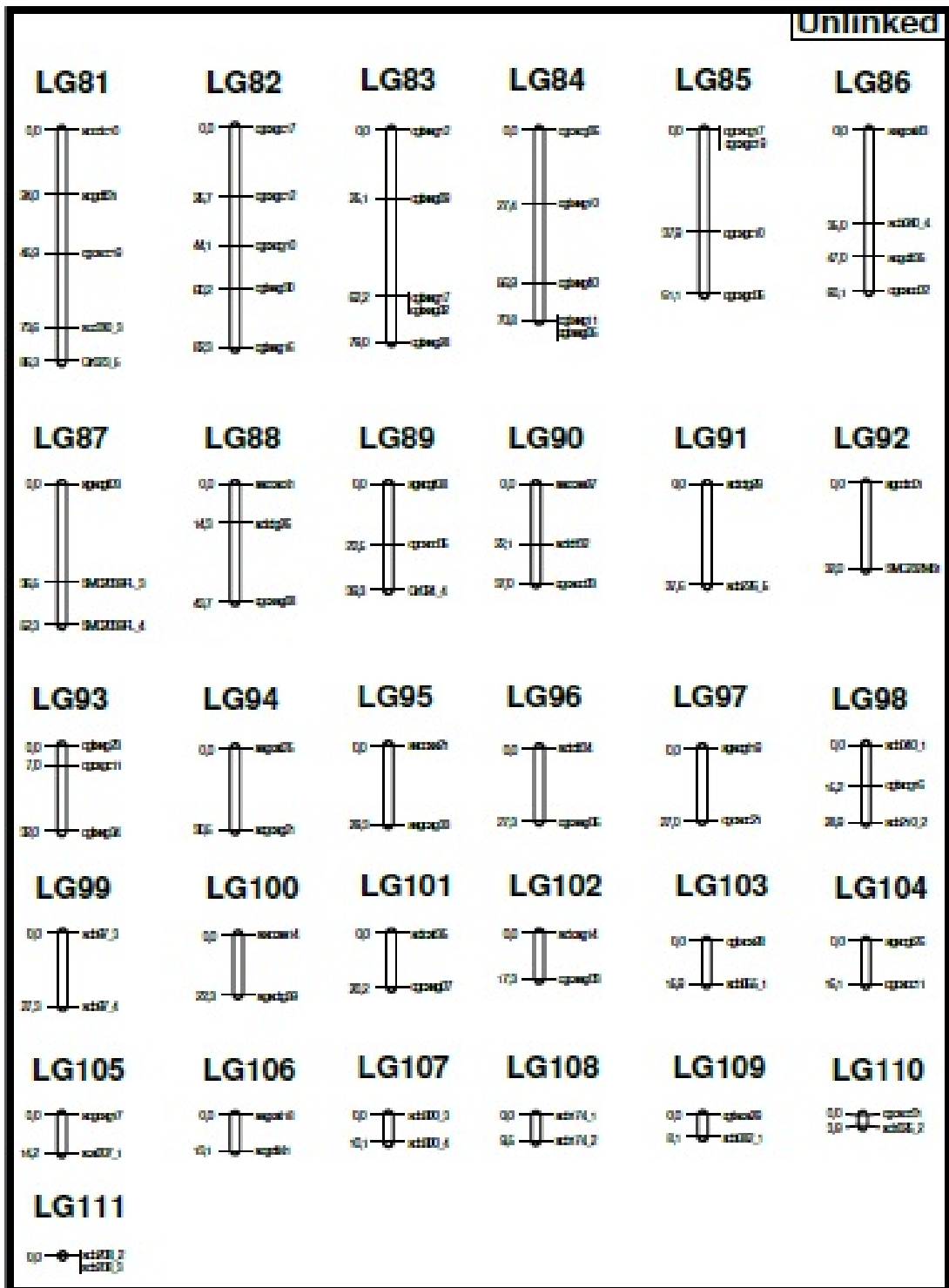
**LG81**



**LG82**



Continued



Continued



## De Novo Assembly and Transcriptome Analysis of Contrasting Sugarcane Varieties

Claudio Benicio Cardoso-Silva<sup>1,3</sup>, Estela Araujo Costa<sup>1,3</sup>, Melina Cristina Mancini<sup>1</sup>, Thiago Willian Almeida Balsalobre<sup>1</sup>, Lucas Eduardo Costa Canesin<sup>1</sup>, Luciana Rossini Pinto<sup>2</sup>, Monalisa Sampaio Carneiro<sup>3</sup>, Antonio Augusto Franco Garcia<sup>4</sup>, Anete Pereira de Souza<sup>1,5</sup>, Renato Vicentini<sup>1\*</sup>

**1** Center for Molecular Biology and Genetic Engineering (CBMEG), University of Campinas (UNICAMP), Campinas, SP, Brazil, **2** Centro Avançado da Pesquisa Tecnológica do Agronegócio de Cana (IAC/Apta), Ribeirão Preto, SP, Brazil, **3** Departamento de Biotecnologia e Produção Vegetal e Animal, Centro de Ciências Agrárias, Universidade Federal de São Carlos, Araras, SP, Brazil, **4** Departamento de Genética, Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba, SP, Brazil, **5** Departamento de Biologia Vegetal, Instituto de Biologia, Universidade Estadual de Campinas (UNICAMP), Campinas, SP, Brazil

### Abstract

Sugarcane is an important crop and a major source of sugar and alcohol. In this study, we performed *de novo* assembly and transcriptome annotation for six sugarcane genotypes involved in bi-parental crosses. The *de novo* assembly of the sugarcane transcriptome was performed using short reads generated using the Illumina RNA-Seq platform. We produced more than 400 million reads, which were assembled into 72,269 unigenes. Based on a similarity search, the unigenes showed significant similarity to more than 28,788 sorghum proteins, including a set of 5,272 unigenes that are not present in the public sugarcane EST databases; many of these unigenes are likely putative undescribed sugarcane genes. From this collection of unigenes, a large number of molecular markers were identified, including 5,106 simple sequence repeats (SSRs) and 708,125 single-nucleotide polymorphisms (SNPs). This new dataset will be a useful resource for future genetic and genomic studies in this species.

**Citation:** Cardoso-Silva CB, Costa EA, Mancini MC, Balsalobre TWA, Canesin LEC, et al. (2014) De Novo Assembly and Transcriptome Analysis of Contrasting Sugarcane Varieties. PLoS ONE 9(2): e88462. doi:10.1371/journal.pone.0088462

**Editor:** Cynthia Gibas, University of North Carolina at Charlotte, United States of America

**Received:** August 15, 2013; **Accepted:** January 7, 2014; **Published:** February 11, 2014

**Copyright:** © 2014 Cardoso-Silva et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** The authors gratefully acknowledge the Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP, <http://www.fapesp.br>) for the financial support grants 2008/52197-4 (AS) and 2008/58031-0 (RV) and for the graduate scholarships to CBCS, EAC, MCM, and TWB, and to the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ, <http://www.cnpq.br>) for the research fellowships to AAG, APS, and RV. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

\* E-mail: shinapes@unicamp.br

These authors contributed equally to this work.

### Background

Sugarcane belongs to the grass family (Poaceae), which is an economically important seed plant family that includes maize, wheat, rice, sorghum and many types of grasses. The sugarcane crop is the main source of both sugar and alcohol, accounting for two-thirds of the world's sugar production [1]. It is estimated that approximately 653.81 million tons of sugarcane will be produced during the 2013/2014 harvest in Brazil, surpassing the production of the last harvest [2].

Modern sugarcane varieties are derived from interspecific hybridization between *Saccharum officinarum* and *Saccharum spontaneum*, resulting in highly polyploid and aneuploid plants. Indeed, the chromosome number of these varieties ranges from 80 to 140. Modern varieties of sugarcane typically exhibit more than eight homologous copies of each basic chromosome from *S. officinarum* and several copies of the homologous chromosomes from *S. spontaneum* [3]. Therefore, sugarcane cultivars are highly heterozygous, presenting several different alleles at each locus, and this high level of genetic complexity creates challenges during conventional and molecular breeding programs.

Recent technological developments have the potential to greatly increase our understanding of sugarcane plants through the

application of emerging genomic technologies, and the use of next-generation sequencing (NGS) technologies could have significant implications for crop genetics and breeding. Although the sequencing of large genomes remains expensive, even using NGS technologies [4], transcriptome sequencing can provide information regarding the gene content of a species and can complement genome sequencing approaches.

RNA sequencing (RNA-Seq) has been applied as a tool for transcriptome analysis in many species, such as *Arabidopsis thaliana* [5], *Brassica* spp. [6], rice [7] and maize [8]. RNA-Seq has several advantages, including (i) allowing more precise measurement of the levels of transcripts and their isoforms than other methods, (ii) presenting the potential for the development of SNPs that can be used to detect allele-specific expression because the same base is sequenced multiple times, (iii) the ability to identify reads containing post-transcriptional modifications or rearranged sequences that cannot be mapped directly to the genome [9] and (iv) allowing the identification of species-specific genes [10]. Moreover, the availability of a large number of genetic markers developed using NGS technologies is facilitating trait mapping and marker-assisted breeding [11].

In plant breeding programs, genotypes of interest to breeders, such as the parental genotypes of mapping populations, can be

sequenced using NGS technologies. More than one genotype can be employed to generate sequence data with these technologies, and these data can be aligned using genome or transcriptome sequencing data for model or major crop species that are closely related to the species of interest [11]. This approach has also been applied for marker discovery in some crop species, such as eucalyptus [12], maize [13] and chickpea [14], and has been used to identify SNPs between the parental genotypes of mapping populations. These SNPs can then be employed to develop markers for marker-deficient crops to allow trait mapping through marker-assisted selection (MAS).

Despite its economic importance, no published genome sequence is currently available for sugarcane. Instead, the basic resource used for the study of sugarcane gene sequences is the substantial expressed sequence tag (EST) information available in public databases. Transcriptome studies in sugarcane began in South Africa [15,16], and the largest EST collection (~238,000 ESTs) was developed through the Brazilian SUCEST project [17,18]. Researchers in Australia [19–21] and the USA [22] have generated three additional libraries containing 10,000 ESTs each. Currently, all of the reported ESTs are collected in the Sugarcane Gene Index, version 3.0, which contains 282,683 ESTs and 499 complete cDNA sequences, resulting in 121,342 unique assembled sequences, or unigenes. There are still more than 10,000 sugarcane coding genes that have yet to be identified [23], highlighting the need for new sequencing efforts in the sugarcane transcriptome. This information would increase the panel of potential molecular markers and sequence information available for sugarcane breeding programs, resulting in biotechnological improvements. In the present study, using the Illumina GA IIx sequencing platform, we performed *de novo* transcriptome sequencing in six sugarcane genotypes that are employed as parents in Brazilian Sugarcane Breeding Programs. We identified conserved genes that have not previously been described in sugarcane, and these data will be useful for future genome assembly and marker identification.

## Materials and Methods

### Ethics Statement

We confirm that no specific permits were required for the described field studies. This work was a collaborative research project developed by researchers from UNICAMP, ESALQ/USP, IAC/Apta (Instituto Agronômico de Campinas) and UFSCAR-RIDESA (Universidade Federal de São Carlos-Rede Interinstitucional de Desenvolvimento do Setor Sucroalcooleiro) (all from Brazil). We also confirm that the field studies did not involve endangered or protected species.

### Plant Materials and RNA Extraction

Six genotypes were included in this study. IACSP96-3046 and IACSP95-3018 are the parents of a mapping population from the Sugarcane Breeding Program at IAC/Apta. IACSP95-3018 is a promising clone that is also used as a parent in the breeding program. IACSP93-3046 is a variety that exhibits good tillering, an erect stool habit [24] and resistance to rust [25].

SP81-3250×RB925345 and SP80-3280×RB835486 are the parents of two different mapping populations from the Sugarcane Breeding Program at UFSCar, which is part of RIDESA. These parents exhibit contrasting properties: SP81-3250 and SP80-3280 are resistant to rust [26,27], whereas RB925345 and RB835486 are susceptible [28]. All of the examined genotypes display high levels of sucrose.

Leaves at the third position [29] were collected from one plant per genotype and immediately frozen, and total RNA was extracted using a modified protocol [30]. The integrity and quantity of the isolated RNA were assessed using a 2100 Bioanalyzer (Agilent). Equal quantities of high-quality RNA from each genotype were pooled for cDNA synthesis.

### mRNA-Seq Library Construction for Illumina Sequencing

Paired-end Illumina mRNA libraries were generated from 4 µg of total RNA in accordance with the manufacturer's instructions for mRNA-Seq Sample Preparation (Illumina Inc., San Diego, CA, USA). The quality of the library was assessed using a 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA).

Cluster amplification was performed using the TruSeq PE Cluster Kit and a cBot (Illumina), and each sample was sequenced in a separate GAIIx lane using the TruSeq SBS 36 Cycle Kit (Illumina). The read length was 72 bp.

### Sequence Data Analysis and Assembly

The raw data generated by Illumina sequencing were converted from the BCL format to qSeq using Off-line Basecaller, v.1.9.4 (OLB) software. The qSeq files were transformed in FastQ files, which contain sequences that are 72 bp in length, using a custom script. Low-quality sequences were removed; these sequences included reads with ambiguous bases, reads with less than 70 bases, and reads with a Phred quality score  $Q \leq 20$  using the NGS QC toolkit [31]. All reads were deposited in the National Center for Biotechnology Information (NCBI) database and can be found under accession number SRA073690.

All datasets were combined, and the sequenced reads were assembled using Trinity (<http://trinityrnaseq.sourceforge.net/>), which is a program developed specifically for *de novo* transcriptome assembly from short-read RNA-Seq data that recovers transcript isoforms efficiently and sensitively using the de Bruijn graph algorithm [32]. The optimal assembly results were chosen according to an evaluation of the assembly encompassing the total number of contigs, the distribution of contig lengths, the N50 statistic and the average coverage. The assembled transcripts were based on the main isoform of each transcript, and only contigs with lengths of greater than 300 bp were included in the downstream analysis.

To identify the genotypic contribution to each transcript, reads from each library were mapped against the assembly generated from all libraries using the bowtie aligner [33]. The BAM files generated by bowtie were then used to estimate the transcript-level abundance for each library using the RSEM (RNA-Seq by Expectation Maximization) software [34].

### Functional Annotation of Sugarcane Transcripts

The assembled sequences were compared against the NCBI non-redundant protein database (NR) using BLASTX with a cut-off E-value of  $10^{-6}$ . To annotate the assembled sequences according to Gene Ontology (GO) terms (The Gene Ontology Consortium, 2000), the above BLAST results were analyzed using Blast2GO [35] to determine and compare gene functions. The GO terms were assigned to the representative transcripts for each sample through an enrichment analysis using Fisher's exact test (p-value  $< 0.01$ ), with a false discovery rate (FDR) correction in terms of biological processes and molecular functions. The transcript sequences were also aligned against the *Viridiplantae*, grass and sorghum protein databases (<http://www.phytozome.org/>) using BLASTX and against the Sugarcane Gene Index (<http://compbio.dfci.harvard.edu/tgi/>) using BLASTN; in both alignments, a cut-off E-value of  $10^{-6}$  was applied. The BLAST search

was limited to the first ten significant query hits, and the gene names were assigned to each query based on the highest score. Transcripts that showed similarity to *Viridiplantae* proteins were aligned against the sorghum genome using sim4 software [36]. Open reading frames (ORFs) were predicted using a script available in the TransDecoder package (<http://transdecoder.sourceforge.net/>), with 300 bp as the minimum ORF length. Those transcripts showing predicted ORFs were aligned against grass proteins using the STRING database, v.9.05 (<http://string-db.org>), to predict Clusters of Orthologous Groups (COG).

To further characterize the subset of unigenes that did not show similarity to any known plant proteins, we applied a computational strategy to mine putative long non-coding RNA (lncRNA) data. We first aligned all 121,342 EST unigenes to *Viridiplantae* proteins and to the GenBank NR database using BLASTX. Those EST unigenes that did not align with any proteins were then mapped to the *Sorghum bicolor* genome, obtaining at least 70% coverage and a maximum intron size of 15 kb. The coding probability of the positively mapped unigenes was then evaluated by removing sequences with potential ORFs longer than 100 aa using ESTScan [37]. We further investigated the functional role of the remaining unigenes and putative lncRNAs by searching for three indirect indications of functionality: we examined the stability of the secondary structure using the Vienna package [38], normalized to the Z-score index [39]; we mapped the small RNAs (sRNAs) [40] against sugarcane unigenes; and we analyzed the sequence similarities between the unigenes and *S. bicolor* ESTs (BLASTN, E-value  $\leq 1e^{-5}$ ). Only EST unigenes with at least one indirect piece of functional evidence were analyzed further. The putative lncRNAs were then aligned to the 18,910 assembled transcripts that showed no similarity to any plant protein but were successfully mapped to *S. bicolor* (Text S4). Only hits with an E-value below  $1e^{-5}$  and coverage higher than 40% were considered positive.

### Putative Molecular Markers

We utilized the MISA program (<http://pgrc.ipk-gatersleben.de/misa/>) to search for simple sequence repeat (SSR) motifs in the unigenes; the MISA script can identify both perfect and compound (interrupted by a certain number of bases) motifs. To identify the presence of SSRs, only motifs of two to six nucleotides were considered, and the minimum repeat unit was defined as six for dinucleotide motifs and five for tri-, tetra-, penta- and hexanucleotide motifs. A compound motif was defined as two or more SSR motifs interrupted by sequences of up to 100 bp.

To identify putative single-nucleotide polymorphisms (SNPs) in the sugarcane transcript assembly, we first separately mapped all of the short reads from each library to the assembly using the Burrows-Wheeler Aligner (BWA). Next, FreeBayes [41] and SAMtools [42] were used to detect the variable positions of SNPs from the consensus sugarcane assembly. The FreeBayes tool allowed us to identify genetic variants in the polyploid organisms. The putative SNPs were then filtered using the varFilter command, where variants were called only for positions with a minimal mapping quality (-Q) and coverage (-d) of 25. To compare the composition of the SNP variation in the parental genotype, unique and shared SNPs were extracted using an in-house script. The transition and transversion ratios were calculated using the tsv tool developed by SnpSift software [43].

## Results and Discussion

### *De novo* assembly of the sugarcane transcriptome

The libraries sequenced using the Illumina platform produced a total of 610,232,490 paired-end (PE) sequence reads, each of

which was 72 bp in length. We filtered the sequence data for low-quality reads, resulting in 445,374,504 high-quality PE trimmed reads (97.67%), which were used to obtain the *de novo* assembly. An overview of the sequencing procedure is presented in Table 1. The *de novo* assembly generated 119,768 transcripts when all isoforms were considered. These transcripts represent a total of 72,269 unigenes that were considered for downstream analysis (Text S1). The length of the unigenes ranged from 300 bp to ~7 kb, with a mean length of 921 bp, an N50 equal to 1,367 bp and 46.39% GC content. The average length of the assembled unigenes was greater than those obtained from chickpea (523 bp) [14], rubber trees (485 bp) [44] and bamboo (736 bp) [45] using similar sequencing technologies. Considering the N50 values, the values for the sugarcane unigenes were greater than those for rubber trees (592 bp), bamboo (1,132 bp) and chili pepper (1,076 bp) [46], which were also assembled using short reads generated by the Illumina platform. In total, we obtained 18,624 (27.21%) unigenes longer than 1 kb and 7,657 (10.6%) unigenes longer than 2 kb. The length distributions of the unigenes are shown in Table 2, revealing that more than 40,000 unigenes (55.76%) were longer than 500 bp. These unigenes were submitted to an ORF predictor using TransDecoder, and we detected 33,673 (46.59%) unigenes with ORFs, with 9,350 (12.94%) presenting complete ORFs.

### Unigene annotation

The 72,269 sugarcane unigenes were analyzed for sequence similarity against the *Viridiplantae* (comprising all green plants) and grass (*S. bicolor*, *Oryza sativa*, *Zea mays*, *Panicum virgatum*, *Setaria italica* and *Brachypodium virgatum*) datasets through BLASTX searches. The unigenes were also compared against the sugarcane EST database via a BLASTN search (Table 3). A total of 35,456 (49.06%) unigenes showed significant similarity to *Viridiplantae*. The high percentage of sugarcane unigenes obtained in this study that did not match the *Viridiplantae* protein database (50.84%) indicates that there is potential for the discovery of as-yet-undescribed and novel genes in sugarcane, although most of these unigenes may encode non-coding RNAs. In fact, more than 26% of the unigenes in this set exhibited high similarity to intergenic regions of the sorghum genome (Figure 1). Additionally, the significance of a BLAST search depends on the length of the query sequence; therefore, short sequences are rarely matched to known genes [12], or these sequences may represent rapidly evolving sequences that have diverged substantially from their homologs [47].

In turn, alignment of the unigenes against the grass protein database returned 34,814 significant hits. When considering the hits by species, 28,788 unigenes showed significant similarity to sorghum, corresponding to 98% of sorghum proteins (Figure 1).

**Table 1.** Summary of Illumina transcriptome sequencing data for the sugarcane varieties included in this study.

Sample	Read length (bp)	Raw data	Trimmed data	GC (%)	Q20 (%)
SP95-3018	72+72	84,105,462	64,906,391	49.04	98.09
SP81-3250	72+72	103,971,718	71,002,186	47.52	97.32
RB925345	72+72	112,124,334	77,476,268	46.91	97.11
SP80-3280	72+72	101,983,186	73,160,814	47.59	97.56
RB835486	72+72	119,280,444	87,873,521	46.62	97.66
SP93-3046	72+72	88,767,346	70,955,324	48.07	98.25

doi:10.1371/journal.pone.0088462.t001

**Table 2.** Summary of the *de novo* assembly results for the sugarcane transcriptome.

Unigene length (bp)	Total unigenes	Percentage
300–500	31,971	44.24%
500–1000	20,634	28.55%
1000–2000	12,007	16.61%
2000–3000	4,827	6.68%
3000–4000	1,790	2.47%
4000–5000	636	0.88%
>5000	404	0.56%
Total length (bp)	66,572,642	-
Unigenes	72,269	-
N50 length	1,367	-
GC (%)	46.39	-

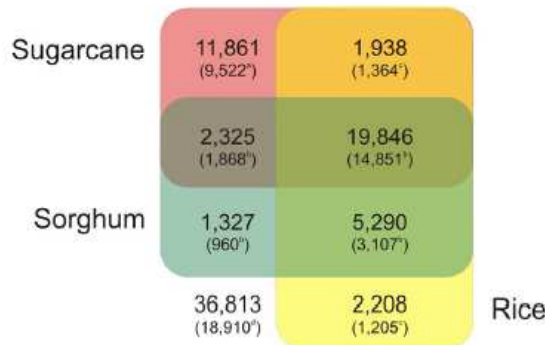
doi:10.1371/journal.pone.0088462.t002

These results were expected, as comparative genomic studies [48] have revealed conservation and synteny among the sugarcane and sorghum genomes. The sugarcane transcriptome also significantly matched that of rice, with approximately 29,285 unigenes (corresponding to 28,732 unique protein accessions) showing significant similarity to rice proteins.

To investigate previously unidentified potential genes in sugarcane, we compared the unigenes against the sugarcane transcripts deposited in public databases and performed BLAST searches to detect possible similarities with the SoGI database (*S. officinarum*). Furthermore, the unigenes that did not show similarity to sugarcane ESTs were compared against sorghum proteins. Approximately 22,171 unigenes exhibited significant similarity to sorghum proteins and sugarcane transcripts (Figure 1). The remaining 5,272 unigenes (Text S3) showed significant similarity to sorghum and rice proteins but not to the sugarcane transcripts that were considered to be putative new sugarcane genes (Figure 1). By examining the presence of candidate coding regions in these unigenes, we identified 4,895 sequences that contained ORFs, with 732 unigenes containing complete ORFs. These unigenes represent genes that have not yet been described for sugarcane.

**Clusters of Orthologous Groups (COG) classification**

COG classification was performed for the transcriptome data, and a total of 7,519 unigenes were identified (Figure 2). These unigenes were classified into 23 COG categories, with the largest



**Figure 1.** Proportions of sugarcane transcripts showing homology to sugarcane unigenes and sorghum and rice proteins. For annotation, the best BLASTX/N hit against the protein or nucleotide sequences of the reference organisms was employed, with an E-value cut-off of  $\leq 10^{-6}$ . The number between the parentheses indicates the number of different proteins/unigenes in each species (sugarcane<sup>a</sup>, sorghum<sup>b</sup> and rice<sup>c</sup>). The number outside of the Venn diagram indicates no-hit transcripts and the number of transcripts<sup>d</sup> that mapped to the sorghum genome. doi:10.1371/journal.pone.0088462.g001

number of unigenes being grouped in the ‘replication, recombination and repair’ cluster (20.49%), followed by the ‘general function prediction only’ cluster (17.05%) and the ‘postranslational modification, protein turnover and chaperones’ cluster (7.39%). These three categories are the same categories that are highly represented in sorghum (Figure 2).

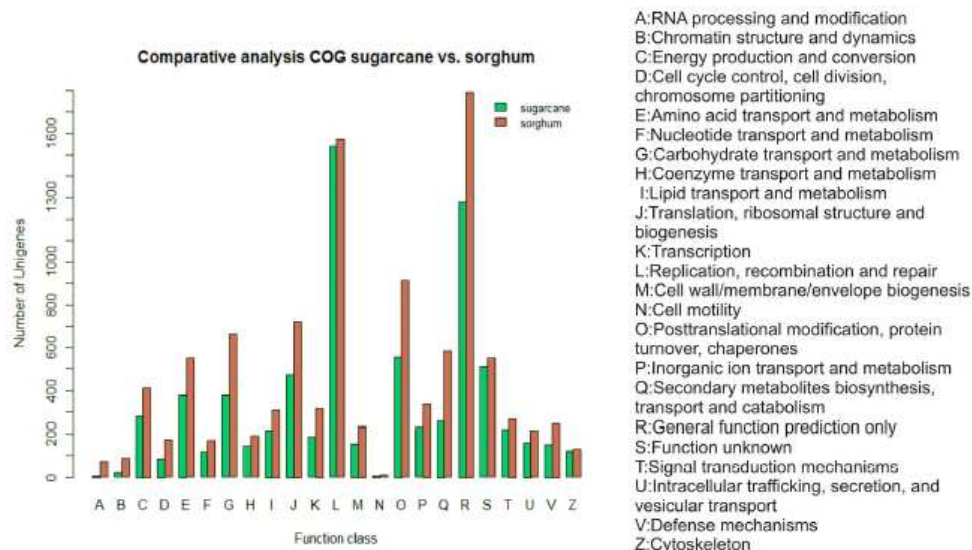
A total of 19 of the 23 COG categories were present in the transcriptome data, and at least 60% of the sugarcane unigenes were annotated when compared with the annotation of sorghum genes in the COG categories.

The categories ‘energy production and conversion’ (3.72%), ‘carbohydrate transport and metabolism’ (5%) and ‘defense mechanisms’ (2%) exhibited at least 56% of the expected genes compared with the sorghum genes. These categories should be considered to represent gene sequences showing a high potential for the development of molecular markers in sugarcane breeding programs. Therefore, the likelihood of these markers being associated with agronomic traits of interest in QTL mapping and marker-assisted selection (MAS) [49] is increased.

**Table 3.** Summary of the annotation of each database.

Database	Number of unigenes	Number of proteins matched	Percentage of unigenes <sup>a</sup>
Viridiplantae proteins	35,456	34,969	49.06%
Grass proteins	34,814	34,304	48.17%
Sorghum proteins	28,788	28,030	39.83%
Hits against sorghum proteins and sugarcane ESTs	22,171	20,969	30.68%
Total of no-hit unigenes	36,813	-	50.94%
No-hit unigenes with high similarity to the sorghum genome	18,910	-	26.16%

<sup>a</sup>Percentage relative to the total number of sugarcane unigenes. doi:10.1371/journal.pone.0088462.t003



**Figure 2. Histogram of the Clusters of Orthologous Groups (COG) classifications of the sugarcane transcripts and sorghum proteins.**

doi:10.1371/journal.pone.0088462.g002

### Gene Ontology enrichment analyses

The identification of functional classes that differ statistically between two lists of terms is a typical data-mining approach applied in functional genomics research [35]. In this work, we were interested in identifying which functions were distinctly represented among the different sugarcane genotypes. A total of 14,983 unigenes (Text S2) were annotated based on BLAST matches to known proteins in the NR database and were assigned to GO classes representing 39 terms, including some (10) that contain important information related to the enriched genotype (Figure 3).

Genes responsible for disease resistance, corresponding to the categories 'signaling,' 'response to stimulus,' 'cellular response to stimulus,' 'response to chemical stimulus' and 'response to auxin stimulus', were enriched in the SP81-3250, SP80-3280 and IACSP93-3046 genotypes, with IACSP93-3046 being represented in all of these categories (Figure 3). These three genotypes exhibit resistance to rust [25–27], whereas the other genotypes, RB925345, RB835486 and IACSP95-3018, are susceptible to rust [24,28]. Common sugarcane rust, caused by the fungus *Puccinia melanocephala*, is a disease that occurs worldwide and can result in large losses of sugar tonnage in susceptible varieties [50]. Rust resistance is generally considered to be a quantitatively inherited trait showing a high degree of heritability and a strong additive genetic variance component [51,52].

The obtained enriched terms suggest that these three genotypes harbor transcripts that are involved in stimulus response pathways and probable disease responses. These results are correlated with the characteristics of resistance and susceptibility in these varieties.

Another important characteristic of sugarcane crops is their accumulation of sucrose. Wild sugarcane species produce less than 4% fresh weight of sucrose, whereas high-yield varieties can produce sucrose contents of up to 20% of their fresh weight [53]. The major differences between these varieties is based on sugar transport and metabolism in storage tissues [54]. The entire network involving sucrose synthesis, accumulation, storage and

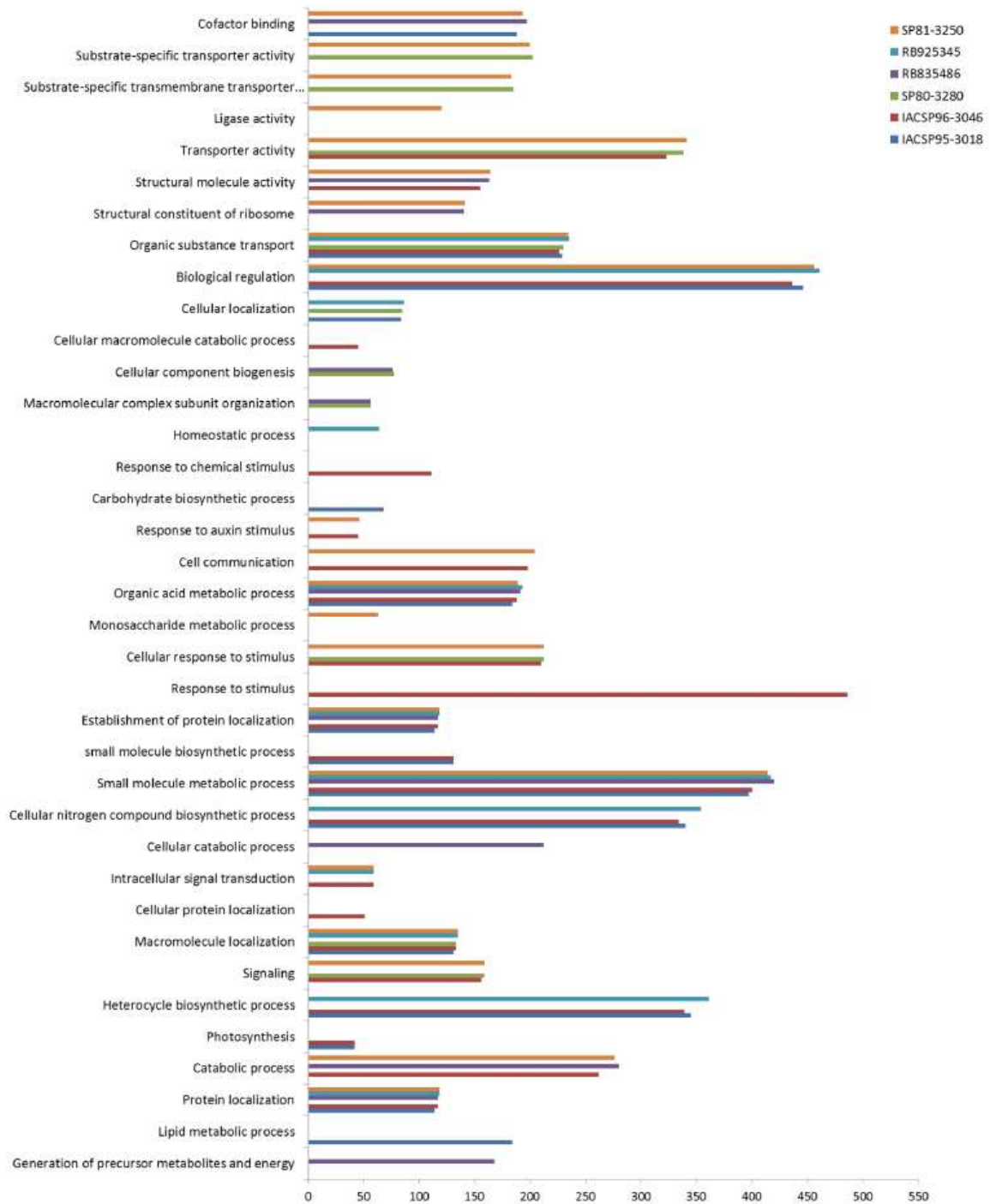
retention is a complex system in which several metabolic pathways interact with each other [55]. The most important aspect of this network is transport, which chiefly involves specific carrier molecules, ion transport and active transport and depends on the amount of available ATP. Within this context, we observed some genotypes that were enriched in categories related to this network, particularly the transport process. These categories included 'organic substance transport' (SP81-3250, RB925345, SP80-3280, IACSP96-3046 and IACSP95-3018), 'substrate-specific transporter activity', 'substrate-specific transmembrane transporter activity' (SP81-3250 and SP80-3280), 'ion transmembrane transport' (SP81-3250 and IACSP93-3046) and 'transporter activity' (SP81-3250, SP80-3280, and IACSP93-3046).

Important categories involved in sugar transport and metabolism in storage tissues include the 'monosaccharide metabolic process,' 'glucose metabolic process,' 'small molecule biosynthetic process' and 'small molecule metabolic process' categories. The terms in the first and second categories were only enriched in the SP81-3250 genotype, whereas the terms in the third category were enriched in both the IACSP93-3046 and IACSP95-3018 genotypes. All genotypes showed enrichment in the last category, although SP80-3280 was the least represented.

All of the genotypes were enriched for transcripts involved in this complex network of sucrose synthesis, accumulation, storage and retention, and these results were corroborated by the agronomic characteristics of the plants. All of these genotypes produce high levels of sucrose, in accordance with the agronomic description of the genotypes SP81-3250 [26], RB925345, RB835486 [28], SP80-3280 [27], IACSP93-3046 [25] and IACSP95-3018 [24].

### Putative lncRNAs

Among the initial set of 121,342 EST retrieved unigenes, 23,529 showed no similarity to any known plant protein. These unigenes were mapped to the *S. bicolor* genome, resulting in 4,476 positive hits, with only 1,884 not exhibiting an ORF or presenting an ORF



**Figure 3. Enrichment of Gene Ontology terms for each sugarcane variety.**  
 doi:10.1371/journal.pone.0088462.g003

shorter than 100 aa. This subset comprised the putative sugarcane lncRNAs that are publicly available. We found that for ~4% of these sequence, there were small RNAs (sRNAs) that mapped to their sequence, with ~59% showing similarity to *S. bicolor* and ~39% showing a highly stable secondary structure. In total, 1,446 non-redundant putative lncRNAs were identified that showed indirect evidence of functionality (Figure S1). We then compared this inclusive set (1,884 sequences) with the 18,910 assembled transcripts that lacked similarity to plant proteins. We observed 358 putative lncRNAs represented among the assembled transcripts, with ~42% of these sequences showing a highly stable secondary structure and ~40% showing evidence of transcription in the *S. bicolor* EST dataset. None of the unigenes to which sRNAs were mapped were similar to any assembled transcript. Finally, we compared the expression profiles of the putative lncRNAs between the different genotypes, which suggested that these transcripts may display genotype-specific expression patterns, as shown in Figure 4. A hierarchical clustering analysis revealed a pattern of separation between the genotypes from the different breeding programs, a result that is in accordance with the observation that the varieties from the same breeding program have the same genetic basis. We observed that the plant lncRNAs may display elevated intraspecific variation in expression, and several recent works have demonstrated that these transcripts exhibit tissue- and cell-specific expression patterns [56–59]. This study adds information regarding the dynamic involvement of these transcripts and reveals putative targets for further investigation [60,61].

#### Marker discovery

**SSR discovery.** Expressed sequence tag/simple sequence repeat (EST-SSR) markers are well established as important tools for researchers assessing genetic diversity and are useful in the development of genetic maps, comparative genomics and MAS breeding. Thus, the unigene sequences were searched for repeat motifs to explore the SSR profiles in the sugarcane transcriptome. A total of 5,106 SSRs were obtained from 4,616 unigene sequences (7.96%), and 576 of the unigenes contained more than one SSR (Text S7). Of these unigenes, 189 exhibited compound SSR formation. Trinucleotide repeat motifs were the most abundant, accounting for 2,585 SSRs (50.63%) in 2,318 unigene sequences; dinucleotide repeat motifs accounted for 1,927 SSRs (37.74%) in 1,732 unigenes; and other motifs accounted for 594 SSRs (11.63%) in 1,708 unigenes (Table 4). The relative percentage of the sequences containing SSRs was higher than that obtained in the SUCEST (Sugarcane Expressed Sequence Tag database) study, in which 2,005 clusters containing SSRs were found among 43,141 clusters (4.64%) [62].

The most abundant motifs included the dinucleotide AG motif (49.9%) and the trinucleotide CCG (17%) and ACC (4.7%) motifs. These results are similar to those of the SSR motif analysis

**Table 4.** Summary of the simple sequence repeat (SSR) types in the sugarcane transcriptome.

Repeat motif	Number <sup>a</sup>	Unigenes <sup>b</sup>	Percentage (%) <sup>c</sup>
<b>Di-nucleotide</b>			
AC/GT	551		
AG/CT	962		
AT/TA	336		
CG/GC	78		
<b>Total</b>	<b>1,927</b>	<b>1,732</b>	<b>37.74</b>
<b>Tri-nucleotide</b>			
AAC/GTT	141		
AAG/CTT	152		
AAT/ATT	60		
AGC/GCT	219		
ACG/CGT	197		
AGT/ACT	62		
ACC/GGT	122		
AGG/CCT	252		
ACA/TGT	97		
AGA/TCT	46		
ATA/TAT	24		
ATC/GAT	42		
ATG/CAT	43		
CAC/GTG	69		
CAG/CTG	228		
CCG/CGG	442		
CGC/GCG	241		
CTC/GAG	148		
<b>Total</b>	<b>2,585</b>	<b>2,318</b>	<b>50.63</b>
<b>Other motifs<sup>d</sup></b>	<b>594</b>	<b>1,708</b>	<b>11.63%</b>
<b>Total</b>	<b>5,106</b>	<b>5,758</b>	<b>-</b>

<sup>a</sup>Number of the total SSRs (di-, tri- and other motifs).

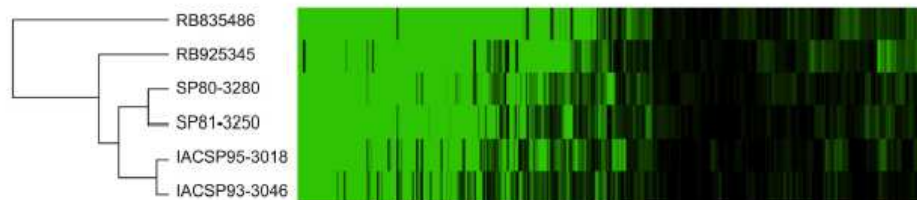
<sup>b</sup>Number of unigene sequences containing SSRs.

<sup>c</sup>The relative percentage of SSRs with different repeat motifs among the total SSRs.

<sup>d</sup>The total number of SSRs of other sizes.

doi:10.1371/journal.pone.0088462.t004

performed in sorghum [63]. Additionally, CCG and ACC were the most commonly found motifs in the SUCEST study [62], and CCG was the motif that was identified most often by Cordeiro *et al.* [64]. The most frequent tetranucleotide motif found in the



**Figure 4.** Hierarchical clustering of the 358 putative sugarcane lncRNAs. The expression patterns allowed the identification of the genotypes based on their ability to store sucrose and according to the bi-parental crosses involved in the different mapping populations. doi:10.1371/journal.pone.0088462.g004

present study was AAAG. The overall frequency of SSRs was observed to be 1/1.6 kb.

The prevalence of trimeric motifs over other SSR repeats may be explained based on the risk of frameshift mutations that may occur when microsatellites alternate in size [65]. Furthermore, a large number of trinucleotide coding repeats appear to be controlled primarily by mutation pressure.

The development of SSR markers associated with important agronomic traits can be used to assist in the selection of varieties during the early stages of MAS breeding programs and can be helpful in the selection of the best parents for crossing [66]. Consequently, the application of such markers supports breeding programs by significantly reducing the time and cost involved in developing new varieties and can help bypass barriers in sugarcane breeding programs.

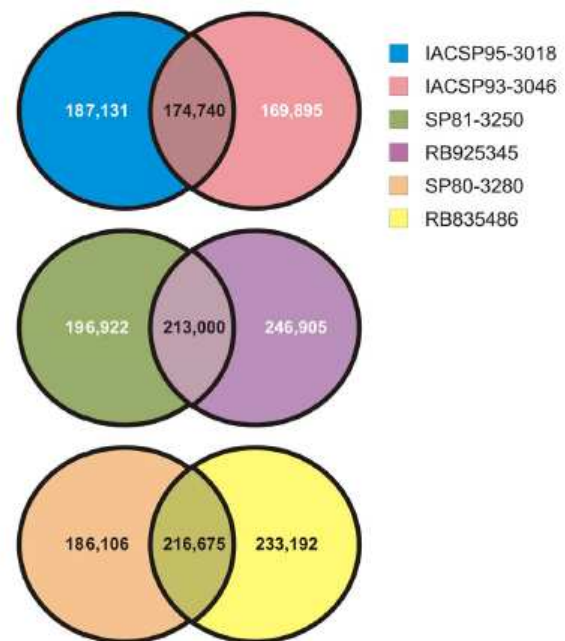
**SNP discovery.** A total of 708,125 putative SNP positions were identified (Text S5), with a density of 1 SNP per 86 bp. The frequency of SNPs found in the sugarcane genes was higher than has been observed in other grasses, such as rice and sorghum, which exhibit a frequency of  $\geq 1$  SNP per 300 bp [67]. The observed number of transitions was 456,666, and 254,658 transversions were detected, with the number of the former being 1.79 times that of the latter. Transitions were most likely more frequent because they are more tolerated by natural selection as the tendency to generate synonymous mutations in coding sequences is related to the number of transversions [68].

We identified SNPs in 58,903 different unigenes, which represent 81.50% of the total unigenes. Considering the number of unigenes without SNPs, we verified that 10,516 (79%) are unigenes with a length of less than 500 bp. Considering only those unigenes with predicted ORFs (33,673 unigenes), we found a total of 289,969 SNPs (37.5% of the total detected SNPs).

To detect different heterozygous SNPs between the parents from each mapping population, the reads from each genotype were mapped against all the unigenes (Text S6). Figure 5 shows the heterozygous SNPs that were detected, and the unique and shared SNPs in each parent from the mapping populations were evaluated. The percentages of SNPs that were common in the three mapping populations, IACSP95-3018  $\times$  IACSP93-3046 (32.86%), SP81-3250  $\times$  RB925345 (32.42%) and SP80-3280  $\times$  RB835486 (34.06%), were similar, and these SNPs may thus be polymorphic between the parents. As sugarcane is a polyploid species, polymorphisms can be generated from a different number of allelic copies present in each genotype. However, such polymorphisms are difficult to validate (Garcia *et al* 2013, *submitted*).

The SNPs that were unique to each genotype (Figure 5) exhibited a higher probability of association with the contrasting agronomic traits of interest. Because polymorphism markers between parents are important for generating saturated genetic mapping in mapping populations, these SNPs are a source of data for generating markers associated with quantitative trait loci (QTLs). Such functional molecular markers have been broadly applied for the genetic improvement of several crops [69].

According to the Gene Ontology annotation, we identified SNPs in 6,712 unigenes with annotation information, representing 44.80% of the unigenes included in the enrichment analyses. Some categories exhibited important results related to the genotype (Figure 3), particularly those associated with disease resistance. In the 'signaling' category, we identified 161 unigene sequences with SNPs, whereas we identified 477 unigenes with SNPs in the 'response to stimulus' category. These unigenes likely represent source data for the development of functional markers related to disease resistance.



**Figure 5. Unique and shared heterozygous putative SNPs in the parental genotypes of the three sugarcane mapping populations.**

doi:10.1371/journal.pone.0088462.g005

When we analyzed the categories related to sucrose synthesis, accumulation, storage and retention, we also observed unigenes with SNPs in the 'organic substance transport' (226), 'substrate-specific transporter activity' (196) and 'ion transmembrane transport' (53) clusters. Equally important categories involving sugar transport and metabolism in storage tissues, such as the 'glucose metabolic process' (43), 'small molecule biosynthetic process' (133) and 'small molecule metabolic process' (414) categories, also containing unigene sequences with SNPs.

All of these unigene sequences with SNPs represent an important source of data. These sequences could be priority candidates for the development of specific functional markers and could be very useful in further genetic or genomic studies in sugarcane.

## Conclusion

This is the first publicly available sugarcane transcriptome sequencing study performed using NGS technology to investigate the entire sugarcane transcriptome, and our data provide the most comprehensive transcriptome resource currently available for sugarcane. In addition, polymorphisms associated with candidate genes potentially involved in the stimulus response, energy production and growth were identified among the contrasting varieties and deserve future investigation. Based on the enrichment analysis, we identified putative genes related to disease and the accumulation of sucrose. Additionally, a large number of SNPs and SSRs were identified, and marker development would be a useful resource for future genetic or genomic studies of this species. Finally, this work contributed information on 5,000 undescribed



genes, which is more than half of the expected sugarcane genes that are missing from sugarcane databases.

### Supporting Information

**Figure S1** Venn diagram showing the classification of the identified putative sugarcane lncRNAs in the EST data (A) and RNA-Seq data (B).  
(TIF)

**Text S1** Unigene sequences in FASTA format.  
(ZIP)

**Text S2** Gene ontology enrichment annotation for the transcripts of each genotype.  
(ZIP)

**Text S3** Putative previously unknown sugarcane transcripts showing the best matches to sorghum proteins.  
(TXT)

**Text S4** List of 18,910 putative sugarcane ncRNAs with high coverage in the sorghum genome.  
(TXT)

**Text S5** List of 708,125 putative SNP positions identified in this study.  
(ZIP)

**Text S6** List of putative SNPs identified in each genotype.  
(ZIP)

**Text S7** List of 5,106 putative SSR positions identified in this study.  
(XLS)

### Author Contributions

Conceived and designed the experiments: AAFG MSC LRP APdS RV. Performed the experiments: EAC MCM TWAB. Analyzed the data: CBCS EAC LECC RV. Contributed reagents/materials/analysis tools: EAC MCM TWAB. Wrote the paper: CBCS EAC RV.

### References

- United States Department of Agriculture (2013) Sugar: World Markets and Trade. Foreign Agric Service. Available: <http://usda01.library.cornell.edu/usda/current/sugar/sugar-11-21-2013.pdf>. Accessed 10 December 2013.
- Ministério da Agricultura (2013) Acompanhamento de safra brasileira: cana-de-açúcar Safra 2012/2013 Terceiro levantamento. Cia Nac Abast. Available: [http://www.conab.gov.br/OlalaCMS/uploads/arquivos/12\\_12\\_12\\_10\\_34\\_43\\_boletim\\_cana\\_portugues\\_12\\_2012.pdf](http://www.conab.gov.br/OlalaCMS/uploads/arquivos/12_12_12_10_34_43_boletim_cana_portugues_12_2012.pdf). Accessed 10 December 2013.
- Ming R, Liu SC, Lin YR, da Silva J, Wilson W, et al. (1998) Detailed alignment of saccharum and sorghum chromosomes: comparative organization of closely related diploid and polyploid genomes. *Genetics* 150: 1663–1682.
- Li S-W, Yang H, Liu Y-F, Liao Q-R, Du J, et al. (2012) Transcriptome and gene expression analysis of the rice leaf folder, *Cnaphalocrossis medinalis*. *PLoS One* 7: e47401.
- Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, et al. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* 133: 523–536.
- Trick M, Long Y, Meng J, Bancroft I (2009) Single nucleotide polymorphism (SNP) discovery in the polyploid Brassica napus using Solexa transcriptome sequencing. *Plant Biotechnol J* 7: 334–346.
- Lu T, Lu G, Fan D, Zhu C, Li W, et al. (2010) Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. *Genome Res* 20: 1238–1249.
- Hansey CN, Vaillancourt B, Selkhor RS, de Leon N, Kaepler SM, et al. (2012) Maize (*Zea mays* L.) genome diversity as revealed by RNA-sequencing. *PLoS One* 7: e33071.
- Marguerat S, Bahler J (2010) RNA-seq: from technology to biology. *Cell Mol Life Sci* 67: 569–579.
- Mozzova O, Marra Ma (2008) Applications of next-generation sequencing technologies in functional genomics. *Genomics* 92: 255–264.
- Varshney RK, Nayak SN, May GD, Jackson Sa (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends Biotechnol* 27: 522–530.
- Novaes E, Drost DR, Farmerie WG, Pappas GJ, Grattapaglia D, et al. (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* 9: 312.
- Barbazuak WB, Emrich SJ, Chen HD, Li L, Schnable PS (2007) SNP discovery via 454 transcriptome sequencing. *Plant J* 51: 910–918.
- Gang R, Patel RK, Tyagi AK, Jain M (2011) De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification. *DNA Res* 18: 53–63.
- Carson DI, Botha FC (2000) Preliminary Analysis of Expressed Sequence Tags for Sugarcane. *Crop Sci* 40: 1769–1779.
- Carson D, Botha F (2002) Genes expressed in sugarcane maturing internodal tissue. *Plant Cell Rep* 20: 1075–1081.
- Vettore AL, Silva FR, Kemper EL, Arruda P (2001) The libraries that made SUCEST. *Genet Mol Biol* 24: 1–7.
- Vettore AL, da Silva FR, Kemper EL, Souza GM, da Silva AM, et al. (2003) Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. *Genome Res* 13: 2725–2735.
- Casu RE, Grof CPL, Rae AI, McIntyre CL, Dimmock CM, et al. (2003) Identification of a novel sugar transporter homologue strongly expressed in maturing stem vascular tissues of sugarcane by expressed sequence tag and microarray analysis. *Plant Mol Biol* 52: 371–386.
- Casu RE, Dimmock CM, Chapman SC, Grof CPL, McIntyre CL, et al. (2004) Identification of differentially expressed transcripts from maturing stem of sugarcane by in silico analysis of stem expressed sequence tags and gene expression profiling. *Plant Mol Biol* 54: 503–517.
- Bower NI, Casu RE, Maclean DJ, Rewerter A, Chapman SC, et al. (2005) Transcriptional response of sugarcane roots to methyl jasmonate. *Plant Sci* 168: 761–772.
- Ma H-M, Schulze S, Lee S, Yang M, Mirkov E, et al. (2004) An EST survey of the sugarcane transcriptome. *Theor Appl Genet* 108: 851–863.
- Vicentini R, Ben LEV., Shyu Ma., Nogueira FTS, Vincentz M (2012) Gene Content Analysis of Sugarcane Public ESTs Reveals Thousands of Missing Coding-Genes and an Unexpected Pool of Grasses Conserved ncRNAs. *Trop Plant Biol* 5: 199–205.
- Mancini MC, Leite DC, Perecin D, Bidóia MaP, Xavier Ma., et al. (2012) Characterization of the Genetic Variability of a Sugarcane Commercial Cross Through Yield Components and Quality Parameters. *Sugar Tech* 14: 119–125.
- Landell MGA, Campana MP, Figueiredo P, Vasconcelos ACM, Xavier MA, Bidóia MaP, Prado H, Silva MA, Miranda ILLD AC (2005) Variedades de cana-de-açúcar para o centro sul do Brasil. *Technical Bulletin IAC* 197: 33.
- Beldi N, Macedo I (1995) Quinta geração de variedades de cana-de-açúcar. COOPERATIVA DOS PRODUTORES DE CANA, AÇÚCAR E ALCOOL DO ESTADO DE SÃO PAULO. *Technical Bulletin*: 16–23.
- Sabino J (1997) Sexta geração de variedades de cana-de-açúcar. COOPERATIVA DE PRODUTORES DE CANA, AÇÚCAR E ALCOOL DO ESTADO DE SÃO PAULO LIDA. *Technical Bulletin*: 1.
- Hoffmann H (2008) Variedades RB de cana-de-açúcar. *CGA/UFSCar Technical Bulletin* 1: 30.
- McCormick AJ, Cramer MD, Watt DA (2006) Sink strength regulates photosynthesis in sugarcane. *New Phytol* 171: 759–770.
- Kistner C, Matamoros M (2005) RNA ISOLATION USING PHASE EXTRACTION AND L I C L. In: Márquez A, editor. *Lotus japonicus Handbook*. Dordrecht, The Netherlands, pp. 123–124.
- Patel RK, Jain M (2012) NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One* 7: e30619.
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29: 644–652.
- Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10: R25–R25.
- Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12: 323.
- Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, et al. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21: 3674–3676.
- Florea I, Hartzell G, Zhang Z, Rubin GM, Miller W (1998) A Computer Program for Aligning a cDNA Sequence with a Genomic DNA Sequence. *Genome Res* 8: 967–974.
- Iseli C, Jongeneel CV, Bucher P (1999) ESTScan: A Program for Detecting, Evaluating, and Reconstructing Potential Coding Regions in EST Sequences. *ISMB-99 Proceedings*. AAAI Press, pp. 138–148.
- Lorenz R, Bernhart SH, Höner Zu Siederdisen C, Tafer H, Flamm C, et al. (2011) ViennaRNA Package 2.0. *Algorithms Mol Biol* 6: 26.
- Clote P, Ferré F, Kravakis E, Krizanc D (2005) Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA* 11: 578–591.

40. Domingues DS, Cruz G MQ, Metcalfe CJ, Nogueira FTS, Vicentini R, et al. (2012) Analysis of plant LTR-retrotransposons at the fine-scale family level reveals individual molecular patterns. *BMC Genomics* 13: 137.
41. Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. *Genomics (q-bioGN); Quant Methods*: 1–9.
42. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078–2079.
43. Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, et al. (2012) Using *Drosophila melanogaster* as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. *Front Genet* 3: 35.
44. Li D, Deng Z, Qin B, Liu X, Men Z (2012) De novo assembly and characterization of bark transcriptome using Illumina sequencing and development of EST-SSR markers in rubber tree (*Hevea brasiliensis* Muell. Arg.). *BMC Genomics* 13: 192.
45. Liu M, Qiao G, Jiang J, Yang H, Xie L, et al. (2012) Transcriptome sequencing and de novo analysis for Ma bamboo (*Dendrocalamus latiflorus* Munro) using the Illumina platform. *PLoS One* 7: e46766.
46. Liu S, Li W, Wu Y, Chen C, Lei J (2013) De Novo Transcriptome Assembly in Chili Pepper (*Capsicum frutescens*) to Identify Genes Involved in the Biosynthesis of Capsaicinoids. *PLoS One* 8: e48156.
47. Vincentz M, Cara FAA, Okura VK, da Silva FR, Pedrosa GI, et al. (2004) Evaluation of monocot and eudicot divergence using the sugarcane transcriptome. *Plant Physiol* 134: 951–959.
48. Grivet L, Hout AD, Dufour P, Hamon P, Roquest D (1994) Comparative genome mapping of sugar cane with other species within the Andropogoneae tribe. *Hereditas* 73: 500–508.
49. Dekkers JCM, Hospital F (2002) The use of molecular genetics in the improvement of agricultural populations. *Nat Rev Genet* 3: 22–32.
50. Dangrois JH, Grivet L, Roques D, Hoarau JY, Lombard H, et al. (1996) A putative major gene for rust resistance linked with a RFLP marker in sugarcane cultivar 'R570'. *Theor Appl Genet* 92: 1059–1064.
51. Tai PYP, Miller JD, Dean JL (1981) INHERITANCE OF RESISTANCE TO RUST IN SUGARCANE. *F Crop Res* 4: 261–268.
52. Hogarth DM, Ryan CC, Taylor PWJ (1993) Quantitative inheritance of rust resistance in sugarcane. *F Crop Res* 34: 187–193.
53. Irvine JE (1975) Relations of Photosynthetic Rates and Leaf and Canopy Characters to Sugarcane Yield. *Crop Sci* 15: 671.
54. Moore PH, Botha F, Furbank R, Grof CP (1996) Intensive sugarcane production: Meeting the challenge beyond 2000. Keating BA and Wilson JR, editor Oxon, UK: CAB International. p544.
55. Henry R, Kole C (2010) Genetics, Genomics and Breeding of Sugarcane. 1st ed. Henry, R. J., Kole C, editor Science Publishers. p300.
56. Guo X, Gao L, Liao Q, Xiao H, Ma X, et al. (2013) Long non-coding RNAs function annotation: a global prediction method based on bi-colored networks. *Nucleic Acids Res* 41: e35.
57. Hangauer MJ, Vaughn IW, McManus MT (2013) Pervasive Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic Noncoding RNAs. *PLoS Genet* 9: e1003569.
58. Derrien T, Johnson R, Bussotti G, Tanzer A, Djehali S, et al. (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* 22: 1775–1789.
59. Liu J, Jung C, Xu J, Wang H, Deng S, et al. (2012) Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in *Arabidopsis*. *Plant Cell* 24: 4333–4345.
60. Sun J, Zhou M, Mao Z-T, Hao D-P, Wang Z-Z, et al. (2013) Systematic analysis of genomic organization and structure of long non-coding RNAs in the human genome. *FEBS Lett* 587: 976–982.
61. Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, et al. (2013) Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. *PLoS Genet* 9: e1003470.
62. Pinto IR, Oliveira KM, Ulian EC, Garcia AAF, de Souza AP (2004) Survey in the sugarcane expressed sequence tag database (SUCEST) for simple sequence repeats. *Genome* 47: 795–804.
63. Ramu P, Kassahun B, Senthilvel S, Ashok Kumar C, Jayashree B, et al. (2009) Exploiting rice-sorghum synteny for targeted development of EST-SSRs to enrich the sorghum genetic linkage map. *Theor Appl Genet* 119: 1193–1204.
64. Cordeiro GM, Casu R, McIntyre GI, Manners JM, Henry RJ (2001) Microsatellite markers from sugarcane (*Saccharum* spp.) ESTs cross transferable to *erianthus* and *sorghum*. *Plant Sci* 160: 1115–1125.
65. Metzgar D, Bytof J, Wills C (2000) Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Res* 10: 72–80.
66. Marconi TG, Costa EA, Miranda HR, Mancini MC, Cardoso-Silva CB, et al. (2011) Functional markers for gene mapping and genetic diversity studies in sugarcane. *BMC Res Notes* 4: 264.
67. Felus FA, Wan J, Schulze SR, Estill JC, Jiang N, et al. (2004) An SNP resource for rice genetics and breeding based on subspecies *indica* and *japonica* genome alignments. *Genome Res* 14: 1812–1819.
68. Wakeley J (1996) The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. *Tree* 11: 158–162.
69. Borevitz JO, Chory J (2004) Genomics tools for QTL analysis and gene discovery. *Curr Opin Plant Biol* 7: 132–136.



## OPEN

## SUBJECT AREAS:

GENOME

POLYPLOIDY

PLANT GENETICS

GENETICS

# SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids

Received  
15 May 2013

Accepted  
15 November 2013

Published  
2 December 2013

Correspondence and requests for materials should be addressed to A.P.S. (anete@unicamp.br)

\*These authors contributed equally to this work.

Antonio A. F. Garcia<sup>1\*</sup>, Marcelo Mollinari<sup>1\*</sup>, Thiago G. Marconi<sup>1,2</sup>, Oliver R. Serang<sup>3</sup>, Renato R. Silva<sup>1</sup>, Maria L. C. Vieira<sup>1</sup>, Renato Vicentini<sup>2</sup>, Estela A. Costa<sup>2</sup>, Melina C. Mancini<sup>2</sup>, Melissa O. S. Garcia<sup>2</sup>, Maria M. Pastina<sup>1</sup>, Rodrigo Gazaffi<sup>1</sup>, Eliana R. F. Martins<sup>4</sup>, Nair Dahmer<sup>4</sup>, Danilo A. Sforça<sup>2</sup>, Claudio B. C. Silva<sup>2</sup>, Peter Bundock<sup>5</sup>, Robert J. Henry<sup>6</sup>, Glaucia M. Souza<sup>7</sup>, Marie-Anne van Sluys<sup>8</sup>, Marcos G. A. Landell<sup>9</sup>, Monalisa S. Carneiro<sup>10</sup>, Michel A. G. Vincentz<sup>2,4</sup>, Luciana R. Pinto<sup>9</sup>, Roland Vencovsky<sup>1</sup> & Anete P. Souza<sup>2,4</sup>

<sup>1</sup>Departamento de Genética, Escola Superior de Agricultura “Luiz de Queiroz” Universidade de São Paulo, Brazil, <sup>2</sup>Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas, SP, Brazil, <sup>3</sup>Department of Neurobiology, Harvard Medical School and Proteomics Center, Children’s Hospital Boston, USA, <sup>4</sup>Departamento de Biologia Vegetal, Instituto de Biologia, Universidade Estadual de Campinas, Campinas, SP, Brazil, <sup>5</sup>Southern Cross University, Lismore, Australia, <sup>6</sup>Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, Brisbane, Australia, <sup>7</sup>Instituto de Química, Universidade de São Paulo, São Paulo, SP, Brazil, <sup>8</sup>Departamento de Botânica, Instituto de Biociências, Universidade de São Paulo, São Paulo, SP, Brazil, <sup>9</sup>Centro de Cana, Instituto Agronômico de Campinas, Ribeirão Preto, SP, Brazil, <sup>10</sup>Departamento de Biotecnologia de Plantas, Centro de Ciências Agrárias, Universidade Federal de São Carlos, São Carlos, SP, Brazil.

Many plant species of great economic value (e.g., potato, wheat, cotton, and sugarcane) are polyploids. Despite the essential roles of autopolyploid plants in human activities, our genetic understanding of these species is still poor. Recent progress in instrumentation and biochemical manipulation has led to the accumulation of an incredible amount of genomic data. In this study, we demonstrate for the first time a successful genetic analysis in a highly polyploid genome (sugarcane) by the quantitative analysis of single-nucleotide polymorphism (SNP) allelic dosage and the application of a new data analysis framework. This study provides a better understanding of autopolyploid genomic structure and is a sound basis for genetic studies. The proposed methods can be employed to analyse the genome of any autopolyploid and will permit the future development of high-quality genetic maps to assist in the assembly of reference genome sequences for polyploid species.

Common marker systems, such as Amplified Fragment Length Polymorphism (AFLP) and Simple Sequence Repeat (SSR), have been successfully used in the last few decades for several types of genetic studies, including diversity analysis, genetic mapping, quantitative trait locus (QTL) mapping, synteny (co-linearity) definition, co-ancestry estimation, and more. However, most of these applications have been developed in diploid plant species in which the theoretical foundation for analysis and interpretation of the results has already been established. These tools are less developed for autopolyploids, i.e., organisms that have more than two sets of chromosomes of the same type and origin<sup>1</sup>. Despite the fact that great progress has been made using marker systems in autotetraploids (e.g., potato), other, more complex polyploid species, such as sugarcane, strawberry, and some forage crops, have not yet fully benefited from molecular marker information.

This is because several unrealistic and simplified assumptions need to be made. AFLP and SSR (and even RFLP) do not allow a straightforward estimation of the number of copies of each allele (dosage) at a given polymorphic locus in complex polyploids (species with more than four chromosomes per homology group). For example, in sugarcane, there are approximately 22 linkage maps<sup>2</sup>, and only a few of these maps include loci with high allelic doses. The scenario is similar for QTL studies<sup>3</sup>. Some models have attempted to consider the effects of QTL dosage<sup>4,5</sup>, but these models still rely on marker data that are not fully informative. In *Saccharomyces cerevisiae*,

microarray studies have demonstrated that gene expression and gene regulation may depend upon the ploidy level<sup>6</sup>, emphasising that allele dosage should be included in marker-assisted selection or genome-wide association studies. Similar conclusions were reached by<sup>7</sup>.

The development of modern genotyping technologies, such as the Sequenom iPLEX MassARRAY<sup>®</sup>, Illumina GoldenGate<sup>™</sup>, and protocols such as Genotyping by Sequencing (GBS<sup>10</sup>) or RAD seq<sup>11</sup>, allows the evaluation of single-nucleotide polymorphisms (SNP) throughout the genome. One interesting feature of these novel approaches is the possibility of evaluating the relative abundance of each allele, i.e., the allelic dosage. This significantly increases the information embodied in each locus and provides several advantages for genetic analysis, such as mapping mutants via quantitative bulked segregant analysis<sup>12</sup> and the possibility of estimating ploidy level for polyploids<sup>13,14</sup>. For complex polyploids, such as sugarcane, this is essential because each marker locus needs to be positioned in a homology group. What is remarkable is that the sugarcane homology groups have different numbers of chromosomes<sup>15</sup>. This makes the estimation of ploidy level for each SNP an essential step for further analysis. Furthermore, less studied polyploid species with unknown ploidy levels could directly benefit from this modern marker approach.

To illustrate one of the advantages of using SNPs for these purposes, let us assume a hypothetical population of an autohexaploid species having the following genotypes for a given locus: *aaaaaa*, *Aaaaaa*, *AAaaaa*, *AAAaaa*, and so on up to *AAAAAA*. Using the *A* allele as reference, these individuals are said to have between zero (nulliplex) and six copies (hexaplex) of the allele. The number of copies of the reference allele is the allele dosage. If the individuals are evaluated with a marker system, such as AFLPs or SSRs, they are scored as 0 (gel band absent) for *aaaaaa* or 1 (gel band present) for all the other individuals due to one intrinsic limitation of the method that is associated with overlooking ploidy level. Thus, a result of “1” in a binary marker system indicates the presence of at least one copy of allele *A*. However, if SNPs are evaluated, the scores will be *0A : 6a*, *1A : 5a*, *2A : 4a*, and so on up to *6A : 0a* (this allelic dosage notation will be used throughout this manuscript). A marker system that allows for the direct observation of all genotypes is therefore much more informative and should be preferred. Nevertheless, this raises new challenges because new statistical methods must be developed to allow for the comprehensive analysis and interpretation of data in this new scenario.

In this work, we have evaluated the use of SNPs and novel statistical methods for SNP calling and ploidy level estimation in sugarcane using mass spectrometry-based procedures and the SuperMASSA software<sup>13,14</sup>. We demonstrate that it is possible to estimate the ploidy level and the dosage of SNPs, providing useful insights into the sugarcane genome interpretation. Sugarcane is an excellent test case because it is a complex polyploid with an unknown ploidy level and frequent aneuploidy<sup>15</sup>. This work will make studies on linkage and QTL mapping, association mapping, and genomic selection possible by bringing the advantages of molecular markers to complex polyploids that, with the exception of a few well-studied autotetraploids (such as potato), have poorly understood genomes. We explored two different scenarios. First, 271 SNPs generated using the Sequenom iPLEX MassARRAY technology<sup>8</sup> were used to analyse a population of 180 individuals from a biparental cross between the varieties IACSP95-3018 and IACSP93-3046. Second, 1034 SNPs were analysed in a panel of 142 relevant sugarcane genotypes. The panel consisted of important commercial varieties in addition to ancestral and parental genotypes that have been frequently used in a wide spectrum of breeding programs.

## Results

Figure 1, panels A.1, B.1, and C.1 show examples of scatter plots of genotypes in the segregating population for a selected SNP (*SugSNP382*). It is clear that there are three clusters of points, each

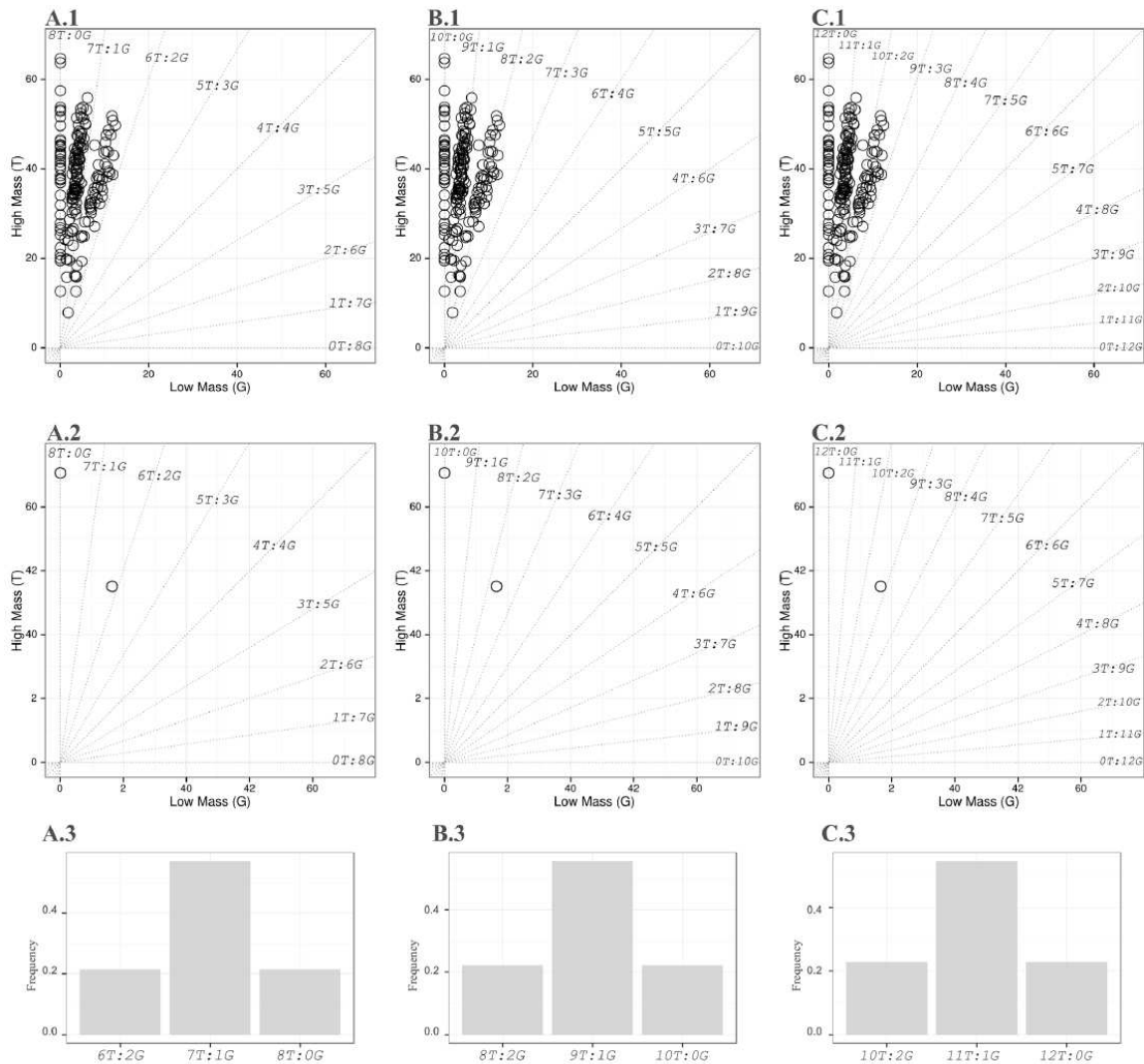
corresponding to one genotype. The data are shown together with dotted lines indicating the expected angles where the individuals would be placed if the ploidy level were 8, 10, and 12. The results suggest that the ploidy level was 10 because the clouds of points deviated slightly from the lines to other ploidy levels. When observing the data from the parents (Figure 1, A.2, B.2, C.2) and considering the closest distance to the expected genotype, the deduced configurations should be  $8T : 0G \times 6T : 2G$ ,  $10T : 0G \times 8T : 2G$ , and  $12T : 0G \times 9T : 3G$ . We must note that to be consistent with the number of observed clusters (three), the expected genotype distributions in the population were set to assume that the locus had a double dosage in one parent and was nulliplex in the other. The deduced value for ploidy 12 was not consistent with the putative number of observed clusters (three) in the progeny because a triple-dosage locus would allow for four clusters in the progeny. The expected population ratios (Figure 1, A.3, B.3, C.3) were slightly different for each ploidy level, with 3 : 8 : 3, 2 : 5 : 2, and 5 : 12 : 5 values for octa-, deca-, and dodecaploidy, respectively. It must be emphasised that it would be extremely difficult to distinguish these levels only by inspection or even by a simple statistical test with reasonable sample sizes.

The results described above help to explain the complex scenarios involved in determining ploidy and dosage. These issues have recently been analysed using the statistical procedures included in the SuperMASSA software<sup>14</sup>. The model simultaneously considers all available information and the genetic constraints that the derived results must fulfil, i.e., the possible genotypes to be observed given the ploidy level and the parental genotypes, the ratio between allele intensities, and the expected complete polysomic segregations. This allowed the exclusion of a triple dosage for ploidy 12. Because the expected segregations are similar, the classification relies on the ratio of the alleles (indicated by dotted lines on Figure 1), and this is one of the reasons why the choice of a technology with less bias for ratios is essential. These issues have been thoroughly discussed in<sup>14</sup>. Those authors analysed how to address situations where some bias is present. In our previous experience with Sequenom and Illumina data<sup>13</sup>, we observed that the former experimental approach is much less likely to produce an allele ratio bias.

We present a deeper analysis of SNPs using SuperMASSA<sup>14</sup> in Figure 2, where the statistical results for three selected SNPs are depicted. For *SugSNP382* (described previously in Figure 1), the results indicate that the *posterior* probability of ploidy 10 is close to 1; all individuals were allocated to clusters with individual *posterior* probabilities no smaller than 0.6 (almost all these probability values were close to 0.9). There was also a good agreement between the observed and expected distribution of the genotypes in the biparental population. In addition, we can deduce that the parental genotypes must have been  $8T : 2G \times 10T : 0G$ . The preliminary visual inspection of the scatter plot described in Figure 1 is consistent with our statistical results.

For *SugSNP151* and *SugSNP715*, the other two SNPs shown in Figure 2, the analysis is more complicated. Although it was possible to find models with high *posterior* probability for ploidy levels 18 and 16, the individual *posterior* probabilities in both cases were all smaller than 0.6. This means that if a small naive *posterior* threshold of 0.65 were used, none of the individuals would be classified as having a specific genotype. This clearly shows that, as reported previously<sup>14</sup>, the *posterior* probability cannot be used as a single criterion to interpret the results. There were also differences between the observed and expected distributions. Although this result may not be considered reliable enough to interpret the available laboratory data, the most likely configuration for these SNPs is ploidy 18 and 16, with parental genotypes of  $15G : 3A \times 12G : 6A$  and  $10T : 6C \times 7T : 9C$ , respectively.

The estimates of ploidy level for the 249 SNPs evaluated in the biparental population fell between 2 and 20 (Figure 3a). An examination of three loci classified as having a ploidy of 2 (*SugSNP\_0004*, *SugSNP\_0033* and *SugSNP\_0036*) showed that these results are

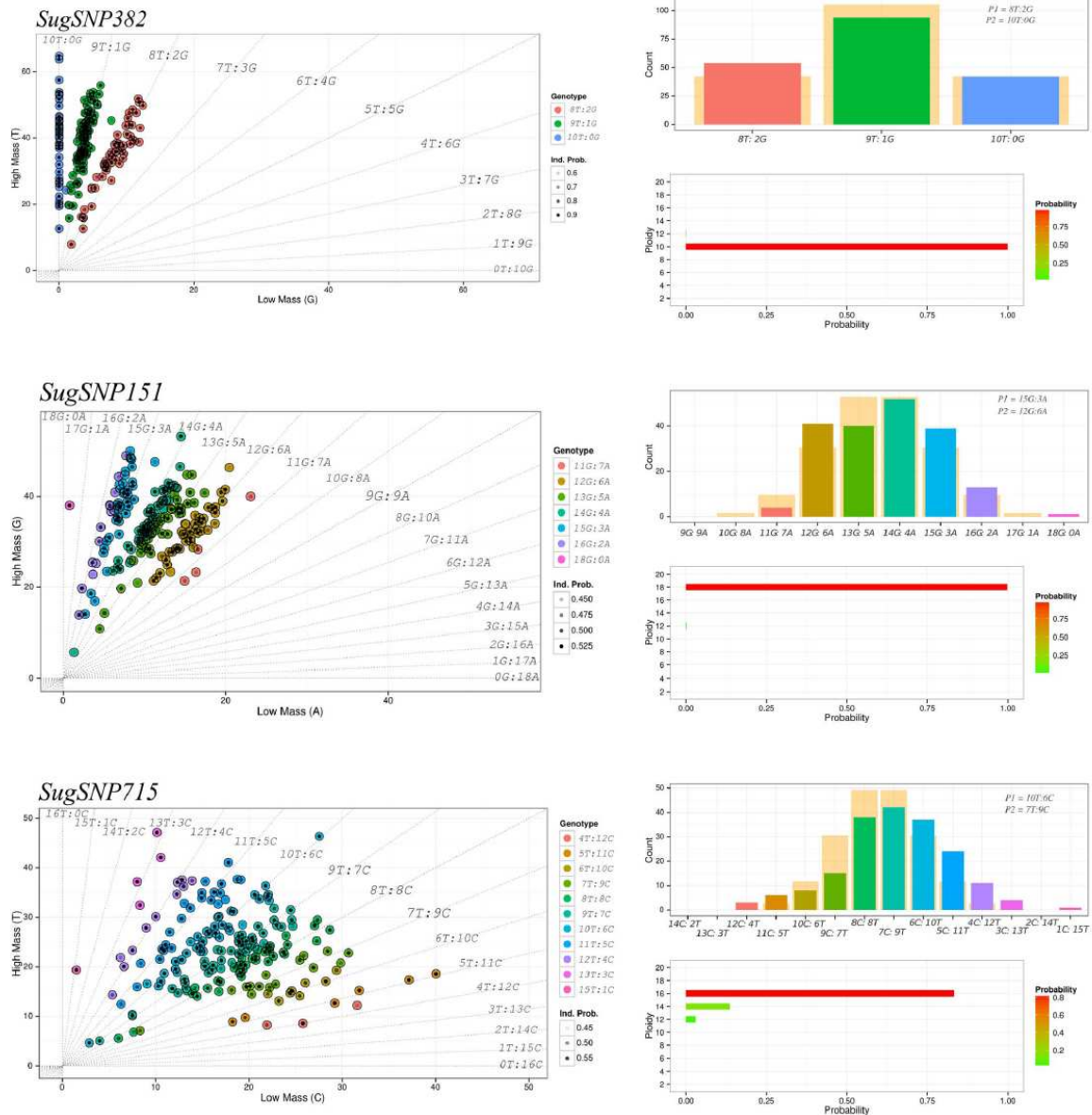


**Figure 1** | A panel of nine graphs, with three columns (A, B, and C) representing ploidy levels of 8, 10, and 12, respectively. Row 1: Raw data of two allele intensities for *Sug*SNP382 in the biparental segregating population. Dotted lines show the possible genotypes and the allele ratios that could be observed for each corresponding ploidy level. Row 2: allele intensities for the parents of the population (the average of 12 replicates), also considering the respective ploidy level. Row 3: expected segregations for the respective ploidy level and assuming parents with genotypes  $8T:0G \times 6T:2G$ ,  $10T:0G \times 8T:2G$ ,  $12T:0G \times 10T:2G$ ; these genotypes were chosen based on a visual inspection of rows 1 and 2.

clearly associated with data of poor quality. The ratios between the masses of these alleles did not follow any expected pattern and were quite different from what was observed for all other SNPs. Therefore, these SNPs were not included in the final presentation of the results (Figure 3); for the same reason, five loci with ploidy 4 were also discarded (*Sug*SNP\_0011, *Sug*SNP\_0017, *Sug*SNP\_0018, *Sug*SNP\_0048 and *Sug*SNP\_0083); note that another two SNPs with ploidy 4 (*Sug*SNP\_0008 and *Sug*SNP\_0061) are presented in Figure 3; both had a single allelic dosage in one of the parents.

The procedure to develop the SNPs must not, in principle, exclude or favour any homology group. In our analysis, only 2 out of 249 loci were classified as having a ploidy of 4 and a single dosage, but there are no reports of such ploidy levels in the sugarcane literature. We must conclude that it is unlikely that sugarcane has homology groups

with four chromosomes (autotetraploid). One possible explanation is that the observed results were caused by some bias in the angles of the scatter plots. If the PCR amplification has a different efficiency for each chromosome, the ratio between the allele intensities may be slightly different from the real ratio and therefore the angles of lines in the scatter plots could be biased by these differences (please see the additional simulations examining this bias in the Supplementary Material). As explained for Figure 1, for small dosages, the differences in the expected segregations are virtually indistinguishable and rely heavily on the scattered plot angle estimation; therefore, if this bias was present, some loci may have been misclassified as autotetraploids. We applied the same reasoning when analysing the association mapping panel; consequently, loci with an estimated ploidy of 2 or 4 were not included in the final results (Figure 3b), and of the 987



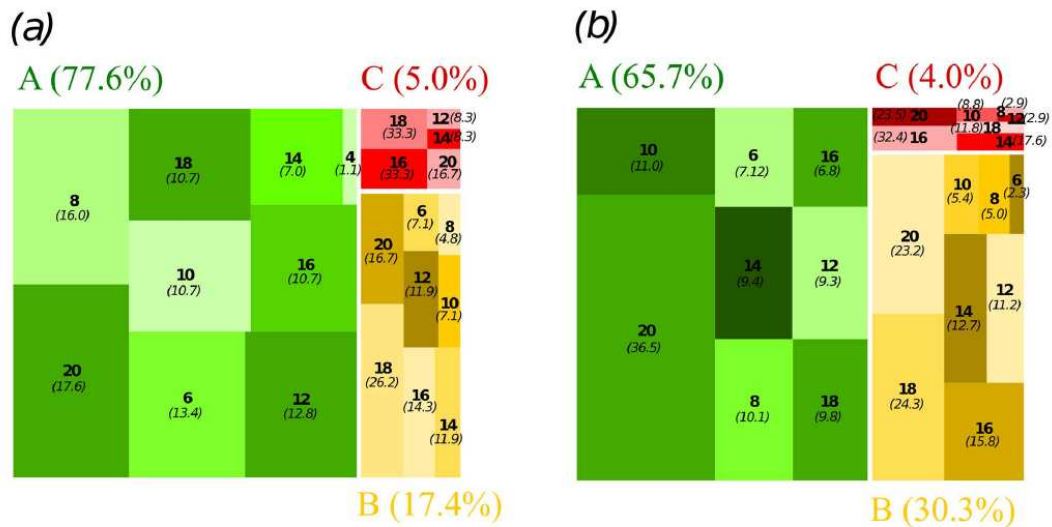
**Figure 2 |** Results of statistical analysis for three selected SNPs in a biparental sugarcane population. Each panel of three graphs correlates to one SNP. The scatter plots show the classification of each individual in a cluster (genotype), indicated with a different colour; the centre of each circle has a small grey dot, whose colour intensity indicates its posterior probability of being allocated in the cluster. Expected (in yellow) and observed distributions for the estimated ploidy level and dosage on the parents are indicated on the histograms; the same colours used on the scatter plots were considered for the observed distribution. The *posterior* probabilities for each ploidy evaluated in the range 2 to 20 (only even numbers) are also indicated.

SNPs that were initially available (after quality control), 855 were taken into account. For all other ploidy levels, the number of loci within each ploidy class suggests that our results are reliable. The ploidy levels fell between 6 and 20, showing that the number of chromosomes within the homologous groups is not constant in sugarcane, which is in agreement with previous results<sup>15</sup>.

The distribution of loci within each ploidy level and category (A, B, and C) was similar for both the biparental population (Figure 3a) and the panel of sugarcane genotypes (Figure 3b), with the exception of

those loci with ploidy 20, which were more frequent in the panel. All of the category A ploidy levels seemed to be present in about the same proportions (except ploidy 4, which was likely to be a misclassification) in both scenarios (Figures 3a and 3b). For category B, there was a trend of having more loci with higher ploidy levels; this was even clearer for loci of category C, particularly for the biparental population (Figure 3a), where none of the loci had a ploidy level smaller than 12.

It is important to mention that the analysis of the 142 sugarcane genotypes within the panel (Figure 3a) was much more complicated



**Figure 3** | Representation of the estimates of ploidy level (in bold font) for the configurations with highest posterior probabilities for the biparental population (a) and association mapping panel (b). The areas of the rectangles are proportional to the number of SNPs that have the same ploidy level, indicated within each rectangle in parenthesis. According to the posterior probabilities calculated for each even-numbered ploidy level in the range 2 to 20, each SNP was classified into one category, using the following *ad hoc* criteria: Category A (green), when the highest posterior probability is greater than or equal to 0.80; Category B (yellow), when no single value of the posterior probability is higher than 0.80 but the sum of the two highest ones is greater than or equal to 0.80; and Category C (red): all other cases. In parentheses: the number of SNPs within the given ploidy level and category. The total SNP number for (a) was 241, and the total SNP number for (b) was 855.

because there were no parental genotypes available to guide the analysis, as there were in the biparental population. For this group, we assumed Hardy-Weinberg equilibrium and that all individuals had the same ploidy level for a given locus. Given the complexity of the sugarcane genome, this assumption may seem rather strong, but the final genetic results are consistent with a ploidy level distribution similar to that of the biparental population. Again, we observed that the number of chromosomes within homologous groups was not constant.

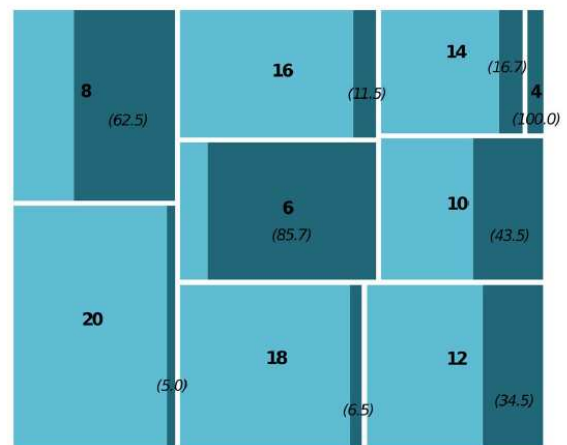
The number of single-dose loci for SNPs in categories A and B (Figure 4) indicates that these are more frequent for ploidy levels up to 12. At the ploidy level of 20, only 5% of the SNPs were single dose. It is remarkable that so few SNPs overall had single-dose alleles. Interestingly, if these SNPs were used to build a linkage map using the conventional approach (single-dose markers), only loci classified as single-dose loci in IACSP93-3046 and as nulliplex in IACSP95-3018 (or vice-versa) or those classified as single-dose loci in both parents would be considered. Only 76 (30.5%) SNPs would meet these criteria if all ploidy levels were considered altogether.

The results presented in Figures 3 and 4 are interesting and informative, but because they are based only on the *posterior* probability of a ploidy level, they need to be interpreted together with individual probabilities<sup>14</sup>. Figure 5 shows that the analysis of ploidy levels 6, 8, and 10 was more reliable, as most of the loci had medians for the individual *posterior* probabilities in the range 0.80 to 1.00. The opposite was observed for ploidy levels 18 and 20, as almost all of the loci (both in the biparental population and the panel) had medians in the range 0.40 to 0.60. Most of the individual medians at ploidy levels 12 and 14 were between 0.60 and 0.80, whereas the individual means at ploidy level 16 was evenly distributed in the ranges of 0.60 to 0.80 and 0.40 to 0.60.

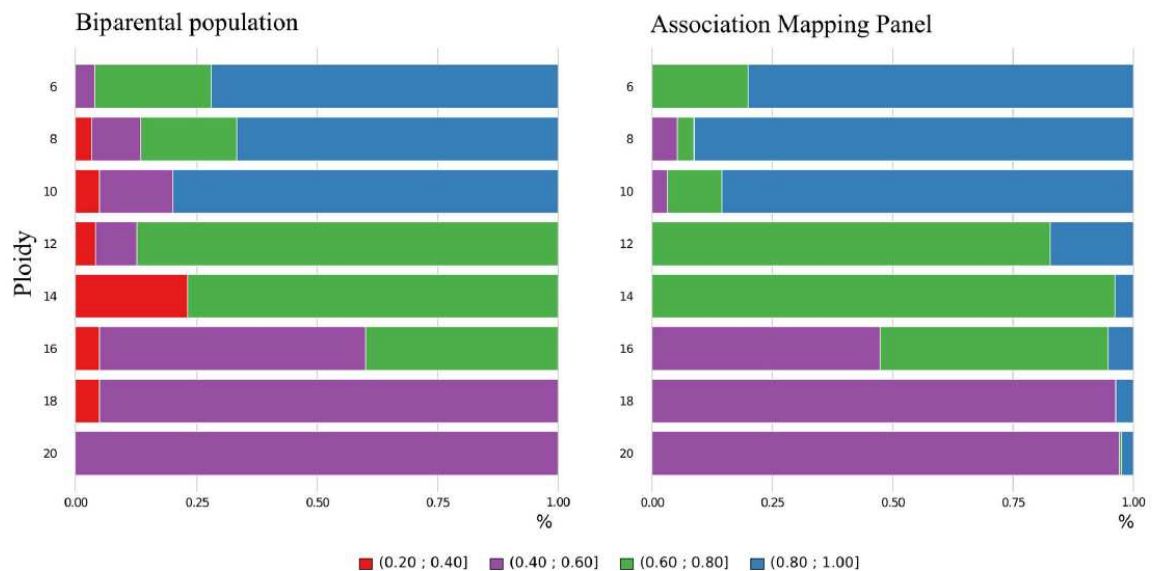
## Discussion

Developing a consistent and self-contained analysis depends on being able to estimate ploidy in species that have complicated genome structures, such as sugarcane. Due to its particular domestication

process that involves the unequal participation of the parental species' genomes (*Saccharum officinarum* and *S. spontaneum*, known to have high chromosome numbers), cytogenetic studies may not be reliable under some circumstances. The approach used in this study combined mass spectrometry and the computer program SuperMASSA<sup>13</sup>



**Figure 4** | Proportion of loci with a single dose in the biparental population. Loci were classified as single dosage when they had a SNP with only a single copy of one of the alleles in one parent, being a nulliplex in the other (thus segregating in a 1 : 1 fashion in the progeny), or when both parents had a single copy of the same allele (segregating in a 3 : 1 ratio). The areas of the rectangles are proportional to the number of SNPs of each ploidy level, indicated in bold font. SNPs with single doses are represented in dark blue, with the proportion of the respective ploidy in parenthesis. Only SNPs within categories A and B (see Figure 3) were considered.



**Figure 5** | Distribution of SuperMASSA individual posterior probabilities. For each locus, the median of all individual posterior probabilities was calculated. For instance, a median of 0.80 indicated that 50% of the individual posterior probabilities were greater than 0.80. The graphs show the distribution of the medians of each SNP locus that were classified with a specific ploidy. Only loci of category A (see Figure 3) were considered in this analysis.

and accomplished this task by simultaneously considering parental information, the number of clusters, the intensities associated with the different alleles, the expected frequencies of individuals in each cluster, and experimental error. During the analysis, each individual was assigned to a single cluster (genotype calling) with a high degree of confidence for several loci (Figures 3 and 5). Thus, we were able to use this technique to suggest a model to explain the complex genome structure of sugarcane.

The primary advantage of the approach used by SuperMASSA is that it makes use of the distribution of alleles in the population in addition to the relative intensities of each allele. Using both types of information is important for resolving cases in which similar relative allele intensities could be produced. For instance, tetraploid and octoploid individuals can both produce relative allele intensities of 0:4, 1:3, 2:4, 3:1, and 4:0; however, if no distinct clusters of individuals with relative intensities near 1:7, 3:5, 5:3, or 7:1 are observed, then it is highly unlikely that the population is octoploid. No two octoploid parents (or, if no parental data are considered, no Hardy-Weinberg allele frequency) can be expected to produce alternating observed and absent genotype classes. Because several ploidies can potentially produce clusters with similar relative allele dosages, exploiting population information is critical when inferring the ploidy level. In sugarcane, this advancement is particularly useful because the exact allele dosage of a locus is frequently unknown. Furthermore, the availability of parental data adds further constraints and increases the accuracy of ploidy estimation.

Most genetic studies of sugarcane have considered only simplex markers<sup>5</sup>, and our results show that the actual portion of the genome explored to date is rather small. Our observations are quite different from previously published results. For example, one study reported that 80% of the AFLP markers in a biparental population occurred at a single dose<sup>16</sup>. This is similar to the values we found for ploidy levels from 6 to 12 for loci with category A, but not for the overall genome, suggesting that the strategy for finding single-dose loci may involve a biased genome sampling. Those authors considered only markers that segregated in only one parent, but there is no biological reason

to support this approach because both parents can have different alleles segregating in the population<sup>17,18</sup>. However, it is important to note that AFLP analysis does not allow the identification of all genotypic classes in a segregating population because all clusters that have at least one copy of the allele will collapse into a single cluster (i.e., a dominant action). This also suggests that the identification of single-dose loci using AFLP is strongly biased.

What can be said about the sugarcane ploidy level? Our results suggest that the most likely ploidy levels are between 6 and 14 (Figure 5), and several lines of evidence support our findings. The genetic maps that have already been published using different sugarcane population types (e.g., biparental crosses, selfings, and others) all have recognised homo(eo)logy groups; interestingly, most homo(eo)logy groups were established with particular numbers of co-segregation groups, which also supports the mixed-ploidy nature of the sugarcane genome, consistent with the results presented here. Our estimates for ploidies 6–14 showed high (or intermediate) individual *posterior* probabilities. Furthermore, the proportion of loci with single dosages for these ploidy levels in the biparental population (Figure 4) is in agreement with previous reported results (e.g.<sup>16</sup>), with the exception of ploidy 6. The proportion of loci with ploidy levels between 6 and 14 was approximately the same for loci within category A, both in the biparental population and in the genotype panel (Figure 3). This was expected because sugarcane chromosomes are approximately the same size and the markers were in principle chosen to evenly cover the genome. There is also biological evidence to support these findings; ploidies 6 to 14 are found in the group of species that contribute to the generation of modern cultivars of sugarcane. *S. officinarum* is the domesticated sugar-producing species that is directly derived from *S. robustum*, which encompasses clones with  $2n = 60$  or  $2n = 80$ . Both species are autopolyploids, and their basic chromosome number is  $x = 10$ , meaning that *S. robustum* has 6 or 8 copies of each chromosome, depending on the genotype analysed<sup>19–21</sup>. A total of 13.4% of the SNPs used to genotype the biparental population and 7.12% of the SNPs used in the panel in this study have their level of ploidy classified as 6.



We speculate that this class of SNPs belongs to the subgenome (or haplotype) of *S. robustum* that persists in the sugarcane genotypes after breeding. The vast majority of *S. officinarum* clones display  $2n = 80$  chromosomes. The species is stated to have eight sets (or copies) of 10 chromosomes ( $x = 10$ ), i.e., octoploid.

Currently, it is supposed that modern sugarcane cultivars could exhibit  $2n$  (*S. officinarum*) +  $n$  (*S. spontaneum*) constitution; when hybrids with *S. spontaneum* are produced, the chromosomes of *S. officinarum* double their number and form pairs of homologues, and those of *S. spontaneum* pair among themselves. This point was considered in classical publications<sup>22,23</sup>. Subsequent *in situ* hybridisation-based studies have confirmed the basic chromosome numbers ( $x$ ) in the genus *Saccharum*<sup>24</sup> and suggested that the genomes of modern hybrids are composed of 10–20% *S. spontaneum* chromosomes, 5–17% recombinant chromosomes and the remainder composed of *S. officinarum* chromosomes<sup>25,26</sup>. Therefore, one would expect to find 8 as the most frequently estimated ploidy level, all derived from *S. officinarum*. This particular value was found in 26.7% of SNPs classified in Category A (considering only ploidies 6–14) in the biparental population (Figure 3a) and 10.1% SNPs used in the panel of genotypes and belonging to category A (Figure 3b). A possible explanation is that almost all genotypes analysed here were commercial varieties (mainly interspecific hybrids) with a modified chromosomal composition from the ancestors as a result of domestication.

For *S. spontaneum*, which displays a wide range of chromosome numbers (from  $2n = 40$  to  $2n = 128$ ), a basic chromosome number of  $x = 8$  was suggested. The five major cytotypes with  $2n = 64, 80, 96, 112,$  and  $128$ <sup>27</sup> have 8, 10, 12, 14 and 16 sets (or copies) of eight chromosomes, respectively. These are consistent with the values observed in this study. We may suppose that all these SNP-containing loci were inherited from *S. spontaneum* (maybe as haplotypes) or that they are located on the chromosomes that were identified as recombinants between the two species *S. officinarum* and *S. spontaneum*. Alternatively, when looking at ploidy level 8, all chromosomes could be inherited only from *S. officinarum*. It is also important to mention that the repeated cycles of backcrosses to *S. officinarum* applied by early breeders, combined with the double transmission phenomenon<sup>22,23</sup>, could result in high ploidy levels because the contribution of the recurrent parent will be prevalent.

Chromosomal rearrangements are reported to be a rapid response to the formation of allopolyploid genomes<sup>28</sup>; intergenomic translocations occur predominantly between homo(eo)logous chromosomes<sup>29</sup>, and homo(eo)logous shuffling and chromosome compensation maintain genome balance in re-synthesised allopolyploids<sup>30</sup>. All the rearrangements may have occurred in the early evolutionary process of modern sugarcane. Supposedly, there is a most regular ploidy level, and all variations represent chromosome rearrangements that were herein observed.

The observation of 18 or 20 copies of a SNP-containing locus does not mean that this extreme figure represents the actual ploidy level. One could suggest reasonable cytological explanations for these high numbers; for example, for at least some of these loci, we may be detecting polysomic loci as a consequence of chromosomal segment copy number due to chromosomal rearrangements. On the other hand, the presence of univalents as a result of intergenomic pairing is well documented in sugarcane varieties. One should assume that bivalent pairing is not random but rather involves the same homo(eo)logous chromosomes<sup>31</sup>; therefore, two (or more) copies of the same univalent can be inherited from ascendants and pair during meiosis. The detection of certain high-copy SNP-containing loci may be a consequence of additional non-homologous pairing. However, it is important to mention that high values of ploidy were associated with some loci that did not have a reliable classification in our study. They were also more frequent in the panel, which is more difficult to analyse. Loci with ploidy 16 fall between these two scenarios (ploidy 6–14 and 18–20).

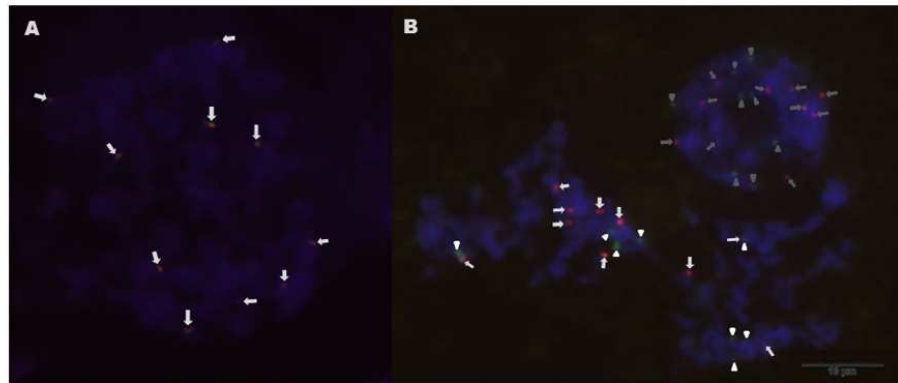
A recent review of the quantitative genotyping of polyploids<sup>14</sup> reported that, even when data are difficult to analyse (i.e., presenting high variance or strong allele-specific bias), the SuperMASSA software can still provide useful information to help to interpret the results and allow the evaluation of the reliability of those results. In that review, the authors evaluated the *posterior* probabilities of extremely high ploidies (in the range 2 to 100). It is obvious that most of these ploidies do not have biological support, but the study revealed that when the locus displays a high variance, the generated model tends to attribute a cluster to each point in the diffuse cloud, resulting in a very high estimate of ploidy level. We have not tried to adjust our models with ploidy levels above 20 due to computing-time limitations, but we have deliberately included ploidy values without biological support (2 and 4) or with weak evidence (18 and 20). The results show that this was a good strategy because the resulting individual *posterior* probabilities were rather small, indicating that our observations of high ploidy values (18 and 20) are likely to be explained as discussed above<sup>18</sup>.

We have also performed some simulations to better understand the SuperMASSA output under normal and extreme situations (Figures S1 to S10, Supplementary Material). We observed that the software performed well when no extreme violations of its underlying assumptions were considered (for example, skew on the expected angles of clusters and segregation distortion). However, in the presence of high segregation distortion (for example, due to preferential pairing at meiosis) or some bias in the allele ratios, the estimated ploidy could be rather high (18 or 20).

The *in situ* hybridisations also helped us interpret the SuperMASSA estimations (Figure 6). The number of observed blocks (or signals) in these hybridisations could be taken as a rough estimate of ploidy level for IACSP 93-3046 (P2 of the biparental population). For *SugSNP382*, which yielded good and reliable results in the ploidy analysis (Figure 3), the number of observed blocks has been 8, which is close enough to the estimate of 10 provided by SuperMASSA. It is important to mention that SuperMASSA uses segregation ratios as an important feature to estimate ploidy; this is not necessarily the same as estimating ploidy by chromosome counting. For example, a homo(eo)logy group could have 10 chromosomes: 6 from *S. officinarum* and 4 from *S. spontaneum*. If there is preferential pairing at meiosis and the polymorphism is present in the genome of *S. officinarum*, the locus will behave like a hexaploid in the segregating population; in contrast, the results from cytological studies revealed a ploidy of 10. For *SugSNP715* and *SugSNP151*, the number of blocks was 10 for both. This is clear evidence that, as previously explained, high ploidy estimates (16 and 18) combined with small individual *posterior* probabilities are likely to be statistical artefacts. Moreover, *SugSNP382* yielded the same estimate for the ploidy level in the biparental population and the panel of genotypes, which was not the case for the other two SNPs analysed in Figure 3.

In conclusion, the results derived from the two different scenarios presented here (a biparental population and a panel of genotypes) provide extremely useful insights. First, as expected, it is clear that the sugarcane genome is complex and that the number of chromosomes in each homo(eo)logy group varies depending on the SNP-containing locus. Second, our results agree with previous sugarcane cytogenetic data<sup>25</sup> and demonstrate the robustness of analysing SNP markers in autopolyploid species. Third, the ploidy level of each SNP locus was also estimated; it must be emphasised that this estimation cannot be performed with common marker systems.

In the light of our results, the ploidy of sugarcane commercial varieties (interspecific hybrids) was estimated to be in the range from 6 to 14 for each homo(eo)logy group; this has biological and statistical support. Several factors may explain the observation of estimates in the 16–20 ploidy range, a) they are actual results; b) they were caused by a combination of preferential pairing at meiosis and a lack of bivalent pairing or segregation distortion; c) there are intrinsic



**Figure 6** | *In situ* hybridisation of IACSP93-3046 chromosomes. (A) *SugSNP715* (10 blocks, arrows); (B) *SugSNP151* (10 blocks; grey arrow, nucleus; white arrow, metaphase nucleus) and *SugSNP382* (8 blocks; grey arrowhead, interphase nucleus; white arrowhead, metaphase nucleus).

difficulties in analysing loci with high ploidy and allelic dosage; or d) MassARRAY technology did not perform well for some loci, causing bias in the allele ratio and/or high variance for clusters. The results reported in the literature and our own *in situ* hybridisations for the three selected SNPs suggest that reason (a) is very unlikely. However, if these high estimates were actual results, further linkage studies will show that these loci with high ploidy will show evidence of linkage with other loci of the same ploidy level and, also, will not be linked with the ones in the ploidy range 6–14. It is important to mention that linkage studies based on genetic maps will require the development of new statistical approaches, such as the ones presented by<sup>32</sup> and<sup>33</sup> for autotetraploids, that would not be straightforward to use for our results. Current ideas that put strong emphasis on single-dose loci are not appropriate. Concerning point (b), this argument may be verified by further cytological information, which will help us understand the meiotic behaviour of this complex species and subsequently make modifications to the underlying assumptions in the statistical model. For explanations (c) and (d), specific procedures should be developed to optimise the methodology for dealing with complex polyploids. It is reasonable to assume that if most of these high ploidy values are true, these loci will co-segregate and will not be linked with loci with small ploidy; this will result in homo(eo)logous groups for the corresponding ploidy level. It is important to perform linkage studies in the biparental population to determine if loci with high/unknown ploidy are co-segregating with others showing high *posterior* probability for ploidy level; then the ploidy of these loci could be indirectly inferred.

None of the other currently available approaches are suitable to investigate polyploid genome structures as comprehensively as this approach. Therefore, we anticipate that the shaping of polyploid genomes by evolutionary processes will be better understood by applying this SNP genotyping method. Considering that most of the angiosperms are polyploid<sup>34</sup> and recent sequenced genomes also suggest a polyploid ancestry for eukaryotes<sup>1</sup>, significant scientific breakthroughs can be achieved using this novel approach.

We strongly believe that the results presented herein will lead to new possibilities for the study of complicated autopolyploid species not only in terms of new genetic understanding, statistical genetic modelling, and prediction capabilities but also in terms of understanding the biological aspects of evolutionary and domestication processes. Finally, it is interesting to note that our study unveiled the genomic structure of a complex polyploid species by exploiting the simplest manifestation of genetic variation, the SNP. This approach should provide an important tool for developing high-quality genetic maps that will assist in QTL mapping and the assembly of reference genome sequences for the large proportion of plant

plants species that are polyploid or have duplicated chromosomal regions.

## Methods

**Molecular and cytological analysis.** Two representative scenarios were considered: a) a progeny of 180 individuals from a sugarcane F1 biparental population derived from the cross between two commercial varieties, IACSP 95-3018 (female, named P1 along the text) x IAC93-3046 (male, named P2); and b) an association mapping panel with 142 relevant sugarcane genotypes (Table 1), representing commercial varieties and important ancestors of modern cultivars. Sugarcane genomic DNA was obtained from young leaves using standard techniques. A total of 1034 sugarcane SNPs were developed; 91 were derived from previously reported sequence data<sup>35</sup> (Table S1), and the remaining 943 were developed from 2908 cluster sequences with differential expression<sup>36</sup> that were selected from the SUCEST database<sup>37</sup> (Table S2). SNPs were discovered using QualitySNP software<sup>38</sup> with minor modifications, and primers were designed using the MassARRAY Assay Design package. All 1034 sugarcane SNPs (Tables S1 and S2) were genotyped in the association mapping panel (iPLEX GOLD chemistry, Sequenom Inc., San Diego/CA, USA) (Table 1), and 271 SNPs from these (SUCEST database, Table S1) were evaluated in the progeny of the biparental population. Due to data quality control (especially due to very low signal), the data from 22 and 47 SNPs were discarded from the biparental population and from the panel of genotypes, respectively. Therefore, for the statistical analysis, 249 and 987 SNPs were used in the biparental population and in the panel, respectively. The SNP genotyping method was based on MALDI-TOF analysis performed on a mass spectrometer platform from Sequenom Inc.<sup>39</sup> Both parents from the biparental population were scored 12 times for each SNP.

The SNP assay is based on the single-base extension of locus-specific primers followed by mass spectrometry to detect polymorphisms, yielding allele-specific information<sup>39</sup>. Assuming equal ionisation efficiency for all alleles, equal PCR amplification of alternate alleles, and equal nucleotide incorporation accuracy/equilibria, the mass intensities should be proportional to the abundances of each allele.

Three selected SNPs (*SugSNP382*, *SugSNP151*, and *SugSNP715*) were analysed with FISH to check their hybridisation with IACSP93-3046. Leaf genomic DNA was isolated using the DNeasy Plant Mini Kit (Qiagen) and amplified using a *Pfu* DNA Polymerase kit (Thermo Scientific) and specific primers (Table S3). The fragments of DNA were cloned using *Escherichia coli* DH10b as host and pGEM-T Vector Systems (Promega) as vector. Colonies containing recombinant plasmids were identified for selection on LB agar medium supplemented with X-gal and IPTG. Recombinant plasmids were isolated using the alkaline miniprep procedure, and the insert nucleotide sequences were determined with an ABI3500 automated DNA sequencer (Applied Biosystems). DNA Sequences were analysed with Lasergene 7 (DNASTar, Madison, WI, USA) and aligned by using the ClustalW option of the MegAlign program. The clones were used to amplify the probes for FISH using *Taq* DNA polymerase (Invitrogen) and purified using Wizard<sup>®</sup> SV Gel and PCR Clean-Up System (Promega). Chromosome preparations were made from root tips collected from culms grown in a plastic box containing filter paper with the regular application of water. Cytological preparations were carried out as previously described<sup>40</sup>. All probes were labeled by nick translation (Invitrogen). *SugSNP715* and *SugSNP151* were labeled with digoxigenin-11-dUTP (Life Technologies) and detected with Anti-DIG-rhodamine; *SugSNP382* was labelled with Biotin-14dATP (Roche) and detected with avidin-FITC. The procedure and conditions for FISH were previously described<sup>40</sup>.

**Statistical analysis.** The output of Sequenom iPLEX MassARRAY technology is a scatter plot  $D$  with quantitative alleles intensities for individuals  $i = 1, 2, \dots, n$ <sup>41,42</sup>. Because each SNP was bi-allelic, two intensities are presented,  $x_i$  and  $y_i$ , usually

**Table 1 | Genotypes from the panel of 142 sugarcane varieties (panel of genotypes)**

Badila	IAC87-3396	RB735275	RB92579
CB36-24	IAC91-1099	RB739359	RB935744
CB40-13	IACSP93-2060	RB739735	RB965902
CB41-76	IACSP93-3046	RB75126	RB965917
CB46-47	IACSP95-3018	RB765418	RB966928
CB47-355	IACSP95-3028	RB785148	SP70-1005
CB53-98	IACSP95-5000	RB815690	SP70-1078
Chunnee	IACSP98-3022	RB825317	SP70-1143
Co290	IN84-58	RB825336	SP70-1284
Co331	L60-14	RB83102	SP70-1423
Co419	Maneria	RB835019	SP70-3370
Co449	NA56-79	RB835054	SP71-1406
Co740	NC0310	RB835089	SP71-6163
Co997	PO88-62	RB835205	SP71-6949
CP51-22	Q165	RB835486	SP71-799
CP52-68	R570	RB845197	SP72-4928
CP70-1547	RB1	RB845210	SP77-5181
CTC15	RB2	RB845257	SP79-1011
CTC2	RB3	RB855002	SP79-2233
CTC9	RB4	RB855035	SP79-2312
EK28	RB5	RB855036	SP79-2313
F31-962	RB6	RB855077	SP79-6134
F36-819	RB7	RB855113	SP79-6192
Ganda Cheni	RB8	RB855156	SP80-1520
H53-3989	RB9	RB855206	SP80-1816
H59-1966	RB10	RB855350	SP80-1836
IAC48-65	RB11	RB855453	SP80-1842
IAC49-131	RB12	RB855463	SP80-3280
IAC50-134	RB721012	RB855511	SP81-3250
IAC51-205	RB72199	RB855536	SP83-2847
IAC52-150	RB72454	RB855563	SP83-5073
IAC64-257	RB725053	RB855595	SP89-1115
IAC82-2045	RB725828	RB867515	SP91-1049
IAC82-3092	RB732577	RB925211	TUC71-7
IAC83-4157	RB735200	RB925268	
IAC86-2210	RB735220	RB925345	

represented in bi-dimensional scatter plots (see Figure 1 for an example of a loci with alleles T and G). For data quality, all data points with small intensities for both alleles were removed; they were located within a circular area on the scatter plots defined by the radius  $(0.10)\min\{x_i, y_i\}$ , centred on the origin of both axes.

All loci were then classified using the statistical method implemented in the SuperMASSA software<sup>13</sup>. A comprehensive review of this method is presented in<sup>14</sup>. In short, rather than iteratively clustering the samples and then predicting the genotype of each cluster, a graphical Bayesian method was used. The model can be described in two parts. First, a Gaussian model based on the relative dosage is used to model the probability that an individual with a known genotype will produce certain intensities for each allele; ideally, the relative intensities would be proportional to the relative dosages of the respective alleles. Second, a multinomial distribution is used to model the probability that a given set of genotypes will occur given the population structure. The population structure is general and can be used to analyse the biparental population (F1 model) and the association mapping panel (Hardy-Weinberg model). For any type of population model, the hidden parameters (i.e., the allele frequency for the Hardy-Weinberg model and the parental genotypes in the F1 model) can be estimated with maximum likelihood. Similarly, the ploidy can be predicted by estimating the genotypes and population parameters for each ploidy level and then selecting the ploidy that yields the highest likelihood. In the case of the F1 model, additional data were provided by the parents, which were scored with 12 replicates; these data can help restrict the set of reasonable parents and ploidies. The primary contribution of this method is that it makes use of the distribution of alleles in the population and the relative intensity of each allele. The use of both types of data is important for resolving cases that could produce similar relative allele intensities.

Following the recommendation reported in<sup>14</sup>, to find the *maximum a posteriori* (MAP) solution for the estimates of the parameters in the model, all even-numbered ploidy levels in the range of 2 to 20 were tested. The SuperMASSA *naive posterior report threshold* was set to 0, and the values of individual *posterior* probability (which indicates the maximum threshold that will allow the individual to be assigned to a given genotype) were also calculated. For example, if two individuals have *posterior* probabilities 0.55 and 0.65 and the *naive posterior report threshold* is set to 0, both of them will be assigned to genotypes; changing the threshold to 0.60, only the latter will be included; with a threshold of 0.90, both will be excluded. This was shown to be important when interpreting the results of the SNP calling.

- Comai, L. The advantages and disadvantages of being polyploid. *Nature Rev. Genet.* 6, 836–846 (2005).
- Alwala, S. & Kimbeng, C. A. *Genetics, Genomics and Breeding of Sugarcane*. Henry, R. J. & Kole, C. (ed.), 69–96 (CRC Press, 2010).
- Pastina, M. M., Pinto, L. R., Oliveira, K. M., Souza, A. P. & Garcia, A. A. F. *Genetics, Genomics and Breeding of Sugarcane*. Henry, R. J. & Kole, C. (ed.) 117–148 (CRC Press, 2010).
- Cao, D., Craig, B. A. & Doerge, R. W. A model selection-based interval-mapping method for autopolyploids. *Genetics* 169, 2371–2382 (2005).
- Doerge, R. W. & Craig, B. A. Model selection for quantitative trait locus analysis in polyploids. *Proc. Natl. Acad. Sci. USA* 97, 7951–7956 (2000).
- Galitski, T., Saldanha, A. J., Styles, C. A., Lander, E. S. & Fink, G. R. Ploidy regulation of gene expression. *Science* 285, 251–254 (1999).
- Osborn, T. C. *et al.* Understanding mechanisms of novel gene expression in polyploids. *TIG* 19, 141–147 (2003).
- Gabriel, S., Ziaugra, L. & Tabbaa, D. SNP genotyping using the Sequenom MassARRAY iPLEX platform. *Curr. Protoc. Hum. Genet.* 60, 2–12 (2009).
- Akhunov, E., Nicolet, C. & Dvorak, J. Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay. *Theor. Appl. Genet.* 119, 507–17 (2009).
- Elshire, R. J. *et al.* A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6, e19379 (2011).
- Baird, N. *et al.* Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* 3, e3376 (2008).
- Liu, *et al.* High-Throughput Genetic Mapping of Mutants via Quantitative Single Nucleotide Polymorphism Typing. *Genetics* 184, 19–26 (2010).
- Serang, O., Mollinari, M. & Garcia, A. A. F. Efficient exact maximum a posteriori computation for bayesian SNP genotyping in polyploids. *PLoS ONE* 7, e30906 (2012).
- Mollinari, M. & Serang, O. Quantitative SNP Genotyping of Polyploids with MassARRAY and Other Platforms. *Methods in Molecular Biology*. Walker, J. M. (ed.) (Humana Press, 2013 - in press).
- Grivet, L. & Arruda, P. Sugarcane genomics: depicting the complex genome of an important tropical crop. *Curr. Opin. Plant Biol.* 5, 122–127 (2001).
- George, A. W. & Aitken, K. A new approach for copy number estimation in polyploids. *J. Hered.* 101, 521–524 (2010).
- Garcia, A. A. F. *et al.* 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 (2006).
- Oliveira, K. M. *et al.* Functional integrated genetic linkage map based on EST-markers for a sugarcane (*Saccharum spp.*) commercial cross. *Mol. Breeding* 20, 189–208 (2007).
- Price, S. Cytology of *Saccharum robustum* and related sympatric species and natural hybrids. *USDA Tech. Bull.* 1337, 1–44 (1965).
- Daniels, J. & Roach, B. T. *Sugarcane Improvement Through Breeding*. Heinz, D.J. (ed.) (Elsevier, Amsterdam, 1987).
- Lu, Y. H., D'Hont, A., Walker, D. I. T. & Rao, P. S. Relationships among ancestral species of sugarcane revealed with RFLP using single copy maize nuclear probes. *Euphytica* 78, 7–18 (1994).
- Brandes, E. W. & Sartoris, G. B. Sugarcane: Its Origin and Improvement. *Yearbook of the United States Department of Agriculture*, 561–623 (USDA, 1936).
- Bremer, G. Problems in breeding and cytology of sugar cane. *Euphytica* 10, 59–78 (1961).
- D'Hont, A., Ison, D., Alix, K., Roux, C. & Glaszmann, J. C. Determination of basic chromosome numbers in the genus *Saccharum* by physical mapping of ribosomal RNA genes. *Genome* 41, 221–225 (1998).
- D'Hont, A. *et al.* Characterisation of the double genome structure of modern sugarcane cultivars (*Saccharum spp.*) by molecular cytogenetics. *Mol. Gen. Genet.* 250, 405–413 (1996).
- D'Hont, A. Unraveling the genome structure of polyploids using FISH and GISH; examples of sugarcane and banana. *Cytogenet. Genome Res.* 109, 27–33 (2005).
- Panje, R. & Babu, C. Studies in *Saccharum spontaneum*. Distribution and geographical association of chromosome numbers. *Cytologia* 25, 152–172 (1960).
- Pontes, O. *et al.* Chromosomal locus rearrangements are a rapid response to formation of the allotetraploid *Arabidopsis suecica* genome. *P. Natl. Acad. Sci. USA* 101, 18240–18245 (2004).
- Chester, M. *et al.* Extensive chromosomal variation in a recently formed natural allopolyploid species, *Tragopogon miscellus* (Asteraceae). *P. Natl. Acad. Sci. USA* 109, 1176–1181 (2012).
- Xiong, Z., Gaeta, R. T. & Pires, J. C. Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid *Brassica napus*. *P. Natl. Acad. Sci. USA* 108, 7908–7913 (2011).
- Pagliarini, M. S., Vieira, M. L. C. & Valle, C. *Meiosis – Molecular Mechanisms and Cytogenetic Diversity*. Swan, A. (ed.), (InTech, 2012).
- Leach, L. J., Wang, L., Kearsey, M. J. & Luo, Z. Multilocus tetrasomic linkage analysis using hidden Markov chain model. *P. Natl. Acad. Sci. USA* 107, 4270–4274 (2010).
- Hackett, C. A., McLean, K. & Bryan, G. J. Linkage Analysis and QTL Mapping Using SNP Dosage Data in a Tetraploid Potato Mapping Population. *PLoS ONE* 8, e63939 (2013).

34. Masterson, J. Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. *Science* **264**, 421–424 (1994).
35. Bundock, P. C. *et al.* Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploidy plant species using 454 sequencing. *Plant Biotech. J.* **7**, 347–354 (2009).
36. Papini-Terzi, F. S. *et al.* Sugarcane genes associated with sucrose content. *BMC Genomics* **10**, 120 (2009).
37. Vettore, A. L. *et al.* Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. *Genome Res.* **13**, 2725–2735 (2003).
38. Tang, J., Vosman, B., Voorrips, R. E., Gerard van der Linden, C. & Leunissen, J. A. M. QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploidy species. *BMC Bioinformatics* **7**, 438 (2006).
39. Moraes, A. P. & Guerra, M. Cytological differentiation between the two subgenomes of the tetraploid *Emilia fosbergii* Nicolson and its relationship with *E. sonchifolia* (L.) DC. Asteraceae. *Plant Syst. Evol.* **287**, 113–118 (2010).
40. Schwarzscher, T. & Helslop-Harrison, J. S. *Practical in situ Hybridization*. (BIOS Scientific Publishers, 2000).

### Acknowledgments

The authors wish to thank Dr Daniel Ugarte for his invaluable suggestions for elaboration of the manuscript. This work was supported by grants from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) - Bioen Program, grant no. 2008/52197-4, and Conselho Nacional do Desenvolvimento Científico e Tecnológico (CNPq) - INCT- Bioetanol and CeProBio Project. CBCS, DAS, EAC, MCM, MM, and MMP received graduate fellowships from FAPESP. TGM and RRS received a fellowship from CNPq, ND, MOSG, and RG

received post-doctoral fellowships from FAPESP. AAFG, APS, ERFM, GMS, MLCV, MAVL, MV, and RVe received research fellowships from CNPq.

### Author contributions

R.Vi. and T.G.M. identified and developed the sugarcane SNPs. A.A.F.G., M.M., M.M.P., O.R.S., R.G. and R.R.S. performed the statistical analysis. M.O.S.G. and T.G.M. carried out the SNP genotyping. L.R.P. and M.G.A.L. carried out the biparental cross. L.R.P., M.G.A.L. and M.S. were responsible for the field trials and the collection of the plant material. E.A.C., C.B.C.S., G.M.S., L.R.P., M.A.V.S., M.C.M., M.S.C., M.V., P.B., R.H. and R.Ve. participated in the molecular genetic studies. E.R.F.M., N.D. and D.A.S. performed *in situ* hybridisation experiments. M.L.C.V. provided the cytogenetic interpretation of the results. A.A.F.G. and A.P.S. conceived the study and participated in its design and coordination. A.A.F.G., M.M. and A.P.S. drafted the manuscript. All authors read and approved the final manuscript.

### Additional information

**Supplementary information** accompanies this paper at <http://www.nature.com/scientificreports>

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Garcia, A.A.F. *et al.* SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. *Sci. Rep.* **3**, 3399; DOI:10.1038/srep03399 (2013).



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported license. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/3.0>

## Anexo V

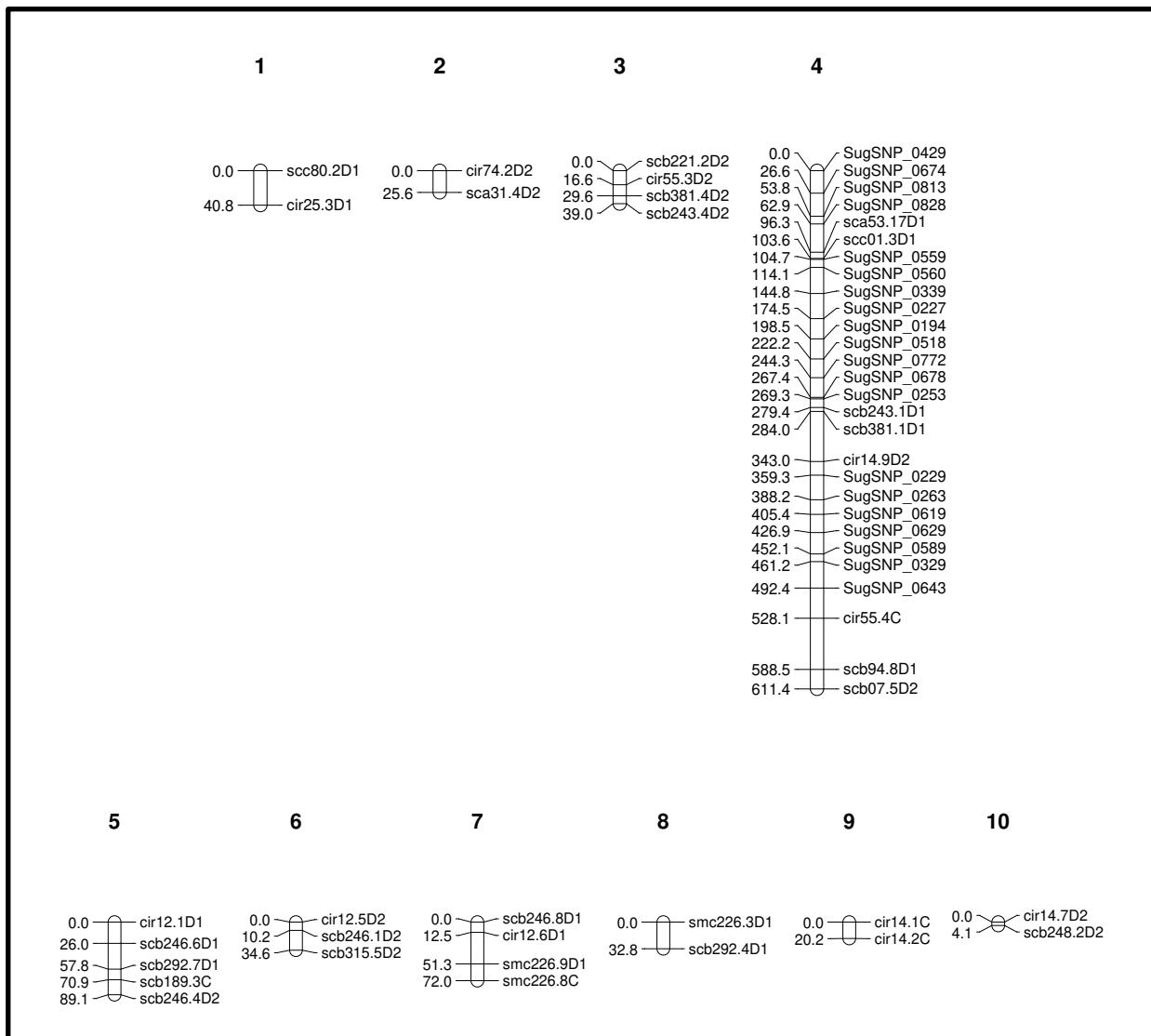


Figura 1. Mapa de ligação de cana-de-açúcar obtido para o cruzamento SP81-3250 e RB925345, utilizando microssatélites e SNPs. Os números superiores referem-se ao grupo de ligação (GL) que pertencem as marcas. Os números à esquerda são as distâncias em centimorgans (cM) entre as marcas. Os nomes das marcas estão à direita, seguidos pela sigla D1 (presente no genitor SP81-3250) D2 (presente no RB925345) ou C (presente em ambos os genitores), todos identificando os microssatélites.

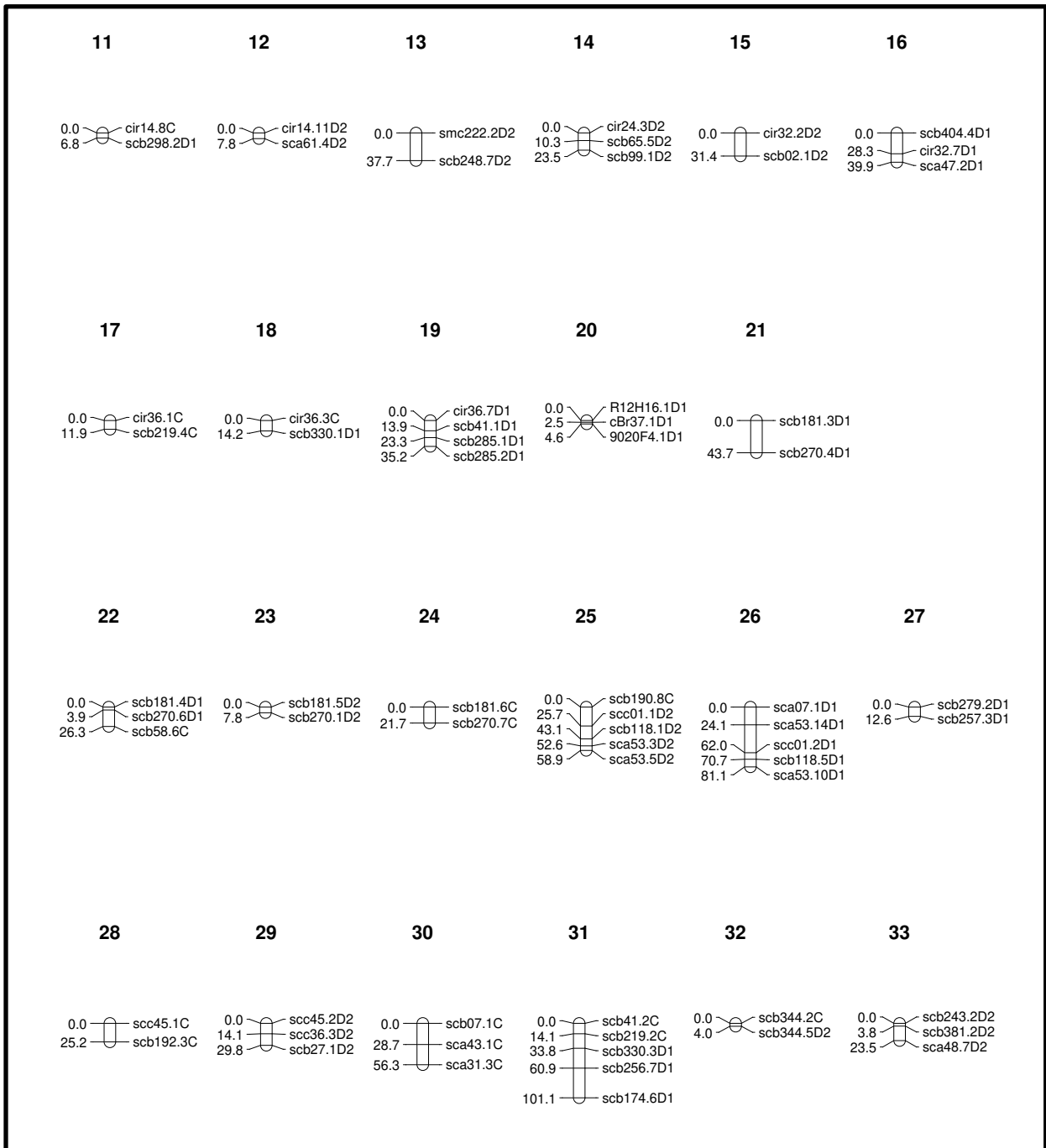


Figura 1. Continuação.

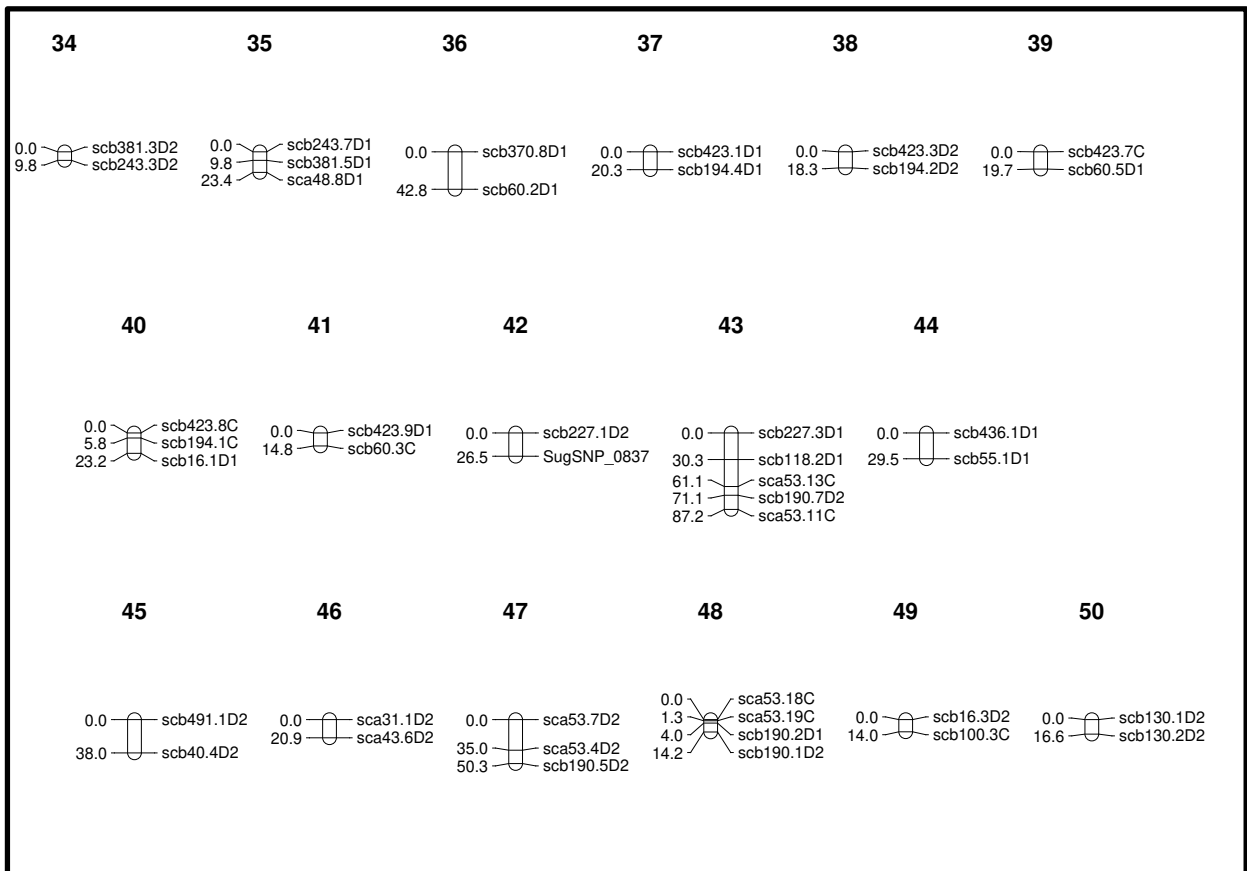


Figura 1. Continuação.





DECLARAÇÃO

Declaro para os devidos fins que o conteúdo de minha dissertação de Mestrado/tese de Doutorado intitulada "**Construção de um mapa funcional e detecção de QTLs de importância econômica em uma população derivada de cruzamento bi-parental entre duas variedades comerciais em cana-de-açúcar**":

( ) não se enquadra no § 4º do Artigo 1º da Informação CCPG 002/13, referente a bioética e biossegurança.

Tem autorização da(s) seguinte(s) Comissão(ões):

(X) CIBio – Comissão Interna de Biossegurança , projeto No. 13/2003 , Instituição: \_\_\_\_\_

( ) CEUA – Comissão de Ética no Uso de Animais , projeto No. \_\_\_\_\_, Instituição: \_\_\_\_\_

( ) CEP - Comissão de Ética em Pesquisa, protocolo No. \_\_\_\_\_, Instituição: \_\_\_\_\_

*\* Caso a Comissão seja externa ao IB/UNICAMP, anexar o comprovante de autorização dada ao trabalho. Se a autorização não tiver sido dada diretamente ao trabalho de tese ou dissertação, deverá ser anexado também um comprovante do vínculo do trabalho do aluno com o que constar no documento de autorização apresentado.*

*Melina Cristina Mancini*

Aluno: Melina Cristina Mancini

*Anete Pereira de Souza*  
Orientador: Anete Pereira de Souza

Para uso da Comissão ou Comitê pertinente:

(X) Deferido ( ) Indeferido

*Ed. Lucia Sartorato*

Profa. Dra. EDI LUCIA SARTORATO  
Lab. Genética Humana  
CBMEG - UNICAMP

Carimbo e assinatura

Para uso da Comissão ou Comitê pertinente:

( ) Deferido ( ) Indeferido

Carimbo e assinatura