Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"

Mapeamento de QTLs em progênie de irmãos-completos de cana-de-açúcar utilizando imputação múltipla

Carina de Oliveira Anoni

Dissertação apresentada para obtenção do título de Mestre em Ciências. Área de concentração: Genética e Melhoramento de Plantas

Piracicaba
2011

Carina de Oliveira Anoni
Engenheira Agrônoma

Mapeamento de QTLs em progênie de irmãos-completos de cana-de-açúcar utilizando imputação múltipla

Orientador:
Prof. Dr. **ANTONIO AUGUSTO FRANCO GARCIA**

Dissertação apresentada para obtenção do título de Mestre em Ciências. Área de concentração: Genética e Melhoramento de Plantas

Piracicaba
2011

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - ESALQ/USP

Anoni, Carina de Oliveira
Mapeamento de QTLs em progênie de irmãos-completos de cana-de-açúcar utilizando imputação múltipla / Carina de Oliveira Anoni. - - Piracicaba, 2011.
100 p. : il.

Dissertação (Mestrado) - - Escola Superior de Agricultura "Luiz de Queiroz", 2011.

1. Cana-de-açúcar 2. Mapeamento genético 3. Marcador molecular 4. Melhoramento genético vegetal 5. Poliploidia I. Título

CDD 633.61
A615m

"Permitida a cópia total ou parcial deste documento, desde que citada a fonte – O autor"

À minha mãe **Evaneide** e ao meu pai **Luiz**, pelo amor incondicional,
por acreditarem em meus sonhos e me ensinarem o valor da vida.
Ao meu irmão **Ricardo**, pelo exemplo de perseverança.
Dedico.

“... Cada um de nós compõe a sua história
cada ser carrega o dom de ser capaz
e ser feliz...”

Renato Teixeira; Almir Sater

Agradecimentos

A DEUS.

Ao Prof. Dr. Antonio Augusto Franco Garcia, pela receptividade inicial, pelos ensinamentos, confiança e amizade, sempre com seu grande incentivo.

Aos meus pais Evaneide e Luiz, por me apoiarem e incentivarem meus sonhos, por me aconchegarem nos momentos que precisei e me ensinarem que com amor e dedicação, alcançamos nossos objetivos. Se cheguei até aqui, foi por eles.

Ao meu irmão Ricardo que mesmo tão distante esteve presente ao longo dessa caminhada, com suas alegrias.

À minha avó Maria, meu exemplo de vida.

Ao Renato Cruz Neves, que com seu amor, paciência, dedicação e imensurável compreensão, me apoiou fortemente ao longo desse trabalho.

Ao Prof. Dr. Luiz Alexandre Peternelli que me apresentou o fantástico mundo científico, pela confiança, pelos ensinamentos e principalmente pela sua amizade. Serei eternamente grata pela sua dedicação que muito contribui em minha caminhada.

Ao Departamento de Genética da ESALQ/USP, pela oportunidade concedida.

Ao CNPq pela bolsa de estudos inicial e à FAPESP, pela bolsa concedida nos demais meses do curso de Mestrado.

À Profª. Drª. Anete Pereira de Souza e à Drª Luciana Rossini Pinto pela parceria, disponibilizando os dados moleculares e fenotípicos. Em especial à Melina C. Mancini, sempre disposta em solucionar minhas dúvidas relacionadas aos dados aqui utilizados.

Ao pessoal do Laboratório de Genética-Estatística: João Ricardo, Marcelo, Gabriel, Maria Marta, Rodrigo, Renato, Edjane, Graciela, Luciano, Guilherme, Adriana e Rodrigo II, pela amizade, companhia, auxílio nas dificuldades e por todos os momentos de descontração vivenciados com alegria. Sem vocês, o mestrado não teria sentido.

Em especial ao Renato, Rodrigo, Maria Marta, Luciano e Marcelo, que muito contribuíram com a execução deste trabalho.

Ao Marcelo, Rodrigo, Maria Marta e Gabriel, que com audácia, desbravaram o LaTeX .

Ao João Ricardo, meu companheiro de disciplinas, muito obrigada pela nossa amizade.

À Graciela, obrigada pela sua constante companhia.

Ao Dr. Marcos Antonio Lopes, por me mostrar desde o primeiro momento, o quanto que a vida pode ser doce.

À toda a minha família, amigos e colegas que estiveram presente em minha vida e contribuiram para meu crescimento pessoal e profissional, em especial ao Luiz, Renata, Bruna, Joana, Leonardo, Natally, Aline, Denise, Aninha, Fernanda, Nathee, Nayara e Melissa, meu muito obrigada.

SUMÁRIO

RESUMO 11
ABSTRACT 13
1 INTRODUÇÃO 15
2 REVISÃO BIBLIOGRÁFICA 17
2.1 Aspectos da Cana-de-Açúcar 17
2.2 Mapeamento Genético 18
2.2.1 Mapeamento Molecular e Mapas Genéticos 18
2.3 Mapeamento de QTLs 21
2.3.1 Métodos para de Mapeamento de QTLs 22
2.3.2 Mapeamento de QTLs em Cana-de-Açúcar 25
2.4 Modelos Para Dados Não Observados 27
2.4.1 Imputação de Dados Não Observados 28
2.4.2 Dados Não Observados no Mapeamento de QTLs 31
2.4.2.1 Probabilidade dos Genótipos Não Observados 32
2.5 Imputação de Dados no Mapeamento de QTLs 34
3 MATERIAL E MÉTODOS 39
3.1 Material 39
3.1.1 Material Vegetal 39
3.1.2 Avaliações Fenotípicas 39
3.1.3 Avaliações Moleculares 40
3.2 Métodos 41
3.2.1 Mapa Genético de Ligação 41
3.2.2 Análise Fenotípica 41
3.2.3 Preditores Genéticos 42
3.2.4 Imputação Múltipla de Pseudomarcadores 44
3.2.5 Modelo de Mapeamento 45
4 RESULTADOS 47
4.1 Mapa Genético de Ligação 47

4.2 Análises Fenotípicas 47
4.3 Mapeamento de QTLs 50
5 DISCUSSÃO 65
6 CONCLUSÃO 73
REFERÊNCIAS 75
APÊNDICE 89

RESUMO

Mapeamento de QTLs em progênie de irmãos completos de cana-de-açúcar utilizando imputação múltipla

O mapeamento de QTLs em cana-de-açúcar (*Saccharum* spp.) é de grande importância para entendimento da arquitetura genética de caracteres quantitativos e aperfeiçoamento dos programas de melhoramento. Entretanto uma situação comum é a ocorrência de genótipos de marcadores não observados, ocasionando viés na estimativa e localização de possíveis QTLs. Portanto, métodos que consideram a informação de genótipos não observados devem ser utilizados. O presente trabalho teve como objetivo detectar QTLs em uma população de irmão completos de cana-de-açúcar utilizando o método de mapeamento por intervalo com a implementação da abordagem da imputação múltipla. A população de mapeamento foi composta por 220 indivíduos oriundos do cruzamento entre os genitores IACSP95-3018 e IACSP93-3046. Os caracteres avaliados foram : percentual de fibra (Fibra), conteúdo de sacarose (POL), produção de cana por parcela (PC) e produção de açúcar por parcela (PP). Foram utilizadas dez imputações para gerar pseudomarcadores a cada 1cM no mapa de ligação que totalizou 4370 cM. Observou-se que ao total, 57 QTLs foram mapeados nos 113 grupos de ligação avaliados, sendo 14 QTLs para o caráter Fibra, 19 para POL, 12 para PC e 12 para PP. O valor de LOD Score e R^2 variaram entre 3,82-7,52 e 6,49%-16,61%, respectivamente. Foi observado que em geral, os efeitos aditivos e de dominância foram significativos, predominando os efeitos aditivos. Além de reduzir o viés ocasionado pelos genótipos não observados, a abordagem da imputação múltipla aumentou o poder de detecção de QTLs, mapeando QTLs com sucesso. Assim, acredita-se que os resultados do presente trabalho, contribuiram para futuros estudos que objetivem a melhor compreensão da arquitetura genética dos caracteres quantitativos da cana-de-açúcar.

Palavras-chave: Mapeamento Genético; Imputação; Genótipos Não Observados; Progênie de irmãos completos

ABSTRACT

QTL mapping in a full-sib family of sugarcane using multiple imputation

QTL mapping in sugarcane (*Saccharum* spp.) is important to understand the genetic architecture of quantitative traits that are important in breeding programs. However, the ocurrence of missing marker phenotypes is common and decrease the power to detect QTL and causes bias in estimates of locations and effects of QTL. Therefore, methods that include missing marker phenotypes should be considered. Our work was aimed at detecting QTL in a full-sib family of sugarcane via interval mapping method using the multiple imputation approach. The mapping population was composed of 220 individuals derived from a biparental cross between IAC95-3018 and IACSP93-3046. The evaluated traits related to yield were: fiber content (Fiber), sugar content (POL), cane yield in kg.plot^{-1} (PC), sugar yield in kg.plot^{-1} (PP). Ten imputation data sets were build using a 1-cM grid to infer pseudomarker genotype along the genetic linkage map. The QTL mapping on the data with imputed pseudomarker genotypes detected 57 QTLs; 14 QTL were obtained for Fiber; 19 for POL; 12 for cane yield and 12 for sugar yied. The LOD Score value and the R^2 proportional to the average weight of all pseudomarker realizations at each grid position ranged from 3.82-7.52 and 6.49%- 16.61%, respectively. In general, it was observed that additive and dominance effects were significant, with predominance of additive effects. The application of multiple imputation approach was successful in reducing the bias and increasing the power of QTL detection. Thus, it is believed that the results of this work contribute to future studies to understand the genetic architecture of quantitative traits in sugarcane.

Keywords: Genetic Mapping; Imputation; Missing Genotypes, Full-sib family

1 INTRODUÇÃO

A cana-de-açúcar (*Saccharum* spp.) é uma cultura de significativa importância econômica e agronômica. Além do açúcar e etanol, que são seus principais subprodutos, gera a partir dos seus resíduos, o etanol de segunda geração convertido em biocombustível, energia elétrica e até mesmo eteno, matéria prima do polietileno para produção de plásticos. O Brasil é o líder mundial em produção (UNICA, 2011). De acordo com o último levantamento da Companhia Nacional de Abastecimento, a produção total em território nacional foi de 641.982 mil toneladas (CONAB, 2011). A cana-de-açúcar apresenta um cenário em expansão, tanto para atender a demanda interna como também aumentar as perspectivas de exportações dos seus subprodutos, principalmente o açúcar e o etanol.

Através dos programas de melhoramento genético é possível a detecção de genótipos superiores, que posteriormente são lançados como variedades contruibuindo assim para o aumento da produção e expansão da cana-de-açúcar. Embora os programas de melhoramento sejam dinâmicos e constantemente aperfeiçoados, o tempo médio necessário para o lançamento de novos cultivares é de 10 a 12 anos (MATSUOKA; GARCIA; ARIZONO, 1999). Otimizar o tempo e os custos dos programas sem que haja comprometimento na avaliação dos materiais genéticos é desejado. Considerando o avanço das técnicas moleculares e que a maioria dos caracteres de importância agronômica e econômica da cana-de-açúcar são de natureza quantitativa, o mapeamento genético de QTLs (*QTL - Quantitative Trait Loci*), é considerado uma ferramenta útil. Além de maior entendimento da arquitetura genética dos caracteres quantitativos, é possível utilizar as informações obtidas com os marcadores moleculares na seleção assistida por marcadores, selecionando os genótipos superiores precocemente e tornando os programas de melhoramento mais eficazes.

Várias metodologias genético-estatísticas estão disponíveis para o mapeamento de QTLs em populações endogâmicas, de retrocruzamentos e F_2 (STUBER; EDWARDS; WENDEL, 1987; LANDER; BOTSTEIN, 1989; ZENG, 1993, 1994; KAO; ZENG, 1997; KAO; ZENG; TEASDALE, 1999; ZENG; KAO; BASTEN, 1999). Para populações heterozigóticas, como é o caso da cana-de-açúcar, poucas estratégias de mapeamento foram desenvolvidas (SONG; SOLLER; GENIZI, 1999; JOHNSON; JANSEN; ARENDONK, 1999; LIN et al., 2003; GAZAFFI, 2009; PASTINA, 2010), dentre elas, análise de marcas individualmente (SM), mapeamento por intervalo (IM) e mapea-

mento por intervalo composto (CIM). Porém, uma ocorrência comum no mapeamento de QTLs é a falta de informação dos dados genotípicos (*missing genotype datas*), pois ao serem realizadas as técnicas de genotipagem dos indivíduos, é difícil encontrarmos marcadores em todas as regiões ao longo do genoma. A ocorrência de marcadores não observados é ocasionada por vários motivos (MARTÍNEZ; CURNOW, 1992). Caso a taxa de marcadores não observados seja alta, o mapeamento de QTLs é afetado e as estimativas obtidas são não confiáveis (ZHANG et al., 2010).

Para contornar a falta de informação do genótipo dos marcadores, algumas abordagem são aplicadas no mapeamento de QTLs. Metodologias genético-estatísticas como estimativa da máxima verossimilhança baseada em um modelo de mistura (BROMAN; SEN, 2009) e regressão de Haley-Knott (HALEY; KNOTT, 1992; MARTÍNEZ; CURNOW, 1992) utilizando aproximações baseadas em quadrados mínimos são utilizadas. Métodos de imputação (SEN; CHURCHILL, 2001) também foram incorporados às análises. Os métodos de imputação visam preencher os genótipos não observados com valores plausíveis baseados nos genótipos observados e após a imputação, os métodos tradicionais para a detecção de possíveis QTLs são aplicados. Deste modo, o mapa genético contendo a informação dos genótipos imputados se torna mais saturado e consequentemente mais informativo (ZHANG et al., 2010) possibilitando então, eventuais melhorias nas estimativas na localização dos QTLs.

A técnica da imputação múltipla (RUBIN, 1987), considera as incertezas associadas às imputações através das variâncias obtidas em cada imputação realizada, apresentando assim, estimativas mais eficientes. Sen e Churchill (2001), sugerem uma abordagem que utiliza a imputação múltipla para estimativa dos genótipos de QTLs não observados. A realização da imputação múltipla de genótipos não observados, tem sido utilizada com sucesso em estudos genéticos (BALL, 2001; SEN; CHURCHILL, 2001; SUN et al., 2006; BROMAN; SEN, 2009). No entanto, as metodologias desenvolvidas até o momento, são aplicadas apenas às populações endogâmicas. A imputação múltipla de genótipos não observados, nunca foi investigada em espécies sem linhagens endogâminas, como é o caso da cana-de-açúcar. Com a implementação da imputação múltipla no mapeamento de QTLs nessas populações, algumas limitações inerentes ao mapeamento seriam contornadas. O presente trabalho teve como objetivo implementar a imputação múltipla de genótipos não observados no mapeamento de QTLs em progênie de irmãos completos em cana-de-açúcar.

2 REVISÃO BIBLIOGRÁFICA

2.1 Aspectos da Cana-de-Açúcar

A cana-de-açúcar (*Saccharum* spp.) é uma das culturas mais importantes no Brasil e no mundo. Além de ser a principal fonte de açúcar e etanol, ocasionando portanto grande importância no setor econômico brasileiro, é uma cultura que apresenta grande demanda por mão de obra. Por apresentar baixo custo de produção, o setor sucroalcooleiro se expandiu nos últimos anos visando atender a demanda de exportação para países desenvolvidos. A produção total de cana-de-açúcar no Brasil na safra 2011/2012 foi em torno de 650 mil toneladas (CONAB, 2011). A crescente demanda por fontes renováveis de energia e necessidade de redução de gases poluentes, faz do etanol, uma importante fonte alternativa de combustível, tanto pelo seu potencial de benefícios econônicos quanto ambientais (UNICA, 2011).

Deseja-se o crescimento da produção da cana-de-açúcar e do etanol, principalmente para atender a demanda interna ocasinada pela produção de veículos com motores *flex fluel* e aumentar as perspectivas de exportação do etanol brasileiro nos próximos anos. Estima-se que após 2020 a área cultivada com cana-de-açúcar deve se estabilizar, no entanto, o volume de produção continuará em crescimento em razão das novas tecnologias (UNICA, 2011). Assim, o melhoramento genético da cana-de-açúcar possibilitaria o aumento da produção por hectare e consequentemente, redução na demanda por novas áreas agrícolas para expansão da cana-de-açúcar no Brasil.

Pertencente à família Poacea e ao gênero *Saccharum*, a cana-de-açúcar é uma gramínea perene amplamente cultivada em áreas tropicais e subtropiais. Acredita-se que atualmente existam seis espécies pertencentes a este gênero: *S. barbari* (2n = 111 - 120), *S. edule* (2n = 60 - 80), *S. officinarum* (2n = 80), *S. robustum* (2n = 60 - 80), *S. sinense* (2n = 80 - 140) e *S. spontaneum* (2n = 40 - 128). O cruzamento das espécies do gênero *Saccharum* com *Erianthus* sect. *Ripidum*, *Miscanthus* sect. *Diandra* Keng, *Sclerostachya* (Hack.) A. Camus e *Narenga* Bor, juntas formam o denomidado “complexo *Saccharum*” (MUKHERJEE, 1957). Porém, técnicas moleculares sugerem que as atuais seis espécies podem ser agrupadas mais precisamente em duas espécies, *S. spontaneum* e *S. officinarum* (IRVINE, 1999). Sendo assim, a cana-de-açúcar apresenta uma natureza híbrida, heterozigótica e com alto nível de ploidia.

Os primeiros trabalhos de melhoramento da cana-de-açúcar baseiam-se no cruzamento entre

Saccharum officinarum (x = 10, 2n = 8x = 80) e outras espécies do gênero *Saccharum* como por exemplo *S. spontaneum* (x = 8, 2n = 5 - 16x = 40 - 128) (IRVINE, 1999). Esses cruzamentos inter-específicos originaram híbridos com elevado grau de ploidia e frequente aneuploidia (HEINZ; TEW, 1987), tornando complexas as etapas dos programas de melhoramento (MATSUOKA; GARCIA; ARIZONO, 1999; SOUZA, 2001; KIDO, 2003).

Os tradicionais métodos de melhoramento genético da cana-de-açúcar consistem na seleção parental de uma população base, seguida do cruzamento dos parentais selecionados, realiza-se então, a seleção e clonagem dos indivíduos selecionados durante vários estágios até que os genótipos superiores sejam lançados como cultivares comerciais. Embora os programas de melhoramento genético sejam dinâmicos e constantemente aperfeiçoados, o tempo médio necessário para o lançamento de novos cultivares é de 10 a 12 anos (MATSUOKA, GARCIA, ARIZONO, 1999). Nas etapas iniciais dos programas de melhoramento da cana-de-açúcar há grande demanda de mão de obra e consideráveis custos na implementação, condução e avaliação do material genético selecionado (HOGARTH et al., 1997; BARBOSA; SILVEIRA, 2000; PETERNELLI; BARBOSA, 2004).

Portanto, técnicas que selecionam genótipos superiores precocemente, tornam os programas de melhoramento mais eficazes e neste contexto, o mapeamento de QTLs é considerado uma ferramenta útil.

2.2 Mapeamento Genético

2.2.1 Marcadores Moleculares e Mapas Genéticos

As características capazes de detectar variações existentes no genoma de diferentes indivíduos são consideradas marcadores genéticos. Esses marcadores são a base de estudos para análises de genomas (AGARWAL; SHRIVASTAVA; PADH, 2008). Os primeiros trabalhos realizados envolvendo marcadores utilizavam as variações morfológicas de características fenotípicas. Posteriormente, houve o desenvolvimento de métodos que aplicavam o uso de marcadores bioquímicos, baseados no polimorfismo entre indivíduos ao nível de proteína, destacando-se o uso de marcadores isoenzimáticos (FERREIRA; GRATTAPAGLIA, 1998). Porém, o uso de marcadores morfológicos, bioquímicos ou isoenzimáticos é limitado. Além de serem influenciados pela ação do ambiente, se apresentam em número reduzido, o que impossibilita a cobertura total do genoma dos indivíduos em estudo (SOUZA, 2001).

Com o desenvolvimento dos marcadores moleculares, essas limitações foram contornadas. As diferentes técnicas envolvendo o uso de marcadores moleculares, possibilitam a detecção do polimorfismo existente ao nível de DNA. Além disso, os marcadores moleculares não apresentam alterações influenciadas pelo ambiente, apresentam mínimos ou nulos efeitos epistáticos e pleitrópicos e quando comparado com os outros tipos de marcadores, contém maior informação genética por loco (FERREIRA; GRATTAPAGLIA, 1998). Dentre os mais utilizados, destacam-se os marcadores RFLP (*Restriction Fragment Lenght Polymorphism*), RAPD (*Random Amplified Polymorphic DNA*), AFLP (*Amplified Fragment Lenght Polymorphism*), SSR (*Simple Sequence Repeaut*), EST (*Expressed Sequence Tags*), EST-SSR, SNP (*Single Nucleotide Polymorphism*) e DArT (*Diversity Array Technology*) (BOTSTEIN et al., 1980; TAUTZ, 1989; WILLIAMS et al., 1990; ZABEAU; VOS, 1993; SAGHAI et al., 1994; SCHLÖTTERER, 2004).

Atualmente uma das mais importantes aplicações dos marcadores moleculares é a construção de mapas genéticos de ligação (BOTSTEIN et al., 1980). Com a estimativa de mapas genéticos é possível a localização de genes, estudos de associação de genes à características quantitativas, estudos comparativos de sintenia genômica, cobertura e análise completa de genomas, além de informações a respeito da arquitetura genética, ligações gênicas completas e pleiotropia (GARCIA et al., 2006) entre outros estudos que são diretamente aplicados aos programas genético de melhoramento (PEREIRA et al., 2009).

As inferências das informações obtidas baseadas nos mapas moleculares devem ser precisas, portanto, corretas estimativas de mapas são necessárias (LEWIN et al., 2009). Os métodos utilizados para ordenação dos marcadores moleculares dentro de mapas genéticos são desenvolvidos em sua maioria, para populações segregantes de espécies diplóides, oriundas do cruzamento entre parentais homozigóticos. Porém, em algumas espécies como o eucalipto, citrus e cana-de-açúcar, a obtenção de linhagens homozigóticas é difícil ou praticamente impossível. Isto porque essas espécies apresentam elevadas taxas de depressão por endogamia ou elevados níveis de ploidia, resultando em alta complexidade do genoma o que impossibilita a obtenção de linhagens homozigóticas. Nestas circunstâncias, o mapeamento baseia-se em populações F_1 de irmãos completos resultantes do cruzamento biparental entre indivíduos heterozigóticos.

No entanto, a contrução de mapas genéticos em populações heterozigóticas é dificultada em razão de alguns fatores. Os métodos estatísticos aplicados às estimativas dos mapas moleculares são mais complexos para essas espécies, principalmente para poliplóides, como é o caso da cana-

de-açúcar. Além disso, diferentes padrões de segregação dos marcadores são esperados, e também o não conhecimento da fase de ligação entre os marcadores dificulta a elaboração de mapas genéticos de ligação (HEINZ; TEW, 1987; LIN et al., 2003; ALWALA; KIMBENG, 2010). Visando contornar este tipo de situação, foi desenvolvida a estratégia denominada duplo *pseudo-testcross*, onde considera-se o cruzamento de parentais heterozigotos e utiliza-se marcadores polimórficos de dosagem única com segregação do tipo 1:1. Esses marcadores segregam para determinado genitor enquanto que para o outro genitor não apresentam segregação. Assim, são obtidos dois mapas de ligação, um para cada genitor de modo que cada mapa possui marcadores segregando para cada genitor (GRATTAPAGLIA; SEDEROFF, 1994) .

A construção de mapas genéticos em cana-de-açúcar só foi iniciada depois que Wu et al. (1992) propôs um método para mapeamento em poliplóides baseado na segregação de marcadores de dosagem única (SDRF - *Single Dose Restriction Fragment*). Neste método, os marcadores de dosagem única devem apresentar segregação do tipo 1:1 ou 3:1, ou seja, segregar para um determinado genitor enquanto para o outro genitor não apresentar segregação. Esta medologia além de possibilitar a construção de mapas genéticos de ligação em espécies poliplóides, como é o caso da cana-de-açúcar, pode ser estendida para outras espécies. Dentre os principais trabalhos envolvendo a construção de mapas genéticos em cana-de-açúcar, alguns serão brevemente comentados.

Visando compreender a organização do genoma da cana-de-açúcar e qual o tipo de população para obtenção de melhores ganhos na utilização dos marcadores, os primeiros mapas genéticos foram construídos por Silva et al. (1993) e Al-Janabi et al. (1993), que utilizaram marcadores do tipo RFLP e RADP para o estudo em questão. Já Grivet et al. (1996), contruíram um mapa genético baseado no cultivar R570, utilizando isoenzimas e marcadores do tipo RFLP. Neste trabalho foi obtido um mapa genético de 2.008 centimorgans (cM) sendo que a disposição de alguns marcadores dentro do mapa genético, forneceu informações sobre a organização do genoma da cana-de-açúcar. Outro estudo, baseado em uma população de 295 indivíduos resultante da autofecundação do cultivar R570, foi realizado por Hoarau et al. (2001), que ao utilizar aproximadamente 1000 marcadores do tipo AFLP, construíram um mapa de ligação com tamanho total de 5.849 cM, o que representou cerca de um terço da cobertura do genoma da cana-de-açúcar.

Já Garcia et al. (2006), contruíram um mapa integrado, ou seja, um único mapa para uma população de mapeamento resultante do cruzamento entre variedades pré-comerciais brasileiras SP80-180 e SP80-4966. Neste mapa integrado foram utilizados 357 marcadores RFLP, SSR e

AFLP, sendo que o tamanho total do mapa estimado foi de 2.602,4 cM. Mais recentemente Andru et al. (2011), desenvolveram um mapa de ligação para o cultivar comercial Louisiana LCP85-384, sendo que o mapa apresentou comprimento de 5.617 cM e serviu como base para estudos de citogenética sobre o comportamento do pareamento preferencial dos cromossomos em LCP85-384.

De acordo com Pastina et al. (2010), embora vários trabalhos de mapeamento genético em cana-de-açúcar foram realizados, apenas 33 a 50 % do genoma desta espécie são representados nestes mapas. Mais da metade da constituição genética da cana-de-açúcar não é conhecida. Portanto há necessidade do desenvolvimento de maior número de marcadores para a análise completa do genoma da cana-de-açúcar e melhor entendimento da natureza poliplóide das bases genéticas desta espécie. A grande quantidade de marcadores SSR e SNPs e o alto nível de polimorfismo desses marcadores, fazem das bibliotecas de ESTs, uma ótima fonte para estudos genéticos da cana-de-açúcar (MARCONI et al., 2011). Sendo assim, o uso de marcadores moleculares para corretas estimativas de mapas genéticos possilita melhorias na detecção de QTLs.

2.3 Mapeamento de QTLs

Com o advento dos marcadores moleculares e o aperfeiçoamento de técnicas genético-estatísticas para a análise dos dados moleculares, foi possível a melhor compreensão da arquitetura dos caracteres quantitativos. A identificação de caracteres quantitativos possui várias aplicações. Dentre elas, o aumento da eficácia dos programas de melhoramento genético, principalmente para caracteres de baixa herdabilidade (LANDER; BOTSTEIN, 1989), melhor entendimento da evolução desses caracteres e a identificação de genes de interesse (FALCONER; MACKAY, 1996).

Através do mapeamento de QTLs é possível fazer inferências ao longo de todo o genoma sobre as relações entre o genótipo e o fenótipo de caracteres que apresentam um padrão de variação contínuo, ou seja, os caracteres quantitativos. Além de fornecerem informações sobre número, posição, efeitos, interações alélicas entre e dentro locos, permitem também o melhor entendimento das bases genéticas da heterose, correlação entre caracteres e a identificação de interações entre QTLs e ambientes (JIANG; ZENG, 1995; ZENG; KAO; BASTEN, 1999; FALCONER; MACKAY, 1996; MALOSETTI et al., 2004, 2006, 2008; GARCIA et al., 2006).

Basicamente, as populações utilizadas no mapeamento de QTLs são classificadas em dois tipos: $i)$ populações experimentais (com linhagens endogâmicas) e $ii)$ populações obtidas através do cruzamento de indivíduos não endogâmicos, como por exemplo, progênies de irmãos completos.

O mapeamento de QTLs para populações heterozigóticas é mais complexo e portanto, sofisticadas metodologias são requeridas (LIN et al., 2003).

2.3.1 Métodos Para Mapeamento de QTLs

Os trabalhos iniciais envolvendo o mapeamento de QTLs eram realizados para cada marca individualmente (*SM - Single Marker*). Nesta abordagem, as existências de associações entre os marcadores e os caracteres quantitativos de interesse, são avaliadas através de testes estatísticos como teste t, análise de variância, regressão linear simples, regressão linear múltipla e teste de razão de máxima verossimilhanças (WELLER, 1986; DOERGE; ZENG; WEIR, 1997; LIU, 1998; LYNCH; WALSH, 1998). Nestes testes, a hipótese nula testa se as médias dos caracteres fenotípicos é independente do genótipo de um determinado marcador. Caso a hipótese nula seja rejeitada, a implicação é que um possível QTL está ligado a aquele marcador (DOERGE, 2002). Análises de marcas individuais permitem a realização dos testes estatísticos utilizando simples softwares para análises estatísticas. Além de ser um método facilmente realizado, não necessita de um mapa de ligação anteriormente estimado. Porém, os resultados dos testes estatísticos apresentam certo viés, uma vez que há confundimento entre o efeito do QTL e a sua distância em relação ao marcador, resultando em baixo poder de detecção de QTLs.

Considerando as limitações do método para análise de marcas individualmente, foi desenvolvida a abordagem denominada Mapeamento por Intervalo (*IM - Interval Mapping*) (LANDER; BOTSTEIN, 1989). Para isso, marcadores moleculares ordenados em um mapa de ligação, são a base para inferir a localização de possíveis QTLs. São considerados intervalos entre pares de marcadores adjacentes, e é testada a existência de um possível QTL dentro do intervalo em estudo. Os testes estatísticos utilizados combinam as análises de marcas individualmente com estimativas de máxima verossimilhança sobre a hipótese nula (não existência de um possível QTL) contra a hipótese alternativa (a existência de um possível QTL). Os resultados dos testes são expressos em LOD Score (*Logarithm of the odds*) e é observado maior poder de detecção dos QTLs quando comparado com as análises de marcas individualmente. No entanto, por não controlar a interferência de QTLs que estão localizados fora do intervalo em estudo, há chances de localizalição de QTLs falsos-positivos (comumente conhecidos por “QTLs fantasmas”), pertencentes ao intervalo adjacente ao estudado.

Baseado no método de Mapeamento por Intervalo e visando contornar suas desvantagens, foi proposta a extensão deste dando origem assim, ao método de Mapeamento por Intervalo Composto

(*CIM - Composite Interval Mapping*) (ZENG, 1993, 1994). Diferentemente do IM, o método CIM inclui marcadores adicionais, localizados fora do intervalo definido, como cofatores do modelo estatístico. Deste modo, é possível remover a variação associada a outro QTL que está localizado fora do intervalo em estudo e assim, a interferência de QTLs falsos-positivos pode ser controlada. Os testes estatísticos são realizados combinando a metodologia do mapeamento por intervalo com análises de regressão múltipla. Pelo fato de considerar múltiplos marcadores no modelo de regressão, a variação genética existente na população pode ser controlada na análise estatística e como resultado, há aumento do poder de detecção dos QTLs.

Visando aperfeiçoar as técnicas para mapeamento de QTLs anteriormente desenvolvidas, Kao e Zeng (1997) e Kao; Zeng e Teasdale (1999), propuseram o método de Mapeamento por Múltiplos Intervalos (*MIM - Multiple Interval Mapping*). Este método utiliza múltiplos intervalos de marcadores simultaneamente para adicionar múltiplos QTLs no modelo de mapeamento. Como vantagem, a abordagem do MIM, apresenta maior precisão e poder de mapeamento de QTLs quando comparada com as abordagens anteriormente descritas (SM, IM, CIM). A possibilidade de estimar efeitos de epistasia (QTLs epistáticos), valores genotípicos dos indivíduos e herdabilidade dos caracteres quantitativos são também vantagens desta metodologia. Uma vez que os componentes da variância genética são estimados para cada QTL individualmente, o melhoramento assistido pode ser considerado quando utilizada esta abordagem.

A grande maioria dos modelos para mapeamento de QTLs considera populações resultantes de linhagens homozigóticas, derivadas de populações F_2, retrocruzamentos e linhagens endogâmicas recombinantes. Quando são consideradas populações heterozigóticas, onde não há possibilidade de obtenção de populações homozigóticas como é o caso da cana-de-açúcar, é observado a presença de QTLs com diferentes padrões de segregação (1:1:1:1, 1:2:1, 3:1, 1:1). Sendo assim, os métodos tradicionais de mapeamento de QTL não devem ser utilizados. Nestas circunstâncias, geralmente a metodologia do duplo *pseudo testcross* é adotada, porém, a abordagem permite apenas a utilização de marcadores com segregações do tipo 1:1 (GRATTAPAGLIA; SEDEROFF, 1994). Marcadores com diferentes padrões de segregação, não podem ser considerados.

Dado o cenário anterior, Lin et al. (2003), propôs um modelo de mapeamento de QTLs em progênies de irmãos completos, baseado no Mapeamento por Intervalo. Este método permite a utilização de marcadores com diferentes padrões de segregação, são também considerados os efeitos e as posições dos QTLs e há a possibilidade de estimativa da fase de ligação entre os QTLs com os

marcadores, o que é de fundamental importância para determinar a origem dos alelos de interesse. No entanto, uma vez que a metodologia de mapeamento considerada é o Mapeamento por Intervalo, exitem algumas limitações inerentes ao modelo, como por exemplo, a não possibilidade de inclusão de cofatores no modelo de mapeamento.

Neste contexto, Gazaffi (2009) expandiu a abordagem proposta por Lin et al. (2003) para o Mapeamento por Intervalo Composto, considerando progênies de irmãos completos. Este modelo, além de localizar QTLs, determina o padrão de segregação e as fases de ligação entre QTLs, permitindo a análise simultânea de marcadores e QTLs com diferentes padrões de segregação (1:1:1:1, 1:2:1, 3:1, 1:1) e adicionalmente, considera a inclusão de cofatores no modelo, o que possibilita estimativas mais precisas de possíveis QTLs localizados no intervalo do genoma em estudo.

Basicamente, o modelo de mapeamento proposto por Gazaffi (2009), é baseado em 3 contrastes ortogonais para estimar os efeitos aditivos de cada genitor e a interação intraloco (dominância) entre os efeitos aditivos, para os alelos dos QTLs. São incluídos também cofatores para controle da variação ocasionada por QTLs que estão localizados fora do intervalo de mapeamento. A aplicação do modelo foi constatada através do mapeamento de QTLs em uma população segregante de cana-de-açúcar. Foi observado que a abordagem possibilitou a detecção de QTLs eficientemente, bem como os respectivos efeitos dos QTLs, as fases de ligação com os marcadores e o padrão de segregação destes. O modelo apresentou-se superior às outras abordagens existentes até o momento.

Entretanto, uma situação inerente aos métodos de mapeamento de QTLs é a presença de genótipos não observados de marcadores, que interferem e ocasionam viés na estimativa e localização de possíveis QTLs (BROMAN; SEN, 2009). Uma vez que os genótipos dos indivíduos são observados apenas em um conjunto discreto de marcadores genéticos, e são consideradas posições entre marcadores para a possível localização de QTLs ao longo de todos os genomas, deve-se utilizar os dados dos genótipos dos marcadores observados para inferir os possíveis genótipos não observados e consequentemente inferir de localização de QTLs com maior exatidão.

São poucas as abordagens que consideram os genótipos não observados nas análises de mapeamento de QTLs (BROMAN; SEN, 2009). Neste contexto, Sen e Churchill (2001), implementaram uma metodologia para mapeamento de QTLs que inclui o problema de genótipos não observados. Considerar a implementação de metodologias que incluem a informação do genótipo não observado aos modelos de mapeamento de QTL, tornam os mesmos mais eficazes (BROMAN; SEN, 2009). Consequentemente, o conhecimento dos genótipos dos marcadores dos indivíduos ao longo

de cada posição no genoma, possibilita a identificação de caracteres quantitativos de interesse mais precisamente.

2.3.2 Mapeamento de QTLs em Cana-de-Açúcar

Ao considerar a natureza poliplóide, a alta complexidade genética, bem como a dificuldade na obtenção de populações homozigóticas para o mapeamento de QTLs em cana-de-açúcar, é observado que os métodos de mapeamento mais utilizados são: $i)$ as análises de marcas individualmente, $ii)$ mapeamento por intervalo e $iii)$ mapeamento por intervalo composto, sendo que nesses dois últimos métodos, os mapas genéticos que servem como base para o mapeamento de QTLs, são geralmente construídos baseados na estratégia do duplo *pseudo-testcross* (PASTINA et al., 2010).

As análises estatísticas são implementadas em softwares que facilitam e simplificam as análises para detecção de possíveis QTLs nas populações de cana-de-açúcar. Entre os softwares utilizados, destacam-se SAS (SAS INSTITUTE, 2001), MAPMAKER/QTL (LINCOLN et al., 1987) e QTLCartographer (BASTEN et al., 2005). Porém, ainda existem muitas limitações inerentes às metodologias para mapeamento de QTLs que precisam ser estudadas e implementadas aos softwares para obtenção de análises mais precisas.

Dentre os principais trabalhos para mapeamento de QTLs em cana-de-açúcar, é observado que caracteres relacionados à produtividade como produção, peso, diâmetro do colmo e número de colmos são corriqueiramente estudados. Além desses, caracteres como porcentagem de fibra e resistência à doenças são considerados (AITKEN et al., 2008; MING et al., 2001, 2002). Os primeiros trabalhos de mapeamento genético de QTLs em cana-de-açúcar surgiram concomitantemente às construções dos mapas genéticos.

Os primeiros resultados de mapeamento de QTLs em cana-de-açúcar foram apresentados por Sills et al. (1995) realizando o cruzamento entre *S.officinarum* (“La Purple”) e *S.robustum* (“Mol 5829”), e utilizando a abordagem de marcas individualmente e marcadores do tipo RAPD. Houve a identificação de 42 associações significativas para os caracteres estudados. Além disso, foram identificadas três interações epistáticas para os caracteres em avaliação, porém apenas uma foi significafiva após análises de regressão múltipla.

Já Ming et al. (2001), por meio da avaliação de dois cruzamentos utilizando progênies derivadas de parentais heterozigóticos, a saber, *S.officinarum* “Green German” $\times$ *S.spontaneum* “IND 81-146” e *S.spontaneum* “PIN 84-1” $\times$ *S.officinarum* “Muntok Java”, realizaram o mapeamento de QTLs

através do mapeamento por intervalo e mapeamento de marcas individualmente. Ao total, considerando os dois cruzamentos, identificaram 36 associações positivas, sendo que destas, 14 QTLs foram identificados para o caractere TPH (Toneladas de Pol Por Hectare) para um dos cruzamentos, enquanto que 22 QTLs estavam relacionados ao caracter TPH do outro cruzamento, respectivamente. Posteriomente, utilizando os mesmos cruzamentos Ming et al. (2002), identificaram 102 associações significativas para caracteres de produtividade de açúcar, POL, teor de fibras, número de colmos, peso de colmos e conteúdo de cinzas. Dessas associações 61 QTLs estavam localizados no mapa de ligação e 41 estavam associados à marcadores não ligados. Para obtenção dessas associações foram utilizados os métodos de mapeamento por intervalo e mapeamento de marcas individualmente.

Aitken et al. (2008), mapearam QTLs utilizando análises de marcas individualmente e mapeamento por intervalo composto, utilizando dados de três cortes da cana-de-açúcar. Para isso, utilizaram marcadores do tipo AFLP e SSR para mapear os caracteres TCH (Toneladas de Cana Por Hectare), peso de colmos, número de colmos, altura de colmos e diâmetro de colmos. Foram identificados 38 QTLs considerando os caracteres em estudo, além disso, interações epistáticas entre os caracteres estudados foram encontradas.

Gazaffi (2009), mapeou QTLs em uma população de cana-de-açúcar utilizando um procedimento baseado no mapeamento por intervalo composto (CIM), que considera a inclusão de marcadores com diferentes padrões de segregação, além de caracterizar as fases de ligação e a segregação de QTLs. Foram mapeados 41 QTLs, sendo 14 para o caráter TPH, oito para o caráter TCH, 12 para a porcentagem de açúcar por cana (PCC) e sete QTLs para porcentagem de Fibra. Além disso, o trabalho permitiu estimar a segregação dos QTLs, sendo que, 33 QTLs apresentaram segregação 1:1, dois QTLs, apresentaram segregação 3:1, cinco QTLs com segregação 1:2:1 enquanto que apenas um QTL, dentre os mapeados, apresentou segregação 1:1:1:1.

Pastina (2010), visando mapear QTLs em uma população de cana-de-açúcar, utilizou uma abordagem baseada em modelos mistos e mapeamento por intervalo. Foram consideradas diferentes estruturas de variância e covariância, as quais permitiram detectar supostas correlações genéticas entre cortes e locais. Foram considerados três locais e dois cortes para a avaliação dos indivíduos da população de irmãos completos. Ao total, foram detectados 50 QTLs considerando os caracteres avaliados. Foi possível também, identificar efeitos de interação entre QTL $\times$ cortes, QTL $\times$ local e QTL $\times$ corte $\times$ local, em todos os caracteres avaliados. Esses resultados, forneceram importantes

informações para entendimento da base genética de caracteres quantitativos em cana-de-açúcar.

Mais recentemente Margarido (2011), visando estudar a interação de QTLs por ambiente, utilizou o método de mapeamento por intervalos múltiplos para detectar QTLs em uma população de irmãos completos de cana-de-açúcar, derivada do cruzamento biparental entre SP80-180 e SP80-4966, avaliada em dois locais e três cortes. Foram utilizadas análises de modelos mistos para a modelagem das variâncias genéticas e correlações entre caracteres, locais e cortes. Dentre os QTLs mapeados, 13 possuiam efeitos principais enquando que oito apresentaram interações epistáticas. Nenhum QTL apresentou efeitos estáveis ao longo de todas as combinações de ambientes. Além disso, 21 dos QTLs mapeados apresentaram algum grau de pleiotropia.

É notável que dentro dos vários trabalhos para mapeamento de QTLs em cana-de-açúcar, grande parte destes, utilizam as mesmas metodologias de mapeamento, a saber, marcas individualmente e mapeamento por intervalo. Sendo assim, há necessidade em explorar outros métodos para mapeamento de QTLs. Além disso, incluir estudos que consideram a informação de genótipos não observados, é de grande importância para melhor exploração dos dados genotípicos e aperfeiçoamento dos métodos para mapeamento de QTLs (BROMAN; SEN, 2009).

2.4 Modelos Para Dados Não Observados

Quando se trabalha com análises de conjunto de dados, uma situação comum é a ocorrência de dados faltantes ou dados não observados (*missing data*). Sendo assim, os conjuntos de dados que contém dados faltantes, não podem ser avaliados com a mesma eficiência e validação que conjuntos com dados completos (RUBIN, 1976). Vários métodos são utilizados para avaliação de conjunto de dados que contenham dados faltantes, entretanto, determinar qual a abordagem mais adequada para tratar e analisar o conjunto que contém os dados não observados é uma tarefa criteriosa uma vez que a utilização de metodologias inadequadas podem resultar em inferências erradas e estimativas ineficientes (ZHANG, 2003).

Os valores não observados não resultam apenas em estimativas menos eficientes em razão da redução do tamanho amostral do conjunto de dados, mas também os métodos padrões para análises de dados completos não podem ser imediatamente utilizados uma vez que possíveis viéses estão inerentes aos dados não observados (RUBIN, 1987). Em muitos trabalhos, onde há ocorrência de dados não observados, o processo que causa a não observação dos dados é erroneamente ignorado. Em outras circunstâncias, este processo é explícito e pode ser ocasionado por vários motivos, dentre

eles, erros inerentes à coleta de informações, descarte de dados, técnicas inadequadas para obtenção de dados, erros de digitação e mensuração ou até mesmo problemas relacionados ao planejamento experimental (RUBIN, 1976).

Para tratar estes problemas, as técnicas ou metodologias existentes que envolvem o tratamento de dados não observados, devem possuir as seguintes característica de acordo com Rubin (1987): $i)$ permitir que um modelo adequado aos dados observados e não observados seja utilizado, $ii)$ o modelo deve ser capaz de produzir inferências válidas apresentando estimativas que se ajustem à diferentes observações entre os dados observados e não observados e $iii)$ o modelo deve mostrar a sensibilidade da inferência aos modelos plausíveis para dados não observados. Várias abordagens estatísticas tem sido desenvolvidas ao longo dos anos para tratar situações que apresentem dados incompletos como por exemplo, o algoritmo EM (DEMPSTER; LAIRD; RUBIN, 1977; LITTLE; SMITH, 1987) e modelos de efeito aleatório (MCCULLAGH; NELDER, 1989). Esses modelos são geralmente estimados através de funções de verossimilhanças baseadas nos dados observados. Sobre certas pressuposições tais modelos produzem inferências válidas dos parâmetros não conhecidos, no entanto, a precisão das inferências são geralmente reduzidas pelo fato da ocorrência dos dados não observados.

Outra abordagem existente para tratar os dados não observados é completar tais dados com valores plausíveis, baseados nos dados observados e posteriormente aplicar o modelo para dados completos ao conjunto de dados completados de forma à realizar inferências válidas e eficientes. Esta técnica é denominada imputação de dados e contorna a complexidade dos dados não observados. Uma das grandes vantagens desta técnica é que após a recomposição dos dados não observados, os métodos tradicionais para análises de dados completos, podem ser utilizados normalmente. No contexto do mapeamento de QTLs isso é particularmente interessante porque, uma vez que os dados estejam completos, modelos sofisticados, podem ser empregados com razoável facilidade.

2.4.1 Imputação de Dados Não Observados

Uma técnica amplamente utilizada quando há ocorrência de dados faltantes ou não observados é a imputação de dados (RUBIN, 1976; 1987; SCHAFER et al., 1999; LITTLE; RUBIN, 2002; CORDELL, 2006; BERGAMO, 2008). Imputação significa preencher os dados não observados com valores plausíveis baseados nos valores observados do conjunto de dados e após a imputação, analisar o conjunto de dados utilizando métodos tradicionais para avaliação de dados completos.

Desta forma os resultados das análises realizadas se tornam válidos, e as inferências obtidas através destes são, eficientes.

Basicamente, as técnicas que utilizam a imputação de dados para recomposição de dados não observados são dividas em dois tipos, imputação única ou simples e imputação múltipla. De acordo com Rubin (1987), a técnica da imputação única visa completar com um único valor cada um dos dados não observados e provavelmente é o método mais utilizado para tratar dados faltantes. Procedimentos simples como o cálculo da média amostral, variância amostral ou amostragens aleatórias são utilizados para completar a informação faltante e obtenção de conjuntos completados. Como vantagens essa técnica apresenta facilidade de execução e há a possibilidade de incorporar o conhecimento do pesquisador às análises. Entretanto, uma vez que um único valor é imputado para cada dado não observado, não há existência de variabilidade na amostragem realizada e consequentemente, a incerteza do modelo não é incluída.

Sendo assim, a incerteza associada à imputação dos dados, necessita ser adicionada às análises, pois os valores que são imputados não são exatamente os valores reais que seriam observados. De acordo com Zhang (2003), existem três tipos de incertezas que estão associadas à imputação dos dados: $i)$ a modelagem da distribuição conjunta das variáveis resposta e as indicadoras de dados faltantes, $ii)$ a amostragem a partir de um dado modelo de imputação, assumindo que os dados observados e os parâmetros do modelo são conhecidos (seleção aleatória dos valores que serão imputados) e $iii)$ os valores dos parâmetros do modelo (a seleção do modelo).

Visando contornar as incertezas associadas à imputação de dados, foi proposta a abordagem da imputação múltipla de dados. A técnica da imputação múltipla corrige as principais desvantagens da imputação única e conserva algumas vantagens desta última abordagem como a incorporação do conhecimento do pesquisador às análises. Inicialmente proposta por Rubin (1977), a ideia existente na imputação múltipla é que para cada dado não observado, são imputados vários valores, ditos m, ao invés de apenas um único valor. Deste modo, as m imputações para cada dado não observado, originam m conjunto de dados completados (contendo dados observados e imputados). Cada conjunto de dado completado é analizado individualmente utilizando-se os procedimentos padrões para análise de dados completos. Posteriormente, os m resultados obtidos são combinados, através de uma média ponderada, em um único resultado.

Como principais vantagens da imputação múltipla sobre a imputação única, observamos o aumento na eficiência das estimativas dos dados não observados, há variabilidade adicional ocasionada

pelo número de m imputações e consequentemente a incerteza associada à imputação é adicionada. Além disso, caso seja utilizado mais que um modelo para realizar as imputações, pode-se comparar a sensibilidade das inferências entre os modelos. Porém, além das incertezas associadas à imputação devem também ser considerados os mecanismos de não observação dos dados, ou seja, o mecanismo que gerou os dados não observados e a associação de tais dados com as variáveis respostas. Little e Rubin (2002), classificaram os mecanismos de não observação dos dados em três categorias:

- **Não Observação Completamente Aleatória** - (*Missing Completely At Random - MCAR*): Quando a distribuição dos dados faltantes de uma variável específica não depende dos dados observados ou dos dados não observados das outras variáveis.

- **Não Observação Aleatória** - (*Missing At Random - MAR*): Quando o padrão de não observação é previsível a partir de outras variáveis do banco de dados e não devido a variável específica na qual os dados foram não observados. Os dados são não observados por um processo aleatório, quando a probabilidade de não resposta depende dos dados observados, mas não dos não observados.

- **Não Observação Não Aleatória** - (*Missing Not At Random - NMAR*): O mecanismo de não observação está relacionado com os valores não observados e outras variáveis que constituem o conjunto de dados.

Uma vez que os conjuntos de dados são organizados em forma de matriz para posterior análise dos mesmos, onde as linhas da matriz correspondem aos indivíduos e as colunas às variáveis, é possível estabelecer o padrão dos dados não observados. Os padrões dos dados podem ser classificados em *Monotônicos* e *Não Monotônicos*. Os dados assumem o padrão Monotônico sempre que um elemento y_{i_j} de uma determinada variável é não observado e o elemento y_{i_k} de outra variável é também não observado para todo $k > j$. Caso os dados não assumam este padrão, são classificados como Não Monotônicos. No padrão Não Monotônico, os elementos correspondentes em variáveis diferentes, não são observados juntos, ou seja, o padrão de não observação dos dados é aleatório (RUBIN, 1987).

A detecção do mecanismo e do padrão dos conjuntos de dados que contém os dados não observados é de fundamental importância e deve ser previamente realizada às imputações. De acordo

com a classificação destes é que determinada técnica de imputação será mais indicada para o conjunto de dados em estudo. Além disso, considerar as incertezas associadas à imputação é necessário, pois deste modo, as inferências baseadas nos resultados obtidos através do conjunto de dados que contenham dados imputados, tornam-se mais precisas.

2.4.2 Dados Não Observados no Mapeamento de QTLs

Mapear QTLs significa encontrar regiões no genoma que estão associadas com a expressão do fenótipo (ZENG et al., 1999). Para a detecção dos QTLs é preciso inicialmente elaborar um mapa de locos marcadores, o qual fornece a estrutura física para a alocação dos QTLs (BEARZOTI, 2000). Porém, a falta de informação dos dados genotípicos é uma das limitações inerentes aos métodos de mapeamento de QTLs (LANDER; BOTSTEIN, 1989). Dados moleculares não observados são oriundos de erros de genotipagem, marcadores não informativo, marcadores perdidos, entre outros.

Na obtenção das informações moleculares após a genotipagem dos indivíduos, são obtidos como resultados conjuntos de marcadores genéticos os quais contém a informação genotípica dos indivíduos em estudo. De acordo com o modelo selecionado para o mapeamento de caracteríscas quantitativas, é esperada a possibilidade de localização dos possíveis QTLs nos marcadores genotipados, como também é desejado detectar QTLs nas posições localizadas entre os marcadores. Sendo assim, os dados dos genótipos dos marcadores devem ser utilizados para inferir a localização dos QTLs (BROMAN; SEN, 2009). Dentre as diferentes metodologias que tratam os dados genotípicos não observados, algumas mecerem destaque.

O mapeamento por intervalo padrão é uma das técnicas que incorpora a informação dos genótipos não observados às análises utilizando estimativas de máxima verossimilhança baseadas em um modelo de mistura (BROMAN; SEN, 2009). Outra metodologia desenvolvida é a regressão de Haley-Knott (HALEY; KNOTT, 1992), a qual faz uso de aproximações baseadas em quadrados míminos. Porém em ambas metodologias, os dados não observados são tratados como parâmetros dos modelos de mapeamento dos QTLs. Uma abordagem alternativa às metodologias anteriormente citadas, a qual trata a falta de informação dos dados genotípicos é a imputação dos genótipos dos marcadores localizados entre os marcadores genotipados (observados). Assim, a informação não observada dos marcadores genéticos se torna completa e é uma ferramenta adicional para a localização dos possíveis QTLs.

Basicamente o processo de imputação dos marcadores consiste em completar a informação do

genótipo não observado, através de determinada técnica de imputação, “adicionando” marcadores no mapa genético anteriormente estimado de modo que este mapa se torne mais saturado. Após a saturação do mapa genético, a técnica para detecção de possíveis QTLs é aplicada normalmente ao conjunto de dados, e então o mapeamento de QTLs é realizado (SEN; CHURCHILL, 2001). Consequentemente a acurácia das técnicas para a detecção dos QTLs se torna maior pois o mapa se apresenta mais saturado e há eventual possibilidade de identificar novas regiões com presença de QTLs que anteriormente seriam desconsideradas. Entretanto, para realizar a imputação dos genótipos dos QTLs localizados entre os marcadores genotipados, é preciso anteriormente estimar a probabilidade de ocorrência dos genótipos não observados condicional aos genótipos dos marcadores observados (JIANG; ZENG, 1997). Então, baseado em tais probabilidades condicionais, a apropriada técnica de imputação dos genótipos é aplicada aos dados e finalmente o mapeamento de QTLs é realizado.

2.4.2.1 Probabilidade dos Genótipos Não Observados

Lander e Green (1987), propuseram inicialmente que baseado em uma determinada ordem de marcadores moleculares é possível estimar frações de recombinação entre os marcadores via cadeia de Markov oculta, utilizando a abordagem multiponto. Tal abordagem permite recuperar a informação não observada ao longo de todo o mapa de ligação. Desto modo, as estimativas obtidas são mais precisas pois ao utilizar a abordagem multiponto, a verossimilhança é recalculada mais eficientemente. Posteriormente, Jiang e Zeng (1997), detalharam as ideias inicialmente propostas por Lander e Green (1987) para vários tipos de populações oriunda do cruzamento entre linhagens endogâmicas, no contexto do mapeamento de QTLs.

A abordagem proposta para o cálculo da probabilidade de genótipos não observados (JIANG; ZENG, 1997), considera que determinado número de marcadores genéticos apresentam genótipos conhecidos e estão ordenado em um mapa de ligação, sendo m os marcadores genéticos e $M_1, \cdots, M_m$ a respectiva ordem dos marcadores. Quando o genótipo do marcador ou do QTL não é observado, seu genótipo não é conhecido e assume determinadas possibilidades de ocorrência. Por exemplo, no caso de uma população resultante de um retrocruzamento, os possíveis genótipos dos marcadores ou dos QTLs nessa população são heterozigotos (AB) ou homozigotos (AA).

De acordo com Jiang e Zeng (1997), caso M_k seja um marcador não observado no mapa genético estimado, x_k o genótipo e z_k o fenótipo do marcador, e que este marcador é flanqueado por dois outros marcadores M_{k-1} e M_{k+1} observáveis, a probabilidade condicional do genótipo do marcador

M_k dependerá do fenótipo de M_k, do genótipo de M_{k-1} e M_{k+1}, e da frequência de recombinação entre M_k e M_{k-1} e entre M_k e M_{k+1}. Então, é possível observar que:

$$P\left(x_k \mid z_1, \cdots, x_{k-1}, z_k, x_{k+1}, \cdots, z_m\right) = P\left(x_k \mid x_{k-1}, z_k, x_{k+1}\right) \tag{1}$$

Caso os marcadores flanqueadores ao marcador não observado também não sejam observados, os próximos marcadores adjacentes aos marcadores flanqueadores podem fornecer alguma informação a respeito do genótipo dos marcadores flanqueadores, aperfeiçoando assim, o cálculo da probabilidade do genótipo do QTL não observado. Esta ideia deve ser estendida para cada direção dos marcadores não observados (tanto à direita quanto à esquerda do marcador), até o loco em que o marcador seja observável. Sendo assim, a probabilidade do genótipo do marcador não observado é expressa pela probabilidade do genótipo deste marcador dado a sequência dos fenótipos dos marcadores ordenados no mapa genético, $P(x_k \mid z_1, \cdots, z_l)$. O Teorema de Bayes observado é:

$$P(x_k \mid z_1, \cdots, z_l) = \frac{P(x_k)P(z_i \cdots z_l \mid x_k)}{\sum_{x_k} P(x_k)P(z_i \cdots z_l \mid x_k)} \tag{2}$$

Ainda, ao considerar a não interferência de crossing-over, para um dado marcador x_k é possível observar que $P(z_i \cdots z_l \mid x_k) = P(z_i \cdots z_k \mid x_k)P(z_{k+1} \cdots z_l \mid x_k)$. Ou seja, a probabilidade da sequência dos fenótipos dos marcadores dado o genótipo do marcador x_k é igual a probabilidade dos fenótipos dos marcadores à esquerda do marcador x_k dado o marcador x_k e a probabilidade dos fenótipos dos marcadores à direita do marcador x_k dado o marcador x_k. Uma vez que o genótipo do QTL não observado não é conhecido, $P(x_k)$ pode assumir duas possibilidades de probabilidade *a priori* caso a população de mapeamento seja resultante de um retrocruzamento, a saber, $P(x_k = AA)$ ou $P(x_k = AB)$, sendo $\mathbf{q_k} = \{P\left(x_k\right)\}_{(2x1)}$.

Então, o vetor de probabilidade *a priori* de $P(x_k)$, considerando a população resultante do retrocruzamento, é expresso por $\mathbf{q'_k} = [P(x_k = AA), P(x_k = AB)]$, onde $'$ denota a transposição. Da mesma forma é observado que: $\mathbf{q_k^R} = \{P(z_{k+1} \cdots z_l \mid x_k)\}_{(2x1)}$, $\mathbf{q_k^L} = \{P(z_i \cdots z_k \mid x_k)\}_{(2x1)}$, $\mathbf{q_k} = \{P(x_k \mid z_i \cdots z_l)\}_{(2x1)}$.

E deste modo, a equação 2, pode ser expressa por:

$$\mathbf{p_k} = \frac{\mathbf{q_k} \circ (\mathbf{p}_k^L \circ \mathbf{p_k^R})}{\mathbf{q'_k}(\mathbf{p}_k^L \circ \mathbf{p_k^R})} \tag{3}$$

Os vetores $\mathbf{p_k^R}$ e $\mathbf{p_k^L}$, podem ser calculados utilizando o processo de Cadeia de Markov Oculta

(*Hidden Markov Models - HMMs*). Uma cadeia de Markov é uma descrição probabilística, que possui um conjunto de estados em cada instante de tempo, sendo que cada estado depende exclusivamente do estado do tempo anterior. Quando os estados correspondem a eventos observados, a cadeia é dita *observável*. Quando na cadeia existem estados que não são observados, a cadeia é dita *oculta* e os estados observados na cadeia auxiliam na inferência dos estados ocultos (RABINER, 1989). Analogamente, ao considerar uma sequência de marcadores ordenados em um mapa molecular, onde alguns marcadores não são observados, os estados ocultos da cadeia, são representados pelo conjunto das probabilidades dos genótipos dos QTLs, cada instante de tempo é representado por cada posição dentro da sequência de marcadores, e a observação da cadeia equivale ao fenótipo de cada marcador.

Assim, utilizando a abordagem anteriormente descrita, é possível calcular a probabilidade condicional do genótipo do QTL não observado. Uma vez que a probabilidade do genótipo do QTL é obtida, é possível estimar qual o genótipo do marcador ou QTL não observado, e então completar a informação não observada onde esta ocorreu, "adicionando" marcadores ao mapa genético através da técnica de imputação mais adequada. Posteriormente, a inferência sobre um possível QTL, pode ser realizada. A abordagem multiponto aqui apresentada tem a grande vantagem de considerar todos os marcadores simultaneamente.

2.5 Imputação de Dados no Mapeamento de QTLs

Uma vez que na detecção de QTLs, há ocorrência de genótipos de QTLs não observados, o uso de métodos de imputação é uma excelente alternativa para tratar este tipo de ocorrência (BROMAN; SEN, 2009). Como ressaltado na sessão 2.4.1, quando se trabalha com conjunto de dados os quais contém dados não observados, o primeiro passo para analisar tais conjuntos de dados, é identificar quais são os mecanismos de não observação e o padrão dos dados não observados para posteriormente aplicar a técnica de imputação mais adequada. Tratando-se do padrão de não observação dos dados, é verificado que dados moleculares dispostos em um mapa de ligação e utilizados para mapeamento de QTLs apresentam o padrão Não Monotônico de não observação (SEN; CHURCHILL, 2001). Segundo Zhang (2003), quando os dados não observados assumem o padrão Não Monotônico, procedimentos computacionais mais complexos precisam ser utilizados. Uma técnica indicada para tratar dados moleculares que apresentam este tipo de padrão é a imputação múltipla (SEN; CHURCHILL, 2001; BROMAN; SEN, 2009). A imputação múltipla apresenta es-

timativas mais eficientes, uma vez que a incerteza associada aos dados imputados é considerada através das variâncias obtidas quando os dados não observados são substituídos.

A técnica da imputação múltipla, é basicamente dividida em três passos: $i)$ cada dado não observado é imputado m vezes, sendo que o tamanho de m dependerá da ferramenta computacional disponível, $ii)$ após, são obtidos m conjuntos de dados completados (valores observados e imputados), e estes conjuntos são analisados individualmente através dos procedimento usuais para análise de conjunto de dados completos e $iii)$ então, os vários resultados obtidos são combinados em um único resultado, através da obtenção de uma média ponderada dos resultados de todas as imputações (RUBIN, 1977). Ao transpor os passos da imputação múltipla para imputação múltipla de dados moleculares é observado que no primeiro passo, o conjunto de dados moleculares é imputado m vezes, sendo que cada dado imputado, neste caso um pseudomarcador, é gerado a partir da probabilidade condicional do genótipo do marcador molecular ou QTL. Após as m imputações, cada m conjunto de genótipos, que corresponde a um mapa genético mais saturado, com pseudomarcadores "adicionados", é então analisado normalmente através de um determinado modelo genético assumido para o mapeamento de QTLs. O último passo, é o passo mais complexo, onde os resultados obtidos no passo anterior são combinados em um resultado final, através de uma média ponderada (SEN; CHURCHILL, 2001).

Como descrito anteriormente na sessão 2.4.2.1, a probabilidade do genótipo do marcador ou QTL não observado é calculada através de técnicas que utilizam a abordagem de Cadeia de Markov Oculta (JIANG; ZENG, 1997). Uma vez obtida as probabilidades dos genótipos dos QTLs em cada posição onde é desejado completar a informação não observada, é necessário gerar esta informação, isto é, gerar os genótipos dos pseudomarcadores. Esta informação pode facilmente ser obtida através de técnicas de amostragens que consideram a probabilidade condicional do genótipo do QTL anteriormente estimada para gerar o genótipo não observado. Neste passo, é importante definir o número das m imputações que será utilizado.

Uma vez que a etapa mais complexa da imputação múltipla dos dados é a combinação final dos resultados, esta deve ser realizada com muita cautela. A única metodologia que aborda a imputação múltipla no mapeamento de QTLs foi desenvolvida por Sen e Churchill (2001), que sugerem uma abordagem para a combinação dos resultados das múltiplas imputações em um único resultado final. A abordagem consiste em três passos básicos. Primeiramente, em cada localização u do mapa genético estimado, correspondente ao loco do marcador, são realizadas m imputações de pseudo-

marcadores. Cada imputação é realizada por amostragem baseada na probabilidade condicional do genótipo do QTL. Posteriormente, para cada localização u, em cada i imputação $(i = 1, \cdots, m)$, é calculado um peso de acordo com o modelo genético de mapeamento G assumido:

$$W_G(r_i(u)) = p(y \mid g = r_i(u)) \tag{4}$$

onde, y corresponde aos valores fenotípicos dos indivíduos que constituem o conjunto de dados genotípicos , e $g = r_i(u)$ corresponde aos genótipos imputados dos indivíduos para cada i imputação em cada u localização.

Entretando, a $p(y \mid g = r_i(u))$ é na verdade uma distribuição de probabilidade geral que dependerá da distribuição que os dados fenotípicos y dos indivíduos da população apresentam. E então para cada imputação i em cada localização u, cada indivíduo apresentará uma função densidade de probabilidade (f.d.p.) específica baseada na distribuição dos dados fenotípicos y. Sendo assim, o peso para cada imputação i em cada localização u será resultado do produtório da f.d.p. de cada j indivíduo $(j = 1, \cdots, n)$ da população:

$$p(y \mid g = r_i(u)) = \prod_{j=1}^{n} fdp_j \tag{5}$$

Aqueles genótipos dos pseudomarcadores que obtivem o maior peso são os mais compatíveis aos fenótipo dos indivíduos. Então, o último passo é a obtenção do peso médio de todas as imputações dos pseudomarcadores para cada localização u (Equação 6).

$$\sum_{i=1}^{m} W_G(r_i(u)) \tag{6}$$

Uma vez obtido o peso médio de todas as imputações, em cada localização, é possível inferir sobre a localização dos possíveis QTLs presentes nos conjuntos de dados moleculares de acordo com o modelo genético de mapeamento especificado.

As ferramentas computacionais existentes hoje, permitem maior exploração dos dados moleculares, e assim, um maior número de realizações de imputações. Entretando é necessário levar em conta o tamanho do mapa genético em que os dados serão imputados, e também, qual será o “grid” de espaçamento entre os marcadores moleculares. A ideia é que o número de imputações a serem realizadas dependerá do tamanho do mapa molecular e do “grid” de espaçamento entre

os marcadores. Quanto maior o mapa e menor o “grid” de espaçamento, mais computações são necessárias para determinado m número de imputações.

Dentre os poucos trabalhos envolvendo imputações múltipla de dados moleculares e mapeamento de QTLs, alguns serão brevemente comentados. Visando avaliar a eficiência do modelo proposto por Sen e Churchill (2001), tais autores reanalizaram um conjunto de dados de um retrocruzamento entre linhagens de camundongos para o caráter pressão sanguínea. Um total de 174 marcadores foram genotipados e foram consideradas 16 imputações para os dados em questão. Estudos para comparar a eficiência de diferentes modelos de mapeamento incluindo efeitos de um, dois ou três QTLs no modelo de mapeamento respectivamente, foram realizados. Os autores concluíram que o modelo que inclui dois QTLs no modelo de mapeamento é o mais indicado para o caso em estudo. Além disso, com a implementação da técnica da imputação múltipla, foi possível detectar a presença de cinco possíveis QTLs como também estimar os efeitos e interações dos QTLs em questão.

Outro estudo envolvendo a utilização de imputações múltiplas e mapeamentos de QTLs foi realizado por Ball (2001). Este, descreveu um método de aproximação para mapeamento de QTLs baseado em um modelo de regressão múltipla para populações de *Pinus Radiata*. Os valores de marcadores não observados foram tratados por meio de imputações múltiplas, baseadas nos marcadores flanqueadores. Foram realizadas 10 imputações no total, sendo que a justificativa para o número de imputações foi baseado no conjunto discreto de marcadores imputados. Baseado nos resultados obtidos, foi possível concluir que considerando o tamanho da amostra da população e a incertaza de localização do QTL, o uso de imputações múltiplas foi adequado.

Em estudos de mapeamento de QTLs de caracteres relacionados à osteoporose em camudongos, Bouxsein et al. (2004), utilizaram imputações múltiplas para realizar o escaneamento do genoma. Foram realizadas 16 imputações para gerar pseudomarcadores a cada 5 cM no mapa de ligação, sendo que o método de mapeamento utilizado foi o mapeamento por intervalo. Já Symonds et al. (2005), em estudos comparativo para detecção de QTLs entre quatro populações de mapeamento em *Arabidopsis thaliana*, utilizaram a técnica da imputação múltipla de genótipos de QTLs, visando a detecção de interações epistáticas após a utilização de tal metodologia. Foi utilizado o mapeamento por intervalo para a detecção de QTLs em cada população e após a detecção dos QTLs em cada população, foi realizado um estudo comparativo entre QTLs e populações. Ainda no contexto de implementar a imputação múltipla de dados moleculares para o mapeamento de QTLs, Sun et al.

(2006), ao mapear QTLs em populações de pepino, utilizaram a abordagem da imputação múltipla para recomposição de marcadores não observados e posteriormente utilizaram o mapeamento por intervalo para a detecção de QTLs.

Dentre os poucos trabalhos que utilizam a técnica da imputação múltipla para recomposição de marcadores não observados e posterior mapeamento de QTLs, é notável que a abordagem de mapeamento utilizada se restringe ao mapeamento por intervalo. Em espécies vegetais, poucos trabalhos utilizando a imputação de dados para mapeamento de QTLs foram observados sendo que em populações derivadas de progênies de irmãos completos, essa questão nunca foi abordada até o presente momento.

3 MATERIAL E MÉTODOS

3.1 Material

3.1.1 Material Vegetal

Os dados utilizados para realizar as análises fenotípicas e moleculares foram obtidos através de uma população resultante de um cruzamento biparental entre o clone elite IACSP95-3018 e o cultivar IACSP93-3046. O experimento foi realizado na estação experimental do Centro de Tecnologia Canavieira (CTC) em Camamu - BA, no ano de 2004. O clone IACSP95-3018 (genitor feminino) apresenta suscetibilidade à ferrugem enquanto que o cultivar IACSP93-3046 (genitor masculino) apresenta resistência à ferrugem, altos níveis de sacarose, bom perfilhamento, além de apresentar colmos eretos, sendo então indicado para a colheita mecanizada.

O material vegetal utilizado para a genotipagem e avaliações fenotípicas foi obtido por meio de uma amostragem aleatória da progênie resultante desse cruzamento, sendo que 220 indivíduos representativos foram amostrados.

3.1.2 Avaliações Fenotípicas

O experimento foi realizado na fazenda Nova Aliança, no município de Salles Oliveira no estado de São Paulo. O plantio foi realizado nos anos agrícolas de 2007/2008 e 2008/2009. Foi utilizado o delineamento de blocos casualizados com quatro repetições. Cada bloco foi constituído por quatro linhas de 118 metros, espaçadas a cada um metro e meio. Cada parcela alocada dentro das linhas do bloco possuiu dois metros lineares de comprimento.

Dentro de cada bloco foram alocados aleatoriamente 220 indivíduos, os dois genitores, além de duas variedades utilizadas como testemunhas, a saber, SP81-3250 e RB83-5486. Cada bloco apresentou como bordadura a variedade IAC86-2480 (infectora da ferrugem marrom) e o experimento todo possuiu como bordadura a variedade IAC86-2210 (infectora do mosaico).

Após decorrido 12 meses do plantio de ambos os anos agrícolas (cana planta e cana soca), de cada material em estudo, considerando as testemunhas e os genitores, foram coletadas e avaliadas amostras de dez canas-de-açúcar para os componentes de produção e os parâmetros de qualidades. Para as quatro repetições, nos dois anos agrícolas, foram avaliados os seguintes componentes de

produção: $i)$ diâmetro do colmo; amostrando-se dez colmos seguidos na linha, mensurado com paquímetro no meio do entrenó localizado no comprimento médio do colmo; $ii)$ peso em kilos (kg) do feixe de dez colmos; $iii)$ altura do colmo em centímetros (cm), medida da base à inserção da folha +3, estimada nos mesmos dez colmos, $iv)$ número de colmos estimado com a contagem dos colmos no sulco de dois metros e $v)$ produção de cana por parcela, estimado com base no número de colmos da parcela e no peso do feixe de dez colmos e $vi)$ produção de açúcar por parcela, estimados com base no valor da produção de cana por parcela e do POL das amostras obtidas.

Os parâmetros de qualidade Fibra, Brix e POL também foram avaliados considerando dez amostras de cana-de-açúcar coletada de cada material em estudo, conforme os métodos descritos no guia CONSECANA (2006). Porém, devido aos altos custos de tais análises, somente dois dos quatro blocos foram avaliados. Brix se refere a estimativa do total de sólidos solúveis que estão presentes no caldo da cana-de-açúcar, POL é definido como a quantidade de sacarose presente em 100mL da solução e Fibra se refere à matéria insolúvel em água que está presente no colmo.

3.1.3 Avaliações Moleculares

Foram utilizados para a genotipagem dos genitores e da progênie 25 marcadores do tipo AFLP (*Amplified Fragment Length Polymorphism*), 35 marcadores do tipo SSR (*Simple Sequence Repeat*) e 105 marcadores do tipo EST (*Expressed Sequence Tag*) derivados de SSR. Os marcadores utilizados foram previamente obtidos e codificados por Cordeiro et al. (2000) e Mancini (2010). A partir dos resultados obtidos nos géis, foram realizados testes de segregação dos marcadores, de modo que os indivíduos em estudo (genitores e progênie) foram codificados com base na presença ou ausência do marcador. As análises dos padrões de segregação de cada marcador foram realizadas por meio do teste estatístico de qui-quadrado (χ^2), em relação às proporções dos marcadores em dose única em apenas um genitor (padrão de segregação 1:1) e em ambos os genitores (padrão de segregação 3:1), considerando o nível de significância $p < 0,05$. Também foram testada segregações para doses duplas em um dos genitores (padrão se segregação 11:3) e doses triplas em um dos genitores (13:1). Foi utilizada a correção de Bonferroni para múltiplos testes. Após tais análises foram considerados apenas locos que apresentavam segregação em dose única (padrão 1:1 e 3:1).

3.2 Métodos

3.2.1 Mapa Genético de Ligação

Foram utilizados para a construção do mapa genético de ligação 165 marcadores moleculares, dos quais, 25 marcadores eram do tipo AFLP, 35 do tipo SSR e 105 do tipo EST-SSR. Tais marcadores geraram 634 locos polimórficos de dosagem única, apresentando o padrão de segregação do tipo 1:1 ou 3:1. Aqueles marcadores que apresentaram segregação para o genitor IACSP95-3018 foram denotados por D_1 e apresentaram configuração alélica "$ao \times oo$". Marcadores que apresentaram segregação para o genitor IACSP93-3046 foram denotados por D_2 e apresentaram configuração alélica "$oo \times ao$". Já aqueles marcadores que apresentaram segregação para os dois genitores simultaneamente, foram denotados por C com configuração "$ao \times ao$", sendo a o alelo dominante e o o alelo nulo, de acordo com a notação proposta por Wu et al. (2002b).

Para a execução do mapa genético de ligação foi utilizado o pacote *OneMap* (MARGARIDO; SOUZA; GARCIA, 2007), que utiliza uma abordagem multiponto proposta por (WU et al., 2002a, 2002b), possibilitando estimativas da ordem e possíveis fases de ligação entre os marcadores. Para a determinação prévia dos grupos de ligação, foram realizados testes de dois-pontos, considerando para tais testes um LOD Score mínimo de $5{,}77$ e a fração de recombinação máxima de $0{,}40$. Para estabelecimento das ordens e das fases de ligação dentro de cada grupo de ligação contendo cinco marcadores ou menos, foram comparadas todas as possíveis ordens, sendo a ordem mais provável a selecionada. Para os grupos de ligação com seis ou mais marcadores, foram selecionados cinco marcadores e foram comparadas todas as possíveis ordens, selecionando a mais provável; posteriormente os demais marcadores foram posicionados sequencialmente a partir da ordem inicial do grupo de ligação em questão, de modo a considerar sempre a posição de maior verossimilhança. Para obtenção das distâncias finais em cM, foi utilizada a função de mapeamento Kosambi (KOSAMBI, 1944).

3.2.2 Análise Fenotípica

Para o presente estudo foram considerados os caracteres de principal importância agronômica e econômica em cana-de-açúcar: Fibra, POL, produção de cana por parcela (PC) e produção de açúcar por parcela (PP). Com o objetivo de avaliar os dados fenotípicos em questão, inicialmente foi ajustado um modelo misto apropriado para tais dados considerando os 220 indivíduos da progênie

resultante do cruzamento entre os materiais genéticos citados anteriormente. O delineamento experimental utilizado para obtenção dos dados fenotípicos foi blocos casualizados. Foram avaliados dados de dois cortes em apenas um local.

Sendo assim, a análise de modelos mistos foi realizada considerando diferentes estruturas para as matrizes de variância e covariância (VCOV) para os efeitos genético ($\mathbf{G}$) e não-genético ($\mathbf{R}$). Para isso, foi considerado o seguinte modelo, ajustado para cada caráter separadamente (termos sublinhados correspondem aos efeitos aleatórios):

$$\underline{y}_{ijk} = \mu + C_j + B_k + CB_{jk} + \underline{G}_{ij} + \underline{\varepsilon}_{ijk} \tag{7}$$

em que $\underline{y}_{ijk}$ é o fenótipo do i-ésimo genótipo, no bloco j e corte k; μ é a média do caráter; C_j é o efeito do corte j; B_k é o efeito do bloco k; CB_{jk} é o efeito da interação entre corte e bloco; $\underline{G}_{ij}$ é o efeito genético do genótipo i no bloco j; e $\underline{\varepsilon}_{ijk}$ é o efeito residual (não-genético).

Os diferentes modelos para estrutura de VCOV foram comparados com base nos valores de AIC (Akaike Information Criterion)(AKAIKE, 1974). A matriz $\mathbf{R}$ foi obtida a partir do produto direto (produto de Kronecker) das matrizes $\mathbf{R^C}$ (efeitos residuais entre cortes) e $\mathbf{R^B}$ (efeitos residuais entre blocos).

Uma vez modeladas as matrizes de variância-covariância $\mathbf{G}$ e $\mathbf{R}$, foi possível a obtenção dos BLUPs (Best Linear Unbiased Predictor) para cada caráter analizado e posteriormente, os componentes da variação foram estimados utilizando o método da máxima verossimilhança restrita. Todas as análises descritas anteriormentes foram realizadas no software Genstat (PAYNE et al., 2009).

3.2.3 Preditores Genéticos

O primeiro passo para a estimativa dos preditores genéticos ao longo de todo o genoma, foi realizado com o cálculo da probabilidade condicional do genótipo do QTL considerando o mapa estimado e os preditores genéticos observados.

Considere uma progênie de irmãos completos resultante do cruzamento entre dois indivíduos não aparentados heterozigotos P e Q e ainda que tais indivíduos possuem dois alelos que segregam para cada loco. De acordo com Lin et al. (2003), ao considerar um intervalo definido por dois marcadores adjacentes m e $m+1$, e que dois alelos segregam para cada loco, sendo 1 e 2 os alelos de determinado loco, é possível representar os genótipos destes locos por $P_m^{\{1,2\}}$, $Q_m^{\{1,2\}}$, $P_{m+1}^{\{1,2\}}$ e $Q_{m+1}^{\{1,2\}}$ (Figura 1). Ainda, assumindo a presença de um QTL localizado entre tais marcadores

$$
\begin{array}{c|c|ccc|c}
 & P & & & Q & \\
P_m^1 & P_m^2 & & Q_m^1 & & Q_m^2 \\
P^1 & P^2 & \times & Q^1 & & Q^2 \\
P_{m+1}^1 & P_{m+1}^2 & & Q_{m+1}^1 & & Q_{m+1}^2
\end{array}
$$

Figura 1 – Representação do cruzamento entre os indivíduos não aparentados heterozigóticos P e Q. $P_m^{\{1,2\}}$, $Q_m^{\{1,2\}}$, $P_{m+1}^{\{1,2\}}$ e $Q_{m+1}^{\{1,2\}}$ são os alelos dos marcadores nos locos m e $m+1$; P^1, P^2, Q^1 e Q^2 são os alelos do QTL

adjacentes, é observada a segregação de quatro classes genotípicas dos QTLs na progênie, a saber, P^1Q^1, P^1Q^2, P^2Q^1 e P^2Q^2 com proporção de segreção igual a 1:1:1:1.

Deste modo, é possível estimar três contrastes ortogonais entre as médias das classes genotípicas esperadas:

$$
\begin{array}{cccccccc}
+ & P^1Q^1 & + & P^1Q^2 & - & P^2Q^1 & - & P^2Q^2 \\
+ & P^1Q^1 & - & P^1Q^2 & + & P^2Q^1 & - & P^2Q^2 \\
+ & P^1Q^1 & - & P^1Q^2 & - & P^2Q^1 & + & P^2Q^2
\end{array}
$$

Sendo que os dois primeiros contrastes representam os efeitos aditivos dos QTLs nos genitores P e Q respectivamente, enquanto que o terceiro contraste representa o efeito de dominância ou interação intraloco entre os efeitos aditivos de cada genitor. Embora tal abordagem foi proposta para espécies diplóides, esta pode ser expandida à espécies poliplóides, como é o caso da cana-de-açúcar e consequentemente ao presente estudo uma vez que foram cosiderados apenas marcadores em dosagem única (GAZAFFI, 2009). Sendo assim, tais contrastes representam a presença ou ausência dos alelos do QTL.

Porém, uma vez que o genótipo do QTL não é observado, foi necessário estimar as probabilidades condicionais dos genótipos dos QTLs não observados baseado nos marcadores genéticos que constituem o mapa genético estimado, o que serviu de base para a realização das imputações. Neste trabalho, as probabilidades condicionais foram estimadas por meio da abordagem multiponto via Cadeia de Markov Oculta (HMM), utilizando o pacote *OneMap* (MARGARIDO; SOUZA; GARCIA, 2007). Considerando o mapa de ligação estimado e os marcadores observados, foram obtidas quatro probabilidades condicionais, a saber, P^1Q^1, P^1Q^2, P^2Q^1 e P^2Q^2, para cada localização no genoma distanciadas a no máximo 1 cM, ao longo de todo o genoma.

3.2.4 Imputação Múltipla de Pseudomarcadores

Uma vez que as quatro possíveis probabilidades dos genótipos dos marcadores não observados foram obtidas em todas as localizações no genoma, como citado anteriormente, foi possível simular o genótipo dos marcadores não observados por meio da imputação de pseudomarcadores.

Ao avaliar o padrão de não observação dos dados moleculares do presente estudo, dispostos em um mapa de ligação, verificou-se que os dados apresentam o padrão não monotônico de não observação, uma vez que tal padrão é dado de forma aleatória. Em relação ao mecanismo de não observação verificou-se que a distribuição dos dados não observados de um determinado marcador, não depende dos demais marcadores (tanto observados como não observados), sendo classificado como mecanismo de não observação completamente aleatório (MCAR). Considerando portanto a classificação do mecanismo e do padrão de não observação dos dados moleculares a técnica da imputação múltipla é indicada para a recomposição de marcadores não observados (SEN; CHURCHILL, 2001).

Para a obtenção de pseudomarcadores por meio das imputações realizou-se amostragens condicionadas às quatro possíveis probabilidades de ocorrência do genótipo não observado, em todas as localizações ao longo do genoma. Para isso, em cada localização realizou-se dez amostragens, totalizando dez conjuntos de pseudomarcadores imputados para cada posição no “grid” de espaçamento ao longo do genoma. Deste modo, obteve-se dez conjunto de dados moleculares completados (contendo pseudomarcadores e marcadores observados), os quais serviram como base para o mapeamento de QTLs de acordo com o modelo genético de mapeamento assumido. Para cada um dos conjuntos de dados moleculares completados, realizou-se o mapeamento de QTLs, obtendo portanto, dez resultados de mapeamento para cada uma das localizações consideradas.

Então, para cada localização u no “grid” do genoma, em cada imputação i ($i = 1, \cdots, 10$), foi calculado um peso de acordo com a seguinte equação:

$$W_G(r_i(u)) = p(y \mid g = r_i(u)) \tag{8}$$

onde, $W_G(r_i(u))$ é o peso da imputação i na localização u; y: o valor fenotípico dos indivíduos que constituem a população de mapeamento e $g = r_i(u)$: o genótipo imputado dos indivíduos da imputação i na localização u no "grid"do genoma.

Porém, a $p(y \mid g = r_i(u))$ é na verdade uma distribuição de probabilidade geral baseada da

distribuição que os dados fenotípicos y dos indivíduos da população apresentam. E então, para cada imputação i em cada localização u, cada indivíduo apresentou uma função densidade de probabilidade (f.d.p.) específica baseada na distribuição dos dados fenotípicos y. Sendo assim, o peso para cada imputação i em cada localização u foi resultado do produtório da f.d.p. de cada j indivíduo $(j = 1, \cdots, n)$ da população:

$$p(y \mid g = r_i(u)) = \prod_{j=1}^{n} f.d.p._j \tag{9}$$

Vale ressaltar que o valor de u variou de acordo com o tamanho de cada grupo de ligação estimado.

Após a obtenção do peso para cada imputação i $(i = 1, \cdots, m)$ em cada localização u, foi estimado o peso médio de todas as imputações para cada localização u, como proposto por Sen e Churchill (2001):

$$\sum_{i=1}^{m} W_G(r_i(u)) \tag{10}$$

Então, a combinação dos dez resultados de mapeamento, considerando todas as imputações para dada localização u foi obtida por meio dos seguintes passos: $i)$ para cada um dos efeitos estimados por meio do modelo de mapeamento, calculou-se a ponderação do efeito para cada i imputação multiplicando-se o valor do efeito pelo peso (Equação 8) de cada imputação, obtendo-se portanto, dez efeitos ponderados para cada localização u. $ii)$ Posteriormente, realizou-se o somatório dos efeitos ponderados de todas as imputações para cada localização u. $iii)$ Por fim, obteve-se a razão entre o somatório dos efeitos ponderados de todas as imputações e o peso médio (Equação 10) de todas as imputações.

Deste modo, foi possível obter a média ponderada dos efeitos considerando todas as imputações para cada localização e consequentemente inferir a existência de possíveis QTLs baseado no modelo genético de mapeamento especificado.

3.2.5 Modelo de Mapeamento

O modelo utilizado para o mapeamento de QTLs foi o mapeamento por intervalo (IM) (LANDER; BOTSTEIN, 1989), baseado na análise de regressão múltipla, o qual considera intervalos entre pares de marcadores adjacentes e é testada a existência de um possível QTL em cada localiza-

ção do intervalo em estudo. O modelo foi uma adaptação do modelo de mapeamento proposto por Gazaffi (2009), sem a inclusão de cofatores na análise, de acordo com a equação a seguir:

$$y_j = \mu + \gamma + \alpha_p^* x_{pj}^* + \alpha_q^* x_{qj}^* + \delta_{pq}^* x_{pj}^* x_{qj}^* + \varepsilon_j \tag{11}$$

em que y_j: é o fenótipo do j-ésimo indivíduo $(j = 1, \cdots, n)$; μ: é o intercepto da regressão linear múltipla; γ: é um vetor de parâmetros contendo os coeficientes de regressão linear múltipla (α_p, α_q, δ_{pq}); x_{pj}^* e x_{qj}^*: são as variáveis indicadoras para o genótipo do QTL na progênie; α_p^*, α_q^*, δ_{pq}^*: são os efeitos aditivos para os genitores P e Q e o efeito de dominância, respectivamente.

Os valores que x_{pj}^* e x_{qj}^* assumiram de acordo com o genótipo do QTL observado na progênie foram:

$$x_{pj}^* = \begin{cases} 1 & se \quad P^1Q^1 \\ 1 & se \quad P^1Q^2 \\ -1 & se \quad P^2Q^1 \\ -1 & se \quad P^2Q^2 \end{cases} \quad ; \qquad x_{qj}^* = \begin{cases} 1 & se \quad P^1Q^1 \\ -1 & se \quad P^1Q^2 \\ 1 & se \quad P^2Q^1 \\ -1 & se \quad P^2Q^2 \end{cases}$$

Em cada localização no “grid” do genoma espaçadas a no máximo 1 cM foram obtidas estatísticas de razão de verossimilhança (LRT) convertidas em LOD Score, afim de verificar a presença do QTL na posição considerada. A hipótese de nulidade testada foi H_0: $\alpha_p^* = \alpha_q^* = \alpha_{pq}^* = 0$. Após a identificação dos QTLs, foram verificadas as significâncias dos efeitos α_p^*, α_q^*, δ_{pq}^*.

Considerando que para a detecção de QTLs houve a realizaçao de múltiplos testes, o valor crítico (*Threshold*), foi obtido de modo a evitar o erro tipo I. Para tanto, foi considerado o número de testes independentes ao longo do genoma, aplicando-se a correção de Bonferroni para obter um nível de 5% de significância global (VIEIRA et al. 2000). Já para a estimativa do valor crítico dos efeitos α_p^*, α_q^*, δ_{pq}^*, foi realizado o teste estatístico de qui-quadrado(χ^2) para um grau de liberdade, considerando 5% como nível de significância. Isso é possível uma vez que tal teste é realizado apenas nas posições com evidência de QTLs e não em todo o genoma, não requerendo portanto correção para múltiplos testes. O limiar obtido foi convertido para LOD Score. A proporção da variância fenotípica (R^2) explicada para cada QTL mapeado individualmente para determinado caráter, foi obtida considerando todos os efeitos simultaneamente. A realização das análises citadas anteriormente foram elaboradas e executadas no ambiente de programação R (R Developmente Core Team, 2009).

4 RESULTADOS

4.1 Mapa Genético de Ligação

O mapa de ligação construído foi constituído por 113 grupos de ligação (APÊNDICE A). Esses grupos de ligação corresponderam à 409 das 634 marcas que apresentaram segregação em dosagem única, ou seja 64,5% do total de marcas analisadas. O comprimento final do mapa de estimado foi 4370 cM, sendo que a distância média entre marcadores adjacentes foi de aproximadamente 10,7 cM.

O maior grupo de ligação possuiu 12 marcas agrupadas e apresentou o comprimento de 142,9 cM, enquanto que o menor grupo de ligação, apresentou duas marcas agrupadas, localizadas na mesma posição dentro do grupo de ligação. 62 dos 113 grupos de ligação, ou seja, 54,9% do total, apresentaram comprimento menor que 35 cM, sendo que tais grupos em sua maioria, foram formados pelo agrupamento de quatro, três ou duas marcas respectivamente. Quatro grupos de ligação apresentaram comprimento maior que 100 cM, representando 11% do comprimento total do mapa de ligação estimado. Esses grupos agruparam 13, 12, 11 e dez marcas respectivamente e foram os grupos que apresentaram o maior número de marcas agrupadas dentro de determinado grupo de ligação.

Como é esperado observar em mapas genéticos de cana-de-açúcar, o presente mapa apresentou baixa saturação e consequentemente baixa cobertura do genoma, o que pode ser ocasionado pela utilização apenas de marcadores em dosagem única.

4.2 Análises Fenotípicas

Baseado na análise de modelos mistos foram comparadas diferentes estruturas das matrizes de variâncias e covariâncias (VCOV) para os efeitos genéticos e não genéticos. Assim, a estrutura de VCOV com menor valor de AIC, corresponde a mais apropriada aos dados (Tabela 1).

Para todos os caracteres avaliados, o melhor modelo para a matriz $\mathbf{G}$ foi o Autoregressivo de Primeira Ordem, considerando heterogeneidade de variâncias e correlações genéticas entre cortes, que decrescem a medida que aumente a distância temporal entre cortes. Para os caracteres Fibra e POL, para ambas matrizes $\mathbf{R^C}$ e $\mathbf{R^B}$ foi considerado o modelo Simetria Composta onde as va-

Tabela 1 – Diferentes modelos para a matriz de variâncias-covariâncias genéticas ($\boldsymbol{G}$) e não-genéticas ($\boldsymbol{R}$) e valores correspondentes de AIC (Critério de Informação de Akaike)

Caráter	Modelo Matriz $\boldsymbol{G}$	Modelo Matriz $\boldsymbol{R}$	n_{PAR}	AIC
Fibra	ID	ID $\otimes$ ID	2	893,18
Fibra	DIAG	DIAG $\otimes$ DIAG	6	883,07
Fibra	AR1$_{Het}$	CS $\otimes$ CS	7	**782,83**
POL	ID	ID $\otimes$ ID	2	922,52
POL	DIAG	DIAG $\otimes$ DIAG	6	922,45
POL	AR1$_{Het}$	CS $\otimes$ CS	7	**859,70**
PC	ID	ID $\otimes$ ID	2	4419,87
PC	DIAG	DIAG $\otimes$ DIAG	6	4395,32
PC	AR1$_{Het}$	CS$_{Het}$ $\otimes$ CS$_{Het}$	9	**4365,50**
PP	ID	ID $\otimes$ ID	2	1340,98
PP	DIAG	DIAG $\otimes$ DIAG	6	1316,95
PP	AR1$_{Het}$	CS$_{Het}$ $\otimes$ CS$_{Het}$	9	**1290, 74**

$\boldsymbol{G}$: matriz de variâncias-covariâncias genéticas; $\boldsymbol{R}$: matriz de variâncias-covariâncias não-genéticas; DIAG: diagonal; CS$_{Het}$: simetria composta heterogênea; AR1$_{Het}$: autoregressivo de primeira ordem; n_{PAR}: número total de parâmetros no modelo. Os menores valores de AIC são evidenciados em **negrito**

riâncias genéticas são homogêneas enquanto que as covariâncias genéticas são comuns. Já para os caracteres PC e PP, foi considerado o modelo Simetria Composta Heterogênea onde as variâncias genéticas são consideradas heterogêneas enquanto que a covariância genética é comum para ambas matrizes $\mathbf{R^C}$ e $\mathbf{R^B}$ (SMITH et al., 2007; MALOSETTI et al., 2008; PASTINA, 2010). Deste modo, a obtenção dos BLUPs para todos os caracteres avaliados individualmente possibilitou a estimativa eficiente dos parâmetros descritos a seguir.

Considerando os quatro caracteres avaliados, relacionados aos parâmetros estimados (Tabela 2), foi verificado que as estimativas das variâncias genotípica e fenotípica apresentaram boa acurácia, uma vez que foram observados intervalos de confiança relativamente pequenos. A partir das estimativas dos coeficientes de herdabilidade no sentido amplo, considerando a média dos indivíduos nos dois cortes avaliados, é notável a presença de valores altos e intermediários que variaram entre 82,36% para Fibra; 59,96% para POL; 19,92% para PC e 23,93% para PP. Os valores observados

para os caracteres POL, PC e PP podem ser atribuídos à baixa precisão experimental uma vez que para tais caracteres, a avaliação experimental foi realizada em apenas um local considerando duas repetições. Já o alto coeficiente de herdabilidade observado para o caráter Fibra pode ser atribuído tanto à natureza do caráter quanto à melhor precisão experimental pois apenas para este caráter em questão, foi considerada a avaliação em um local e quatro repetições, reduzindo assim, o erro experimental. Uma vez que dentre os caracteres avaliados, Fibra e POL foram os que apresentaram maiores valores de coeficiente de herdabilidade, é esperado que um maior número de QTLs seja detectado para esses caracteres.

Tabela 2 – Estimativas dos parâmetros: coeficiente de herdabilidade no sentido amplo (h^2); variância genotípica ($\hat{\sigma}^2_G$); variância fenotípica ($\hat{\sigma}^2_P$), com respectivos intervalos de confiança a 95% (IC)

	Caráter[a]			
Parâmetro	Fibra	POL	PC	PP
h^2	82,36	59,96	19,92	23,93
IC	86,28;78,24	62,96;56,96	20,92;18,92	25,13;22,73
$\hat{\sigma}^2_G$	0,52	0,43	11,31	0,35
IC	0,54;0,49	0,45;0,41	11,87;10,72	0,36;0,34
$\hat{\sigma}^2_P$	0,63	0,71	56,75	1,45
IC	0,66;0,60	0,75;0,71	59,59;53,91	1,52;1,37
CV(%)[b]	5,20	3,46	5,98	23, 59
Média	12,86	15,81	11,67	1,88

[a] Fibra: porcentagem de fibra; POL: conteúdo de sacarose; PC: produção de cana por parcela (kg parcela^{-1}); PP: produção de açúcar por parcela (kg parcela^{-1}). [b] Coeficiente de Variação

Em relação os coeficientes de variação estimados, foi observado que em geral, apresentam valores baixos variando entre 3,46% à 23,59%. Apenas o caráter PP apresentou desvio significativo do coeficiente de variação estimado em relação aos demais caracteres, porém, o valor observado é aceitável. Em relação às estimativas da variância genética obtidas para os caracteres avaliados, foi verificado que para o caráter PC, o valor observado foi alto quando comparado com os demais caracteres. O mesmo comportamento para este caráter é observado quando avaliada a variância fenotípica existente. Tal comportamento é esperado e pode ser justificado pela natureza quantitativa do caráter, pois o caráter PC, dentre os caracteres avaliados, é o que mais sofre influência dos efeitos ambientais. Como era esperado, o caráter Fibra foi o que apresentou menor valor de variância fenotípica.

4.3 Mapeamento de QTLs

O valor limiar para o LOD Score (*Threshold*) baseado na estimativa que considerou o número de testes independentes ao longo do genoma e a correção de Bonferroni, foi semelhante para todos os caracteres avaliados sendo que, o valor assumido foi 3,90 considerando 5% como nível de significância. Já o valor crítico para a significância dos efeitos foi igual a 0,83, de modo que LODs marginais acima deste limiar foram considerados significativos, indicando significância dos efeitos correspondentes.

O resultado de mapeamento, foi obtido eficientemente por meio da média ponderada, usada como peso dos resultados das dez imputações múltiplas realizadas. Ao total, 57 QTLs foram mapeados para os caracteres Fibra, POL, PC e PP (Tabela 3, Figuras 2-5). O número de QTLs mapeados foi considerado em geral alto e é um valor esperado uma vez que o mapa genético estimado apresentou comprimento final de 4370 cM e 113 grupos de ligação. O maior número de QTLs foi identificado para o caráter POL, enquanto que o menor número de QTLs foi identificado para os caracteres PC e PP. Esses resultados eram esperados de acordo com os valores obtidos na estimativa dos coeficientes de herdabilidade.

Para o caráter Fibra foram mapeados 14 QTL, sendo estes localizados nos grupos de ligação 20, 1, 40, 28, 42, 16, 12, 5, 95, 66, 55, 108, 38 e 93 respectivamente (Figura 2). O QTL I localizado no grupo de ligação 20, foi o que apresentou maior LOD Score (5,56) e maior valor de R^2 (12,6%). O menor valor de LOD Score para o caráter em questão foi 3,82, observado nos QTLs II e XIV localizados nos grupos de ligação 1 e 5, respectivamente. Embora tais QTLs não apresentaram valores de LOD Score superior ao limite crítico, foram incluídos nos resultados uma vez que a curva de mapeamento apresentada por esses QTLs, evidencia a existência de um QTL putativo para as correspondentes localizações. Tal inclusão pode ser realizada uma vez que o presente estudo trata-se de uma análise inicial. Sendo assim, a presença de QTLs falsos-positivos não ocasiona grandes problemas. Os QTLs localizados nos grupos de ligação 40, 16, 95 e 93, estão localizados nas regiões proximais dos grupos de ligação enquanto que, nos grupos de ligação 66 e 42, os QTLs V e X estão localizados na região distal. O valor de R^2 explicado para cada QTL individualmente apresentou variação entre 12,6%-8,9%. Foi observado o predomínio de efeitos aditivos quando comparados com os efeitos de dominância.

O caráter POL foi o que apresentou o maior número de QTLs mapeados (19 QTLs). Os QTLs I ao XIX foram mapeados nos grupos de ligação 20, 44, 33, 65, 40, 70, 35, 28, 42, 3, 32, 69, 14,

57, 66, 75, 92, 38 e 73 respectivamente (Figura 3). Os valores de LOD Score considerando os efeitos aditivos e de dominância simultaneamente, variaram entre 6,42-3,81. O maior valor de LOD Score foi associado ao QTL XIII localizado à 10 cM no grupo de ligação 10. Já o menor LOD foi observado no QTL XVI localizado no grupo de ligação 29. Embora o LOD do QTL XVI não tenha ultrapassado o limiar critíco de 3,90; a curva de mapeamento apresentada para este grupo de ligação evidencia a presença de um possível QTL localizado a 29 cM do início do grupo de ligação, sendo portanto, uma referência para estudos futuros visando a identificação de um QTL putativo nesta localização. Interessantemente, foi observado que os QTLs localizados nos grupos de ligação 3, 65, 70, 35, 28 ,3, 69, 14, 57, 75, 92 e 38 estavam situados a menos de 5 cM de marcadores observados no mapa de ligação. O QTL I, localizado no grupo de ligação 20 (48 cM) juntamente com o QTL VIII no grupo de ligação 28 (2 cM) foram os únicos que apresentaram significância estatística para os três efeitos em estudo (α_p^*, α_q^* e δ_{pq}^*). De modo geral, houve predominância dos efeitos aditivos quando comparados com os efeitos de dominância, sendo que tais efeitos aditivos variaram entre -0,16 e 0,18.

O caráter PC apresentou o menor número de QTLs mapeados, totalizando 12 QTLs (Figura 4). Os QTLs I ao XII foram mapeados nos grupos de ligação 6, 67, 42, 23, 18, 7, 66, 26, 50, 59, 61 e 37 respectivamente. O maior valor do LOD Score para os QTLs em questão foi de 6,32, referente ao QTL VI mapeado no grupo de ligação 7. O valor de R^2 para esse QTL foi de 14,17%. O menor valor de LOD Score e R^2 foram referentes aos QTLs IX e X sendo 3,94 para LOD Score em ambos os QTLs e 9,13 e 9,12, respectivamente para R^2. Igualmente ao observado para os caracteres anteriormente descritos, houve significância de efeitos aditivos e de dominância, havendo predominância dos efeitos aditivos de ambos os genitores avaliados. O QTL X localizado à 9 cM no grupo de ligação 59, foi o único QTL a apresentar significância em todos os três efeitos avaliados.

Finalmente, para o caráter PP foram localizados 12 QTLs, sendo estes localizados nos grupos de ligação 67, 42, 22, 12, 48, 7, 43, 95, 66, 84, 52 e 51. Tal caráter foi o que apresentou QTL com o maior valor de LOD Score (7,52), localizado no grupo de ligação 7. Ao observar a curva de mapeamento (Figura 5) do grupo de ligação 7 fica evidenciada a presença de um QTL putativo localizado a 17 cM. Os valores de R^2 variaram entre 16,61%-6,49%, sendo o menor valor de R^2 associado ao QTL XII. O menor valor de LOD Score (3,99) está associado ao QTL X; localizado à 25 cM no grupo de ligação 84. Houve predominância de efeitos aditivos significativos, sendo que estes variaram entre -0,23 e 0,24. É importante ressaltar que o alto coeficiente de variação associado

ao caráter em questão pode ocasionar a redução do poder estatístico da detecção de QTLs.

Foi verificado o mapeamento de QTLs em localizações próximas para os diferentes caracteres, dentro do mesmo grupo de ligação. Essas detecções foram observadas no grupo de ligação 42 onde, para Fibra foi mapeado o QTL V (56 cM); POL, o QTL IX (70 cM); PC, o QTL III (57 cM); PP, o QTL II (54 cM). No grupo de ligação 66 foi mapeado para o caráter Fibra o QTL X (32 cM); POL, o QTL XV (23 cM); PC, o QTL VIII (29 cM) e PP, o QTL IX (29 cM). No grupo de ligação 7 foram detectados dois QTLs para os caracteres PC e PP, ambos localizados na posição 17 do grupo de ligação. No grupo de ligação 12 foram mapeados 2 QTLs para Fibra e PP localizados na posição 20 e 28 cM, respectivamente. Já no grupo de ligação 95, dois QTLs foram identificados para o caráter Fibra e PP, localizados nas posições 2 e 5 cM. Essas associações são de grande importância uma vez que há indicativos da existência de correlações entre caracteres para aqueles QTLs localizados dentro do mesmo grupo de ligação, citados anteriormente. Estudos posteriores precisam ser conduzidos afim de verificar se tais QTLs estão ligados ou trata-se de QTLs pleoitrópicos.

Tabela 3 – Estimativas dos efeitos dos QTLs considerando o modelo de mapeamento por intervalo utilizando a abordagem da imputação múltipla (Fibra: porcentagem de fibra; POL: conteúdo de sacarose; PC: produção de cana por parcela; PP: produção de açúcar por parcela; α_p^*, α_q^* e δ_{pq}^* são os efeitos aditivos dos genitores P e Q e o efeito de dominância, respectivamente, da média ponderada das imputações, específicos para cada QTL. Valores em negrito foram significativos)

(Continua)

Posição					Efeito							
Caráter	QTL	GL	cM	LOD	μ	α_p^*	LOD	α_q^*	LOD	δ_{pq}^*	LOD	$R^2\%$
	I	20	14	5,56	12,87	**0,19**	4,10	0,03	0,14	**0,15**	2,66	12,6
	II	1	31	3,82	12,87	**0,19**	3,99	-0,05	0,26	-0,04	0,19	8,9
	III	40	3	4,14	12,91	**-0,14**	2,00	**-0,17**	2,94	0,08	0,64	9,6
	IV	28	16	4,05	12,89	**0,17**	3,05	**-0,12**	1,61	0,02	0,05	9,4
	V	42	56	4,37	12,86	0,07	0,70	-0,06	0,41	**0,20**	4,23	10,1
	VI	16	3	4,48	12,87	-0,07	0,59	**0,19**	4,14	-0,06	0,40	10,3
Fibra	VII	12	20	4,34	12,88	-0,05	0,29	0,02	0,07	**0,20**	4,41	10,0
	VIII	5	36	4,28	12,85	-0,05	0,31	**0,17**	3,43	**-0,09**	0,99	9,9
	IX	95	2	3,93	12,87	-0,05	0,30	**-0,19**	3,82	**0,11**	1,21	9,1
	X	66	32	4,30	12,83	-0,04	0,38	**0,11**	1,41	**-0,16**	2,89	9,9
	XI	55	10	4,42	12,86	0,05	0,27	0,03	0,12	**0,21**	4,80	10.2
	XII	108	7	4,64	12,81	0,08	0,75	**0,18**	3,45	**-0,15**	2,41	10.6
	XIII	38	8	4,24	12,81	**0,21**	4,52	**0,11**	1,24	-0,03	0,07	9.8
	XIV	93	5	3,82	12,88	**0,10**	1,07	**-0,17**	2,99	0,04	0,25	8.9
	I	20	48	3,99	15,79	**-0,11**	1,83	**0,09**	1,47	**-0,08**	0,97	9,22
	II	44	72	5,19	15,80	0,06	0,65	**0,13**	2,89	**0,09**	1,47	11.83
	III	33	47	4,77	15,81	-0,04	0,27	**-0,15**	3,73	**0,08**	1,11	10,93
	IV	65	81	4,11	15,80	-0,03	0,14	-0,06	0,55	**0,15**	3,88	9,49
	V	40	36	5,31	15,77	**0,14**	3,28	0,04	0,32	**-0,16**	4,00	12,06
	VI	70	41	4,00	15,81	-0,03	0,17	0,03	0,12	**-0,16**	4,13	9,25
	VII	35	16	3,98	15,79	0,01	0,01	-0,01	0,02	**-0,16**	4,44	9,22
	VIII	28	2	5,16	15,81	**-0,08**	1,18	**0,09**	1,28	**-0,15**	3,65	11,76
	IX	42	70	4,29	15,82	-0,03	0,21	**-0,08**	1,06	**-0,15**	3,76	9.88
POL	X	3	21	4,65	15,82	-0,04	0,28	**0,15**	3,75	**0,09**	1,38	10,67
	XI	32	21	3,98	15,79	**0,14**	3,01	**-0,07**	0,93	0,06	0,58	9.22
	XII	69	22	4,69	15,80	**0,18**	5,25	-0,02	0,05	0,01	0,02	10,75
	XIII	14	10	6,42	15,82	**-0,12**	2,39	**0,16**	4,46	0,06	0,67	14,38
	XIV	57	12	5,83	15,78	**-0,10**	1,59	0,01	0,01	**0,16**	4,51	13,17
	XV	66	23	4,05	15,83	**0,16**	4,28	-0,03	0,18	**-0,08**	0,99	9,37
	XVI	75	29	3,81	15,81	0,02	0,09	0,03	0,19	**0,16**	4,17	8,84
	XVII	92	2	3,97	15,82	-0,05	0,44	-0,03	0,17	**0,16**	4,00	9,19
	XVIII	38	19	4,43	15,85	**-0,10**	1,58	**-0,17**	4,37	0,03	0,21	10,19
	XIX	73	7	3,94	15,82	**-0,16**	3,80	0,05	0,43	0,03	0,11	9,13
	I	6	15	4,23	11,65	-0,04	0,07	**0,19**	1,33	**0,30**	3,18	9,77
	II	67	54	4,88	11,66	0,01	0,01	**0,27**	2,70	**-0,31**	3,62	11,15
PC	III	42	57	4,22	11,68	0,14	0,74	0,08	0,26	**-0,29**	3,01	9,73
	IV	23	0	4,06	11,71	**-0,17**	1,03	**-0,33**	3,83	0,05	0,54	9,39
	V	18	3	4,67	11,65	0,11	0,50	**0,18**	1,20	**0,30**	3,26	10,71

Tabela 3 – Estimativas dos efeitos dos QTLs considerando o modelo de mapeamento por intervalo utilizando a abordagem da imputação múltipla (Fibra: porcentagem de fibra; POL: conteúdo de sacarose; PC: produção de cana por parcela; PP: produção de açúcar por parcela; α_p^*, α_q^* e δ_{pq}^* são os efeitos aditivos dos genitores P e Q e o efeito de dominância, respectivamente, da média ponderada das imputações, específicos para cada QTL. Valores em negrito foram significativos)

(Conclusão)

	Posição				Efeito							
Caráter	QTL	GL	cM	LOD	μ	α_p^*	LOD	α_q^*	LOD	δ_{pq}^*	LOD	$R^2\%$
	VI	7	17	6,32	11,69	0,89	0,32	**-0,38**	5,41	**0,20**	1,59	14,17
	VII	66	29	4,75	11,68	**0,38**	5,20	-0,08	0,23	-0,09	0,32	10,88
	VIII	26	23	5,98	11,69	**-0,20**	1,50	0,13	0,68	**-0,38**	5,44	13,48
PC	IX	50	24	3,94	11,71	**-0,28**	2,92	0,04	0,06	**0,20**	1,42	9,13
	X	59	9	3,94	11,63	**0,20**	1,53	**0,17**	1,09	**-0,27**	2,64	9,12
	XI	61	6	4,02	11,66	-0,07	0,20	**0,18**	1,14	**-0,32**	3,69	9,29
	XII	37	1	4,70	11,66	0,02	0,06	0,03	0,07	**0,38**	5,32	10,78
	I	67	26	4,44	1,88	0,03	0,10	0,09	0,75	**-0,22**	4,42	10,21
	II	42	54	4,27	1,88	-0,07	0,40	0,03	0,10	**0,21**	3,81	9,84
	III	22	19	4,65	1,88	**-0,20**	3,53	**0,11**	1,14	**-0,11**	1,19	10,66
	IV	12	28	4,18	1,87	**0,14**	1,88	**-0,17**	2,73	-0,02	0,06	9,64
	V	48	36	4,21	1,90	**-0,18**	2,95	**-0,10**	0,85	-0,07	0,40	9,71
PP	VI	7	17	7,52	1,89	0,09	0,88	**-0,23**	5,21	**0,16**	2,64	16,61
	VII	43	1	4,51	1,90	-0,05	0,26	**0,17**	2,70	**0,12**	1,28	10,36
	VIII	95	5	4,04	1,88	-0,02	0,06	**-0,22**	4,48	0,01	0,01	9,34
	IX	66	29	4,51	1,88	**0,24**	5,07	0,01	0,01	-0,01	0,01	10,36
	X	84	25	3,99	1,87	0,04	0,12	0,01	0,01	**-0,22**	4,52	9,23
	XI	52	18	4,30	1,88	**-0,13**	1,53	**-0,16**	2,35	**0,10**	1,03	9,90
	XII	51	14	5,03	1,86	0,02	0,07	**0,18**	2,94	-0,05	0,34	6,49

GL: grupo de ligação; cM: distância do QTL considerando o primeiro marcador do grupo de ligação

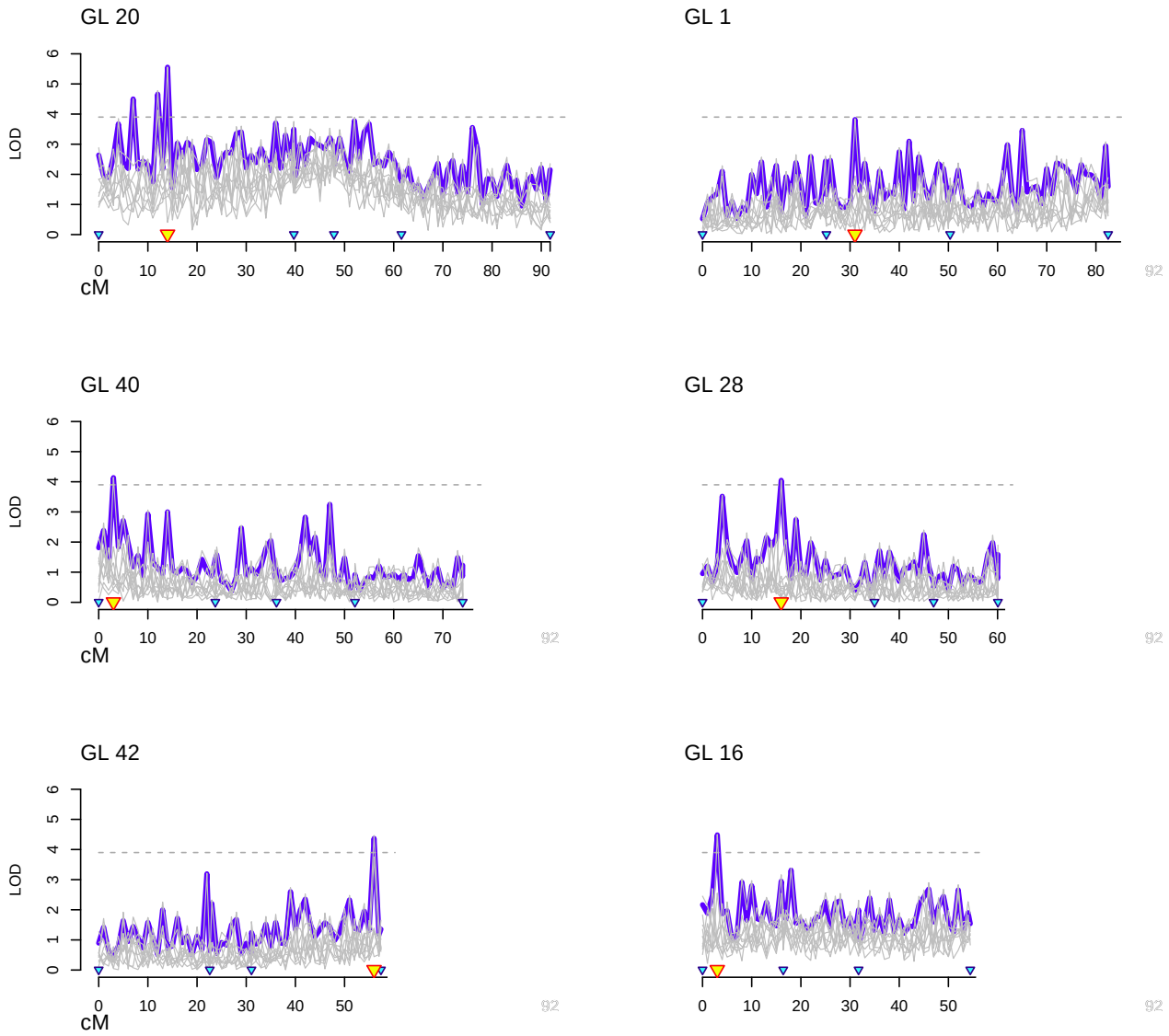


Figura 2 – Resultados do método de mapeamento por intervalo utilizando a abordagem da imputação múltipla para o caráter Fibra. Perfis em cinza indicam cada uma das imputações. Perfis azuis indicam a média ponderada de todas as imputações. Triângulos amarelos: posições dos QTLs. Triângulos azuis: posições dos marcadores moleculares no mapa de ligação. Valores das distâncias são expressos em cM

(Continua)

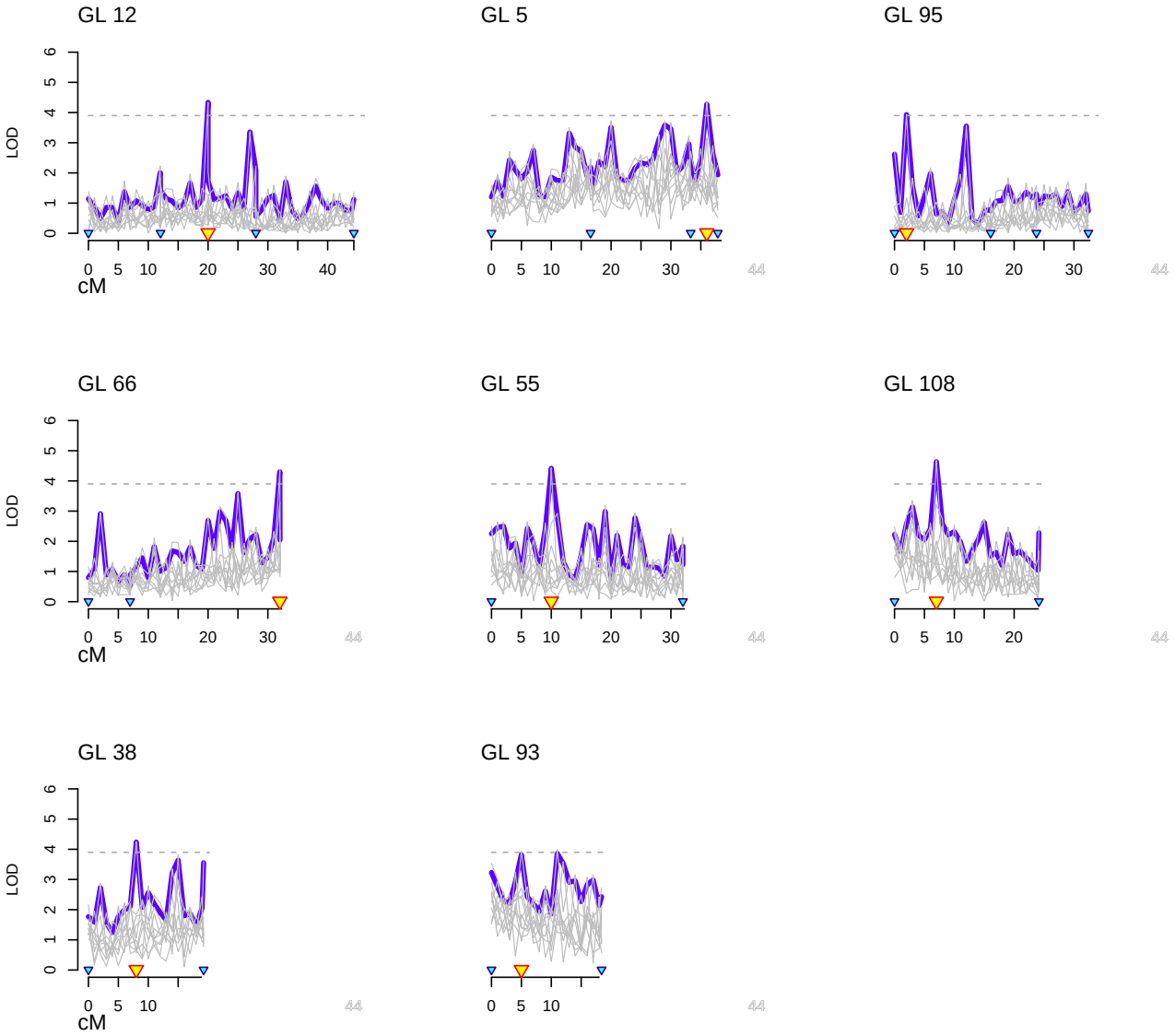


Figura 2 – Resultados do método de mapeamento por intervalo utilizando a abordagem da imputação múltipla para o caráter Fibra. Perfis em cinza indicam cada uma das imputações. Perfis azuis indicam a média ponderada de todas as imputações. Triângulos amarelos: posições dos QTLs. Triângulos azuis: posições dos marcadores moleculares no mapa de ligação. Valores das distâncias são expessos em cM

(Conclusão)

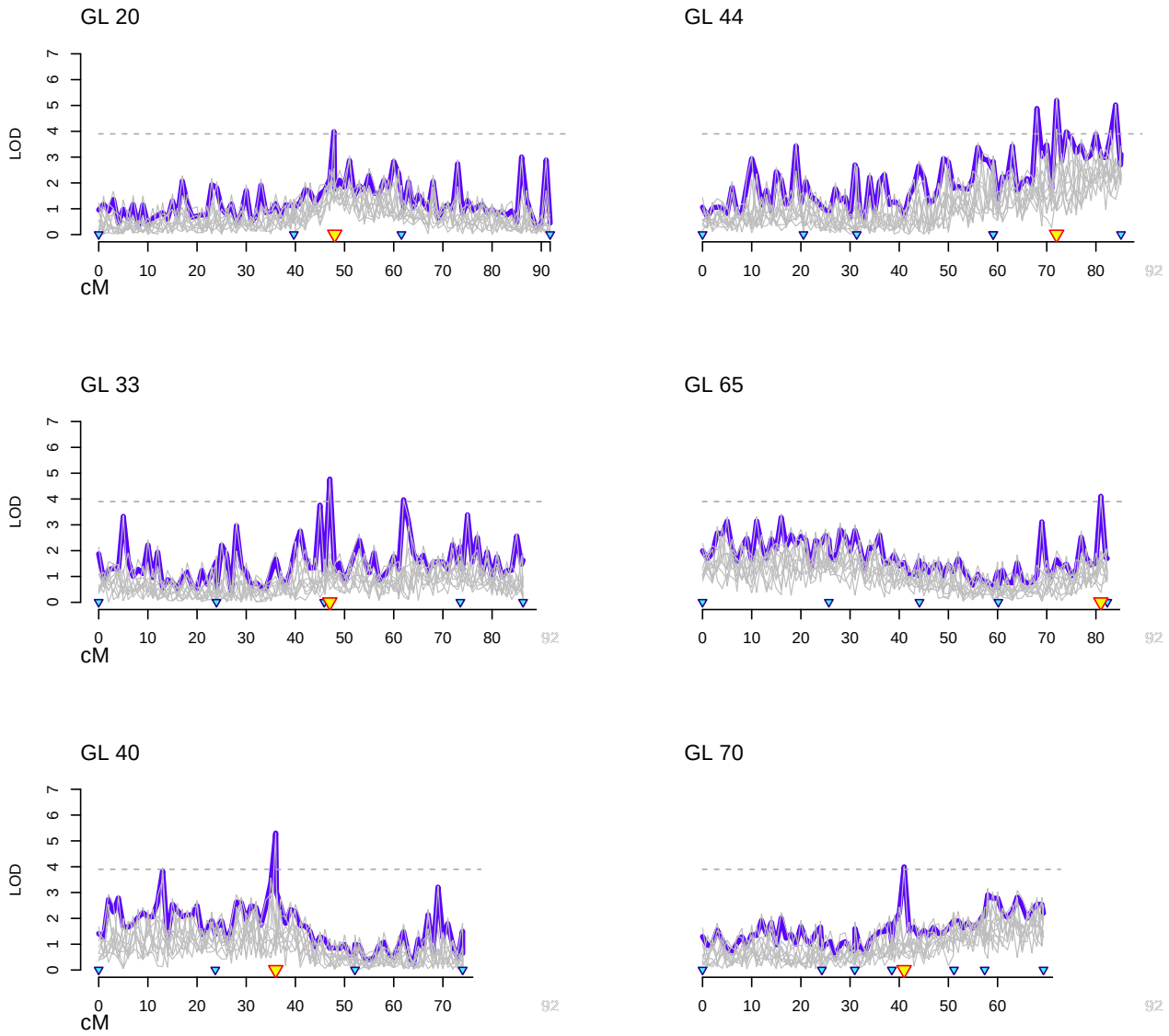


Figura 3 – Resultados do método de mapeamento por intervalo utilizando a abordagem da imputação múltipla para o caráter POL (conteúdo de sacarose). Perfis em cinza indicam cada uma das imputações. Perfis azuis indicam a média ponderada de todas as imputações. Triângulos amarelos: posições dos QTLs. Triângulos azuis: posições dos marcadores moleculares no mapa de ligação. Valores das distâncias são expressos em cM

(Continua)

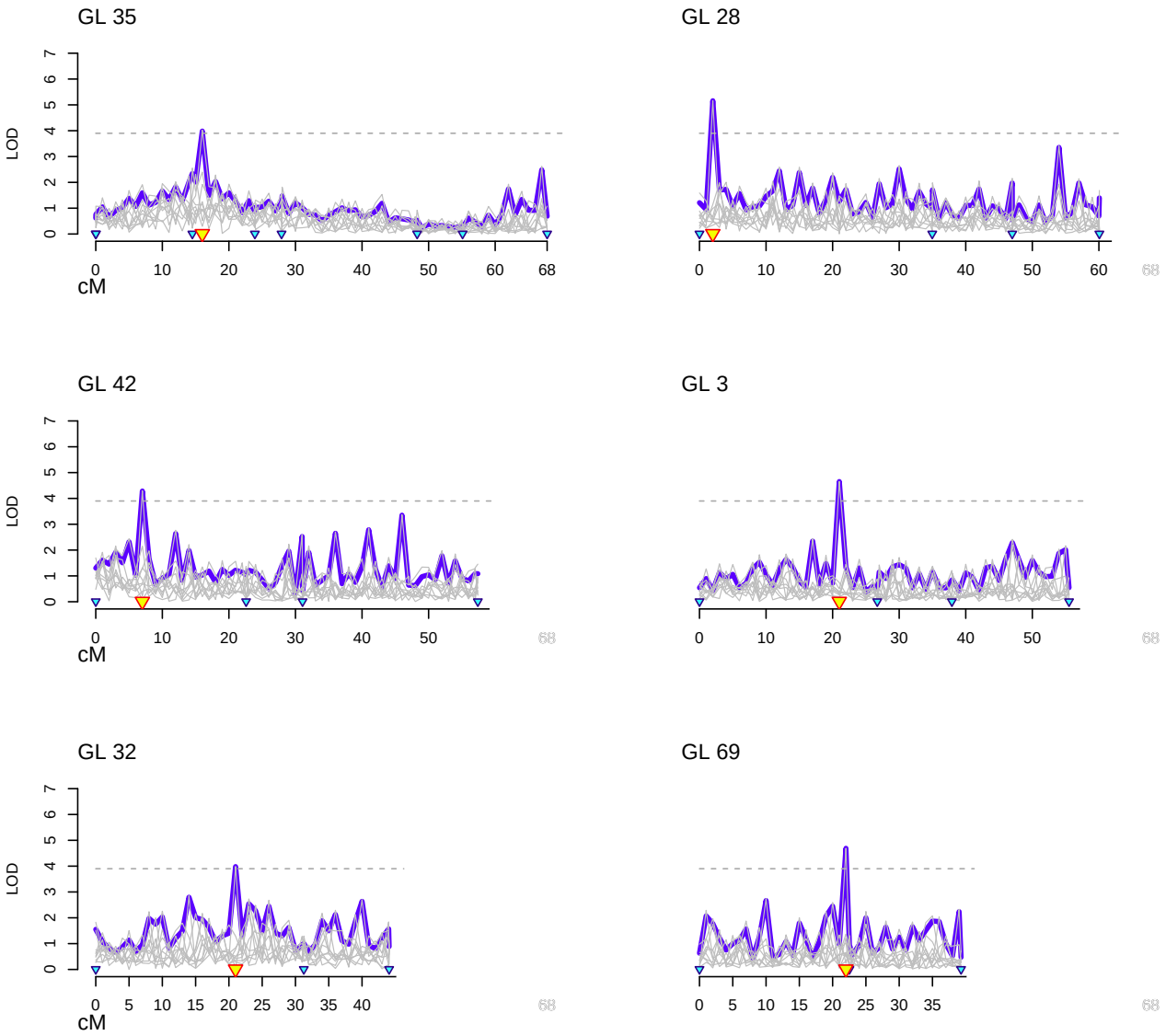


Figura 3 – Resultados do método de mapeamento por intervalo utilizando a abordagem da imputação múltipla para o caráter POL (conteúdo de sacarose). Perfis em cinza indicam cada uma das imputações. Perfis azuis indicam a média ponderada de todas as imputações. Triângulos amarelos: posições dos QTLs. Triângulos azuis: posições dos marcadores moleculares no mapa de ligação. Valores das distâncias são expressos em cM

(Continuação)

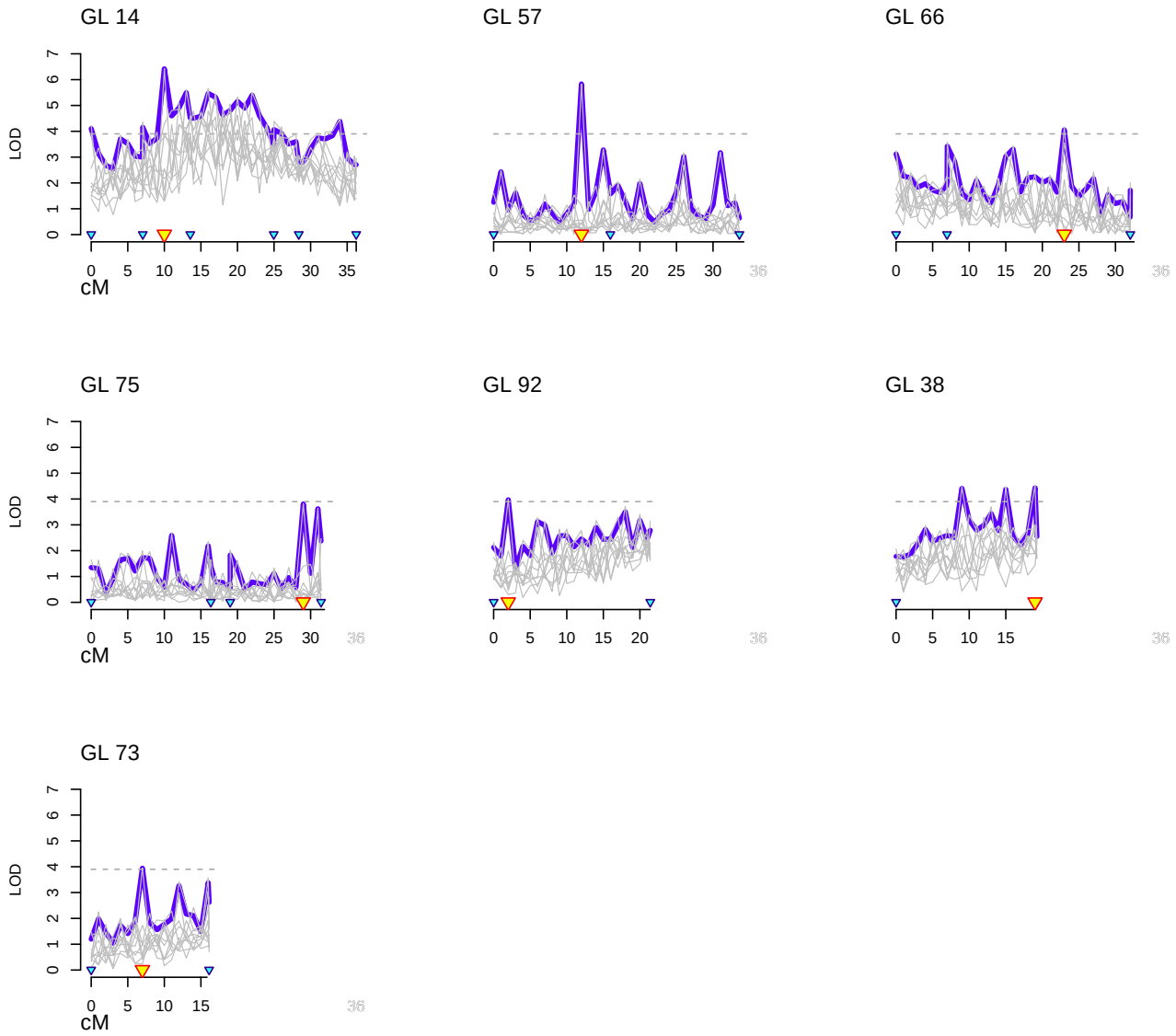


Figura 3 – Resultados do método de mapeamento por intervalo utilizando a abordagem da imputação múltipla para o caráter POL (conteúdo de sacarose). Perfis em cinza indicam cada uma das imputações. Perfis azuis indicam a média ponderada de todas as imputações. Triângulos amarelos: posições dos QTLs. Triângulos azuis: posições dos marcadores moleculares no mapa de ligação. Valores das distâncias são expressos em cM

(Conclusão)

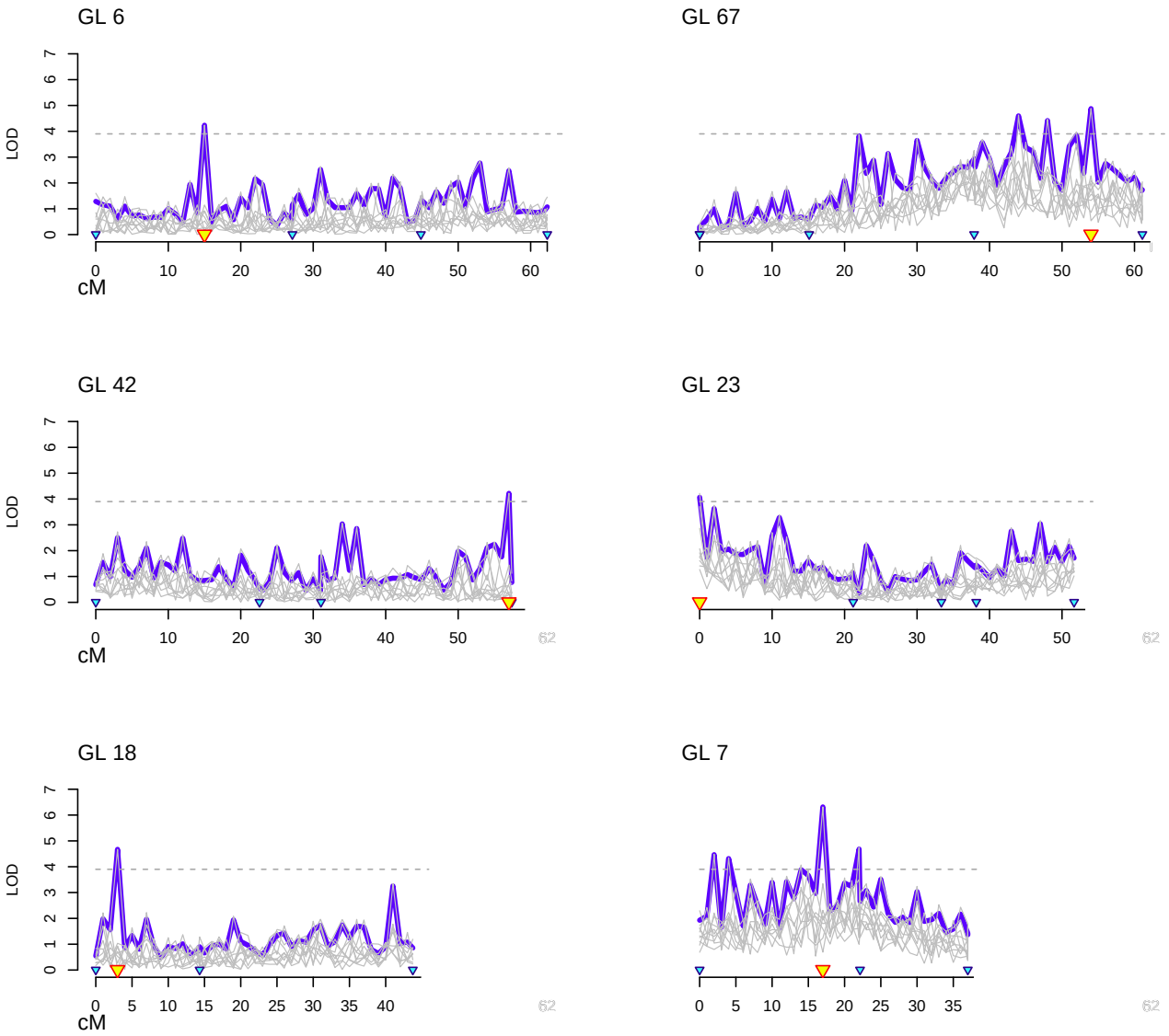


Figura 4 – Resultados do método de mapeamento por intervalo utilizando a abordagem da imputação múltipla para o caráter PC (produção de cana por parcela). Perfis em cinza indicam cada uma das imputações. Perfis azuis indicam a média ponderada de todas as imputações. Triângulos amarelos: posições dos QTLs. Triângulos azuis: posições dos marcadores moleculares no mapa de ligação. Valores das distâncias são expressos em cM

(Continua)

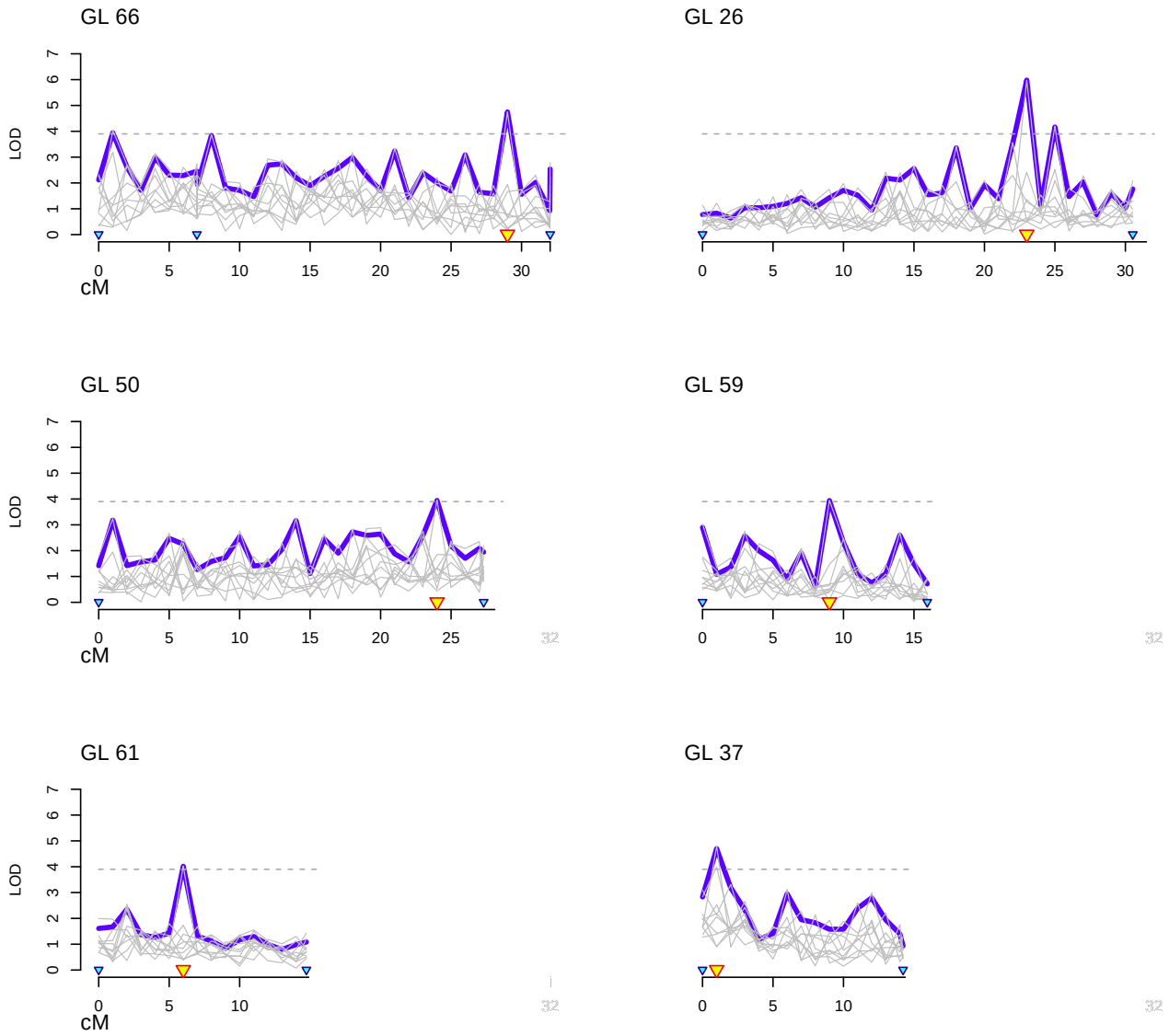


Figura 4 – Resultados do método de mapeamento por intervalo utilizando a abordagem da imputação múltipla para o caráter PC (produção de cana por parcela). Perfis em cinza indicam cada uma das imputações. Perfis azuis indicam a média ponderada de todas as imputações. Triângulos amarelos: posições dos QTLs. Triângulos azuis: posições dos marcadores moleculares no mapa de ligação. Valores das distâncias são expressos em cM

(Conclusão)

Figura 5 – Resultados do método de mapeamento por intervalo utilizando a abordagem da imputação múltipla para o caráter PP (produção de açúcar por parcela). Perfis em cinza indicam cada uma das imputações. Perfis azuis indicam a média ponderada de todas as imputações. Triângulos amarelos: posições dos QTLs. Triângulos azuis: posições dos marcadores moleculares no mapa de ligação. Valores das distâncias são expressos em cM

(Continua)

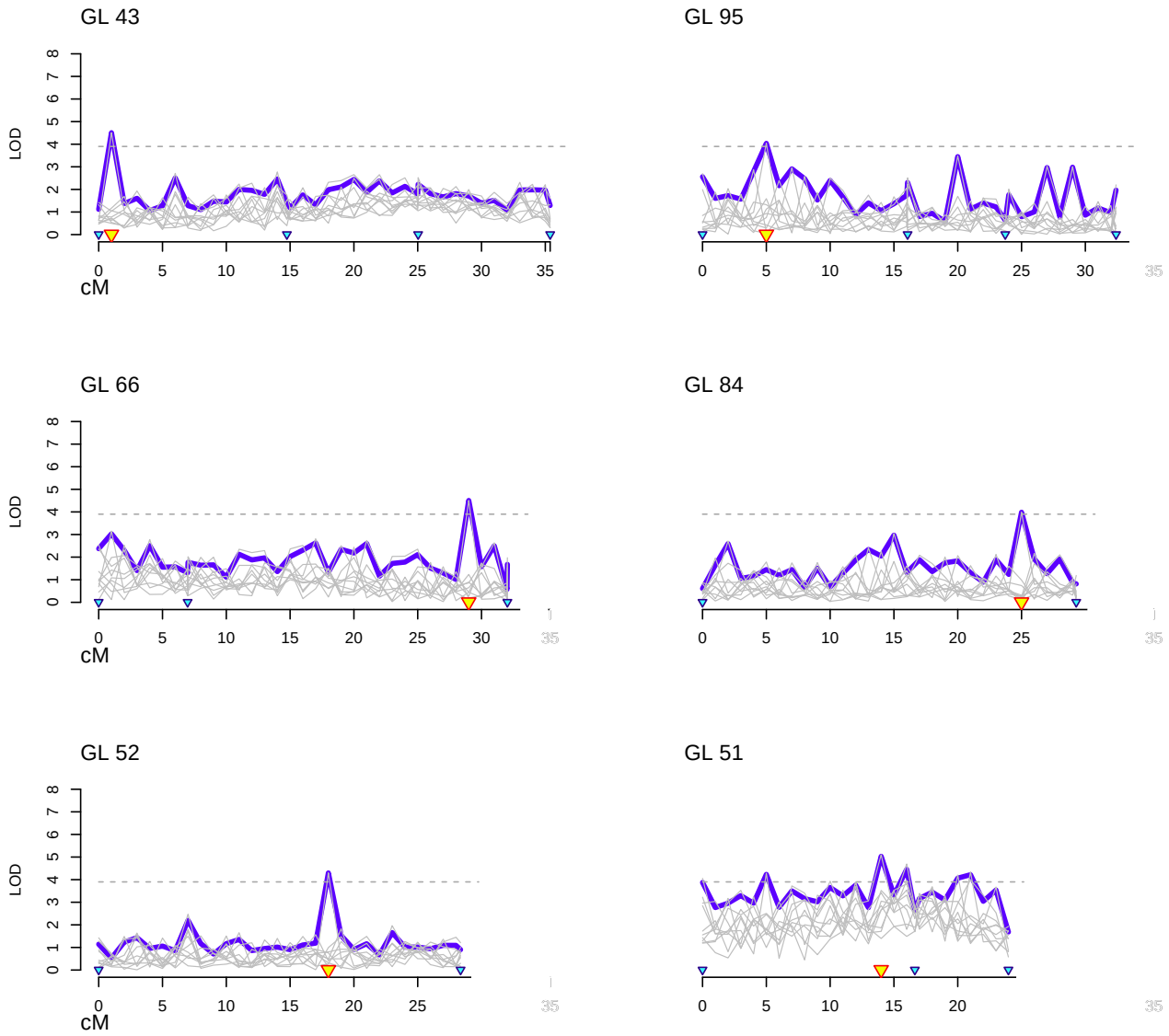


Figura 5 – Resultados do método de mapeamento por intervalo utilizando a abordagem da imputação múltipla para o caráter PP (produção de açúcar por parcela). Perfis em cinza indicam cada uma das imputações. Perfis azuis indicam a média ponderada de todas as imputações. Triângulos amarelos: posições dos QTLs. Triângulos azuis: posições dos marcadores moleculares no mapa de ligação. Valores das distâncias são expressos em cM

(Conclusão)

5 DISCUSSÃO

A maioria dos caracteres de importância agronômica e econômica da cana-de-açúcar possui natureza poligênica. Deste modo, métodos quantitativos, dentre eles o mapeamento de QTLs, têm sido corriqueiramente utilizados para avaliar os genótipos de interesse e buscar maior entendimento da arquitetura genética de tais caracteres quantitativos. Porém, no mapeamento de QTLs, a falta de informação dos genótipos dos QTLs não observados é uma das grandes dificuldades encontradas em função da baixa saturação dos mapas genéticos. Para contornar esse problema, métodos de imputação de genótipos não observados podem ser empregados empregados visando a obtenção de estimativas mais eficientes (BROMAN; SEN, 2009).

No presente trabalho, ao utilizar o método de mapeamento por intervalo com implementação da imputação múltipla, foram detectados 57 QTLs para os quatro caracteres de interesse: Fibra, POL, PC e PP. Foram mapeados 14 QTLs para o caráter Fibra, 19 para o caráter POL, 12 para o caráter PC e 12 para o caráter PP. Observa-se que o menor número de QTLs mapeados foi relacionado aos caracteres PC e PP, respectivamente. Tal observação pode ser justificada pela natureza quantitativa de tais caracteres, que são bastantes influenciados pelo ambiente e consequentemente apresentam os coeficientes de herdabilidade intermediários (Tabela 2). Vale ainda ressaltar que o caráter PP foi o que apresentou o maior CV(%) dentre os caracteres analisados, ocasionando portanto a diminuição do poder de detecção de QTLs para esse caráter. Já o caráter POL foi o que apresentou maior número de QTLs mapeados. Pelo fato do caráter POL apresentar maiores valores de coeficientes de herdabilidade, é esperado que o poder de detecção de QTLs para esse caráter em questão, seja maior. O mesmo raciocínio, pode ser estendido ao caráter Fibra, que apresentou o maior coeficiente de herdabilidade, dentre os caracteres avaliados.

Em relação a todos os caracteres avaliados, o número total de QTLs mapeados foi considerado alto. Embora o mapa genético final tenha totalizado 4370 cM e 113 grupos de ligação, observou-se que as condições experimentais para avaliação dos fenótipos dos indivíduos não foram ideais, uma vez que considerou-se para a maioria dos caracteres avaliados, um único local e apenas duas repetições. Dado este fato, era esperada a redução na detecção de QTLs, o que não foi observado. Porém, é importante ressaltar que para a estimativa do limiar (*Threshold*), utilizou-se o método proposto por Vieira et al. (2000), que embora seja uma metodologia corriqueiramente utilizada, possi-

bilita a admissão de QTLs falsos-postivos, o que pode ter ocasionado o alto número de QTLs mapeados. Deste modo, sugere-se que em estudos futuros sejam realizados métodos baseados em testes de permutação (CHURCHILL; DOERGE, 1994) para a determinação do limiar, afim de diminuir a ocorrência do erro tipo I. Apesar disso, acredita-se que a utilização da imputação múltipla para a identificação de QTLs traga vantagens ao mapeamento quando situações experimentais não ideais são utilizadas.

O único trabalho que avalia a eficiência da abordagem da imputação múltipla no mapeamento de QTLs foi relatado por Sen e Churchill (2001), que reanalisaram um conjunto de dados implementando a imputação múltipla de pseudomarcadores. Esses autores relatam a detecção de um maior número de QTLs após a utilização da imputação múltipla, visando a recomposição de genótipos não observados, do que quando realizado o mapeamento sem a implementação desta. O aumento do número de QTLs detectado, segundo esses autores, pode ser ocasionado pela variação existente nos genótipos imputados de cada pseudomarcador em cada localização pois, uma vez que são gerados vários conjuntos de dados moleculares completados (contendo marcadores observados e pseudomarcadores), a variação existente em cada uma das imputações é adicionada como estimativa da incerteza relacionada à imputação (a incerteza de que o pseudomarcador não seria realmente aquele observado), aumentando assim, a eficiência na localização dos QTLs.

Observou-se que os valores de LOD Score obtidos para os quatro caracteres variaram entre 7,52 e 3,81. Determinados QTLs apresentaram o valor de LOD Score inferior ao limiar de 3,90, porém, tais QTLs foram incluídos nos resultados, apenas de forma especulativa. Isto porque, ao observar as cuvas de mapeamento dos QTLs II e XIV para o caráter Fibra e o QTL XVI para o caráter POL, é observado um pico característico, indicativo de um QTL putativo para as respectivas localizações citadas. Esses QTLs servem como um ponto inical para estudos futuros, visando a confirmação de tais QTLs como significativos ou falsos-positivos. Em relação os efeitos de cada QTL mapeado, de um modo geral foram detectados efeitos aditivos e/ou de dominância, significativos em todos os QTLs mapeados. Houve predominância dos efeitos aditivos, de ambos os parentais, para todos os caracteres avaliados, indicando portanto que os genitores utilizados neste trabalho, foram capazes de fornecer um número significativo de alelos favoráveis para a progênie.

Verificando ainda o comportamento das curvas de mapeamento de determinados QTLs para os diferentes caracteres, nota-se que, para o caráter Fibra, o grupo de ligação 20, apresenta evidências de possíveis QTLs localizados próximos à região de 14 cM. Supõe-se que tais evidências sejam

resultado de efeitos do QTL II localizado na posição 14 cM. O mesmo comportamento foi comumente observado em outros grupos de ligação e em outros caracteres como por exemplo, para o caráter POL, no grupo de ligação 44 e 14 são verificados efeitos ocasionado pelos QTLs II (72 cM) e III (21 cM), respectivamente. Já ao considerar o caráter PC, nos grupos de ligação 67, 7 e 26, há evidências de efeitos ocasionados por QTLs fora do intervalo de mapeamento. Observou-se similares efeitos nos grupos de ligação 42, 22 e 51 do caráter PP. Dadas essas observações, é possível afirmar que embora o método de mapeamento por intervalo apresente algumas vantagens como a fácil implementação e execução, apresenta como desvantagem a impossibilidade de controlar a interferência de QTLs localizados fora do intervalo em estudo com consequente localização de QTLs falsos-positivos. Métodos que consideram a inclusão de efeitos ocasionados por QTLs localizados em intervalos adjacentes, como sugerido por Zeng (1993, 1994) e Kao e Zeng (1997), podem ser utilizados visando a identificação de QTLs mais precisamente.

Porém, embora o modelo de mapeamento utilizado para busca por QTLs tenha sido o mapeamento por intervalo, o qual possui algumas desvantagens como citado anteriormente, é notável que a técnica da imputação múltipla incorpora vantagens ao modelo de mapeamento por incluir a incerteza associada às imputações, através da variância obtida por meio da realização das m imputações (RUBIN, 1977). Essa variância é observada tanto dentro de cada imputação quanto entre as imputações e portanto as incertezas dos dados imputados são corretamente incorporadas na inferência final, resultante da média ponderada de todas as imputações.

De acordo com Rubin (1987), o número das m imputações geradoras da variância, é variável de acordo com cada cenário em estudo. Em geral quando se utiliza a técnica da imputação múltipla, simular entre três e cinco versões de bancos de dados completados é suficiente para obter uma estimativa de variação eficiente. Porém, uma vez que no presente estudo, foi considerado um “grid” de 1 cM ao longo do genoma, optou-se por realizar dez versões de imputação para cada localização ao longo dos grupos de ligação, objetivando assim, gerar a correta incerteza relacionada à cada imputação em cada localização. É importante ressaltar que o número de imputações realizados ($m = 10$) levou em consideração o pequeno “grid” de espaçamento, a ferramenta computacional existente para execução das análises e o modelo genético assumido para o mapemento de QTLs. Em estudos futuros, sugere-se que um maior número de imputações seja realizado pois acredita-se que com a utilização de maior número de imputações, a detecção de possíveis falsos-positivos ao utilizar o método de mapeamento por intervalo, seja reduzida.

Ainda relacionado ao mapeamento de QTLs, observou-se que determinados QTLs considerados no presente estudo, foram mapeados em regiões genômicas muito próximas dentro do mesmo grupo de ligação, porém para caracteres diferentes. Esses resultados sugerem a existência de uma possível correlação genética entre tais caracteres. No entanto, é de grande relevância notar que a presença de QTLs referentes à diferentes caracteres, localizados em intervalos próximos evidenciam além da correlação entre caracteres, a presença de efeitos de pleiotropia ou QTLs ligados, sendo portanto interessante, estudos adicionais para a confirmação dessas suposições.

Um detalhe interessante observado no presente trabalho é que, ao avaliar os perfis das curvas de mapeamento que representam a média ponderada de todas as imputações (curvas em azul, Figuras 2-5), verifica-se que em todos os grupos de ligação onde foram detectados QTLs, tais curvas seguem aproximadamente o perfil da curva da imputação individual (curva em cinza, Figuras 2-5), que apresentou maiores valores de LOD Score. Esse comportamento é esperado uma vez que para o cálculo da média ponderada de todas as imputações são considerados os pesos de cada uma das imputações. Consequentemente, aquelas imputações que receberam os maiores pesos são aquelas que contribuem em maior parte para a estimativa da média ponderada. No entanto, a estimativa dos pesos de cada imputação é baseada na estimativa da verossimilhança de cada imputação para cada localização, de acordo com o modelo genético de mapeamento assumido. Para aquelas imputações que apresentaram maior verossimilhança, foram atribuídos maiores pesos e consequentemente contribuiram em maior parte para a estimativa da média ponderada de todas as imputações. Uma vez que a estimativa do LOD Score apresentado por cada curva de mapeamento possui relação com a estimativa de verossimilhança, fica claro que as maiores estimatimas de verossimilhança, apresentarão maior valor LOD e concomitantemente contribuirão em maior grandeza para a estimativa da média ponderada de todas as imputações, justificando portanto os perfis apresentados.

Ainda em relação aos perfis da curva de mapeamento, em todas as curvas apresentadas neste trabalho, o perfil das mesmas é dado em forma de picos agudos, diferentemente do que é observado quando se trabalha com outras metodologias de mapemento por intervalo, como por exemplo, a abordagem proposta por Haley e Knott (1992). Supõe-se que tal perfil é observado pelo fato de que em cada localização, onde é realizada a imputação, há a probabilidade de ocorrência de quatro classes genotípicas: P^1Q^1, P^1Q^2, P^2Q^1 e P^2Q^2, sendo que determindas classes são mais prováveis do que outras de acordo com a localização e a probabilidade multiponto obtida. Porém, para a localização adjacente à anteriormente imputada existem quatro novas probabilidades de ocorrência

das classes genotípicas. Supondo-se que para uma determinada localização foi imputado o genótipo P^1Q^1, o qual condiciona os preditores genéticos 1,1,1 e que na posição adjacente foi imputado o genótipo P^2Q^1, o qual condiciona os preditores -1,1,-1. De acordo com o exposto, é observado que os preditores genéticos estimados, possuirão valores diferentes, resultando portanto em estimativas com valores distintos para cada posição analisada. Quando tais estimativas são transpostas em forma de curvas, é então observada claramente a formação de curvas apresentando a característica de picos agudos, pois o padrão da imputação de pseudomarcadores não é contínuo ao longo de todo o genoma, ou seja, nas posições adjacentes de uma determinada localização é esperado que diferentes genótipos sejam imputados, condicionando a diferentes preditores genéticos. Comportamento similar, em que as curvas de mapeamento resultante das médias das imputações ponderadas apresentou picos agudos, foi observado por Sen e Churchill (2001).

Outro detalhe observado é que quando se trabalha com o mapeamento de QTLs em cana-de-açúcar, especificamente utilizando apenas marcadores com padrão se segregação 1:1 e 3:1, é comum observar que nem sempre os marcadores moleculares fornecem informação suficiente para a distinção das quatro classes genotípicas para a estimativa dos preditores genéticos. Geralmente nessas circunstâncias, as matrizes dos preditores genéticos podem ser singular, de modo que tais preditores podem ser obtidos por meio da combinação linear de outros preditores. Porém, é comum observar problemas de multicolinearidade que geram complicações na estimação dos parâmetros em um modelo de regressão, ocasionando redução da precisão na obtenção das estimativas (BELSLEY; KUE; WELSCH, 1980). Por exemplo, aqueles grupos de ligação contendo apenas marcadores do tipo D_1, fornecem informação suficiente apenas para a estimativa do contraste referente ao efeito aditivo do genitor P; o mesmo ocorre para marcadores do tipo D_2, onde há possibilidade de estimativas apenas do efeito aditivo do genitor Q. Margarido (2011), afirma que mesmo naqueles grupos de ligação onde são observados três tipos de marcadores, há possibilidade da existência de colinearidade. Porém, com a técnica da imputação múltipla, para cada localização ao longo do genoma, foi observada a não existência de colinearidade em todos os grupos de ligação. Isto porque foi considerado que a probabilidade baseada na abordagem multiponto calculada via HMM, é condicional aos marcadores observados. Sendo assim, caso o grupo de ligação fosse composto apenas por marcadores do tipo D_1 e fosse observada a “presença do alelo”, as possíveis classes genotípicas observadas seriam P^1Q^1 ou P^1Q^2 (com iguais probabilidades). Caso fosse observada a “ ausência do alelo” para esse mesmo tipo de marcador, as possíveis classes genotípicas observadas seriam

P^2Q^1 e P^2Q^2 (com iguais probabilidades).

Uma vez que no processo da imputação múltipla são consideradas as probabilidades condicionais das quatro classes fenotípicas dentro de cada grupo de ligação, e que dentro de todos os grupos de ligação em estudo foram observados tanto presença como ausência dos alelos para os determinados marcadores que compõem os grupos em questão, foi verificado que em determinadas imputações onde ocorreu a "presença do alelo", foram imputados em dadas localizações o genótipo P^1Q^1 enquanto que em outras localizações, foi imputado o genótipo P^1Q^2. Já em determinadas imputações, dentro do mesmo grupo de ligação observou-se a " ausência do alelo" e consequentemente foram imputados em determinadas localizações o genótipo P^2Q^1 enquanto que em outras localizações, P^2Q^2. Sendo assim, dentro de um mesmo grupo de ligação foram imputadas as quatro classes genotípicas. Observou-se tal comportamento em todos os grupos de ligação em estudo. Consequentemente foi possível a estimativa dos preditores genéticos sem a existência da colinearidade e deste modo, todos os contrastes foram incluídos na análise de QTLs, demostrando assim, que a utilização da imputação múltipla para recomposição de genótipos não observados facilitou o processo de mapeamento de QTLs em espécies poliplóides.

É importante ressaltar que QTLs não identificados no presente trabalho, porém citados em outros trabalhos envolvendo o mapeamento genético em cana-de-açúcar, podem não ter sido identificados por algumas razões, tais como estarem localizados em regiões do genoma que não foram mapeadas ou ainda serem QTLs de pequeno efeito. É válido observar que a comparação de resultados entre trabalhos que envolvam o mapeamento de QTLs principalmente em cana-de-açúcar é dificultada, pois nem sempre são utilizados os mesmos genitores para gerar a população de mapeamento, sendo assim, diferentes populações de mapeamento são avaliadas, otendo-se resultados diferentes. Além disso, diferentes metodologias genético-estatística são utilizadas para avaliar diferentes conjuntos de dados, resultando portanto em observações intrínsecas à cada trabalho.

Verifica-se que algumas limitações inerentes ao delineamento experimental, ocasionaram a queda da precisão experimental, uma vez que para a maioria dos caracteres as estimativas fenotípicas foram avaliadas em apenas um local considerando duas repetições. Além disso, o delineamento utilizado impossibilitou estimativas de componetes de produção da cana-de-açúcar de grande importância, como é o caso dos caracteres TCH (toneladas de cana por hectare) e TPH (toneladas de POL por hectare), justificando portanto, o motivo pelo qual tais caracteres de importante grandeza não foram avaliados diretamente, mas sim a partir de valores de medidas ao nível de parcela, que

não podem facilmente ser extrapolados para áreas maiores. Sugere-se o uso de parcelas contendo dimensões representativas, maior número de repetições, e maior número de locais avaliados.

Observa-se também que os métodos utilizados tanto para a construção do mapa genético de ligação quanto para realização do mapeamento de QTLs, considerou apenas a inclusão de marcadores que apresentaram segregação em dosagem única (1:1 e 3:1). Deste modo, verifica-se a necessidade de aperfeiçoamento de técnicas genético-estatísticas que incluam diferentes tipos de segregação, permitindo assim, a utilização de marcadores que forneçam maior informação genética, a contrução de mapas genéticos mais saturados e consequentemente o aumento do poder na detecção de QTLs.

Em relação à abordagem da imputação múltipla, verifica-se algumas limitações como por exemplo, nunca foram realizados estudos que demonstrem qual a magnitide da saturação do genoma em relação à proporção de marcadores observados ao longo do genoma, ou seja, o quanto o tamanho de determinado "grid" de espaçamento em relação a quantidade total de marcadores observados, interfere nas estimativas de QTLs. Além disso, a técnica da imputação múltipla utiliza um tempo maior para realização das análises devido à complexidade de execução das m análises de QTLs.

Contudo, métodos que incluem a abordagem da imputação múltipla objetivando tratar a informação de marcadores não observados nunca haviam sido investigados em espécies heterozigóticas e poliplóides, como é o caso da cana-de-açúcar. A utilização do método de mapeamento por intervalo com a implementação da imputação múltipla de pseudomarcadores foi considerada uma ferramenta adicional para a localização de possíveis QTLs e permitiu a detecção de QTLs com sucesso, em regiões que anteriomente seriam desconsideradas. Deste modo, acredita-se que os resultados do presente trabalho foram válidos e contribuam de forma substancial para futuros estudos que objetivem melhor compreensão da arquitetura genética dos caracteres quantitativos da cana-de-açúcar.

6 CONCLUSÃO

O método de mapeamento por intervalo utilizando a abordagem da imputação múltipla foi eficiente para o mapeamento de caracteres quantitativos da cana-de-açúcar.

REFERÊNCIAS

AL-JANABI, S.M.; HONEYCUTT, R.J.; McCLELLAND, M.; SOBRAL, B.W.S. A genetic linkage map of *Saccharum spontaneum* L. 'SES 208'. **Genetics**, Bethesda, v. 134, p. 1249-1260, 1993.

ANDRU, S.; PAN, Y.B.; THONGTHAWEE, S.; BURNER, D.M.; KIMBENG, C.A. Genetic analysis of the sugarcane (*Saccharum* spp.) cultivar 'LCP 85-384'. I. Linkage mapping using AFLP, SSR, and TRAP markers. **Theoretical and Applied Genetics**, New York, v. 123, p. 77-93, 2011.

AGARWAL, M.; SHRIVASTAVA, N.; PADH, H. Advances in molecular marker techniques and their applicationsin plant sciences. **Plant Cell Rep**, Berlin, v. 27, p. 617–631, 2008.

AITKEN, K.S.; HERMANN, S.; KARNO, K.; BONNETT, G.D.; McINTYRE, C.L.; JACKSON, P.A. Genetic control of yield related stalk traits in sugarcane. **Theoretical and Applied Genetics**, New York, v. 117, p. 1191-1203, 2008.

AKAIKE, H. A new look at the statistical model identification. **The Institute of Electrical and Electronics Engineers Transactions on Automatic Control**, Notre Dame, v. 19, p. 716-723, 1974.

ALWALA, S.; KIMBENG, C.A. Molecular genetic linkage mapping in *Saccharum*: strategies, resources and achievements. In: HENRY, R.J. (Ed.) **Genetics, Genomics and Breeding of Sugarcane**. New Hampshire: Science Publishers, 2010. 1 v.

BALL, R.D. Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the bayesian information criterion. **Genetics**, Bethesda, v. 159, p. 1351-1364, 2001.

BARBOSA, M.H.P.; SILVEIRA, L.C.I. Metodologias de seleção, progressos e mudanças no programa de melhoramento genético da cana-de-açúcar da Universidade Federal de Viçosa. **STAB - Açúcar, Álcool e Subprodutos**. Piracicaba, v.18, n.3, p. 30-32, 2000.

BASTEN, C.J.; WEIR, B.S.; ZENG, Z.B. **QTL Cartographer**: a reference manual and tutorial for QTL mapping. Raleigh, 2005. Disponível em: <http://statgen.ncsu.edu/qtlcart>. Acesso em: 25 maio 2011.

BEARZOTI, E. Mapeamento de QTL. In: PINHEIRO, J.B.; CARNEIRO, I.F. (Ed.). **Análise de QTL no melhoramento de plantas**. Goiânia: FUNAPE, 2000. p. 63-223.

BELSLEY, D.A.; KUH, E.; WELSCH, R.E. **Regression diagnostics**: identifying influential data and sources of collinearity. New York: John Wiley & Sons, 1980. 320 p.

BERGAMO, G.C.; DIAS, C.T.S.; KRZANOWSKI, W.J. Distribuição-free multiple imputation in an interaction matrix through singular value decomposition. **Scientia Agricola**, Piracicaba, v. 65, n.4, p. 422-427, 2008.

BOTSTEIN D.; WHITE R.L.; SKOLNICK M.; DAVIS R.V. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. **The American Journal of Human Genetics**, Bethesda, v. 32, p. 314-331, 1980.

BOUXSEIN, M.L.; UCHIYAMA, T.; ROSEN, C.J.; SHLTZ, K.L.; DONAHUE, L.R.; TURNER, C.H.; SEN, S.; CHURCHILL; G.A.; MULLER, R.; BEAMER, W.G. Mapping quantitative trait loci for vertebral trabecular bone volume fraction and microarchitecture in mice **Journal of bone and mineral research**, Washington, v.19, n.4, p. 587-599, 2004.

BROMAN, K.W. SEN, Ś. **A guide to QTL mapping with R/qtl**. New York: Springer, 2009. 396 p.

BROMAN, K.W.; WU, H.; SEN, Ś.; CHURCHILL, G.A. R/qtl: QTL mapping in experimental crosses. **Bioinformatics**, Oxford, v. 19, p. 889-890, 2003.

COMPANHIA NACIONAL DE ABASTECIMENTO. **Acompanhamento da safra brasileira: cana-de-açúcar, primeiro levantamento, maio/2011**. Disponível em: <http://www.conab.gov.br/OlalaCMS/uploads/arquivos> Acesso em: 21 agosto 2011.

CONSELHO NACIONAL DOS PRODUTORES DE CANA-DE-AÇÚCAR, AÇÚCAR E ÁLCOOL DO ESTADO DE SÃO PAULO. **Manual de instruções – CONSECANA-SP**. Piracicaba, 2006. 112p.

CORDEIRO, G. M.; TAYLOR, G. O.; HENRY, R. J. Characterization of microsatellite markers from sugarcane *(Saccharum sp.)* a highly polyploid species. **Plant Science**, Washington, v.155, p.161-168, 2000.

CORDELL, H.J. Estimation and testing of genotype and haplotype effects in case-control studies: comparison of weighted regression and multiple imputation procedures. **Genetic Epidemiology**, Cranford, v. 30, p. 259–275, 2006.

CHURCHILL, G.A.; DOERGE, R.W. Empirical threshold values for quantitative trait mapping. **Genetics**, Austin, v. 138, n. 3, p. 963-971, 1994.

DEMPSTER, A.P.; LAIRD, N.M.; RUBIN, D.B. Maximum likelihood from incomplete data via the EM algorithm. **Journal of the Royal Statistical Society B**, London, v. 39, p. 1-38, 1977.

DOERGE, R.W.; ZENG, Z.B.; WEIR, B.S. Statistical Issues in the search for genes affecting quantitative traits in experimental populations. **Statistical Science**, Ithaca, v. 12, n. 3, p. 195-219, 1997.

DOERGE, R.W. Mapping and analysis of quantitative trait loci in experimental populations. **Nature Reviews Genetics**, New York, v. 3, p. 43-52, 2002.

FALCONER, D.S.; MACKAY, T.F.C. **Introduction to Quantitative Genetics**. 4.ed. Essex, UK: Longman, 1996. 464 p.

FERREIRA, M.E.; GRATTAPAGLIA, D. **Introdução ao uso de marcadores moleculares em análise genética**. 3.ed. Brasília: EMBRAPA-CENARGEN, 1998. 220 p.

GARCIA, A.A.F.; KIDO, E.A.; MEZA, A.N.; SOUZA, H.M.B.; PINTO, L.R.; PASTINA, M.M.; LEITE, C.S.; SILVA, J.A.G. da; ULIAN, E.C.; FIGUEIRA, A.; SOUZA, A.P. 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. **Theoretical and Applied Genetics**, New York, v. 112, p. 298-314, 2006.

GAZAFFI, R. **Desenvolvimento de modelo genético-estatístico para mapeamento de QTLs em progênie de irmãos completos, com aplicação em cana-de-açúcar**. 2009. 103 p. Tese (Doutorado em Genética e Melhoramento de Plantas) - Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2009.

GRATTAPAGLIA, D.; SEDEROFF, R. Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross: mapping strategy and RAPD markers. **Genetics**, Bethesda, v. 137, p. 1121-1137, 1994.

GRIVET, L.; D’HONT, A.; ROQUES, D.; FELDMANN, P.; LANAUD, C.; GLASZMANN, J.C. RFLP mapping in cultivated sugarcane (*Saccharum* spp.): genome organization in a highly polyploid and aneuploid interspecific hybrid. **Genetics**, Bethesda, v. 142, p. 987-1000, 1996.

HALEY, C.S.; KNOTT, S.A. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. **Heredity**, London, v. 69, p. 315-324, 1992.

HEINZ, D.J.; TEW, T.L. Hybridization procedures. In: HEINZ, D.J. (Ed.) **Sugarcane improvement through breeding**. Amsterdam: Elsevier. 1987. p. 313-342.

HOARAU, J.Y.; OFFMANN, B.; D'HONT, A.; RISTERUCCI, A.M.; ROQUES, D.; GLASZMANN, J.C.; GRIVET, L. Genetic dissection of a modern sugarcane cultivar (*Saccharum* spp.). I. Genome mapping with AFLP markers. **Theoretical and Applied Genetics**, New York, v. 103, p. 84-97, 2001.

HOGARTH, D.M.; COX, M.C.; BULL, J.K. Sugarcane improvement: past achievements and future prospects. In: KANG, M.S. (Ed.). **Crop Improvement for the 21st Century**. Baton Rouge: Louisiana State University, 1997. p.29-55.

IRVINE, J.E. *Saccharum* species as horticultural classes. **Theoretical and Applied Genetics**, New York, v. 98, p. 186-194, 1999.

JIANG, C.; ZENG, Z.B. Multiple trait analysis of genetic mapping for quantitative trait loci. **Genetics**, Bethesda, v. 140, p. 1111-1127, 1995.

JIANG, C.; ZENG, Z.B. Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. **Genetica**, Berlin, v. 101, p. 47-58, 1997.

JOHNSON, D.L.; JANSEN, R.C.; ARENDONK, J.A.M. van. Mapping quantitative trait loci in a selectively genotyped outbred population using a mixture model approach. **Genetical Research**, London, v. 73, p. 75-83, 1999.

KAO, C.H.; ZENG, Z.B. General formulas for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm. **Biometrics**, Washington, v. 53, p. 653-665, 1997.

KAO, C.H., ZENG, Z.B.; TEASDALE, R D. Multiple interval mapping for quantitative trait loci. **Genetics**, Bethesda, v. 152, p. 1203-1216, 1999.

KIDO, E.A. **Mapeamento de marcadores moleculares AFLP em população derivada de clones elite de cana-de-açúcar (*Saccharum* spp.) e suas associações com caracteres agronômicos**. 2003. 164 p. Tese (Doutorado em Energia na Agricultura) – CENA-Centro de Energia Nuclear na Agricultura, Universidade de São Paulo, Piracicaba, 2003.

KOSAMBI, D.D. The estimation of map distances from recombination values. **Annual of Eugenics**, London, v. 12, p. 172-175, 1944.

LANDER, E.S.; GREEN, P. Construction of multilocus genetic linkage maps in humans. **Proceedings of tthe National Academy of Sciences of the United States**, Washington, v. 84, p. 2363-2367, 1987.

LANDER; E.S.; BOTSTEIN, D. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. **Genetics**, Bethesda, v. 121, p. 185-199, 1989.

LEWIN, H.A.; LARKIN, D.M.; PONTIUS, J.; O'BRIEN, S.J. Every genome sequence needs a good map. **Genome Research**, Woodbury, v. 19, p. 1925-1928, 2009.

LIN, M.; LOU, X.Y.; CHANG, M.; WU, R. A general statistical framework for mapping quantitative trait loci in nonmodel systems: issue for characterizing linkage phases. **Genetics**, Bethesda, v. 165, p. 901-913, 2003.

LINCOLN, S.E.; NEWBURN, L. MAPMAKER: an interactive computer package for constructing primary genetic linkage maps os experimental and natural populations. **Genomics**, Cambridge v.1, p. 174-181, 1987.

LITTLE, R.J.A.; RUBIN,D.B.**Statistical Analysis with Missing Data**. New Jersey: John Wiley and Sons, 2002. 380p.

LITTLE, R.J.A.; SMITH, P. J. Edit and imputation for quantitative survey data. **Journal of the American Statistical Association**, Alexandria, v. 82, p. 58-68, 1987.

LIU, B.H. **Statistical genomics:** linkage, mapping, and QTL analysis. Boca Raton: CRC Press, 1998. 611p.

LYNCH, M.; WALSH, B. **Genetics and analysis of quantitative traits**. Sunderland: Sinauer Associates, 1998. 980p.

MALOSETTI, M.; VOLTAS, J.; ROMAGOSA, I.; ULLRICH, S.E.; EEUWIJK, F.A. van. Mixed models including environmental covariables for studying QTL by environment interaction. **Euphytica**, Wageningen, v. 137, p. 139-145, 2004.

MALOSETTI, M.; BOER, M.P.; BINK, M.C.A.M.; EEUWIJK, F.A. van. Multi-trait QTL analysis based on mixed models with parsimonious covariance matrices. In: WORLD CONGRESS ON GENETICS APPLIED TO LIVESTOCK PRODUCTION, 8., 2006, Belo Horizonte. **Proceedings...**, Belo Horizonte: WCGALP, 2006. Abstract 25-04.

MALOSETTI, M.; RIBAUT, J.M.; VARGAS, M.; CROSSA, J.; EEUWIJK, F.A. van. A multi-trait multi-environment QTL mixed model with an application to drought and nitrogen stress trials in maize (*Zea mays* L.). **Euphytica**, Wageningen, v. 161, p. 241-257, 2008.

MANCINI, M.C. **Mapeamento genético eidenticação de marcadores AFLP, microssatélites genômicos e funcionais associados a parâmetros agroindustriais em cana-de-açúcar**. 2010. 98 p. Dissertação (Mestrado em Genética e Melhoramento de Plantas) - Faculdade de Ciências Agrárias e Veterinárias - Campus de Jaboticabal, Universidade Estadual do Estado de São Paulo, Jaboticabal, 2010.

MARCONI, T.G.; COSTA, E.A.; MIRANDA, H.R.C.A.N.; MANCINI, M.C.; MOLLINARI, M.; GARCIA, A.A.F.; SOUZA, A.P. Functional markers for gene mapping and genetic diversity studiesn in sugarcane. **BMC Research Notes**, London, v. 4, n. 264, p. 1-8, 2011.

MARGARIDO, G.R.A. **Mapeamento de QTLs em múltiplos caracteres e ambientes em um cruzamento comercial de cana-de-açúcar usando modelos mistos**. 2011. 107 p. Tese (Doutorado em Genética e Melhoramento de Plantas) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2011.

MARGARIDO, G.R.A.; SOUZA, A.P.; GARCIA, A.A.F. OneMap: software for genetic mapping in outcrossing species. **Hereditas**, Lund, v. 144, p. 78-79, 2007.

MARTÍNEZ, O.; CURNOW, R.N. Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. **Theoretical and Applied Genetics**, New York, v. 85, p. 480-488, 1992.

MATSUOKA, S.; GARCIA, A.A.F.; ARIZONO, H. Melhoramento da cana-de-açúcar. In: BORÉM, A. (Ed.) **Melhoramento de espécies cultivadas**. Viçosa: UFV, 1999. p. 205-252.

MCCULLAGH, P.; NELDER, J.A. **Generalized Linear Models**. Boca Raton: Chapman and Hall/CRC, 1989, 511 p.

MING, R.; LIU, S.C.; MOORE, P.H.; IRVINE, J.E.; PATERSON, A.H. QTL analysis in a complex autopolyploid: genetic control of sugar content in sugarcane. **Genome Research**, London, v. 11, p. 2075-2084, 2001.

MING, R.; LIU, S.C.; BOWERS, J.E.; MOORE, P.H.; IRVINE, J.E.; PATERSON, A.H. Construction of a *Saccharum* consensus genetic map from two interspecific crosses. **Crop Science**, Madison, v. 42, p. 570-583, 2002.

MUKHERJEE, S.K. Origin and distribution of Saccharum. **Bot Gaz**, Bangladesh, v.119, p.55-56, 1957.

PASTINA, M.M. **Mapeamento de QTLs e estudo da interação entre QTLs, ambientes e cortes em cana-de-açúcar, usando a abordagem de modelos mistos**. 2010. 89 p. Tese (Doutorado em Genética e Melhoramento de Plantas) - Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2010.

PASTINA, M.M.; PINTO, L.R.; OLIVEIRA, K.M.; SOUZA, A.P.; GARCIA, A.A.F. Molecular mapping of complex traits. In: HENRY, R.J. (Ed.) **Genetics, Genomics and Breeding of Sugarcane**. New Hampshire: Science Publishers, 2010. 1 v.

PAYNE, R.W.; MURRAY, D.A.; HARDING, S.A.; BAIRD, D.B.; SOUTAR, D.M. Introduction to GenStat 12 for Windows. **VSN International**, Hemel Hempstead, 2009. 37 p.

PEREIRA, M. G.; PEREIRA, T.N.S.; COSTA, F.R.da. Marcadores Moleculares no Pré-Melhoramento. In: BORÉM, A; CAIXETA, E.T. (Ed.). **Marcadores moleculares**, Viçosa: UFV, 2009. 2 v.

PETERNELLI, L.A.; BARBOSA, M.H.P. Desafios na condução e planejamento de experimentos de avaliação de clones de cana-de-açúcar. In: SIMPÓSIO DE ATUALIZAÇÃO EM GENÉTICA E MELHORAMENTO DE PLANTAS: DESAFIOS E PERSPECTIVAS DA INTERAÇÃO BIOMETRIA E MELHORAMENTO DE PLANTAS, 7., 2004, Lavras **Anais...** Lavras: UFLA, 2004. p 17-27.

R Developmente Core Team. R: a language and environment for statistical computing. **R Foundation for Statistical Computing**, Vienna, 2009. Disponível em: <http://www.r-project.org/> Acesso em: 16 mar. 2010.

RABINER, L.R. A tutorial on hidden Markov models and selected applications in speech recognition. **Proceedings of the Institute of Electrical and Eletronics Engineers**, London, v. 77, p. 257-286, 1989.

RUBIN, D.B. Inference and missing data. **Biometrika**, Oxford, v. 63 ,p. 581–592, 1976.

RUBIN, D.B. Formalizing subjective notions about the effect of nonrespondents in sample surveys. **Journal of the American Statistical Association**, Alexandria, v. 72, p. 538–543, 1977.

RUBIN, D.B. **Multiple Imputation for Nonresponse in Surveys**. New York: John Wiley and Sons, 1987. 287 p.

SAGHAI, M.M.A.; BIYASHEV, R.M.; YANG, G.P.; ZHANG, Q.; ALLARD, R.W. Extraordinarily polymorphic microsatellite DNA in barley: Species diversity, chromosomal locations, and population dynamics. **Proceedings of the National Academy of Sciences** Washington, v.91, p.5466–5470, 1994.

SAS Institute. **SAS/STAT user's guide v.8.2. SAS Institute**, Cary, 2001. 1028 p.

SCHAFER, J.L. Multiple imputation: a primer. **Statistical Methods in Medical Research**, London, v. 8, p. 3–15. 1999.

SCHLÖTTERER, C. The evolution of molecular markers - just a matter of fashion? **Nature Reviews Genetics**, New York, v. 5, p. 63-69, 2004.

SEN, S.; CHURCHILL, G.A. A statistical framework for quantitative trait mapping. **Genetics**, Bethesda, v. 159, p. 371-388, 2001.

SILLS, G.R.; BRIDGES, W.; AL-JANABI, S.M.; SOBRAL, B.W.S. Genetic analysis of agronomic traits in a cross between sugarcane (*Saccharum officinarum* L.) and its presumed progenitor (*S. robustum* Brandes & Jesw. ex Grassl). **Molecular Breeding**, Berlin, v. 1, p. 355-363, 1995.

SILVA, J.A.G. da; SORRELLS, M.E.; BURNQUIST, W.; TANKSLEY, S.D. RFLP linkage map and genome analysis of *Saccharum spontaneum*. **Genome**, Ottawa, v. 36, p. 782-791, 1993.

SMITH, A.B.; STRINGER, J.K.; WEI, X.; CULLIS, B.R. Varietal selection for perennial crops where data relate to multiple harvests from a series of field trials. **Euphytica**, Wageningen, v. 157, p. 253-266, 2007.

SONG, J.Z.; SOLLER, M.; GENIZI, A. The full-sib intercross line (FSIL): a QTL mapping design for outcrossing species. **Genetical Research**, London, v. 73, p. 61-73, 1999.

SOUZA, A.P. Biologia molecular aplicada ao melhoramento. In: NASS, L.L.; VALOIS, A.C.C.; MELO, I.S.; VALADARES-INGLIS, M.C. (Ed.) **Recursos genéticos e melhoramento de plantas**. Rondonópolis: Fundação MT. 2001. p. 939-965.

STUBER, C.W.; EDWARDS, M.D.; WENDEL, J.F. Molecular-marker-facilitated investigations of quantitative trait loci in maize. II. Factors influencing yield and its component traits. **Crop Science**, Madison, v. 27, p. 639-648, 1987.

SUN, Z.; STAUB, J.E.; CHUNG, S.M.; LOWER, R.L. Identification and comparative analysis of quantitative trait loci associated with parthenocarpy in processing cucumber. **Plant Breeding**, Berlin, v.25, p.281-287, 2006.

SYMONDS, V.V; GODOY, A.V.; ALCONADA, T. BOTTO, J.F.; JUENGER, T.E.; CASAL, J.J.; LLOYD, A.M. Mapping quantitative trait loci in multiple populations of arabidopsis thaliana identifies natural allelic variation for trichome density. **Genetics**, Bethesda, v. 169, p. 1649-1659, 2005.

TAUTZ, D. Hypervariability of simple sequences as a source for polymorphic DNA markers. **Nucleic Acids Research**, Oxford, v. 17, n. 16, p. 6463-6471, 1989.

UNIÃO DA INDÚSTRIA DA CANA-DE-AÇÚCAR. **Setor Sucroenergético - Histórico**. Disponível em: <http://www.unica.com.br>. Acesso em: 23 ago. 2011.

VIEIRA, C.; PASYUKOVA, E.G.; ZENG, Z.B.; HACKETTI, J.B.; LYMAN; R.F.; MACKAY, T.F. Genotype-environment interaction for quantitative trait loci affecting life span in *Drosophila melanogaster*. **Genetics**, Bethesha, v.14, p. 213-227, 2000.

WILLIAMS, J.G.K.; KUBELIK, A.R.; LIVAK, K.J.; RAFALSKI, J.A.; TINGEY, S.V. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. **Nucleic Acids Research**, Oxford, v. 18, p. 6531-6535, 1990.

WELLER, J.I. Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. **Biometrics**, Washington, v. 42, p. 627-640, 1986.

WU, K.K.; BURNQUIST, W.; SORRELLS, M.E.; TEW, T.L.; MOORE, P.H.; TANKSLEY, S.D. The detection and estimation of linkage in polyploids using single-dose restriction fragments. **Theoretical and Applied Genetics**, New York, v. 83, p. 294-300, 1992.

WU, R.; MA, C.X.; WU, S.S.; ZENG, Z.B. Linkage mapping of sex-specific differences. **Genetical Research**, London, v. 79, p. 85-96, 2002a.

WU, R.; MA, C.X.; PAINTER, I.; ZENG, Z.B. Simultaneous maximum likelihood estimation of linkage and linkage phases in outcrossing species. **Theoretical Population Biology**, New York, v. 61, p. 349-363, 2002b.

ZABEAU, M.; VOS, P. **European Patent Office. Selective restriction fragment amplification : a general method for DNA fingerprinting 0 534 858 A1**, 1993. 2013. Disponível em: < http://www.freepatentsonline.com/EP0534858.pdf>. Acesso em: 21 nov. 2011.

ZENG, Z.B. Theoretical basis of separation of multiple linked gene effects on mapping quantitative trait loci. **Proceedings of the National Academy of Science of the United States of America**, Washington, v. 90, p. 10972-10976, 1993.

ZENG, Z.B. Precision mapping of quantitative trait loci. **Genetics**, Bethesda, v. 136, p. 1457-1468, 1994.

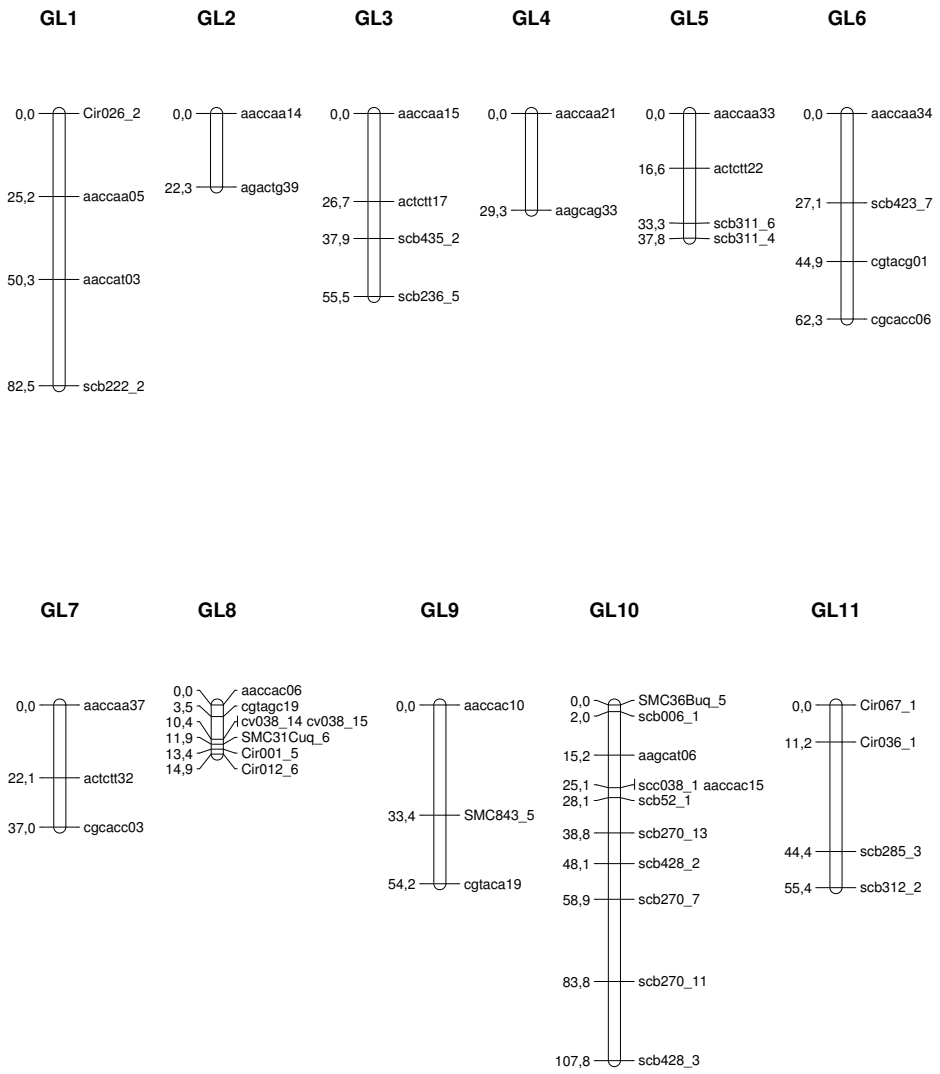
ZENG, Z.B., KAO, C.H.; BASTEN, C.J. Estimating the genetic architecture of quantitative traits. **Genetical Research**, London, v. 74, p. 279-289, 1999.

ZHANG, P. Multiple imputation: Theory and method. **International Statistical Review**, The Hague, v.71, n.3,p.581-92, 2003.

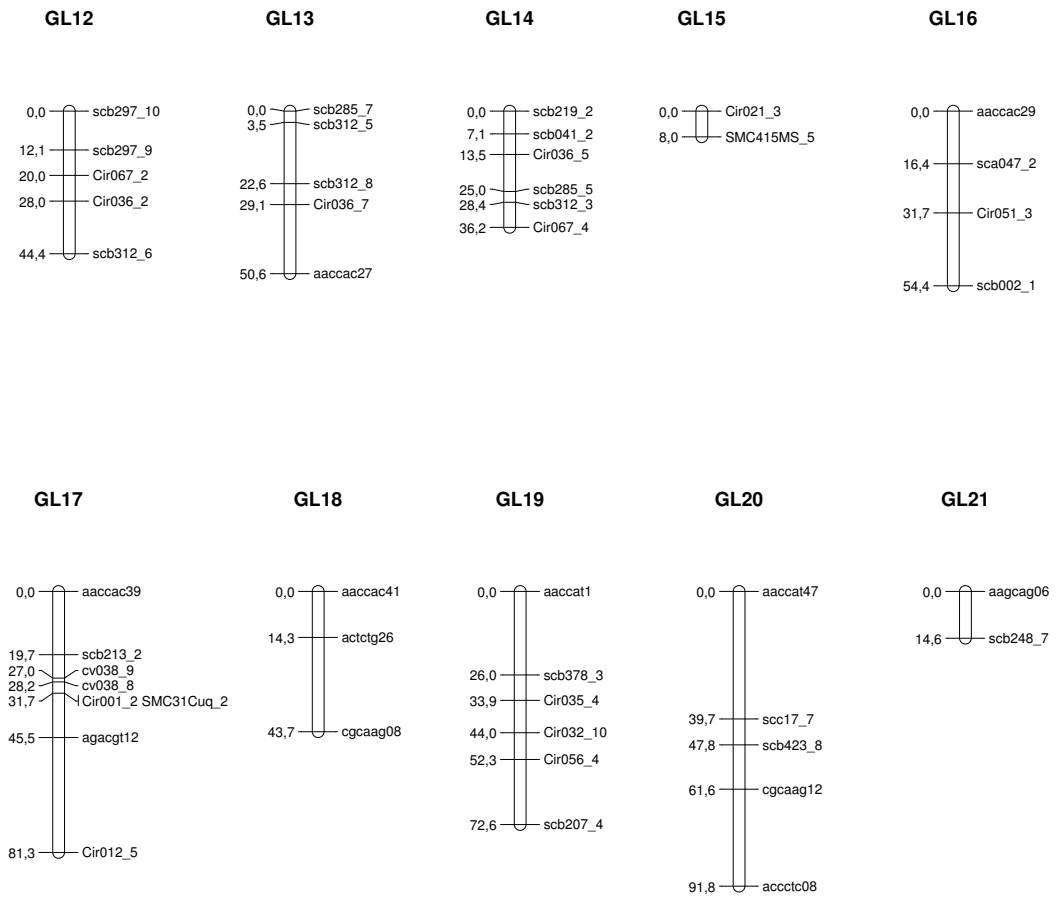
ZHANG, L.; WANG, S.; LI, H.; DENG, Q.; ZHENG, A.; LI, S.; LI, P.; LI, Z.; WANG, J. Effects of missing marker and segregation distortion on QTL mapping in F^2 populations. **Theoretical and Applied Genetics**, New York, v. 121, p. 1071-1082, 2010.

APÊNDICE

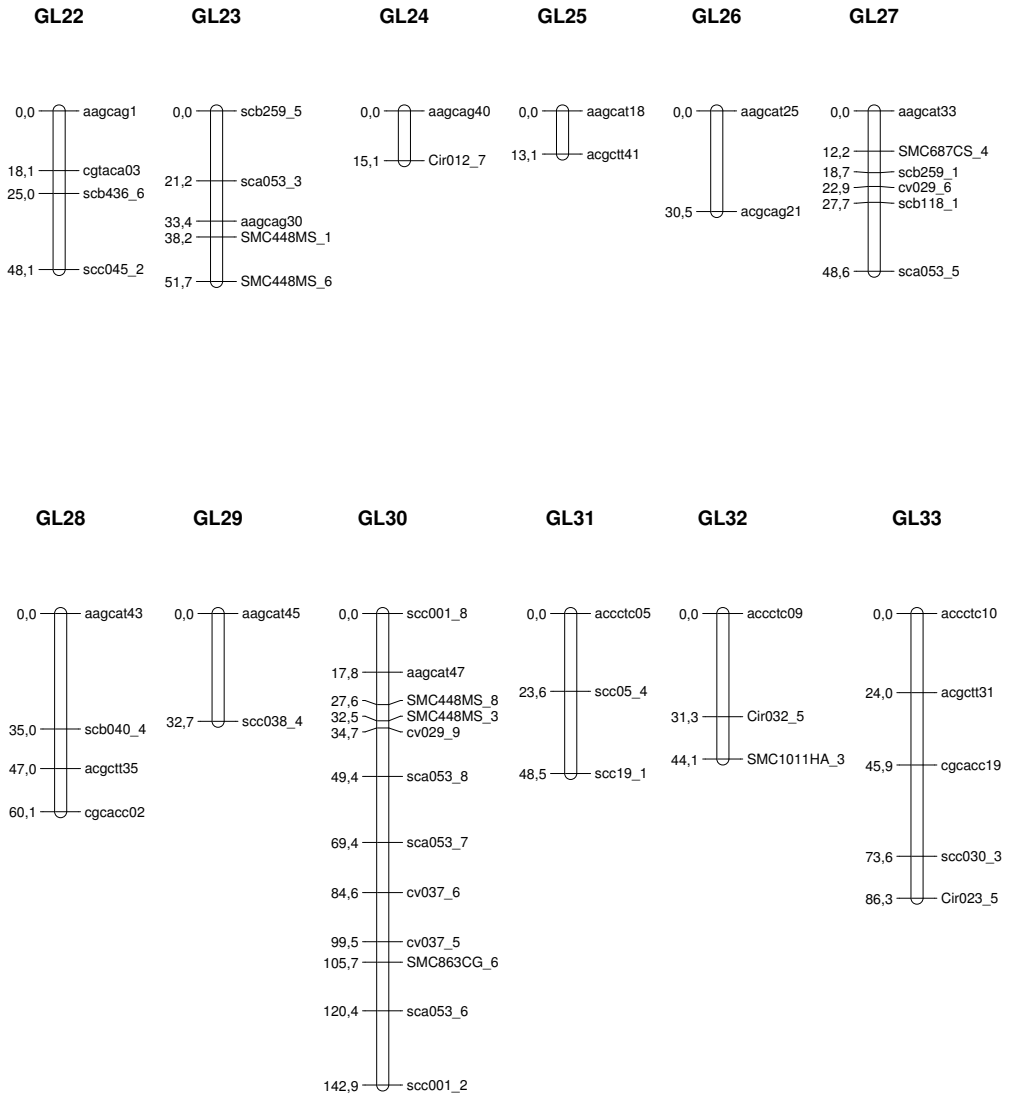
APÊNDICE A - Mapa genético de ligação do cruzamento entre os genitores de cana-de-açúcar IACSP95-3018 e IACSP93-3043 baseado em 220 indivíduos



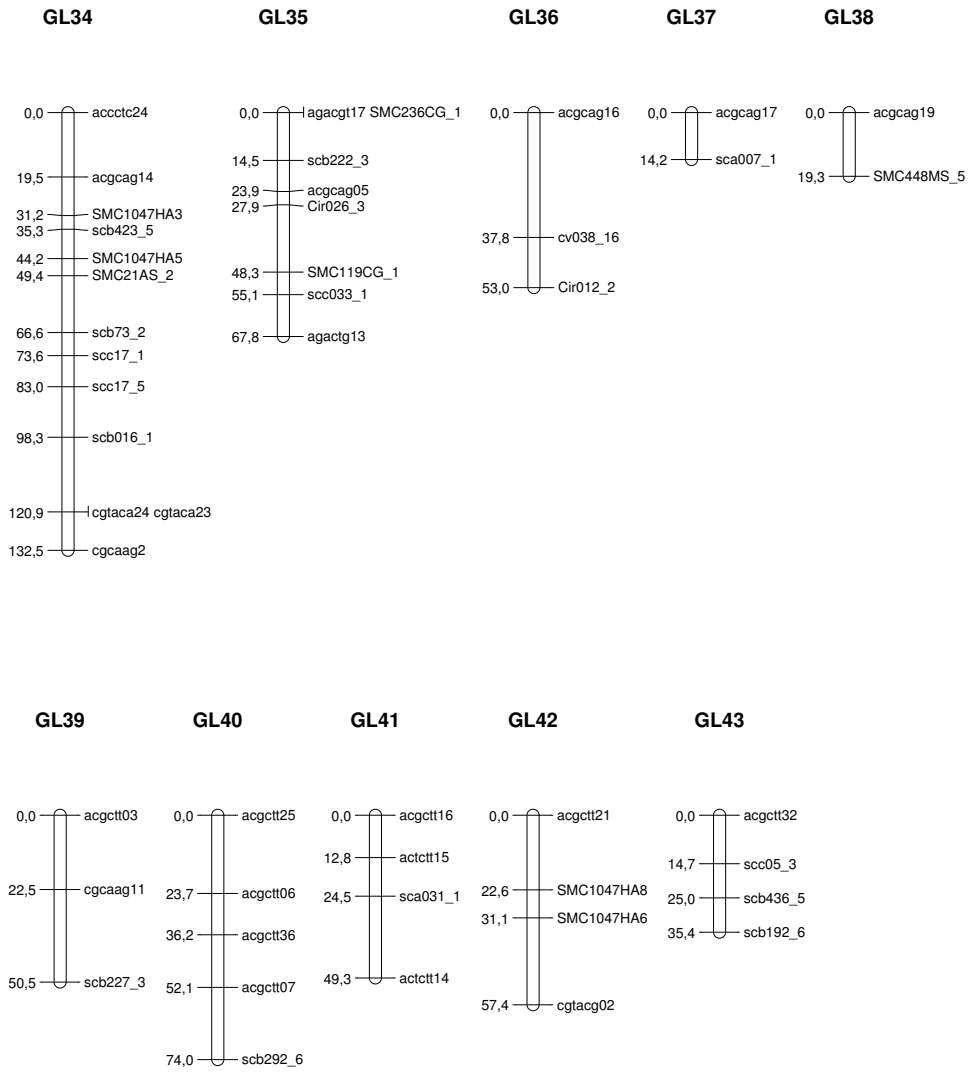
APÊNDICE A - Mapa genético de ligação do cruzamento entre os genitores de cana-de-açúcar IACSP95-3018 e IACSP93-3043 baseado em 220 indivíduos



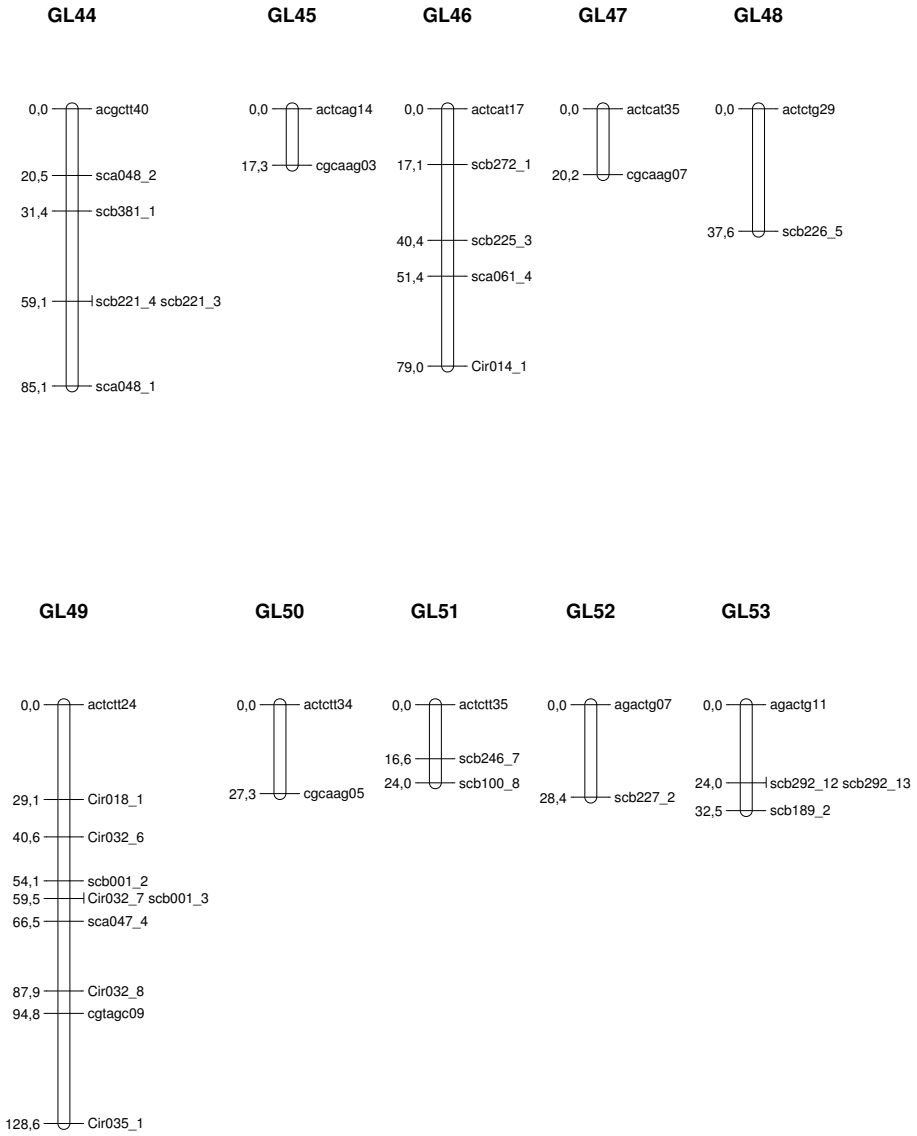
APÊNDICE A - Mapa genético de ligação do cruzamento entre os genitores de cana-de-açúcar IACSP95-3018 e IACSP93-3043 baseado em 220 indivíduos



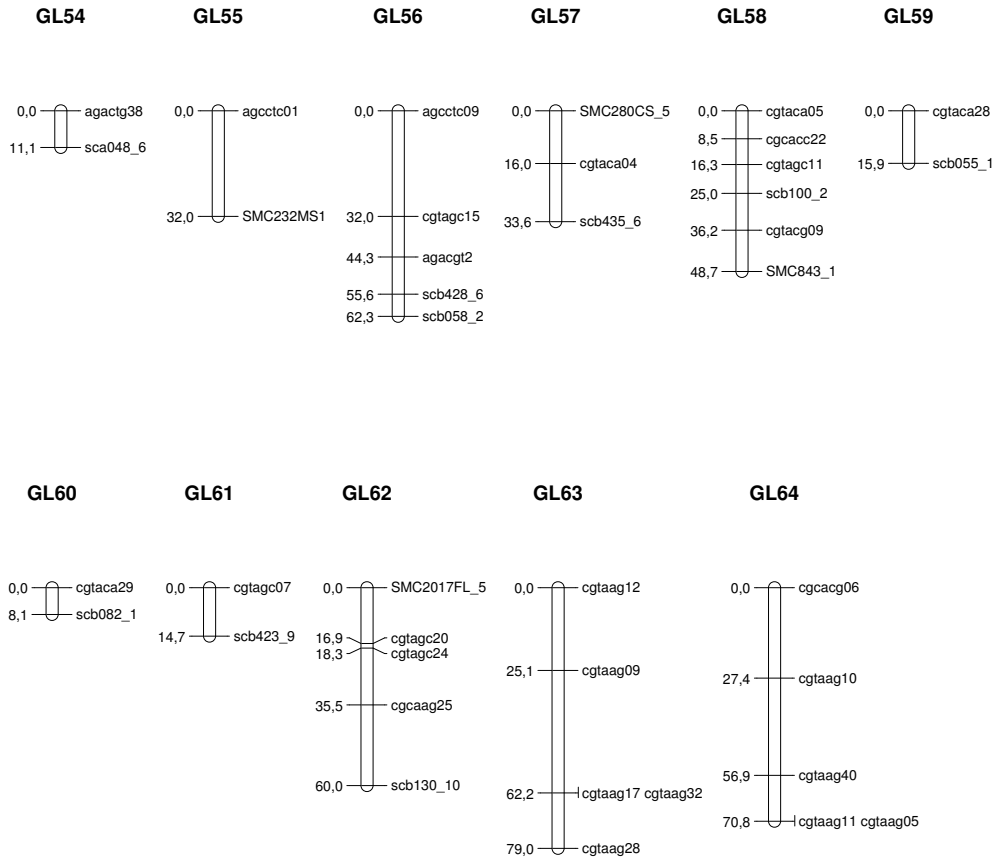
APÊNDICE A - Mapa genético de ligação do cruzamento entre os genitores de cana-de-açúcar IACSP95-3018 e IACSP93-3043 baseado em 220 indivíduos



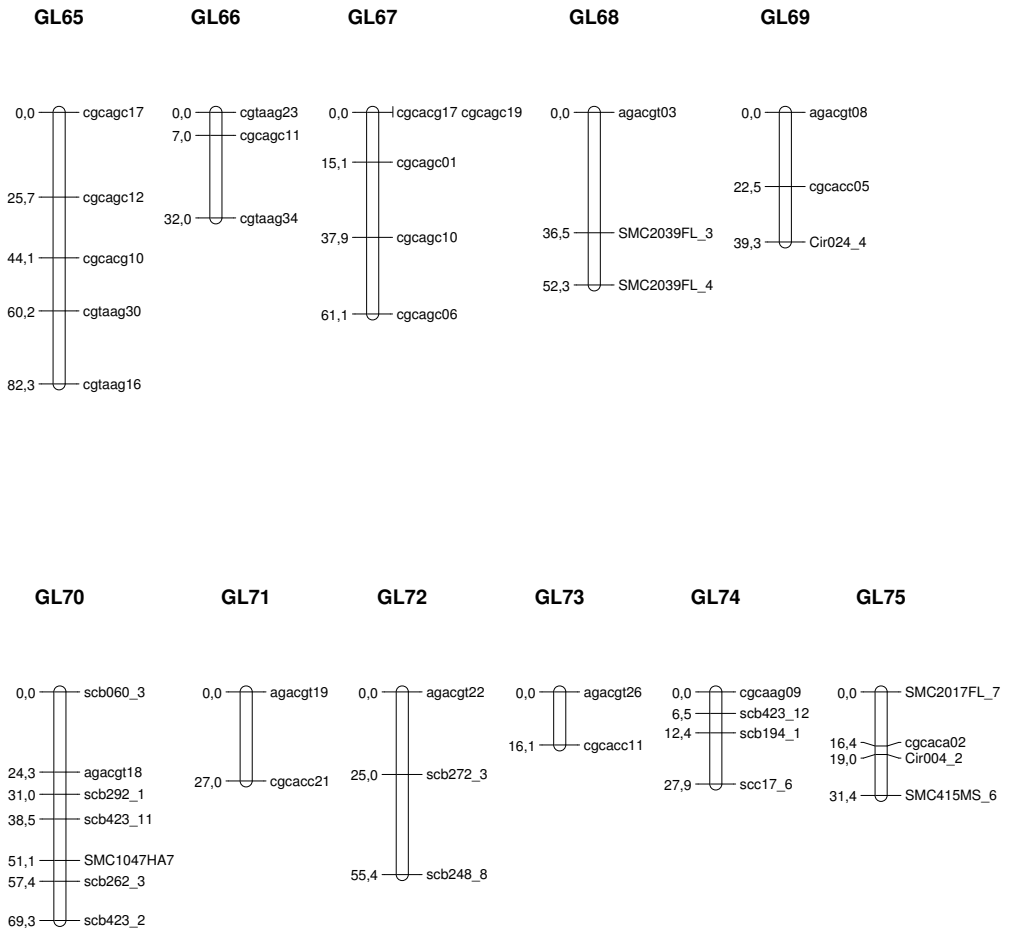
APÊNDICE A - Mapa genético de ligação do cruzamento entre os genitores de cana-de-açúcar IACSP95-3018 e IACSP93-3043 baseado em 220 indivíduos



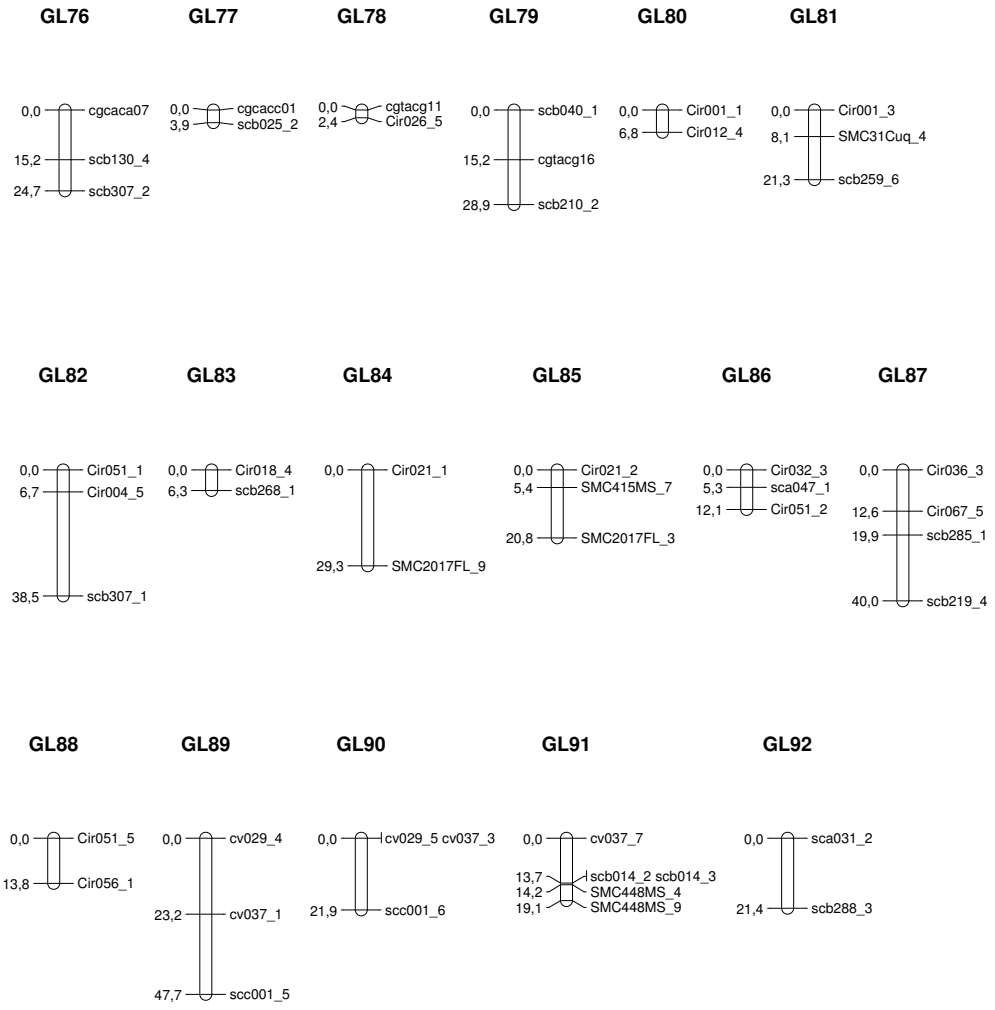
APÊNDICE A - Mapa genético de ligação do cruzamento entre os genitores de cana-de-açúcar IACSP95-3018 e IACSP93-3043 baseado em 220 indivíduos



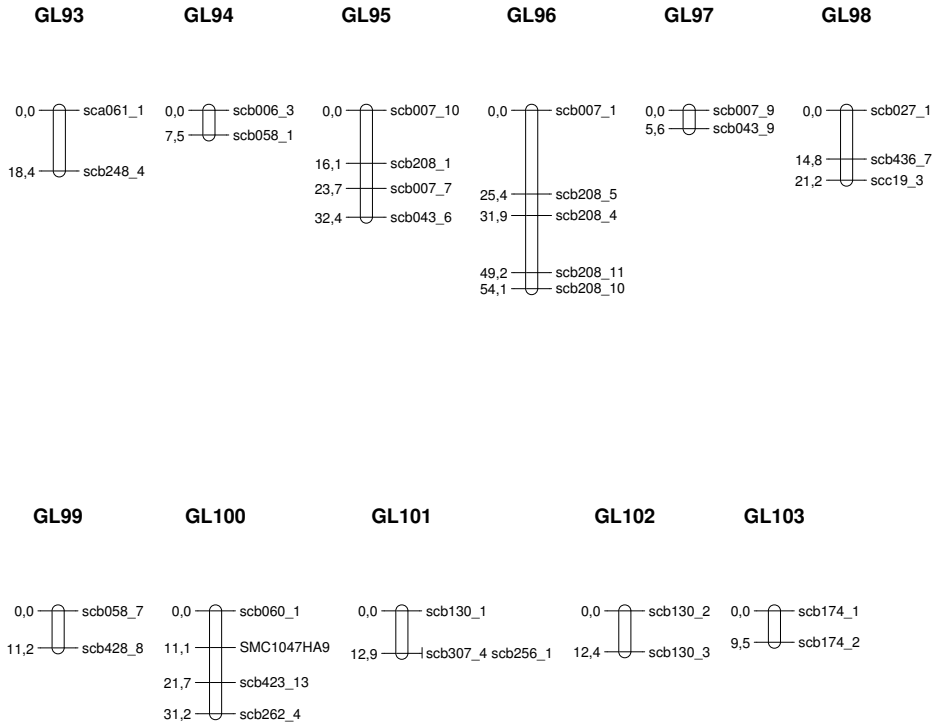
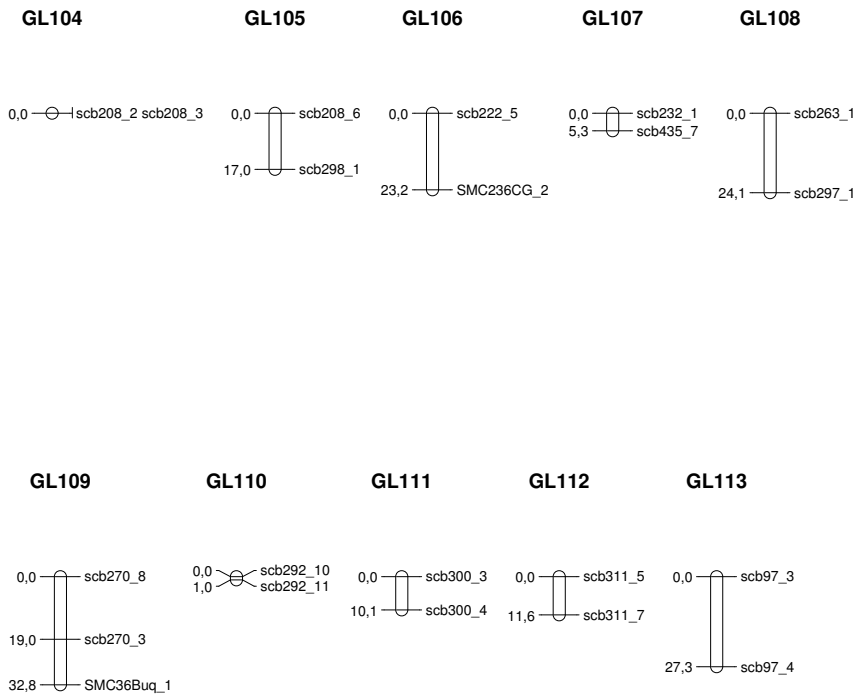
APÊNDICE A - Mapa genético de ligação do cruzamento entre os genitores de cana-de-açúcar IACSP95-3018 e IACSP93-3043 baseado em 220 indivíduos



APÊNDICE A - Mapa genético de ligação do cruzamento entre os genitores de cana-de-açúcar IACSP95-3018 e IACSP93-3043 baseado em 220 indivíduos



APÊNDICE A - Mapa genético de ligação do cruzamento entre os genitores de cana-de-açúcar IACSP95-3018 e IACSP93-3043 baseado em 220 indivíduos



APÊNDICE A - Mapa genético de ligação do cruzamento entre os genitores de cana-de-açúcar IACSP95-3018 e IACSP93-3043 baseado em 220 indivíduos